The Equalization Losses: Gradient-Driven Training for Long-tailed Object Recognition

Jingru Tan, Bo Li, Xin Lu, Yongqiang Yao, Fengwei Yu, Tong He, and Wanli Ouyang, Senior Member, IEEE

Abstract—Long-tail distribution is widely spread in real-world applications. Due to the extremely small ratio of instances, tail categories often show inferior accuracy. In this paper, we find such performance bottleneck is mainly caused by the imbalanced gradients, which can be categorized into two parts: (1) positive part, deriving from the samples of the same category, and (2) negative part, contributed by other categories. Based on comprehensive experiments, it is also observed that the gradient ratio of accumulated positives to negatives is a good indicator to measure how balanced a category is trained. Inspired by this, we come up with a gradient-driven training mechanism to tackle the long-tail problem: re-balancing the positive/negative gradients dynamically according to current accumulative gradients, with a unified goal of achieving balance gradient ratios. Taking advantage of the simple and flexible gradient mechanism, we introduce a new family of gradient-driven loss functions, namely equalization losses. We conduct extensive experiments on a wide spectrum of visual tasks, including two-stage/single-stage long-tailed object detection (LVIS), long-tailed image classification (ImageNet-LT, Places-LT, iNaturalist), and long-tailed semantic segmentation (ADE20 K). Our method consistently outperforms the baseline models, demonstrating the effectiveness and generalization ability of the proposed equalization losses.

Index Terms—Image classification, long-tailed object recognition, object detection, semantic segmentation.

I. INTRODUCTION

Object recognition is one of the most fundamental tasks in computer vision. It is an important step in a host of visual challenges, including object detection, semantic segmentation, and object tracking. Despite this fact, the task remains an open problem, not least due to the discrepancy among the proportions of different categories. Current benchmarks such as ImageNet [1], PASCAL VOC [2], COCO [3], and Cityscapes [4] are carefully collected with balanced annotations for each category, which contradicts the long-tailed Zipfian distribution in natural images. Although existing methods have achieved impressive results, we still can observe performance bottlenecks [5], [6], [7] on various benchmarks, especially in the non-dominant classes with fewer samples. As substantiated by recent literature [8], [9], [10], tail categories are easily overwhelmed by the head categories while learning on a dataset distributed off balance and diversely.

Previous approaches can be roughly categorised into two groups: data resampling [6], [11], [12] and cost-sensitive learning [13], [14]. These methods address the above problems by either designing complex sampling strategies or adjusting loss weights. Although promising, most of these methods are designed based on categories’ frequency and often suffer from several drawbacks: (1) those frequency-based methods are not robust enough due to widespread easy negative samples [15] and redundant positive samples [14]. (2) The accuracy is sensitive to the predefined hyper-parameters.

In this paper, we tackle the problem of long-tailed recognition from a novel perspective. We start by analyzing the distribution of the accumulated gradients across different categories. Specifically, the gradients of one category consist of two parts: (1) the positive part, derived from the samples of the same category, and (2) the negative part, contributed by other categories. Since the tail categories have limited positive samples, their positive gradients can be easily overwhelmed by the negative part. As illustrated in Fig. 1 (bottom row), the gradient ratio of positive to negative examples is distributed off balance when trained on long-tailed datasets such as LVIS [6], ImageNet-LT [5], and ADE20K-LT [7]. The gradient ratio is close to 1 for the head categories but is close to 0 for the tail categories. We hypothesize that such gradient imbalance in training is the main obstacle impeding tail classes from obtaining satisfactory performance. Besides, we also conduct experiments on the well-balanced datasets such as COCO [3], ImageNet [1], and ADE20K [7]. It can be observed that all categories have a gradient ratio close to 1 without introducing any bias toward positives or negatives. Therefore, we believe that the gradient ratio can serve as a significant indicator of how balanced a category is trained.

Such a gradient-based indicator provides useful guidance to adjust gradients of the positive and negative parts, which can be...
Fig. 1. Gradient observation on four different types of tasks. Each column is responsible for a specific task. We define the gradient of a sample as the derivative of the loss function with respect to its network output logits. For each category, we demonstrate its instance number (row 1) and the accumulated gradients (when training is finished) of positive samples (row 2), negative samples (row 3), and the ratio of positive samples to negative samples (row 4). We sorted the category index in decreasing order of its instance number. And, we align COCO 80 categories with LVIS 1203 categories. The left and right y-axes are for long-tailed (e.g., LVIS, ImageNet-LT, ADE20K-LT) and balanced datasets (e.g., COCO, ImageNet, ADE20K), respectively.

easily plugged into different classifiers. To this end, we adapt it to various loss functions. (1) For binary cross-entropy loss (BCE), we introduce a gradient-driven re-weighting mechanism and propose the sigmoid equalization loss (Sigmoid-EQL). It treats the overall classification problem as a set of independent binary classification tasks. Then the accumulative gradient ratio is used to up-weight the positive gradients and down-weight the negative gradients accordingly, aiming to balance the gradients of the two parts. (2) For cross-entropy loss (CE), we propose the softmax equalization loss (Softmax-EQL), which calibrates the decision boundary dynamically based on the statistics of the gradients. (3) For focal loss (FL) [15], we come up with the equalized focal loss (EFL) by decoupling the coefficients in [15] into category-agnostic and category-specific parts. By introducing the gradient into the category-specific parts, the model is able to focus more on the learning of rare categories. Those losses do not rely on the pre-computed data statistics to determine the rebalancing terms. Instead, they control the training process in a dynamic way. This data-distribution agnostic property makes them more suitable for streaming and realistic data.

To demonstrate the effectiveness of our proposed method, we conducted comprehensive experiments on various datasets and tasks. For object detection on the challenging LVIS [6] benchmark, our proposed Sigmoid-EQL and Softmax-EQL outperform Mask R-CNN [16] by about 6.4% and 5.7% in terms of AP, respectively. Without introducing extra computation overhead, our approach improves the performance substantially. With the help of equalization losses, we won the first place both in COCO-LVIS challenge 2019 and 2020 [17]. We also validate the effectiveness of our proposed EFL on the task of single-state object detection. Our method achieves 29.2% AP, delivering significant improvements over state-of-the-art results. In addition to the effectiveness, the equalization losses also show strong generalization ability when transferring to other datasets and visual tasks. For example, equalization losses maintain huge improvements when moving from LVIS to Openimages [18] without further hyper-parameter tuning. In Openimages, Sigmoid-EQL and EFL outperform the baseline CE method by 9.1% AP and 6.6% AP, respectively. For the image classification task, Softmax-EQL achieves superior results compared to sample number based methods on three long-tailed image classification datasets (ImageNet-LT, Place-LT, and iNaturalist2018). We also evaluate our method on semantic segmentation using ADE20K [7]. Our proposed Sigmoid-EQL improves the
powerful baseline, DeepLabV3+ [19], by 1.56% and 2.27% in terms of mIoU and mAcc, respectively, showing strong generalization of gradient-driven losses to varying tasks.

II. RELATED WORK

Long-Tailed Object Recognition: Common solutions for long-tailed image recognition are data re-sampling and loss re-weighting. Re-sampling methods under-sample the head categories [20], [21] or over-sample the tail categories [11], [12], [22], [23]. Re-weighting methods assign different weights to different categories [14], [24], [25], [26] or instances [15], [27], [28]. Decoupled training methods [29], [30] address the classifier imbalance problem with a two-stage training pipeline by decoupling the learning of representation and classifier. In addition, margin calibration [10], [13], [31], [32] inject category-specific margins into the CE loss to re-balance the logits distribution of categories. Recently, other works address the long-tailed problem from different perspectives such as transfer learning [5], [33], [34], supervised contrastive learning [35], [36], [37], [38], ensemble learning [39], [40], [41], [42], [43], and so on. Some works adapt those ideas to object detection [6], including data re-sampling [6], [44], loss re-weighting [8], [9], [45], decoupled training [30], [44], [46], margin calibration [10], [47], incremental learning [48] and causal inference [49]. Despite the efforts, most of them somehow utilize the sample number as the indicator of imbalance to design their algorithms. In contrast, we use the gradient as the indicator. It is more stable and precise thus reflects the models’ training status better. Another line of work involves retrieval-augmented methods, including those based on KNN and text descriptions [50] and attention mechanisms [51]. Such approaches have demonstrated very promising results and are playing an increasingly important role in the era of large-scale foundational models. Our method is orthogonal to theirs to some extent. For instance, our method can still be incorporated into their framework to further enhance performance.

