IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 13 May 2024, accepted 7 July 2024, date of publication 11 July 2024, date of current version 23 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3426280

RESEARCH ARTICLE

Consensus and Risk Aversion Learning in Ensemble of Multiple Experts for Long-Tailed Classification

TAEGIL HA AND JIN YOUNG CHOI, (Member, IEEE)
Automation and Systems Research Institute (ASRI), Department of Electrical and Computer Engineering, Seoul National University, Seoul 08826, South Korea
Corresponding author: Jin Young Choi (jychoi@snu.ac.kr)

This work was supported by the Artificial Intelligence Graduate School Program (Seoul National University).

ABSTRACT Recent expert ensemble methods for long-tailed recognition encourage diversity by maximizing KL divergence between the predictions of experts. However, the excessive diversity using KL divergence, which has no upper bound, induces inaccurate predictions of experts. To address this issue, we propose a new learning method for expert ensemble, which obtains the consensus by aggregating the predictions of experts (Consensus) and maximizes the expected prediction accuracy of each expert without excessive diversity from the consensus (Risk Aversion). To implement this learning scheme, we propose a new loss derived from Rényi Divergence. We provide both empirical and theoretical analysis of the proposed method along with a stability guarantee, which is not guaranteed at the existing methods. Thanks to this stability, the proposed method continues to improve performance as the number of experts increases, while the existing methods do not. The proposed method achieves state-of-the-art performance for any number of experts. Furthermore, the proposed method operates robustly even when evaluated by varying the imbalance factor.

INDEX TERMS Long-tailed object recognition, classification, ensemble of multiple experts, consensus, risk aversion.

I. INTRODUCTION

In real-world scenarios, the distribution of object classes often shows a long-tailed property [1], [2]. Long-tailed distribution means that the some classes have a wealth of data samples, called head classes, while some classes have extremely limited number of data samples, called tail classes. The long-tailed distribution, originally introduced in economics as the Pareto distribution [3], is noteworthy for its presence not only in economic contexts but also in various natural phenomena [4]. This broad applicability has attracted research interest, especially in the field of visual recognition [2], [5], [6], [7]. The central challenge in long-tailed recognition arises from the disparity between training dataset designed to have long-tailed distribution and test dataset which is not necessarily being long-tailed. Conventional learning methods lead to a strong bias towards the head classes and suppress the performance in tail classes since they assume a balanced class distribution of training dataset. To solve the long-tail problems, most studies focus on balancing strategies such as re-sampling [8], [9], [10], adjusting logit or reweighted loss [6], [11], [12], [13], augmentation [14], [15], [16], [17], contrastive learning [18], [19], [20], [21], and so on.

Recently, expert ensemble methods using multiple experts have been proposed [20], [22], [23], [26], [27], [28], [29], [30]. The expert ensemble methods reduce the prediction variance of the tail class that suffers from large variance due to the lack of training samples [27]. Some expert ensemble methods apply specialized training scheme for each expert to make diversity among experts [22], [23], [26] as shown in Figure 1 (a). However, this approach requires man-made heuristics with pre-determined number of experts, which is not scalable.

The associate editor coordinating the review of this manuscript and approving it for publication was Asadullah Shaikh.

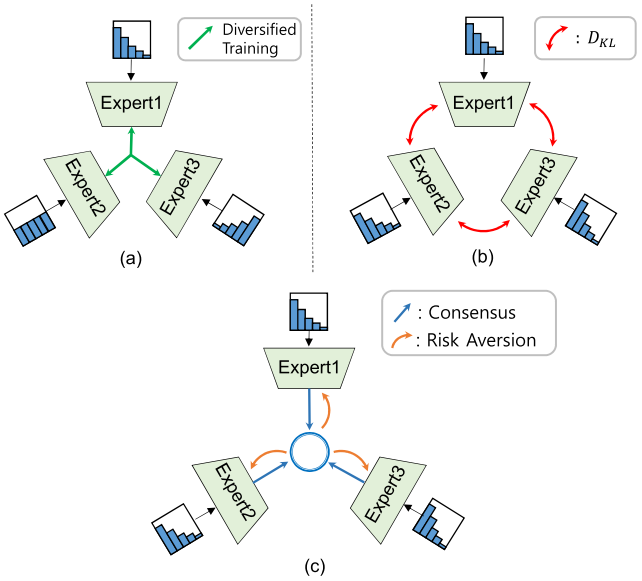

FIGURE 1. Ensemble schemes using multiple experts. (a) Ensemble scheme that trains each expert with its own specialized training scheme [22], [23], [24], [25], [26]. (b) Ensemble scheme that encourages each expert to be diversified [27], [28], [29], or distills knowledge from the others [20]. (c) The proposed consensus and risk aversion (CRAL) scheme that makes consensus and encourages diversified prediction in stable manner.

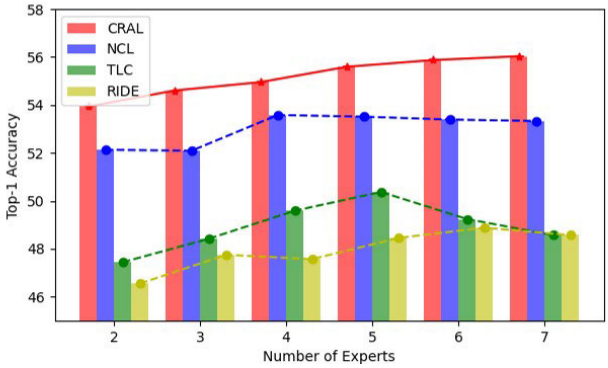

FIGURE 2. Maximizing D_{KL} between experts for diversity(RIDE [27], TLC [28]) results in performance degradation as more experts are involved due to their instability. NCL [20], which minimizes D_{KL} to distill knowledge, shows performance saturating since it encourages experts to predict similarly. The proposed method(CRAL) guarantees stability under consensus and controls the degree of diversity between experts by a risk aversion parameter, which shows continuous performance improvement as the number of experts increases. For a fair comparison, all methods are trained with CIFAR100-LT (imbalance ratio = 100) [5], [6] along with the same data augmentation, training schemes, and environment.

Expert ensemble methods employ Kullback-Leibler divergence D_{KL} [31] as shown in Figure 1 (b). [20], [32] minimizes D_KL between experts to distill knowledge. However, minimizing D_{KL} between experts discourages experts to make diverse predictions, which results in saturated performance even the number of experts increases, as shown in Figure 2. On the other hand, the diversity loss maximizes D_{KL} [27], [29] between experts to encourage diverse prediction. However, maximizing D_{KL} makes experts be diversified excessively because D_{KL} has no upper bound, which induce inaccurate prediction of other experts. As a result, increasing the number of experts leads to performance degradation as shown in Figure 2.

To address the aforementioned limitations, we propose Consensus and Risk Aversion Learning (CRAL) that trains expert ensemble for long-tailed classification. CRAL obtains the consensus that aggregates the predictions of experts, and makes each expert be optimized following risk aversion behavior given the consensus, as shown in Figure 1 (c). Risk aversion behavior maximizes an expected prediction accuracy of each expert without excessive diversity from the consensus. To implement the risk aversion behavior, we propose a Negative Expected Rate of Return (NERR) loss derived from derived from **Rényi Divergence** [33]. Finally, we further enhances our model incorporating Class Prior Adjustment (CPA).