Gradient as Indicator: There are some works [15], [27], that attempt to solve the imbalance problems from the gradient view. They use the instant gradient to reflect the learning difficulty of a sample at a certain moment and determine its loss contribution dynamically, which can be viewed as an online version of hard negative example mining [52]. Those methods are designed for the serious foreground-background imbalance problem. Different from them, we use accumulative gradients to reflect the imbalanced training status of categories. Our method is designed for long-tailed object recognition. Meanwhile, our method is complementary to theirs. We can solve the foreground-background imbalance problem and foreground-foreground (long-tailed) problem simultaneously by combining instant gradients and accumulative gradients indicators.

III. GRADIENT IMBALANCE PROBLEM

In this section, we introduce the imbalanced gradients, which we believe are responsible for the inferior performance of the tail categories. It comes from the entanglement of instances and categories, which we will describe next. Building upon comprehensive experiments, we argue that such gradient statistics can serve as an effective indicator to show the status of category classifiers.

Notation Definition: Suppose we have a training set $\mathcal{X} = \{x_i, y_i\}_N$ with $C$ categories. Let the total instance number over the dataset be $N$ and $N = \sum_j n_j$, where $n_j$ is the instance number of category $j$. For each iteration, we have a batch of instances $\mathcal{I}$ with a batch size of $B$. $\mathcal{Y} \in \mathbb{R}^{B \times C}$ represents the one-hot labels of the batch. We adopt a CNN $h$ with parameter $\theta$ as the feature extractor. Then the feature representations of the batch could be computed by $f(\mathcal{I}; \theta)$. A linear transformation is used as the classifier to output the logits: $Z \in \mathbb{R}^{B \times C}$, $Z = W^T h(\mathcal{I}; \theta) + b$, where $W$ denotes the classifier weight matrix and $b$ is the bias.

We denote each image as an instance. $W$ can be regarded as $C$ classifiers, each responsible for bi-categorizing instances as one class. Each instance can be regarded as a positive sample for one specific category and a negative sample for the remaining $C - 1$ categories. We denote $y_i^j \in \{0, 1\}$ as the label, which equals to 1 if the $i$-th instance belongs to the $j$-th category.

A. Entanglement of Instances and Categories

The total number of positive samples $M^{pos}_j$ and negative samples $M^{neg}_j$ for the $j$-th classifier can be easily obtained:

$$M^{pos}_j = \sum_{i \in \mathcal{X}} y_i^j, \quad M^{neg}_j = \sum_{i \in \mathcal{X}} (1 - y_i^j)$$

(1)

The ratio of the number of positive samples to the negative samples over the dataset is then:

$$\frac{M^{pos}_j}{M^{neg}_j} \propto \frac{n_j}{N - n_j} \propto \frac{1}{n_j - 1}$$

(2)

From (2), we observe that $M^{pos}_j \ll M^{neg}_j$ for the tail categories that have a very limited number of instances, indicating these categories often suffer from an extremely imbalanced ratio of positive to negative samples. Previous methods [13], [14] address the problem by applying different loss weights or decision margins to different categories according to their sample numbers. However, they often fail to generalize well to other datasets because the sample numbers cannot reflect the training status of each classifier well. For example, a large number of easy negative samples and some redundant positive samples hardly contribute to the learning of the model. In contrast, we propose to use gradient statistics as our metric to indicate whether a category is in balanced training status. We conjecture that there is a similar positive-negative imbalance problem in the gradient ratios of rare categories.

B. Gradient Computation

We define the gradient over the batch $\mathcal{I}$ as the derivative of objective cost function $L$ with respect to their logits $Z$. The gradient $g = \frac{\partial L}{\partial Z} \in \mathbb{R}^{B \times C}$ is corresponding to the gradients of all samples belonging to $C$ categories. We denote the gradient of a certain sample as $g_i^j$. Then the positive gradients $g(t)_j^{pos}$ and negative gradient $g(t)_j^{neg}$ of the category $j$ at iteration $t$ can
be computed as follows:

\[ g(t)_{j}^{\text{pos}} = \sum_{i \in I} y_{i}^{1} | g_{i}^{1} |, \quad g(t)_{j}^{\text{neg}} = \sum_{i \in I} (1 - y_{i}^{1}) | g_{i}^{1} | \]  

The accumulated positive gradients \( G(T)_{j}^{\text{pos}} \) and negative gradients \( G(T)_{j}^{\text{neg}} \) at iteration \( T \) could be defined as:

\[ G(T)_{j}^{\text{pos}} = \sum_{t=0}^{T} g(t)_{j}^{\text{pos}}, \quad G(T)_{j}^{\text{neg}} = \sum_{t=0}^{T} g(t)_{j}^{\text{neg}} \]  

For simplicity, we ignore \( T \) in the notations and directly adopt \( G_{j}^{\text{pos}} \) and \( G_{j}^{\text{neg}} \) as the accumulated positive and negative gradients, respectively. The accumulated gradient ratio could be calculated by \( G_{j} = \frac{G_{j}^{\text{pos}}}{G_{j}^{\text{neg}}} \).

C. Gradient Observation

To validate our hypothesis that rare categories suffer from gradient imbalance problems, we collect gradient statistics during the training process across a wide spectrum of recognition tasks and datasets, including long-tailed image classification (ImageNet-LT [5]), two-stage/single-stage long-tailed object detection (LVIS [6]), and long-tailed semantic segmentation (ADE20K [7]). The results are shown in Fig. 1. We consistently observe four key phenomena: (1) The positive gradients \( G_{j}^{\text{pos}} \) follow a long-tailed distribution. (2) The negative gradients \( G_{j}^{\text{neg}} \) follow a long-tailed distribution. (3) The gradient magnitudes of positive and negative over categories are different, leading their ratio \( G \) also have the property of long-tailed distribution. (4) The gradient ratio \( G \) of head categories is close to 1 while the ratio of tail categories is close to 0.

The x-axis of Fig. 1 is sorted by category instance numbers \( n_{j} \). We notice that the positive gradients have a positive correlation with the positive sample number, while the negative gradients do not have a positive correlation with the negative sample number. This is because tail categories are scarcely trained by the model. The model hardly predicts those tail categories to be positives, so most of their negative samples are easy samples. Although easy negative samples have small gradients, the effect accumulated from a large amount of them is not negligible. The observation of the gradient ratio proves that the category classifiers with fewer samples suffer from a more serious gradient imbalance problem between positives and negatives, which validates our conjecture. For the head categories with abundant training samples, the received gradient ratio of the corresponding classifier is close to 1, indicating the classifier gives no inclination to positives or negatives, which we refer to as a balanced training status. For the tail classifier, the gradient ratio is close to 0, indicating the classifier is heavily biased towards negative, which we refer to as an imbalanced training status. It is worth noting that there is a vast number of background negative samples in single-stage object detection, as discussed in [15]. We find that the head categories still have a gradient ratio close to 1 in a well-trained model, as shown in Fig. 1 (column 3). This proves that the gradient indicator is more stable and reliable than the sample number.

As illustrated in Fig. 1, more experiments are conducted on several datasets that have similar number of instances among different categories, including ImageNet [1] and COCO [3]. We observe that the gradients are balanced distributed and the category classifiers have a balanced gradient ratio closing to 1.

By observing the gradient statistics under imbalanced and balanced data distribution, we conclude that the gradient (i.e., \( G_{j}^{\text{pos}} \), \( G_{j}^{\text{neg}} \)) and the gradient ratio (i.e., \( G_{j} \)) could serve as an important indicator of the training status of categories.

IV. THE EQUALIZATION LOSSES

Our central idea is to utilize gradient statistics as an indicator to reflect the training status of category classifiers and then adjust their training process dynamically. In this section, by applying this idea to several loss functions, such as binary cross-entropy loss, cross-entropy loss, and focal loss, we introduce a new family of gradient-driven loss functions, namely the equalization losses.

A. Sigmoid Equalization Loss

1) Binary Cross-Entropy: Binary cross-entropy (BCE) loss estimates the probability of each category independently using \( C \) sigmoid loss functions. Specifically, in a batch of instances \( I \), the classifier outputs the estimated probability \( P \in \mathbb{R}^{B \times C} \) by applying a sigmoid activation function to the logits \( Z \) and \( P = \sigma(Z) \). We define the estimated probability of a certain sample as \( p_{i}^{j} \in [0, 1] \). For this sample, the loss term is computed as:

\[ \text{BCE}(p, y) = \begin{cases} -\log(p) & \text{if } y = 1 \\ -\log(1 - p) & \text{otherwise} \end{cases} \]  

Follow the notation in [15], we define \( p_{k} \) as:

\[ p_{k} = \begin{cases} p \quad & \text{if } y = 1 \\ 1 - p & \text{otherwise} \end{cases} \]  

then we can rewrite \( \text{BCE}(p, y) = \text{BCE}(p_{k}) = -\log(p_{k}) \). And the final loss contribution could be calculated by summing up the loss values from all samples:

\[ L(P, Y) = \sum_{i \in I} \sum_{j=1}^{C} \text{BCE}(p_{k}) \]  

The probability of each category in the BCE is estimated independently without cross normalization. This property makes the binary cross-entropy suitable for tasks that consist of a set of independent sub-tasks, such as object detection, and multi-label image classification.

2) Gradient-Driven Re-Weighting: Under long-tailed distribution, models are in an unbalanced training status. As mentioned in Section III-C, the accumulated gradient ratio \( G_{j} \) can reflect the training status of that category. Therefore we adopt it to adjust the training process for each sub-task in BCE independently and equally. Concretely, we propose a gradient-driven
re-weighting mechanism in which we up-weight the positive gradients and down-weight negative gradients for each classifier dynamically. This re-weighting strategy aims to make the gradient ratio as close to 1 as possible.

We denote \( q_j \) as the weight term for positive samples of category \( j \) and \( r_j \) for negative samples. We propose the formulation as:

\[
q_j = f(G_j) \\
r_j = 1 + \alpha(1 - r_j)
\]

where \( f(\cdot) \) is a mapping function that remaps the value of gradient ratio to a more controllable range, typically \([0, 1]\). Basically, for a small gradient ratio with imbalanced training status, we will have a small \( r_j \) for negative gradients and a big \( q_j \) for positive gradients to re-balance the training status of the current category. Simple mapping functions\(^2\) like linear mapping: \( f(x) = x \), exponential mapping: \( f(x) = x^2 \) or square root mapping: \( f(x) = \sqrt{x} \) could be chosen. More sophisticated mapping functions are also feasible, like the sigmoid-like mapping function: \( f(x) = \frac{1}{1 + e^{-\gamma(x-r)}} \).

We name this novel gradient re-weighting mechanism Sigmoid Equalization Loss (Sigmoid-EQL), and the formula of the loss is:

\[
L(\mathcal{P}, \mathcal{Y}) = \sum_{i \in \mathcal{Z}} \sum_{j=1}^{C} (q_j y_i^j + r_j (1 - y_i^j)) \text{BCE}(p_i)
\]

### B. Softmax Equalization Loss

1) **Cross-Entropy:** Cross-entropy (CE) loss is also a widely-used loss function in recognition tasks. The loss term of CE for a sample is:

\[
\text{CE}(p, y) = -y \log(p)
\]

Different from BCE, CE uses a cross normalization operation softmax to derive the estimated probability:

\[
p_i^j = \frac{e^{z_i^j}}{\sum_{k=1}^{C} e^{z_i^k}}
\]

It is worth noting that although the CE only calculates the loss of the positive sample for an instance, the gradient will flow back to logits of negative samples because of the softmax function. This operation introduces explicit competition between categories. So it is useful for tasks that require a single output category, such as image classification and semantic segmentation. During inference, for instance \( i \), we choose its predicted output category \( y_i^j \) by argmax operation:

\[
y_i^j = \text{argmax}_j p_i^j
\]

2) **Gradient-Based Margin Calibration:** In the situation where the tasks require a single output category, we can no longer treat category classifiers as independent sub-tasks as BCE does. The accurate ranking between categories is crucial, so the learning of all samples in an instance should be adjusted jointly. Under the long-tailed situation, the CE loss prefers to predict instances as head categories because they receive more gradients. The optimization direction is dominated by the head categories thus they occupy a broad area in the decision space.

As discussed in [32], the neural network models the output \( p \) as a posterior probability \( p(y|x) \) of \( y \) given image \( x \). According to Bayes’ theorem:

\[
p(y|x) \propto p(y)p(x|y)
\]

\( p(y) \) is called prior probability. For category \( j \), its prior probability is defined as \( p(y_j) \) which is proportional to the sample number \( n_j \) and the accumulative positive gradients \( G_j^{\text{pos}} \):

\[
p(y_j) \propto n_j \propto G_j^{\text{pos}}
\]

Obviously, head categories have a higher prior probability, so the model is inclined to predict head categories according to the decision function (14). This prediction preference harms the performance of tail categories.

To alleviate the influence of the prior probability, we propose gradient-based margin calibration, a strategy that injects accumulated gradient to conventional softmax so that model is able to be aware of observed data distribution. As a result, the model can output a fair prediction result by adopting the dynamical calibration strategy.

The probability of a sample in our gradient-based margin calibration method could be calculated by:

\[
p_i^j = \frac{(G_j^{\text{pos}})^\pi e^{z_i^j}}{\sum_{k=1}^{C} (G_k^{\text{pos}})^\pi e^{z_i^k}}
\]

Where \( \pi \in [0, \infty) \) is a hyperparameter that controls the degree of calibration. The larger \( \pi \) is, the stronger the effect of gradient calibration. When \( \pi = 0 \), the effect of gradient calibration is completely removed. It is worth noting that we apply the accumulative positive gradient \( G_j^{\text{pos}} \) as the gradient indicator because it could indicate the prior probability better than the gradient ratio \( G_j \). By replacing the conventional softmax function with our gradient-based margin calibration method, we propose the Softmax Equalization Loss (Softmax-EQL). Softmax-EQL shares a similar idea with Balance Meta-Softmax Loss [31], Logit Adjustment [32] Loss, and Seesaw Loss [10], but they use the sample number as the indicator. We will prove that the gradient indicator reflects the relation between categories more precisely and achieves better results.

### C. Equalized Focal Loss

In this section, we will demonstrate that our gradient indicator can be applied not only to the simple BCE and CE losses but also to more complicated losses (e.g., focal loss [15] and quality focal loss [53]) in more difficult tasks (single-stage long-tailed object detection). Single-stage detection has a simple and fast pipeline that is very prevalent in real-world applications. In contrast to two-stage methods that incorporate a region proposal network (RPN [54]) to filter out most background samples before feeding proposals to the final classifier, single-stage detectors directly detect objects over a regular, dense set of candidate locations.
Due to this dense prediction schema, the extreme foreground-background imbalance is introduced. This means that, under the long-tailed distribution, single-stage detectors have to solve the imbalance problem between foreground-background instances and between foreground categories’ instances simultaneously. 

1) Focal Loss: Focal loss [15] is a conventional solution to the foreground-background imbalance problem. Similar to definitions of BCE and CE, the formula of focal loss for a certain sample is:

$$\text{FL} (p_t) = -\alpha_t (1 - p_t)^\gamma \log (p_t)$$

(17)

As declared in [15], $\alpha_t$ is a balancing factor that balances the importance of positive and negative samples. The modulating factor $(1 - p_t)^\gamma$ is the key component of focal loss. It down-weights the loss of easy samples and focuses on the learning of hard samples by the predicted $p_t$ and the focusing parameter $\gamma$. Since most samples of background instances are easy to classify, focal loss greatly weakens the influence of these background instances, thus focusing more on the learning of foreground instances. It could be concluded from (17) that the larger the $\gamma$, the more focus on the learning of hard positive samples, which is suitable for more serious positive-negative imbalance problems.

2) Gradient Driven Modulating: However, focal loss fails to solve the foreground categories imbalance problem under the long-tailed data distribution (see Table IX). In long-tailed dataset (i.e., LVIS), as described in Section III-C, different categories suffer from different degrees of positive-negative gradient imbalance problems. Rare categories have more severe imbalance problems than frequent ones. Therefore, focal loss with the constant modulating factor may be not appropriate for all these imbalance problems. Based on these analysis, we propose Equalized Focal Loss (EFL) by designing a gradient-driven modulating mechanism for the focal loss. Concretely, we introduce two category-relevant factors, i.e., the focusing factor and the weighting factor. Those two factors are controlled by our gradient indicator to dynamically handle positive-negative imbalance problems of different degrees.