Unlike KL-divergence that has no upper bound, the proposed NERR loss has upper bound mathematically, which is proved in this paper. This property guarantees stable learning. Thanks to this property, the performance of CRAL improves continuously as the number of experts increases, while existing methods using ensemble of multiple experts do not as shown in Figure 2. Furthermore, extensive experiments are conducted to show the validity and outperforming accuracy of CRAL in popular long-tailed classification benchmarks [2], [5], [6], [7].

Our contributions are summarized as follows:

- We propose a new expert ensemble learning that gives stable predictions by regulating diversity unlike existing methods suffering from excessive diversity.
- We achieve the diversity regulation by consensus of multiple experts and risk aversion behavior that regulates the prediction diversity from consensus. To this end, we develop the NERR loss.
- We prove that the NERR loss is bounded to ensure stable training unlike existing methods using D_{KL}.

II. RELATED WORKS

A. BALANCING STRATEGIES FOR LONG TAILED VISUAL RECOGNITION

To mitigate the negative effects of class imbalance, researchers have designed on balancing strategies such as re-sampling of training dataset [8], [9], [10]. Subsequently, methods adjusting logit or balancing loss have been proposed to regulate the influence of each class, or samples [6], [11], [12], [13], [34], [35], [36]. The presence of class imbalance within the training data leads to an in ill-conditioned decision boundaries, and predictions are severely biased to the head classes [15], [37], [38]. Balanced losses improves performance of tail classes by effectively balancing imbalanced factors. Reference [13] proposes sample-level re-weighting to control the influence of each sample on training phase. Reference [34] proposes effective number of each class as class-level re-weighting. Reference [12] is another class-level re-weighting method that compensate

class distributions' disparity between training and test dataset. References [11] and [36] introduces adjustment terms of the prediction. Additionally, [6] proposes loss function that induces larger margins for minority classes, thereby rectifying the ill-conditioned decision boundary. On the other hand, the methods to augment samples is explored to mitigate the risk of overfitting to the limited tail class samples [14], [15], [16], [17]. Additionally, contrastive learning have been introduced to extract features with powerful discriminativity [18], [19], [20], [21].

B. ENSEMBLE OF MULTIPLE EXPERTS IN LONG-TAILED CLASSIFICATION

Methods have emerged to improve predictions for tail classes in long-tailed classification [20], [22], [23], [24], [26], [27], [28], [29], [30]. These methods primarily focus on acquiring experts with diversity or distilling knowledge from each other. BBN [22] employs an ensemble of two experts, one trained on a long-tailed distribution dataset and the other on a re-balanced data distribution. Similar to [22], SADE [26] is trained using a dataset with different class data distributions. Furthermore, [26] proposed strategies to aggregate predictions of multiple experts through linear interpolation with varying weights, and this process involves the use of the test dataset. ACE [24] is designed to allocate each expert to perform inferences for specific classes, with multiple experts contributing to the tail class predictions. LGLA [30] employs different logit adjustment strategies for each expert and for aggregated logits. RIDE [27], Mutual Learning [29], and TLC [28] aim to maximize the relative entropy between experts, facilitating a wider range of predictions. In particular, [27] introduces a router that determines which expert to use during the inference phase. TLC [28] employs evidence theory to enhance the predictions of tail classes and present dynamic engagement strategy of each expert. In contrast from [27], [29], and [28], LFME [25] tries to distill other experts' knowledge by minimizing cross extropy of logits. Similary, NCL [20] focuses on knowledge distillation by minimizing relative entropy between experts while incorporating hard category mining.

C. ECONOMICS WITH INFORMATION THEORY

Information theory has been applied to economics to analyze one's decision according to the expected return. When it comes to the horse-racing game [39], [40], D_{KL} between the probability distribution that gambler believes and official odds eqauls the expected rate of return of the gambler under the assumption that the gambler shows growth-optimizing behavior [39]. Furthermore, expected utility hypothesis considers different risk attitudes of individuals [41]. Individuals refuse to take part in fair gamble whose expected return is zero. At the same time, individual shows risk-averse behavior rather than the option with higher risk-higher return. [40] generalized the relationship between expected rate of return with relative risk aversion coefficient [42], [43] using generalized divergence [33]. Moreover, deriving consensus from multiple experts is researched [44], [45].

III. METHODOLOGY

This section presents details of the proposed Consensus and Risk Aversion Learning (CRAL). In Section III-A, we introduce Rényi divergence, a key element in subsequent formulas, along with brief outline about the expected rate of return from economics, that leverages Rényi divergence. In Section III-B, we propose CRAL and its training objective, Negative Expected Rate of Return (NERR), bridging economics to ensemble of multiple experts. In Section III-C, we present an strategy to obtain consensus of multiple experts, and in-depth analysis to prove the stability of CRAL under the proposed consensus function. Finally, in Section III-D, we explain strategies to further improve NERR, along with individualized training scheme to train each expert.

A. PRELEMINARIES

Rényi Divergence [33], which is also known as α-divergence is defined as follows:

$$D_\alpha(P||Q) = \frac{1}{\alpha - 1} \log\left(\sum_{i=1}^{C} \frac{p_i^\alpha}{q_i^{\alpha-1}}\right), \tag{1}$$

where $P = \{p_1, \ldots, p_C\}$ and $Q = \{q_1, \ldots, q_C\}$ are probability distributions and α is a non-negative coefficient. D_α at $\alpha = 1, \infty$ is defined by taking the limit of (1), where D_1 corresponds to the well-kwown Kullback–Leibler divergence D_{KL} [31]. D_α is a generalized, smoothed statistic divergence function of D_{KL} [46]. D_α is asymmetric, meaning that $D_\alpha(P||Q)$ is not equal to $D_\alpha(Q||P)$ in general. Furthermore, D_α is an unbounded function with respect to the probability distributions P and Q.

Divergence as a Measure of Expected Return. Divergence serves as a measure of dissimilarity between two probability distributions, and [39] have applied it to betting theory assuming the gambler who takes part in the horse racing game. In the game, there exists a probability distribution m believed by market, referred to as the official odds that sum up to 1. Then, the amount bet on each horse is proportional to m. However, if the gambler were to bet in accordance with m strictly, expected return that the gambler receive becomes zero. As a result, the gambler seeks to invest based on own belief distribution b. Kelly [39] proved that, under the assumption of growth-optimizing behavior, the gambler's expected rate of return is proportional to $D_{KL}(b||m)$.

Reference [40] further generalized expected rate of return $E_{R_w}(b, m)$ using the generalized divergence, D_α. Reference [40] introduced relative risk-aversion coefficient R_w and formulated expected rate of return as follows:

$$\mathbb{E}_{R_w}(b, m) = \frac{1}{R_w} D_1(b||m) + \frac{R_w - 1}{R_w} D_{1/R_w}(b||m). \tag{2}$$

Here, R_w is defined in [42] and [43]:

$$R_w = -\frac{wu''(w)}{u'(w)}, \tag{3}$$

where $u(w)$ is the utility function associated with the wealth w, and $'$ stands for the derivative operator with respect to w. R_w serves as an indicator of risk attitude, where a low value of R_w implies a high elasticity of return [47], showing a risk-taking behavior.