**Focusing Factor.** We assign large focusing factors to rare categories to alleviate their serious positive-negative gradient imbalance problems. For frequent categories with slight imbalance problem, a small focusing factor is proper. We formulate the loss of a sample belonging to the $j$-th category as:

$$\text{EFL} (p_t) = -\alpha_t (1 - p_t)^\gamma \log (p_t)$$

(18)

Specifically, we decouple the gradient-driven focusing factor $\gamma_j$ into two components: a categories-agnostic parameter $\gamma_b$ and a categories-specific parameter $\gamma'_b$:

$$\gamma_j = \gamma_b + \gamma'_b$$

$$= \gamma_b + s (1 - G_j)$$

(19)

where $\gamma_b$ controls the basic behavior of the loss. And the parameter $\gamma'_b \geq 0$ is a variable associated with the imbalance degree of the $j$-th category. We reflect the imbalance degree (training status) of the $j$-th category by the accumulated gradient ratio $G_j$. We directly apply the linear mapping function $1 - G_j$ to make the focusing factor keep a negative correlation to the imbalance degree. The hyper-parameter $s$ is a scaling factor that determines the upper limit of $\gamma_j$ in EFL. Compared with focal loss, EFL can handle the positive-negative imbalance problem of each category independently and dynamically.

**Weighting Factor:** Even with the focusing factor $\gamma_j^\gamma$, there is still an obstacle degrading the performance: For a binary classification task, a larger $\gamma_j$ is suitable for a more severe positive-negative imbalance problem. However, in the multi-class case (see (18)), for two samples of different classes that have the same logits, the larger the value of $\gamma_j$, the smaller the loss. It leads to the fact that when we want to increase the concentration on learning a category, we have to sacrifice part of its loss contribution in the overall training process. Such a dilemma prevents rare categories from achieving excellent performance. Basically, we expect rare hard samples to make more loss contributions than frequent hard ones.

We propose the weighting factor to alleviate the problem by re-weighting the loss contribution of different categories. Similar to the focusing factor, indicated by the gradient, the weighting factor is assigned a large value for rare categories to raise their loss contributions while keeping close to 1 for frequent categories. Specifically, we set the weighting factor of the $j$-th category as $w^j$ and the final formula of EFL is:

$$\text{EFL} (p_t) = -w^j (1 - p_t)^\gamma \log (p_t)$$

(20)

where $w^j$ shares the same variable $\gamma_b$ and $\gamma'_b$ as the focusing factor:

$$w^j = \frac{\gamma_b + \gamma'_b}{\gamma_b}$$

(21)

The focusing factor and the weighting factor make up the categories-relevant modulating factor in EFL. It enables the classifier to dynamically adjust the loss contribution of a sample based on our gradient indicator. We will show later in experiments that both the focusing factor and the weighting factor play significant roles in EFL. Meanwhile, in the balanced data distribution, all categories in EFL will have balanced training status with all $G_j$ close to 1. Then the $\gamma_j = \gamma_b$ makes the EFL equivalent to focal loss. Such an appealing property makes EFL could be applied well with various data distributions.

What’s more, the gradient-driven modulating factor is also applicable in other focal loss series such as the quality focal loss (QFL) [53]. The novel loss is denoted as the equalized quality focal loss (EQFL), and the formulated of a sample in EQFL of the $j$-th category is:

$$\text{EQFL}(p) = -m^j (y' \log(p) + (1 - y') \log (1 - p))$$

(22)

where $m^j$ is the specific form of the gradient-driven modulating factor in EQFL:

$$m^j = w^j (|y' - p|)^\gamma$$

(23)

The weighting factor $w^j$ and the focusing factor $\gamma_j$ are the same as in EFL. It should be noticed that $y' \in [0, 1]$ is the IoU score for a positive sample and 0 for a negative sample as declared in [53]. As we will show in Section V-B4, combined with the advanced YOLOX [55] detectors, our proposed EQFL
achieves impressive performance improvement in solving the long-tailed problem.

V. EXPERIMENTAL RESULTS

In this section, we conduct extensive experiments on the equalization losses. To show the versatility of the gradient indicator, we present experiments on four popular recognition tasks: two-stage long-tailed object detection, single-stage long-tailed object detection, long-tailed image classification, and long-tailed semantic segmentation. For a specific vision task, we first introduce the dataset setup and the evaluation metric. Then, we describe the important implementation details. Finally, we present the main results and analysis.

A. Two-Stage Long-Tailed Object Detection

Object detection requires detecting all possible objects within a set of pre-defined categories. The evaluation metric (AP) is evaluated class-wise. Therefore, we can model the detection task as a set of independent sub-tasks, each for one specific category, whose probability is estimated by a sigmoid loss function. In large-scale object detection, the concepts of categories are not always mutually exclusive but show the federated characteristic. Compared to the softmax loss function, the sigmoid loss function gives more flexible control of the relationship between categories. Thus, in this object detection task, we mainly study the Sigmoid-EQL.

1) Datasets and Evaluation Metric: LVIS [6] is a challenging benchmark for long-tailed object recognition. It provides precise bounding box and mask annotations for various categories with a long-tailed distribution. We mainly perform experiments on the recently released challenging LVIS v1.0 dataset. It consists of 1203 categories. We train our models on the train set, which contains about 100 k images with 1.3 M instances. In addition to widely-used metric AP across IoU threshold from 0.5 to 0.95, LVIS also reports AP_r (rare categories with 1-10 images), AP_c (common categories with 11-100 images), and AP_t (frequent categories with > 100 images). Since LVIS is a federated dataset, categories are not annotated exhaustively. Each image has two more types of labels: pos_category_ids and neg_category_ids, indicating which categories are or are not present in that image. Detection results that do not belong to those categories will be ignored for that image. We report results on the val set of 20 k images.

2) Implementation Details: Models are trained using SGD with a momentum of 0.9 and a weight decay of 0.0001. The ResNet [56] backbones are initialized by ImageNet pre-trained models. Following the convention, scale jitter and horizontal flipping are used in training and no test time augmentation is used. We use a total batch size of 16 on 16 GPUs (1 image per GPU) and set the initial learning rate to 0.02. According to [6], the maximum number of detection per image is up to 300, and the minimum score threshold is reduced to 0.0001 from 0.01. For Sigmoid-EQL, we initialize the bias of the last fully-connected classification layer (fc-cl) with values of 0.001 to stabilize the training at the beginning. We also add a branch for detecting objectiveness instead of the concrete category to reduce false positives. In the training phase, this branch treats all other categories’ positive samples as its positive samples. In the inference phase, the estimated probability of other categories become: \( p'_j = p_j * p_{obj} \). Where \( p_{obj} \) is the probability of a proposal being an object. Note that, the proposed gradient-driven mechanisms (re-weighting or calibration) are not applied in the objectiveness branch.

3) Main Results: We compare our method with several leading methods on LVIS v1.0 val. All models are trained in an end-to-end way with random sampler by a 2x schedule. The results are shown in Table I. On the R50 Mask R-CNN framework, the Sigmoid-EQL outperforms the CE loss and Grad-Blocking\(^3\) [8] by a large margin, 6.3% AP and 3.9% AP respectively. Note that RFS [6] repeats images that contain tail categories in each epoch, so it increases the total training time. Instead, our method only uses a random sampler and does not increase the training time, and achieves better results, 25.5% AP vs. 24.4% AP. Compared with the sample-number-based methods BAGS [46] and Seesaw Loss [10], Sigmoid-EQL outperforms them with a non-trivial improvement, 0.4% AP and 0.8% AP respectively. We also provide the result of Softmax-EQL and compare it with Seesaw Loss. Without the hard logit mining, i.e., compensation factor like in Seesaw Loss, our Softmax-EQL achieves better performance than it, demonstrating that the accumulated gradient ratio is a superior indicator than the accumulated sample number. Meanwhile, under a stronger backbone (e.g., ResNet-101), our proposed method still achieves higher overall AP across different algorithms.

4) Analysis: We conduct extensive experiments to provide a thorough analysis of sigmoid-EQL. We use a 1x training schedule with a random sampler if not mentioned.

Component Effect: The effect of each component in Sigmoid-EQL is shown in Table II. The baseline model performs poorly on the tail classes, resulting in 0% AP and 12.0% AP for rare and common categories. And the performance gaps between head and tail classes are very large. Adding an objectness branch helps all categories to some extent, improving the overall AP by 2.0% but not very much for the rare categories since the main problem for them is the unbalanced positive and negative gradients, i.e., their positive gradients are overwhelmed by negative gradients caused by a vast number of negative samples. By down-weighting the influence of negative gradients, their accuracy is boosted significantly (5.4% AP for rare categories). Further up-weighting the positive gradient helps to achieve a more balanced ratio of positive to negative gradients. It brings a 7.6% AP performance boost for rare categories. With these three components, we achieve a 23.7% AP, outperforming the baseline model 16.1% AP by a large margin without any re-sampling techniques. It is worth noting that only up-weighting positive gradients already achieved a significant improvement but the AP of tail categories is limited by the massive discouraging negative gradients. These ablation experiments verified the effectiveness of Sigmoid-EQL.