B. CONSENSUS AND RISK AVERSION LEARNING

Leveraging the strength of multiple experts without heuristics requires methods that maximize divergence between them. However, existing approaches such as [27], [29], and [28] exhibit performance degradation as the involved number of experts increases. We attribute this to undesirable mathematical properties of the employed divergence measures. In response, we propose our scheme, Consensus and Risk Aversion Learning (CRAL) as shown in Figure 1 (c). We design CRAL so that it obtains consensus from all experts and makes each expert yield prediction to maximize its expected accuracy without excessive deviation from the consensus. For further explanation, we introduce the notations used throughout the paper.

Notations. Let's define notations for further explanation as followings:

- C: Number of classes of dataset
- E_i: i-th expert where $i = 1, \ldots, n$
- $P_{i,c}(x) = P(y = c|x, E_i)$: probability that the output y of input data x belongs to class c, which is predicted by E_i. We get probability prediction with softmax function of logit $l_{i,c}$
- $P_i = \{P_{i,c}|c = 1, \ldots, C\}$: the probability mass function predicted by E_i
- $f : \mathbb{R}^{C\times n} \rightarrow [0, 1]^C$: consensus function that receives belief probability distributions P_i, $(i = 1, \ldots, n)$ and yields consensus probability distribution

Using (2), we define the training objective function for E_i, called Negative Expected Rate of Return (NERR) loss, as

$$\mathcal{L}_{NERR}(P_i, f; R_w) = -\mathbb{E}_{R_w}(P_i, f), \tag{4}$$

where f indicates the consensus function $f(P_1, .., P_n)$.

Remark: Comparing the parameters between (2) and (4), we can draw analogies between betting theory and the multiple-expert ensemble. The gambler's belief b is replaced by the individual expert's prediction P_i. The consensus function f can be used instead of official odds m in (2).

For simplicity, we adopt the isoelastic function [48] as for the utility function in (3), i.e.,

$$u(w) = \frac{w^{1-\rho} - 1}{1 - \rho}, \tag{5}$$

where ρ is a constant parameter. By (5), (3) is reduced to $R_w = \rho$. From now on, we substitute R_w to constant coefficient ρ. Then the following **Theorem 1** gives intuition about the behavior of $\mathcal{L}_{NERR}(P_i, f; \rho)$.

Theorem 1: For any probability mass functions P_i, f,

$$-\mathcal{L}_{NERR}(P_i, f; \rho) \geq 0 \ \text{ if } \ \rho \geq 1, \tag{6}$$

whereas $-\mathcal{L}_{NERR}(P_i, f; \rho) = 0$ when $P_i = f$ or if $\rho \rightarrow \infty$ under the condition $P_i, c \neq 0 \ \forall c$.

Proof: For $\rho \geq 1$, $-\mathcal{L}_{NERR}(P_i, f; \rho)$ is linear interpolation of divergence metrics $D_{KL}(P_i||f), D_{1/\rho}(P_i||f)$, hence $-\mathcal{L}_{NERR}(P_i, f; \rho)$ is also non-negative divergence metric.

The condition

$$D_{KL}(P_i||f) = D_{1/\rho}(P_i||f) = 0 \tag{7}$$

is equivalent to $P_i = f$. Furthermore, if $P_{i,c} \neq 0$ for all $c = \{1, \ldots, C\}$, then

$$\lim_{\rho\rightarrow\infty} -\mathcal{L}_{NERR}(P_i, f; \rho) = D_0(P_i||f) = -\log 1 = 0, \tag{8}$$

Theorem 1 indicates that the proposed training objective $-\mathcal{L}_{NERR}(P_i, f; \rho)$ is a divergence metric under $\rho \geq 1$. Letting a hyper-parameter $\rho \in [1, \infty)$, minimizing $-\mathcal{L}_{NERR}(P_i, f; \rho)$ with respect to P_i, f encourages diversity between P_i and f.

Unlike methods that maximize D_{KL} [27], [28], [29], the proposed loss employ the consensus f and control the degree of risk aversion [46] with ρ. Assigning a small value to ρ encourages risk-taking behavior, encouraging every expert to yield prediction maximizing its own accuracy without excessive deviation from the consensus. On the other hand, assigning a large value to ρ enforces risk aversion behavior. Increasing ρ results in less diversity between P_i and f and taking $\rho \rightarrow \infty$ makes the loss converge to 0. Note that we cannot guarantee $-\mathcal{L}_{NERR}(P_i, f; \rho)$ is non-negative for $0 < \rho < 1$, and we cannot ensure NERR works properly when ρ approaches zero.

C. STABILITY GUARANTEE

Minimizing $-D_{KL}(P_i, P_j)$, $i \neq j$ to train multiple experts [27], [29] is unstable since $-D_{KL}(P_i, P_j)$ has no lower bound with respect to (P_i, P_j). This instability lead to performance degradation (Fig. 2), and even gradient explosion is observed when employing many experts [28]. To guarantee stability, our scheme gets consensus f and encourages each prediction P_i to be diversified from consensus f. We define the consensus function as $f = \frac{1}{n}\sum_{k=1}^{n} P_i$. Then, the following **Theorem 2** ensures that $-\mathcal{L}_{NERR}(P_i, f; \rho)$ is upper-bounded.

Theorem 2: For given natural number of experts n, $-\mathcal{L}_{NERR}(P_i, f; \rho)$ is upper-bounded for any positive ρ, i.e,

$$\begin{aligned} -\mathcal{L}_{NERR}(P_i, f; \rho) &\leq n \cdot JSD(P_1, \ldots, P_n)) \\ &\leq n \log n, \end{aligned} \tag{9}$$

where $JSD(\cdot)$ stands for Jensen-Shannon divergence [49].

To give proof of Theorem 2, we first prove following Lemma 1.

Lemma 1: For any probability mass function P, Q, $D_\alpha(P||Q)$ is non-decreasing function of α for all nonnegative real number α.

Proof: Here, we extend the proof for the probability density function presented in [50] to the probability mass function. $D_\alpha(P||Q)$ is non-decreasing function of α for all positive α. Denote $p_c = P(y = c)$, $q_c = Q(y = c)$ for $c = 1, \ldots, C$.

According to [50], for $0 \leq \alpha < \beta$, the function $f(x) = x^{\frac{\alpha-1}{\beta-1}}$ defined on $x \geq 0$ is strictly convex if $\alpha < 1$ and strictly concave if $\alpha > 1$.

Case I: $0 \leq \alpha < 1$. By Jensen's inequality, following inequality holds:

$$\left(\sum_{c=1}^{C}\left(\frac{p_c}{q_c}\right)^{(\beta-1)} p_c\right)^{\frac{\alpha-1}{\beta-1}} \leq \sum_{c=1}^{C}\left(\frac{p_c}{q_c}\right)^{\alpha-1} p_c. \tag{10}$$

Taking logarithm of (10), and divide both side by the negative real number $\alpha - 1$, we obtain $D_\alpha \leq D_\beta$.

Case II: $\alpha > 1$. By Jensen's inequality, following inequality holds:

$$\left(\sum_{c=1}^{C}\left(\frac{p_c}{q_c}\right)^{(\beta-1)} p_c\right)^{\frac{\alpha-1}{\beta-1}} \geq \sum_{c=1}^{C}\left(\frac{p_c}{q_c}\right)^{\alpha-1} p_c. \tag{11}$$

Taking logarithm of (11), and divide both side by the positive real number $\alpha - 1$, we obtain $D_\alpha \leq D_\beta$.

Case III: $\alpha = 1$. By the definition of Rényi divergence, D_α is continuous at $\alpha = 1$.

By Case I, II, and III, Lemma 1 holds.

Proof of Theorem 2: By Lemma 1, following equation holds.

$$\begin{aligned} &-\mathcal{L}_{NERR}(P_i, f; \rho) - D_{KL}(P_i||f) \\ &= \left(\frac{\rho - 1}{\rho}\right)\left(D_{1/\rho}(P_i||f) - D_{KL}(P_i||f)\right) \qquad (12) \\ &\leq 0. \\ &\Rightarrow -\mathcal{L}_{NERR}(P_i, f; \rho) \leq D_{KL}(P_i||f) \qquad (13) \end{aligned}$$

According to [51], following inequality holds.

$$JSD(P_1, \ldots, P_n) \leq \log n \tag{14}$$

Then, using (13) and (14), we get following.

$$\begin{aligned} D_{KL}(P_i||f) &\leq \sum_{i=1}^{n} D_{KL}(P_i||f) \\ &= n \cdot JSD(P_1, .., P_n) \\ &= n \cdot JSD(P_1, .., P_n) \\ &\leq n \log n. \end{aligned} \tag{15}$$

By (13) and (15), Theorem 2 holds for all $\rho > 0$.

The consensus is equally contributed by E_i, $(i = 1, \ldots, n)$, which is suitable to CRAL that trains all experts under the same training strategy. As a result, while [27] and [28] shows worse accuracy when employing more experts, CRAL improves accuracy as the number of experts increases, as shown in Figure 2. Also, we use consensus as the final predictor of ensemble at inference phase.

D. INDIVIDUALIZED TRAINING OF EACH EXPERT

We perform CRAL using $\mathcal{L}_{NERR}(P_i, f; \rho)$ to maximize return of E_i, along with classification loss [12] denoted by $\mathcal{L}_{cls,i}$ to train E_i. Instead of aggregating the losses of all experts, we train E_i individually. This individualized training allows each expert to be stand-alone classifier [27], and well-trained experts generate good consensus in collaborative manner [52]. Training loss for E_i is given by

$$\mathcal{L}_{cls,i} + \lambda \mathcal{L}_{NERR}(P_i, f; \rho). \tag{16}$$

In our implementation, the output logit ($l_{i,c} = W_{i,c}x$) of each expert is normalized as

$$l'_{i,c} = \frac{W_{i,c}}{||W_{i,c}||_2}\frac{x}{||x||_2}, \tag{17}$$

where x is a feature vector, $W_{i,c}$ is a classifier weight vector for class c from E_i.

In $\mathcal{L}_{NERR}(P_i, f; \rho)$, class imbalance is not considered. Thus to accelerate the effect of $\mathcal{L}_{NERR}(P_i, f; \rho)$, we adjust the normalized logit ($l'_{i,c}$) using class prior probability ($P(y = c|X_{\text{train}})$) to compensates $l'_{i,c}$ to reflect difference of distributions between training and test datasets [12]. To this end, we apply **Class Prior Adjustment** (CPA) strategy given by

$$l''_{i,c} = P(y = c|X_{\text{train}})l'_{i,c} = \frac{n_c}{\sum_{k=1}^{C} n_k} l'_{i,c}, \tag{18}$$

where X_{train} is the set of all training samples.

IV. EXPERIMENTS AND RESULTS

A. IMPLEMENTATION DETAILS

We implemented all our methods with PyTorch [53] library. Our implementation is based on official implementation of [14] and [27], but we do not use unfair additional data augmentation when comparing performance to other methods. We use weight decay of $2e^{-4}$ and batch size is set to 256. The base learning rate is set to be 0.1, with warm-up epochs following convention [14], [27]. f is frozen during each iteration of training all experts E_is $(i = 1, .., n)$ individually. We set $\lambda = 0.3$ in (16), and $\rho = 5.0$ for all experiments if not mentioned. We evaluate the proposed method on the most commonly used benchmarks: CIFAR100-LT [5], [6] with various imbalance factors(IF = 100, IF = 50, IF = 10) along with large-scaled benchmarks, ImageNet-LT [2] and iNatualist2018 [7]. To ensure a fair comparison, We make an effort to reproduce or gather results of other methods using the same backbone baseline, data augmentation strategies, and training schemes. We used 2 RTX 2080Ti to train CIFAR-100LT, and 4 RTX 4090 to train on ImageNet-LT and iNaturalist2018, which is large-scaled dataset. We use stochastic Gradient Descent as the optimizer with the momentum of 0.9 [54]. Following [27], early layers of backbone are shared.

ImageNet-LT: The ImageNet-LT dataset can be obtained from the published source described in [2]. We conduct experiments using both ResNet-50 [55] and ResNext-50 [56].

TABLE 1. Comparison between CRAL and other multi-experts methods by increasing the number of experts. All models are trained for 200 epochs at CIFAR-100LT (IF = 100). *n* stands for number of experts. We add performance difference from previous *n* in parenthesis. Note that we cannot get the result for the cases written as 'N/A' due to their unstability. All results are reproduced by us under the identical training strategies.

n	RIDE [27]	TLC [28]	NCL [20]	CRAL
2	46.6 (-)	47.4 (-)	52.1(-)	53.9(-)
3	47.8(+1.2)	48.4(+1.0)	52.1(+0.0)	54.6(+0.7)
4	47.6(-0.2)	49.6(+1.2)	53.6(+1.5)	55.0(+0.4)
5	48.5(+0.9)	50.4(+0.6)	53.5(-0.1)	55.6(+0.6)
6	48.9(+0.4)	49.2(-1.2)	53.4(-0.1)	55.9 (+0.3)
7	48.6(-0.7)	48.6(-0.6)	53.3(-0.1)	56.0 (+0.1)
8	49.0(+0.4)	50.3(+1.7)	53.4(+0.1)	56.3(+0.3)
10	44.5(-4.5)	N/A(-)	54.1(+0.7)	56.7 (+0.4)
15	45.0(+0.5)	N/A(-)	53.8(-0.3)	57.2 (+0.5)

TABLE 2. Ablation study to verify the effects of NERR, CPA in CRAL. *n* stands for the number of experts and Med stands for medium. When we do not use NERR, we have trained with $\lambda = 0$ in (16). All models are trained for 200 epochs at CIFAR = 100LT (IF = 100). The best results are marked in bold.

n	NERR	CPA	All	Many	Med	Few
3			52.8	68.5	54.8	32.0
3	✓		54.6	69.9	56.3	**34.8**
3		✓	53.4	68.4	54.7	34.4
3	✓	✓	**55.2**	**70.5**	**57.6**	34.5

Following common learning rate schedule [14], [27], we have trained the proposed model for 200 epochs with learning rate decay by a factor of 0.1 at the 120th and 160th epoch. We also conduct 400-epoch training following recent conventions [20], [21], [57] with learning rate decay by a factor of 0.1 at the 320th and 360th epoch. We use Randaug [58] along with random horizontal flip.