Mapping Functions: In Table III, we show the ablation study of the mapping function in Sigmoid-EQL. All these mapping functions improve the performances while sigmoid-like function

\(^3\)To avoid name conflict, we have renamed the original equalization loss [8] to Grad-Blocking.
\[ f(x) = \frac{1}{1 + e^{-\gamma(x - \mu)}} \]

stands out from the others. It has a value like an inflection point \( \mu \), which we can treat as a high enough gradient ratio. As the gradient ratio increases, this function first changes very slowly when the ratio is far smaller than the inflection point. Then it increases dramatically when the ratio gets close to the inflection point. And it slows down its changes again after becoming larger than the inflection point. The sigmoid-like mapping functions consistently achieve higher performance on both overall AP and \( AP_r \). By default, we choose the inflection point \( \mu \) as 0.8 and set \( \gamma = 12 \).

The Importance of Accumulation: In Table IV, we compare different gradient-based indicators with or without accumulation. When using the instant gradient as an indicator, the overall improvement is very small, only increasing from 18.1 to 19.1. Moreover, it can be observed that this inaccurate gradient significantly affects the performance of frequent categories, and limits the improvement of rare and common categories. This shows the importance of accumulated gradient, which can reflect the data distribution precisely and can serve as a reliable and stable indicator for long-tail problems.

Stability Under Longer Training: To verify the stability across different backbones and training schedules. We conduct experiments with larger models by a 3x schedule. The results are present in Table V. Note that training Mask R-CNN baseline (CE loss) with a longer schedule does not help rare categories a lot (Table I vs. Table V), the AP of rare categories is still bad because rare categories are heavily suppressed by the negative gradients caused by the entanglement of instances and tasks. In contrast, with the proposed Sigmoid-EQL, the performance can be further improved from 25.5% AP to 26.2% AP on the R50 backbone. We do not observe the over-fitting of tail classes in such a long schedule. Furthermore, when using a large ResNet-101 backbone, the gap between Mask R-CNN and Sigmoid-EQL still holds.

Comparison With Decoupled Training algorithms: We mainly compare our method with three decoupled training methods (cRT [29], LWS [29], and BAGS [46]). The results are present in Table VI. The decoupled training models are first initialized from...
naive softmax baseline (CE loss), then re-train their classifier layer (fc-tcls) for another 12 epochs with other layers frozen, resulting in a total 24 epoch training. Those decoupled training methods all improve the AP, mainly for tail classes. The improvements brought by LWS are limited. We conjecture it is because LWS only learns a scaling factor to adjust the decision boundary of the classifier but the classifier itself is not good and imbalanced. Our Sigmoid-EQL achieves a overall 23.7% AP, outperforming the LWS, cRT, and BAGS by 6.7%, 1.6% and 0.6%, respectively. It is worth noting that Sigmoid-EQL does not require an extra fine-tuning stage, and the representation and classifier are learned jointly.

Do We Have a More Balanced Gradient Ratio? We visualize the gradient ratio of our method Sigmoid-EQL and BCE baseline model during the training process, see Fig. 2. The baseline model does not have a balanced ratio for all categories. The positive gradients are overwhelmed by the negative gradients, especially for tail classes, which makes it hard to detect them. Training for a longer period does not help much. In contrast, Sigmoid-EQL preserves a more balanced gradient ratio throughout the entire training phase.

Do We Have a Better Representation? To evaluate the quality of representations trained with our method. We adopt models trained with the Sigmoid-EQL and the CE loss as the pre-trained models. Then, we follow the classic decoupled training recipe to re-train the classifier with frozen representations. The results are shown in Table VII. There are two main observations: Firstly, models initialized with Sigmoid-EQL always achieve a higher AP, resulting in 22.4% AP vs. 22.0% AP for cRT, 23.1% AP vs. 17.0% AP for LWS, 24.0% AP vs. 23.1% AP for BAGS. It verifies that we obtain a better representation by adopting Sigmoid-EQL compared to standard training. This result doubts the claim [29] that re-weighing will hurt the representation. Secondly, the models get marginal improvements or even worse performance after decoupled training. Compared with the end-to-end Sigmoid-EQL (Table VI) with 23.7% AP, the AP only increase 0.3% after using re-training on BAGS, and the AP drops 1.3% and 0.6% after the re-training on cRT and LWS respectively. It shows that decoupled training is not always necessary. We can train models with both a balanced classifier and better representations in an end-to-end manner.

Generalization to Other Datasets: To verify the generalization ability to other datasets, we conduct experiments on the OpenImages [18]. OpenImages is another large-scale object detection dataset with long-tailed distributed categories. We use the data split of challenge 2019, which is a subset of OpenImages V5. The train set consists of 1.7 M images of 500 categories. We evaluate our models on the 41 k val set. In addition to the standard mAP@IOU=0.5 metric, we also group categories into five groups (100 categories per group) according to their instance numbers and report the mAP within each group respectively. The results are shown in Table VIII. Sigmoid-EQL reaches an AP of 52.6%, outperforming the baseline model and Grad-Blocking [8] by 9.5% AP and 7.3% AP respectively. For the tail group (AP1), the Sigmoid-EQL increases the AP by 22.3 points, which is much more than the improvement of Grad-Blocking (6.4% AP). Sigmoid-EQL also outperforms Grad-Blocking considerably on

**TABLE VI**

| strategy | method            | AP   | APₕ | APₜ | APₖ | APₚ |
|----------|------------------|------|-----|-----|-----|-----|
| Baseline | CE               | 16.1 | 0.0 | 12.0| 27.4| 16.7|
|          | BCE              | 16.5 | 0.0 | 13.1| 27.3| 17.2|
| Decoupled| LWS [29]         | 17.0 | 2.0 | 13.5| 27.4| 17.5|
|          | cRT [29]         | 22.1 | 11.9| 20.2| 29.0| 22.2|
|          | BAGS [46]        | 23.1 | 13.1| 22.5| 28.2| 23.7|
| End-to-End| Softmax-EQL       | 23.1 | 15.3| 22.1| 27.6| 23.3|
|          | Sigmoid-EQL      | 23.7 | 14.9| 22.8| 28.6| 24.2|

For cRT and LWS, they use the class-balance sampler to fine-tune their model in the second stage, and BAGS uses a random sampler following the original paper.

**TABLE VII**

| method      | EQL   | AP   | APₕ | APₜ | APₖ | APₚ |
|-------------|-------|------|-----|-----|-----|-----|
| cRT [29]    |       | 22.0 | 13.5| 20.8| 27.1| 22.1|
| cRT [29]    | ✓     | 22.4 | 13.3| 21.7| 27.3| 22.4|
| LWS [29]    |       | 17.0 | 2.0 | 13.5| 27.4| 17.5|
| LWS [29]    | ✓     | 23.1 | 13.8| 22.2| 28.1| 23.2|
| BAGS [46]   |       | 23.1 | 13.1| 22.5| 28.2| 23.7|
| BAGS [46]   | ✓     | 24.0 | 14.6| 23.8| 28.5| 24.5|

EQL indicates that the pre-trained models are trained with EQL, otherwise with standard training. Only random samplers are used.

**TABLE VIII**

| method                | AP   | AP₁ | AP₂ | AP₃ | AP₄ | AP₅ |
|-----------------------|------|-----|-----|-----|-----|-----|
| Faster-R50            | 43.1 | 26.3| 45.2| 45.2| 48.2| 52.6|
| Grad-Blocking [8]     | 45.3 | 32.7| 44.6| 47.3| 48.3| 53.1|
| Sigmoid-EQL           | 52.6 | 48.6| 52.0| 53.0| 53.4| 55.8|
| Faster-R101           | 46.0 | 29.2| 45.5| 49.3| 50.9| 54.7|
| Grad-Blocking [8]     | 48.0 | 36.1| 47.2| 50.5| 51.0| 55.0|
| Sigmoid-EQL           | 55.1 | 51.0| 55.2| 56.6| 55.6| 57.5|

The model Faster R-CNN [58] with ReaNet-FPN is trained with a schedule of 120 k/160 k/180 k. Categories are grouped into five groups according to the instance number. AP₁ is the AP of the first group, where categories have the least annotations, and AP₅ is the AP of the last group, where categories have the most annotations. We directly use the hyper-parameters searched in LVIS.
the larger ResNet-101 backbone. For both baseline and Grad-Blocking models, there is still a large performance gap between head and tail classes. Sigmoid-EQL brings all categories into a more equal status. It achieves similar accuracy for all category groups. It is worth noting that we tune the hyper-parameter \( \lambda \) in Grad-Blocking which puts 250 categories into the tail group for OpenImage. In contrast, the hyper-parameters of Sigmoid-EQL are kept the same as that on LVIS. These experiments not only show the effectiveness but also the good generalization ability of Sigmoid-EQL.