CIFAR-100 LT: CIFAR-100 is obtained by the method described in [5]. To obtain CIFAR-100 LT, the long-tailed version of CIFAR 100, we follow the method described in [34]. We use ResNet-32 [55] as the backbone when we train and evaluate on CIFAR-100 LT dataset. We follow common learning rate schedule following [14] and [27] for 200 epochs training with learning rate decay by a factor of 0.1 at the 160th and 180th epoch. We also provide results of 400 epochs training following recent conventions [20], [21], [57] with learning rate decay by a factor of 0.1 at the 320th and 360th epoch. We use Autoaug [59] along with random horizontal flip.

iNaturalist2018: iNaturalist2018 is acquired as described in [7]. This data is from the natural world, and naturally, the number of training data sample shows long-tailed characteristics. We conduct experiment using both ResNet-50 [55]. Following [14], we have trained our model for 200 epochs with learning rate decay by a factor of 0.1 at 75th and 160th epoch. We use Randaug [58] with random horizontal flip.

B. EVALUATION METRIC

To provide a quantitative comparison, we report top-1 classification accuracy in percent(%) following conventions [20], [27], [28]. Following [2], we report the accuracy for

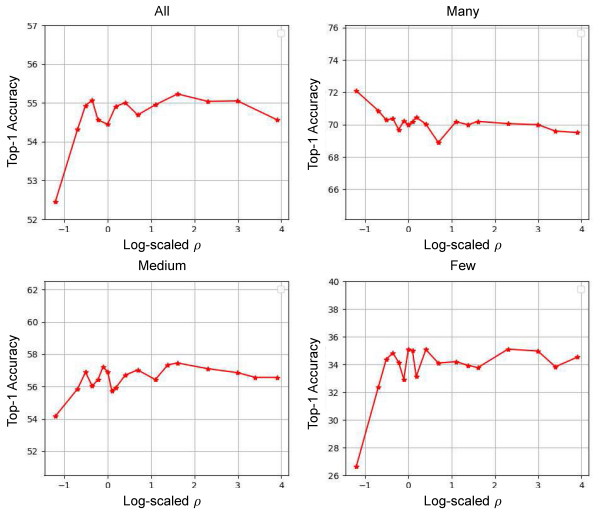

FIGURE 3. Accuracy graph by changing the value of ρ (x-axis, log-scale with base 2). All models are trained for 200 epochs at CIFAR = 100LT (IF = 100).

three disjoint class subsets: Many classes with more than 100 training samples, Medium classes with 20 to 100 training samples, and Few classes with less than 20 samples. For CIFAR-100 LT [34], we can control the imbalance factor (IF), defined as the ratio of the number of training samples of the largest class to that of the smallest class. We design the number of training samples of i-th class n_i ($i \in \{0, 1, \ldots, 99\}$) of CIFAR-100 LT following [34]. In [34], $n_i = 500\mu^i$, where $\mu = (\text{IF})^{-1/99}$. The number of samples of each class in the test dataset remains the same as the CIFAR-100 data. We report the results for commonly-used three IFs (IF = 100, IF = 50, and IF = 10) for CIFAR-100 LT dataset.

C. ABLATION STUDIES

1) STABILITY AND SCABILITY IN NUMBER OF EXPERTS

To verify the stability and scalability of the proposed CRAL, we compare its performance by increasing the number of experts in the model, as shown in Figure 2 and Table 1. For quantitative comparison, we compare the performance with other methods (RIDE [27], TLC [28], and NCL [20]), that provieds official implementation and that can increase the number of experts of the model. For fair comparison, all results in Table 1 are reproduced. They are all trained using CIFAR-100LT(IF = 100) [5], [6] under the same augmentation strategy and training schemes. As pointed out in Sec I and III, the performance of [27], a method of maximizing D_{KL} among experts, fluctuates as the number of experts increases. Additionally, in the case of TLC [28], when the number of experts was 10 and 15, training was not possible due to the exploding gradient despite multiple trials. NCL [20], which minimizes divergence between experts, shows saturation of performance even if the number of experts is increased. However, the performance of the proposed CRAL can be seen improving as the number of experts increases. Through this, the stability of the proposed

TABLE 3. Comparisons with state-of-the-art methods on ImageNet-LT [2]. † denotes results reproduced using the identical environment, data augmentation strategy, and training schedule by us based on their official implementations. * indicates that mixup-based augmentation such as [67] is used. ** uses the test dataset during the training phase. ^{e} indicates that it trained for shorter epochs. ‡ means we re-calculate 'All' performance using 'Many, Medium, Few' performances which is copied from its original paper, since they are slightly inconsistent. Bold text indicates best results among the same training scheme. '-' indicates the original paper has not released the result of corresponding experiment.

Methods	Published	ResNet-50 Backbone				ResNext-50 Backbone			
		All	Many	Medium	Few	All	Many	Medium	Few
CC-SAM [60]	CVPR 2023	52.4	61.4	49.5	37.1	55.5‡	63.1	53.4	41.1
GLMC*e [61]	CVPR 2023	-	-	-	-	56.2‡	70.1	52.4	30.4
SBCL [18]	ICCV 2023	53.4	63.8	51.3	31.2	-	-	-	-
ACE* (3 experts) [24]	ICCV 2021	54.7	-	-	-	56.6	-	-	-
GLMC*^{e}+MaxNorm [61]	CVPR 2023	-	-	-	-	56.7	60.8	55.9	**45.5**
GLMC*^{e}*+BS [61]	CVPR 2023	-	-	-	-	57.2	64.8	55.7	42.2
CR*+RIDE (4 experts) [62]	CVPR 2023	-	-	-	-	57.6‡	68.5	54.2	38.8
FCC+NCL [63]	CVPR 2023	-	-	-	-	60.5	-	-	-
RIDE (3 experts)† [27]	ICLR 2021	52.7	61.7	50.5	34.9	57.3	67.9	54.7	36.1
RIDE (4 experts)† [27]	ICLR 2021	53.7	62.9	51.4	35.8	58.2	68.8	55.7	35.9
FCC+SADE** [63]	CVPR 2023	55.7	-	-	-	-	-	-	-
TLC (3 experts)† [28]	CVPR 2022	56.2	62.5	53.7	**42.4**	-	-	-	-
TLC (4 experts)† [28]	CVPR 2022	54.9	61.5	53.3	41.9	-	-	-	-
SBCL+RIDE (3 experts) [18]	ICCV 2023	56.8	69.2	52.4	36.9	-	-	-	-
MDCS [57]	ICCV 2023	59.3	-	-	-	60.2	-	-	-
AREA^e [64]	ICCV 2023	59.5	-	-	-	-	-	-	-
SHIKE (3 experts) [65]	CVPR 2023	59.7	-	-	-	59.6	-	-	-
BalPoE [66]	CVPR 2023	59.7	-	-	-	**61.6**	-	-	-
CRAL (3 experts)	-	59.7	69.6	57.1	40.0	60.9	70.6	58.8	41.0
CRAL (4 experts)	-	**60.3**	**70.1**	**58.0**	40.4	61.3	**71.8**	**58.5**	41.2
Longer Training Epochs									
Paco+GML [23]	CVPR 2023	-	-	-	-	55.6	-	-	-
Paco+GML (Ensemble) [23]	CVPR 2023	-	-	-	-	57.2	-	-	-
Mutual (3 experts [29])	WACV 2023	59.4	-	-	-	59.5‡	70.2	56.7	39.1
NCL (3 experts) [20]	CVPR 2022	59.5	-	-	-	60.5	-	-	-
LGLA [30]	ICCV 2023	59.7	-	-	-	60.9	-	-	-
MDCS [57]	ICCV 2023	60.7	-	-	-	61.8	-	-	-
BalPoE [66]	CVPR 2023	60.8	-	-	-	62.0	-	-	-
CRAL (3 experts)	-	60.7	70.6	58.5	40.3	61.2	72.3	58.0	40.6
CRAL (4 experts)	-	**60.9**	**71.9**	**58.6**	**40.4**	**62.3**	**73.4**	**59.4**	**41.4**

CRAL and the scalability of the number of experts can be empirically verified.

2) EFFECTIVENESS OF COMPONENTS

We verify the effectiveness of CRAL's each component: $\mathcal{L}_{NERR}$ (NERR) and Class Prior Adjustment (CPA) through an ablation study. As shown in Table 2, NERR enhances the performance of multiple experts for All, Many, Medium, and Few cases. Moreover, the addition of CPA further enhances overall accuracy, compensating the disparity of distributions between training and test dataset.

3) EFFECT OF RISK AVERSION COEFFICIENT ρ

To observe changes in CRAL according to risk aversion ρ, we have trained models changing ρ on a log scale from $\rho = 0.3$ to $\rho = 100$, as shown in Figure 3. For $\rho > 1$, if we assign large value to ρ the performance of CRAL shows saturation from some point. This is because risk-aversion behavior results in less diversity from the consensus. As we have mentioned in Sec. III-B, the performance is unstable when $\rho < 1$, since we cannot guarantee the $\mathcal{L}_{NERR}(P_i, f; \rho)$ is divergence metric. The performance drops most in Few classes with unstable ρ as predicted by **Theorem I**. Therefore, risk-taking behavior with limited number of training samples is not recommended. Meanwhile, we find maximum of overall accuracy among the trained models at $\rho = 5.0$.

D. COMPARISON WITH STATE-OF-THE-ARTS

1) EVALUATION ON ImageNet-LT

Table 3 presents the performance evaluations on ImageNet-LT [2] along with state-of-the-art methods published in 2023 along with representative methods with multiple expert ensemble. For fair comparison, we only collect results that use RandAug [58], and gathers them by their backbone(ResNet-50 [55] and ResNext-50 [56]), or that uses other augmentation such as [67] which is not adopted by others if not specified. We further presents results of CRAL trained for 400 epochs training scheme following [20], [21], [29], and [30]. We compare state-of-the-art methods that uses the same training epochs. The proposed CRAL outperforms state-of-the-art methods in accuracy of all, many, and medium when trained for 400 epochs, while shows competitive performance when trained for 200 epochs.

2) EVALUATION ON CIFAR-100 LT

Table 4 displays the performance evaluation on CIFAR-100LT [6] along with state-of-the-art methods published in 2023, and representative multiple expert ensemble methods. For fair comparison, we only include methods that utilize

TABLE 4. Comparisons with state-of-the-art methods on CIFAR100-LT [5], [6]. † denotes reproduced results with the identical environment, data augmentation, and training schedule to us. * indicates that mixup-based augmentation such as [67] has been employed. ** indicates that the method use test dataset during training phase. bold text indicates best results, and underlined text indicates the second best results. '-' indicates the original paper has not released the result of corresponding experiment.

Methods	Published	200 Epochs			400 Epochs		
		IF=100	IF=50	IF=10	IF=100	IF=50	IF=10
FCC [63]	CVPR 2023	41.1	45.2	-	-	-	-
ACE* (3 Experts) [24]	ICCV 2021	-	-	-	49.4	50.7	-
ACE* (4 Experts) [24]	ICCV 2021	-	-	-	49.6	51.9	-
RIDE (3 Experts) [27] †	ICLR 2021	47.2	49.9	54.0	50.2	52.4	56.3
RIDE (4 Experts) [27] †	ICLR 2021	48.1	50.7	54.5	50.4	52.6	57.2
TLC (4 Experts) [28] †	CVPR 2022	48.2	-	-	45.9	-	-
TLC (3 Experts) [28] †	CVPR 2022	48.8	-	-	46.7	-	-
FCC [63]+SADE** [26]	CVPR 2023	49.4	-	-	-	-	-
GML [23]+PaCo [21]	CVPR 2023	-	-	-	49.8	-	-
GML [23]+PaCo [21] (Ensembled)	CVPR 2023	-	-	-	50.5	-	-
ResLT* (3 Experts) [68]	TPAMI 2022	49.7	54.5	63.7	-	-	-
CR [62]+RIDE [27] (4 Experts)*	CVPR 2023	49.8	<u>59.8</u>	59.5	-	-	-
CC-SAM [60]	CVPR 2023	50.8	53.9	-	-	-	-
NCL (3 Experts) [20]†	CVPR 2022	52.1	-	-	54.2	58.2	-
AREA [64]+PaCo [21]	ICCV 2023	-	-	-	52.4	56.6	65.1
AREA [64]+NCL [20]	ICCV 2023	-	-	-	55.2	58.7	65.8
MDCS [57]	ICCV 2023	53.2	57.2	-	<u>56.1</u>	60.1	-
NCL (4 Experts) [20]†	CVPR 2022	53.6	-	-	55.1	-	-
FCC [63]+NCL (3 Experts) [20]	CVPR 2023	54.5	58.4	-	-	-	-
BalPoE [66]	CVPR 2023	54.7	58.7	66.3	55.9	60.1	68.1
CRAL (3 Experts)	-	<u>55.2</u>	59.3	<u>67.6</u>	<u>56.1</u>	<u>60.3</u>	<u>68.3</u>
CRAL (4 Experts)	-	**55.7**	**60.0**	**68.2**	**56.7**	**61.2**	**69.1**

TABLE 5. Comparisons with state-of-the-art methods on iNaturalist2018 [7]. * indicates that mixup-based augmentation such as [67] is used. ** indicates trained for 400 epochs, while others are trained 200 epochs. ‡ stands for that we re-calculate 'All' performance using 'Many, Medium, Few' performances that is copied from its original paper, since they are slightly inconsistent. '-' indicates the original paper has not released the result of corresponding experiment.