### B. Single-Stage Long-Tailed Object Detection

In this section, we explore how to train a high-performance single-stage object detector successfully in the long-tailed setting. We mainly study the gradient-indicated focal loss, i.e., Equalized Focal Loss.

1) **Datasets and Evaluation Metric:** We perform experiments on the LVIS v1.0 dataset. We report box AP instead of mask AP because most single-stage detectors do not have a mask head. Since most experimental results of two-stage methods are based on the Mask R-CNN [16] framework, their released detection performance AP\(_5\) is affected by the segmentation performance. We report the detection results of those works by rerunning their code with the Faster R-CNN [54] framework for a fair comparison.

2) **Implementation Details:** Most training settings are the same as that in two-stage object detection experiments. As one-stage detectors often predict boxes with low scores, we do not filter out any predicted box before NMS (set the minimum score threshold to 0). The top 300 confident boxes per image are selected as the final detection results. All models are trained with the repeat factor sampler (RFS) by a 2x schedule. For our proposed EFL, we set the balanced factor \( \alpha_b = 0.25 \) and the base focusing parameter \( \gamma_b = 2.0 \) which are same as those in focal loss [15]. The scaling hyper-parameter \( s \) is set to 8.

3) **Main Results:** To show the effectiveness of our proposed method, we compare EFL with other works that report state-of-the-art performance. As demonstrated in Table IX, with ResNet-50-FPN backbone, our proposed method achieves an overall 27.5% AP, which improves the proposed Baseline\(^*\) by 1.8% AP, and even achieves 5.9 points improvement on rare categories. Compared with other end-to-end methods like Grad-Blocking [8] and Seesaw Loss [10], our proposed method outperforms them by 2.4% AP and 1.1% AP, respectively. And compared with decoupled training approaches like cRT [29] and BAGS [46], our approach surpasses them with an elegant end-to-end training strategy (by 2.7% AP and 1.5% AP). Meanwhile, compared with post-hoc calibration method Nor-Cal [47], EFL still achieves non-trivial improvement (+0.9% AP). Besides the high performance, we keep the advantages of one-stage detectors like simplicity, rapidity, and ease of deployment. It is worth noting that the improved baseline only has an AP of 14.3 for the rare category, which is worse than most two-stage detectors. With EFL, we achieve the best 20.2 AP for rare categories, showing that EFL can handle rare categories’ extreme positive-negative imbalance problem very well.

With the larger ResNet-101-FPN backbone, our approach still performs well on the Baseline\(^*\) (+2.2% AP). Meanwhile, our approach maintains stable performance improvements compared with all existing methods, whether they are end-to-end or decoupled. Without bells and whistles, our method achieves 29.2% AP that establishes a new state-of-the-art. Notably, the performance of rare categories on the Baseline\(^*\) does not gain too much performance improvement from the larger backbone while EFL does (from 20.2% AP to 23.5% AP). It indicates that our proposed method has a good generalization ability across different backbones. Additionally, we show some qualitative analysis in Fig 3. EFL outperforms the focal loss by focusing more on the learning of rare and common categories (i.e., the red boxes in the figure).

4) **Analysis. Influence of components in EFL:** There are two components in EFL, which are the focusing factor and the weighting factor. We demonstrate the effect of each component in Table X. Both the weighting factor and the focusing factor play significant roles in EFL. For the focusing factor, it achieves an improvement from 25.7% AP to 26.2% AP. Meanwhile, it brings a significant gain of rare categories with 3.4% AP improvement, indicating its effectiveness in alleviating the severe positive-negative imbalance problems. And for the weighting factor, we investigate its influence by setting the focusing factor always be \( \gamma_b \) among all categories in EFL. Thus the function of the weighting factor could also be regarded as a re-weighting approach combined with focal loss. Without the focusing factor, the weighting factor only outperforms the Baseline\(^*\) slightly by 0.4% AP. With the synergy of these two components, EFL dramatically improves the performance of the Baseline\(^*\) from 25.7% AP to 27.5% AP.

**Influence of the Hyper-Parameter:** Too many hyper-parameters will affect the generalization ability of a method. In this paper, our proposed EFL only has one hyper-parameter \( s \) which is also one of the advantages of our work. We study the
**Table IX**

| backbone     | method          | strategy | sampler   | epoch | AP   | AP_r | AP_c | AP_l |
|--------------|-----------------|----------|-----------|-------|------|------|------|------|
| ResNet-50    | two-stage       | end-to-end | RFS       | 24    | 24.1 | 14.7 | 22.2 | 30.5 |
|              | Faster R-CNN [54]| end-to-end | RFS       | 24    | 25.1 | 15.7 | 24.4 | 30.1 |
|              | Grad-Blocking [8]| end-to-end | RFS       | 24    | 26.4 | 17.5 | 25.3 | 31.5 |
|              | Seasaw Loss [10]| end-to-end | RFS       | 24    | 24.8 | 15.9 | 23.6 | 30.1 |
|              | cRT [29]        | decoupled | RFS+CBS   | 24+12 | 26.0 | 17.2 | 24.9 | 31.1 |
|              | BAGS [46]       | decoupled | RFS+CBS   | 24+12 | 26.6 | 18.7 | 25.6 | 31.1 |
|              | NorCal [47]     | post-hoc  | RFS       | 24    | 18.5 | 9.6  | 16.1 | 25.0 |
|              | RetinaNet [15]  | end-to-end | RFS       | 24    | 25.7 | 14.3 | 23.8 | 32.7 |
|              | Baseline†       | end-to-end | RFS       | 24    | 27.5 | 20.2 | 26.1 | 32.4 |

| backbone     | method          | strategy | sampler   | epoch | AP   | AP_r | AP_c | AP_l |
|--------------|-----------------|----------|-----------|-------|------|------|------|------|
| ResNet-101   | two-stage       | end-to-end | RFS       | 24    | 25.7 | 15.1 | 24.1 | 32.0 |
|              | Faster R-CNN [54]| end-to-end | RFS       | 24    | 27.8 | 18.7 | 27.0 | 32.8 |
|              | Seasaw Loss [10]| end-to-end | RFS       | 24    | 27.6 | 18.7 | 26.5 | 32.6 |
|              | BAGS [46]       | decoupled | RFS+CBS   | 24+12 | 27.8 | 19.4 | 26.9 | 32.5 |
|              | NorCal [47]     | post-hoc  | RFS       | 24    | 19.6 | 10.1 | 17.3 | 26.2 |
|              | RetinaNet [15]  | end-to-end | RFS       | 24    | 27.0 | 14.4 | 25.7 | 34.0 |
|              | Baseline†       | end-to-end | RFS       | 24    | 29.2 | 23.5 | 27.4 | 33.8 |
|              | EFL (Ours)      | end-to-end | RFS       | 24    | 26.5 | 19.9 | 24.6 | 31.6 |

Baseline† indicates the improved baseline. RFS and CBS indicate the repeat factor sampler and the class balanced sampler, respectively. All end-to-end methods are trained by a schedule of 2x with the RFS while the decoupled methods have an additional 1x schedule with the CBS during the fine-tuning stage.

Fig. 3. Equalized Focal Loss results on LVIS val set. The top row and bottom row are for focal loss and EFL respectively. The red, blue, and yellow boxes indicate the detected rare, common, and frequent categories, respectively.

**Table X**

| WF | FF | AP   | AP_r | AP_c | AP_l |
|----|----|------|------|------|------|
|    |    | 25.7 | 14.3 | 23.8 | 32.7 |
| ✓  |    | 26.1 | 15.6 | 24.5 | 32.6 |
| ✓  | ✓  | 26.2 | 17.7 | 24.7 | 31.5 |
| ✓  | ✓  | 27.5 | 20.2 | 26.1 | 32.4 |

WF and FF indicate the weighting factor and the focusing factor, respectively.

| s  | AP   | AP_r | AP_c | AP_l |
|----|------|------|------|------|
| 0  | 25.7 | 14.3 | 23.8 | 32.7 |
| 1  | 26.3 | 16.3 | 24.6 | 32.6 |
| 2  | 26.6 | 17.6 | 24.6 | 32.7 |
| 4  | 27.3 | 19.9 | 25.5 | 32.6 |
| 8  | 27.5 | 20.2 | 26.1 | 32.4 |
| 12 | 26.5 | 19.9 | 24.6 | 31.6 |

s = 8 is adopted as the default setting in other experiments.