Method	All	Many	Med	Few
AREA [64]	68.4	-	-	-
SBCL [18]	70.7‡	73.3	71.9	68.6
CC-SAM [60]	70.8‡	65.4	70.9	72.2
SuperDisco [69]	72.2‡	72.3	72.9	71.3
ResLT* (3 experts) [68]	72.9	73.0	72.6	73.1
ACE* (3 experts) [24]	72.9	-	-	-
RIDE (4 experts) [27]	73.1	-	-	-
BalPoE [66]	73.5	-	-	-
CR+RIDE (4 experts) [62]	73.7‡	71.0	73.8	74.3
NCL (3 experts)** [20]	74.9	72.7	75.6	74.5
Mutual (3 experts) [29]	74.9	-	-	-
SHIKE (3 experts) [65]	75.4	-	-	-
CRAL (3 experts)	75.4	74.3	75.7	75.0
CRAL (4 experts)	**75.7**	**75.0**	**76.4**	**75.1**

the same data augmentation, and they are grouped based on their training epochs, either 200 or 400 in Table 4. For some representative multiple experts ensemble methods [20], [27], [28] that do not report their performance with Autoaug [59] or both 200 and 400 epochs training, we have reproduced the results and marked the methods as † in Table 4. The proposed CRAL outperforms in all IF = 100, IF = 50, and IF = 10 for both trained with 200 epochs and 400 epochs, using $n = 4$.

3) EVALUATION ON INATURALIST2018

Table 5 presents the performances on iNaturalist 2018 [7] with state-of-the-art methods. For fair comparison, we only collect models whose backbone is ResNet50 [55] and trained for 200 epochs except NCL [20]. The proposed CRAL shows competitive performance in the dataset. The proposed CRAL outperforms state-of-the-art methods. CRAL with 4 experts outperforms state-of-the-art methods in accuracy of All, Many, Medium, and Few. Also, CRAL with 3 experts show competitive result to the [65], the best methods among the state-of-the-art methods.

V. CONCLUSION

In this paper, we proposed Consensus and Risk Aversion Learning (CRAL) scheme to train ensemble of multiple experts. Unlike former methods, CRAL obtains the consensus that aggregates the predictions of experts first. Given the consensus, CRAL maximizes expected prediction accuracy of each expert following the risk aversion behavior implemented by the proposed Negative Expected Rate of Return (NERR) loss. We provided in-depth analysis on the behavior and mathematical stability of the NERR loss with the consensus function defined in the paper. Furthermore, we applied Class Prior Adjustment that makes NERR to use compensated logits. Through ablation studies, we verified the effectiveness of CRAL and its scalability in the number of experts. Furthermore, CRAL outperformed state-of-the-art methods on popular long-tailed classification benchmarks.

Limitation: Though we proposed scalable multiple experts ensemble, increasing number of experts still requires more computational resource to train.

VI. FUTURE WORKS

Design of Consensus Function We have used average of the experts' predictions as consensus, but it can be improved further. For example, we may use weighted average instead

of simple average because Theorem II still holds for the weighted average. However, the method of determining weights requires an elaborate design, so we leave this as future work.

Designing Adaptive ρ: We have kept the value of risk aversion factor ρ as constant throughout all experts throughout the training. We expect that designing ρ to adaptively change depending on each expert's accuracy during the training can improve performance rather than using a fixed, unified value. We leave it to future work.

Application to Other Domains: Though our method designed to solve long-tailed image classification problem, it can be applied to other domain such as event-triggered control with multiple experts [70], cross-domain semi-supervised classification in remote sensing images [71], or hyperspectral image classification [72], and so on.

REFERENCES

[1] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.

[2] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu, "Large-scale long-tailed recognition in an open world," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2532–2541.

[3] V. Pareto, *Cours D'économie Politique*, vol. 1. Geneva, Switzerland, 1964.

[4] Y. Zhang, B. Kang, B. Hooi, S. Yan, and J. Feng, "Deep long-tailed learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10795–10816, Sep. 2023.

[5] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Univ. Toronto, Toronto, ON, Canada, Tech. Rep. TR-2009, 2009.

[6] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.

[7] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie, "The iNaturalist species classification and detection dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8769–8778.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[9] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Trans. Syst., Man, Cybern., B*, vol. 39, no. 2, pp. 539–550, Apr. 2009.

[10] A. Estabrooks, T. Jo, and N. Japkowicz, "A multiple resampling method for learning from imbalanced data sets," *Comput. Intell.*, vol. 20, no. 1, pp. 18–36, Feb. 2004.

[11] A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, and S. Kumar, "Long-tail learning via logit adjustment," 2020, *arXiv:2007.07314*.

[12] M. Li, Y.-M. Cheung, and J. Jiang, "Feature-balanced loss for long-tailed visual recognition," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2022, pp. 4175–4186.

[13] S. Park, J. Lim, Y. Jeon, and J. Y. Choi, "Influence-balanced loss for imbalanced visual classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 715–724.

[14] S. Park, Y. Hong, B. Heo, S. Yun, and J. Y. Choi, "The majority can help the minority: Context-rich minority oversampling for long-tailed classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 6877–6886.

[15] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker, "Feature transfer learning for face recognition with under-represented data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5697–5706.

[16] J. Kim, J. Jeong, and J. Shin, "m2m: Imbalanced classification via major-to-minor translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13893–13902.

[17] J. Wang, T. Lukasiewicz, X. Hu, J. Cai, and Z. Xu, "RSG: A simple but effective module for learning imbalanced datasets," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3783–3792.

[18] C. Hou, J. Zhang, H. Wang, and T. Zhou, "Subclass-balancing contrastive learning for long-tailed recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 5395–5407.

[19] J. Zhu, Z. Wang, J. Chen, Y. P. Chen, and Y.-G. Jiang, "Balanced contrastive learning for long-tailed visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 6898–6907.

[20] J. Li, Z. Tan, J. Wan, Z. Lei, and G. Guo, "Nested collaborative learning for long-tailed visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 6939–6948.

[21] J. Cui, Z. Zhong, S. Liu, B. Yu, and J. Jia, "Parametric contrastive learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 695–704.

[22] B. Zhou, Q. Cui, X.-S. Wei, and Z.-M. Chen, "BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9716–9725.

[23] Y. Du and J. Wu, "No one left behind: Improving the worst categories in long-tailed learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 15804–15813.

[24] J. Cai, Y. Wang, and J.-N. Hwang, "Ace: Ally complementary experts for solving long-tailed recognition in one-shot," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 112–121.

[25] L. Xiang, G. Ding, and J. Han, "Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification," in *Computer Vision—ECCV*. Glasgow, U.K.: Springer, 2020, pp. 247–263.

[26] Y. Zhang, B. Hooi, L. Hong, and J. Feng, "Self-supervised aggregation of diverse experts for test-agnostic long-tailed recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 34077–34090.

[27] X. Wang, L. Lian, Z. Miao, Z. Liu, and S. X. Yu, "Long-tailed recognition by routing diverse distribution-aware experts," 2020, *arXiv:2010.01809*.

[28] B. Li, Z. Han, H. Li, H. Fu, and C. Zhang, "Trustworthy long-tailed classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 6960–6969.

[29] C. Park, J. Yim, and E. Jun, "Mutual learning for long-tailed recognition," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 2674–2683.

[30] Y. Tao, J. Sun, H. Yang, L. Chen, X. Wang, W. Yang, D. Du, and M. Zheng, "Local and global logit adjustments for long-tailed learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 11783–11792.