Combined With Other One-Stage Detectors: To demonstrate the generalization ability of EFL across different one-stage detectors, we combine it with FCOS*, PAA, ATSS, and our

FCOS* indicates that the reported FCOS result is trained with a center-sampling strategy [57]
TABLE XII
RESULTS OF EFL COMBINED WITH OTHER ONE-STAGE OBJECT DETECTORS

| method      | loss       | AP  | AP5 | APm | AP1 |
|-------------|------------|-----|-----|-----|-----|
| FCOS* [58]  | FL         | 22.6| 12.7| 20.9| 28.9|
|             | FL+EQL     | 23.0 (+0.4)| 14.1| 21.3| 28.7|
|             | EFL        | **23.4 (+0.8)**| **14.9**| **21.9**| **28.7**|
| PAA [59]    | FL         | 23.7| 14.2| 21.6| 30.2|
|             | FL+EQL     | 24.1 (+0.4)| 16.5| 22.1| 29.8|
|             | EFL        | **25.6 (+1.9)**| **19.8**| **23.8**| **30.2**|
| ATSS [57]   | FL         | 24.7| 13.7| 23.4| 31.1|
|             | FL+EQL     | 25.2 (+0.5)| 15.0| 24.3| 30.8|
|             | EFL        | **25.8 (+1.9)**| **18.1**| **24.5**| **30.6**|
| Baseline†   | FL         | 25.7| 14.3| 23.8| 32.7|
|             | FL+EQL     | 26.8 (+0.9)| 17.7| 25.3| 32.6|
|             | EFL        | **27.5 (+1.8)**| **20.2**| **26.1**| **32.4**|

The bold values represent the best result (number) among its metric column (like \(AP_r, AP_p\)).

TABLE XIII
RESULTS OF THE YOLOX [55] DETECTORS COMBINED WITH EQL AND EQFL ON THE LVIS V1

| model       | loss       | YOLOX* | AP  | AP5 | APm | AP1 |
|-------------|------------|--------|-----|-----|-----|-----|
| Sigmoid     | FL         | ✓      | **15.2**| 2.9 | 11.6| 24.7|
|             | EFL(Ours)  | ✓      | 18.5| 3.6 | 15.7| **28.2**|
|             | QFL        | ✓      | **23.3**| **18.1**| **21.2**| 28.0|
|             | EQFL(Ours) | ✓      | 22.5| 11.0| 20.6| **29.7**|
|             |            | ✓      | **24.2**| **16.3**| **22.7**| 29.4|
| medium      | Sigmoid    | ✓      | 20.9| 5.3 | 17.6| 31.5|
|             | FL         | ✓      | 25.0| 7.1 | 23.5| 34.4|
|             | EFL(Ours)  | ✓      | 30.0| **23.8**| **25.2**| **34.7**|
|             | QFL        | ✓      | 28.9| 16.8| 27.2| 36.1|
|             | EQFL(Ours) | ✓      | **31.0**| **24.0**| **29.1**| **36.2**|

All experiments are trained from scratch by 300 epochs with Res50. YOLOX* indicates the enhanced YOLOX detector with our improved settings (see the supplementary material, available online).

Baseline†, separately. We also use the combination of Sigmoid-EQL and Focal loss, referred to as FL+EQL. As presented in Table XII, EFL and FL+EQL both performs well with all those one-stage detectors. We notice that EFL consistently outperforms FL+EQL due to its tight bond with focal loss. EFL maintains a stable large performance gain on overall AP and larger improvement on rare categories. These experiments show EFL’s strength in settling the long-tailed distribution problem. What’s more, we also investigate whether our proposed EFL and EQFL could work well with the advanced YOLOX [55] detectors (more implementation details could be found in our supplementary material, available online). As presented in Table XIII, combined with the YOLOX* medium model, EFL and EQFL reach the overall AP of 30.0% and 31.0% respectively, outperforming the baseline FL and QFL by a large margin and bringing great improvement on the rare categories (about +16.7% AP from FL to EFL). The results demonstrate that our proposed method is a very practical approach that could greatly alleviate the long-tailed imbalance problem for almost all one-stage detectors.

Generalization on OpenImages: As presented in Table XIV, our proposed improved baseline (Baseline†) greatly improves the performance of one-stage detectors. It brings 11.2% AP improvement compared with the widely used RetinaNet. Combined with the Baseline†, our proposed EFL achieves an overall AP of 51.5% with the ResNet-50 backbone, which outperforms the two-stage baseline Faster R-CNN and the Baseline† by 8.4% AP and 8.2% AP, respectively. What’s more, EFL significantly improves the performance of the rare categories with an improvement of 33.4% AP on the AP1 split compared to the Baseline†. With the larger ResNet-101 backbone, our proposed method still performs well and brings significant AP gains. Meanwhile, it maintains an excellent performance on rare categories. All experimental results demonstrate the strength and generalization ability of our method.

Performance on Balanced Dataset: As we claimed, EFL is equivalent to the focal loss in the balanced data scenario. To verify this analysis, we conduct experiments on the MS-COCO dataset with balanced data distribution. As presented in Table XV, the scaling factor \(s\) has little effect in the COCO dataset, and all results with EFL achieve comparable performance with the focal loss. This indicates that our proposed EFL could maintain good performance under the balanced data distribution. This distribution-agnostic property enables EFL to work well with real-world applications in different data distributions.

C. Long-Tailed Image Classification.

To demonstrate the generalization ability of equalization losses when transferring to other tasks, we also evaluate our method on long-tailed image classification datasets. Different from object detection, the image classification task requires a single predicted category as output, in which case we should take the cross-category rank into account. We found that the sigmoid...
loss is not suitable for this task, so we use the Softmax-EQL in image classification.

1) Datasets and Evaluation Metric: We conduct experiments on three major benchmarks to evaluate the effectiveness of our proposed method.

**ImageNet-LT** [5] is a long-tailed version of ImageNet-2012 [1], which contains 1000 categories with images number ranging from 1280 to 5 images for each category. There are 116 k images for training and 50 k images for testing.

**iNaturalist2018** [62] is a real-world large-vocabulary dataset. There are 437.5 K images from 8,142 categories, which suffer from severe long-tailed imbalances.

**Place-LT** [5] is also a commonly used benchmark for long-tailed recognition. It is a subset artificially truncated from Places2 [63]. Its number of images per class ranges from 4980 to 5 among 365 categories.

For all datasets, besides the metric of total average accuracy, we present the average accuracy of many-shot (≥100 images), medium-shot (20 ∼ 100 images), and few-shot (≤20 images) splits to better understand the improvement of tail classes.

2) Implementation Details: We implement our proposed equalization losses based on the settings of Seesaw Loss [10] and cRT [29]. We notice that some methods have their own codebase (e.g., [30]) and training settings (e.g., [32], [61]). To ensure a fair comparison, we re-implement all compared methods following the common practice settings of cRT [29]. In specific, we use the SGD optimizer with momentum 0.9, batch size 512 (256 for Place-LT), and weight decay 5e-4 (1e-4 for iNaturalist2018). We adopt the cosine learning schedule where the learning rate gradually decays from 0.2 to 0. Random resized crop, horizontal flip, and color jitter (iNaturalist2018 except for this item) are used as data augmentation. We report 90 epochs experimental results on ImageNet-LT (ResNet-50 and ResNeXt-50 [65]) and iNaturalist2018 (ResNet-50). For Place-LT, we choose ResNet-152 (pre-trained on the full ImageNet2012) as the backbone network. We train Place-LT for 30 epochs with everything frozen except the last layer and classifier, which are all aligned with [29].

3) Main Results: Firstly, we present the results of ImageNet-LT in Table XVI. As the results show, our proposed Softmax-EQL achieves the highest accuracy on different backbones with significant improvements on the tail categories, outperforming current state-of-the-art methods. In particular, Softmax-EQL surpasses the performance of number-based adjustment methods (e.g., Seesaw Loss [10], LADE [61], Logit Adjustment [32], and Balanced Softmax [31]), proving that the gradient-based
adjustment method is a reliable choice for long-tailed recognition. Meanwhile, Softmax-EQL also outperforms other multi-training-stage methods (e.g., cRT [29], LWS [29], and DisAlign [30]) are finetuned from the CE pre-trained model, while DIVE [60] mimics the features from the Balanced Softmax).

Meanwhile, Sigmoid-EQL outperforms the BCE baseline by a significant margin, especially on tail categories. However, we can see that the sigmoid-based losses are not comparable with softmax-based losses. This is because, for each instance, the sigmoid-based losses do not take the cross-category competition

We also report grouped mIoU and grouped mACC, where the last value is for tail categories and the first value for head categories.
into account. Thus they are not a good choice for long-tailed classification tasks.