[31] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.

[32] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4320–4328.

[33] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Symp. Math. Statist. Probab., Contrib. Theory Statist.*, vol. 4. Berkeley, CA, USA: Univ. California Press, 1961, pp. 547–562.

[34] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9260–9269.

[35] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.

[36] Y. Hong, S. Han, K. Choi, S. Seo, B. Kim, and B. Chang, "Disentangling label distribution for long-tailed visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6622–6632.

[37] J. Liu, Y. Sun, C. Han, Z. Dou, and W. Li, "Deep representation learning on long-tailed data: A learnable embedding augmentation perspective," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2967–2976.

[38] T. Wang, Y. Zhu, Y. Chen, C. Zhao, B. Yu, J. Wang, and M. Tang, "C2AM loss: Chasing a better decision boundary for long-tail object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 6970–6979.

[39] J. L. Kelly, "A new interpretation of information rate," *Bell Syst. Tech. J.*, vol. 35, no. 4, pp. 917–926, Jul. 1956.

[40] A. N. Soklakov, "Economics of disagreement—Financial intuition for the Rényi divergence," *Entropy*, vol. 22, no. 8, p. 860, Aug. 2020.

[41] J. Werner, "Risk aversion," in *The New Palgrave Dictionary of Economics*. London, U.K.: Palgrave MacMillan, 2008.

[42] K. J. Arrow, "Aspects of the theory of risk-bearing," Cowles Found. Res. Econ., Yale Univ., New Haven, CT, USA, 1965.

[43] J. W. Pratt, "Risk aversion in the small and in the large," in *Uncertainty in Economics*. Amsterdam, The Netherlands: Elsevier, 1978, pp. 59–79.
[44] Y. Ji and Y. Ma, "The robust maximum expert consensus model with risk aversion," *Inf. Fusion*, vol. 99, Nov. 2023, Art. no. 101866.
[45] H. Zhang, Y. Ji, S. Qu, H. Li, and R. Huang, "The robust minimum cost consensus model with risk aversion," *Inf. Sci.*, vol. 587, pp. 283–299, Mar. 2022.
[46] J.-B. Regli and R. Silva, "Alpha-beta divergence for variational inference," 2018, *arXiv:1805.01045*.
[47] A. N. Soklakov, "Elasticity theory of structuring," 2013, *arXiv:1304.7535*.
[48] C. P. Simon and L. Blume, *Mathematics for Economists*, vol. 7. New York, NY, USA: Norton, 1994.
[49] D. M. Endres and J. E. Schindelin, "A new metric for probability distributions," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1858–1860, Jul. 2003.
[50] T. Van Erven and P. Harremos, "Rényi divergence and Kullback–Leibler divergence," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 3797–3820, Jul. 2014.
[51] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Trans. Inf. Theory*, vol. 37, no. 1, pp. 145–151, Jan. 1991.
[52] T. Hartnett, *Consensus-Oriented Decision-Making: The CODM Model for Facilitating Groups to Widespread Agreement*. Gabriola Island, BC, Canada: New Society, 2011.
[53] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *Proc. NIPS*, 2017, pp. 1–4.
[54] N. Qian, "On the momentum term in gradient descent learning algorithms," *Neural Netw.*, vol. 12, no. 1, pp. 145–151, Jan. 1999.
[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[56] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5987–5995.
[57] Q. Zhao, C. Jiang, W. Hu, F. Zhang, and J. Liu, "MDCS: More diverse experts with consistency self-distillation for long-tailed recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 11597–11608.
[58] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 3008–3017.
[59] E. D. Cubuk, B. Zoph, D. Mané, V. Vasudevan, and Q. V. Le, "AutoAugment: Learning augmentation strategies from data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 113–123.
[60] Z. Zhou, L. Li, P. Zhao, P.-A. Heng, and W. Gong, "Class-conditional sharpness-aware minimization for deep long-tailed recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 3499–3509.
[61] F. Du, P. Yang, Q. Jia, F. Nan, X. Chen, and Y. Yang, "Global and local mixture consistency cumulative learning for long-tailed visual recognitions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 15814–15823.
[62] Y. Ma, L. Jiao, F. Liu, S. Yang, X. Liu, and L. Li, "Curvature-balanced feature manifold learning for long-tailed classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 15824–15835.
[63] J. Li, Z. Meng, D. Shi, R. Song, X. Diao, J. Wang, and H. Xu, "FCC: Feature clusters compression for long-tailed visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 24080–24089.
[64] X. Chen, Y. Zhou, D. Wu, C. Yang, B. Li, Q. Hu, and W. Wang, "AREA: Adaptive reweighting via effective area for long-tailed classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 19277–19287.
[65] Y. Jin, M. Li, Y. Lu, Y.-M. Cheung, and H. Wang, "Long-tailed visual recognition via self-heterogeneous integration with knowledge excavation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 23695–23704.
[66] E. S. Aimar, A. Jonnarth, M. Felsberg, and M. Kuhlmann, "Balanced product of calibrated experts for long-tailed recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 19967–19977.
[67] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," 2017, *arXiv:1710.09412*.
[68] J. Cui, S. Liu, Z. Tian, Z. Zhong, and J. Jia, "ResLT: Residual learning for long-tailed recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3695–3706, Mar. 2023.
[69] Y. Du, J. Shen, X. Zhen, and C. G. M. Snoek, "SuperDisco: Super-class discovery improves visual recognition for the long-tail," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 19944–19954.
[70] N. Zhang, Q. Sun, L. Yang, and Y. Li, "Event-triggered distributed hybrid control scheme for the integrated energy system," *IEEE Trans. Ind. Informat.*, vol. 18, no. 2, pp. 835–846, Feb. 2022.
[71] W. Teng, N. Wang, H. Shi, Y. Liu, and J. Wang, "Classifier-constrained deep adversarial domain adaptation for cross-domain semisupervised classification in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 5, pp. 789–793, May 2020.
[72] C.-I. Chang, "Statistical detection theory approach to hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 4, pp. 2057–2074, Apr. 2019.

TAEGIL HA received the B.S. degree in electrical and computer engineering from Seoul National University, Seoul, South Korea, in 2017, where he is currently pursuing the M.S./Ph.D. degree in electrical and computer engineering. His research interests include object tracking and long-tailed visual recognition.

JIN YOUNG CHOI (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in control and instrumentation engineering from Seoul National University, Seoul, South Korea, in 1982, 1984, and 1993, respectively. From 1984 to 1989, he was with the Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea, where he was involved in the Project of Switching Systems. From 1992 to 1994, he was with the Basic Research Department, ETRI, where he was a Senior Member of the Technical Staff involved in the neural information processing systems. From 1998 to 1999, he was a Visiting Professor with the University of California at Riverside, Riverside, CA, USA. Since 1994, he has been with Seoul National University, where he is currently a Professor with the School of Electrical Engineering. He is also with the Automation and Systems Research Institute, the Engineering Research Center for Advanced Control and Instrumentation, and the Automatic Control Research Center, Seoul National University. His current research interests include adaptive and learning systems, visual surveillance, motion pattern analysis, object detection, object tracking, and pattern recognition.