We also conduct experiments on Place-LT and iNaturalist2018. The results are demonstrated in Tables XVII and XVIII, respectively. Softmax-EQL still achieves impressive performance on these datasets, demonstrating the generalization ability of our equalization losses on different recognition benchmarks. Furthermore, we propose to mimic the balanced Softmax features using the Softmax-EQL (called DIVE-Softmax-EQL) following the DIVE strategy. As presented in Table XVIII, DIVE-Softmax-EQL outperforms DIVE by 1.3 points of accuracy, with an improvement of 3.0 points on tail categories.

4) Experiments on Foundation Model: To illustrate the effectiveness of our method in large-scale foundation model scenarios, we conducted another experiment by using the CLIP [66] model as the backbone. As we can see from table XIX, using EQL on top of the powerful CLIP backbone still obtains significant performance improvements over CE and Logit Adjustment CE (LACE). On Place-LT, EQL achieves comparable results with RAC [50], a strong retrieval augmented method with large foundation models for long-tail classification.

D. Long-Tailed Semantic Segmentation

Semantic segmentation is another important visual recognition task in computer vision. Semantic segmentation is doing a dense per-pixel classification, which also involves cross-category competition. There is a long-tailed problem in semantic segmentation since the head categories often occupy much more pixels. To show the versatility of equalization losses, we conduct experiments on large-scale semantic segmentation datasets with various models.

1) Datasets and Evaluation Metric: ADE20K [7] is a large scale dataset for semantic segmentation. It has 20,196 training images and 2 k testing images. There is a total of 150 categories whose annotations are not carefully picked. So there naturally exists an imbalanced pixel ratio between categories.

Although the pixel ratio is imbalanced between categories, the least category still has annotations in over 400 images. To simulate a more severe long-tailed distribution. We sample the original ADE20 K dataset to construct the ADE20K-LT dataset. Specifically, we first calculate the sample ratio for each category with a exponential distribution: $e^{-0.01i}$, where $i \in [0, 150)$ is the category index. For instance, the last category, index 149, has the sample ratio of 22%; the first category, index 0, has a sample ratio of 0%; We start the sampling process from the last category to the first category. With the calculated sample ratio of a given category $j$, we randomly sample images from all images that contain category $j$. More details about the dataset construction process can be found in the appendix, available online. Finally, we generate the long-tailed version of ADE20 K, namely ADE20K-LT. It contains 11,578 training images and the same 2 k test images as the original ADE20 K. Besides the metric mIoU and mAcc, we evenly divide 150 categories into three groups and report the mIoU and mAcc of each group.

2) Implementation Details: We conduct experiments on two strong models: PSPNet [64] and DeeplabV3+ [19]. We use a crop size of $512 \times 512$ in training. Synchronized Batch Normalization is adopted in the ResNet backbone. Models are trained with a total of 80 k iterations for ADE20 K (or 40 k iterations for ADE20k-LT) using the poly learning rate scheduler. The total batch size is 16 and the initial learning rate is 0.01.

3) Main Results: First of all, we present the results of ADE20K-LT in Table XX. We choose two popular and powerful models, PSPNet [64] and DeeplabV3+ [19]. We choose the CE and BCE loss as baseline loss functions. The CE loss achieves higher results than the BCE loss. This is because that semantic segmentation requires a single predicted category per pixel, and the CE loss aligns with the requirement better. Replacing with Sigmoid-EQL or Softmax-EQL both improves the performance significantly. The Softmax-EQL outperforms the baseline CE loss by a large margin in mAcc. This demonstrates that gradient-based margin calibration can benefit tasks that require inter-category competition. However, the Sigmoid-EQL and Softmax-EQL give similar improvements on mIoU, showing that a higher accuracy does not mean a higher mIoU. An auxiliary loss, such as IOU-based loss may mitigate this problem. Next, we report the experimental result of ADE20 K in Table XXI. Since ADE20 k naturally exists the long-tailed problem, we can see that the later groups have lower mIoU and mAcc (i.e., mIoU_3 and mAcc_3). The Sigmoid-EQL and Softmax-EQL still outperform the baseline loss functions. For the largest model, DeeplabV3+ with ResNet-101, the Sigmoid-EQL achieve non-trivial improvement, 1.5 points on mIoU and 2.3 points on mAcc, respectively.

Integration With Maskformer: We integrated EQL into the transformer-based and much stronger framework, Maskformer [67]. Specifically, we replaced the softmax in the query classification branch with softmax-EQL. It is worth noting that in the original Maskformer, there is an imbalance problem because most queries are background ones. To alleviate this problem, they used a loss weight of 0.1 for the background. When incorporating softmax-EQL, there is no longer a need to manually set weight for the background class, as the gradient-based training

| Method               | Dataset      | mAcc | mAcc_1 | mAcc_2 | mAcc_3 | mIoU | mIoU_1 | mIoU_2 | mIoU_3 |
|----------------------|--------------|------|--------|--------|--------|------|--------|--------|--------|
| MaskFormer [67]      | ADE20K       | 57.0 | 68.5   | 53.7   | 48.9   | 44.2 | 54.1   | 41.1   | 37.4   |
| MaskFormer + EQL     | ADE20K       | 60.5 | 69.2   | 58.1   | 54.1   | 44.6 | 54.4   | 41.0   | 38.4   |
| MaskFormer [67]      | ADE20K-LT    | 46.2 | 64.7   | 46.4   | 27.6   | 35.6 | 49.7   | 34.4   | 22.9   |
| MaskFormer + EQL     | ADE20K-LT    | 50.7 | 64.7   | 51.7   | 35.8   | 36.1 | 49.1   | 34.4   | 24.7   |
mechanism automatically adjusts for the background class. Table XXII shows that the introduction of gradient-driven training significantly improves the mACC and mIoU of Maskformer, particularly for tail classes.

VI. CONCLUSION

In this work, we systematically study the gradient imbalance problem in long-tailed data distribution. We show that this problem exists consistently in different visual tasks and datasets. We find the gradient statistic is able to serve as a stable and precise indicator for the imbalance status of models. Finally, a new family of gradient-driven loss functions is proposed, namely equalization loss. Extensive experiments demonstrate its effectiveness and generalization ability. We believe that this gradient-driven training idea will serve as an important concept when designing specific algorithms for long-tailed object recognition.

REFERENCES

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2009, pp. 248–255.

[2] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The Pascal visual object classes (VOC) challenge,” Int. J. Comput. Vis., vol. 88, no. 2, pp. 303–338, 2010.

[3] T.-Y. Lin et al., “Microsoft COCO: Common objects in context,” in Proc. Eur. Conf. Comput. Vis., 2014, pp. 740–755.

[4] M. Cordts et al., “The cityscapes dataset for semantic urban scene understanding,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 3213–3223.

[5] 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 Conf. Comput. Vis. Pattern Recognit., 2019, pp. 2537–2546.

[6] A. Gupta, P. Dollar, and R. Girshick, “LVIS: A dataset for large vocabulary instance segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 5356–5364.

[7] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, “Scene parsing through ADE20K dataset,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 633–641.

[8] J. Tan et al., “Equalization loss for long-tailed object recognition,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 11 662–11 671.

[9] J. Tan, J. Wu, and Z.-H. Zhou, “Exploratory undersampling for class imbalance problem in convolutional neural networks,” 2020, pp. 8577–8584.

[10] X. Zhang, Z. Fang, Y. Wen, Z. Li, and Y. Qiao, “Range loss for deep face recognition with long-tailed training data,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2018, pp. 9268–9277.

[11] L. Shen, Z. Lin, and Q. Huang, “Relay backpropagation for effective learning of deep convolutional neural networks,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 467–482.

[12] D. Mahajan et al., “Exploring the limits of weakly supervised pretraining,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 181–196.

[13] K. Cao, C. Wei, A. Gaidon, N. Areshchiga, and T. Ma, “Learning imbalanced datasets with label-distribution-aware margin loss,” in Proc. Adv. Neural Inf. Process. Syst., 2019, pp. 1567–1578.

[14] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, “Class-balanced loss for deep neural networks,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 9268–9277.

[15] 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., 2017, pp. 2900–2908.

[16] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 2961–2969.

[17] J. Tan et al., “1st place solution of LVIS challenge 2020: A good box is not a guarantee of a good mask,” 2020, arXiv:2009.01559.

[18] A. Kaznetsova et al., “The open images dataset V4: Unified image classification, object detection, and visual relationship detection at scale,” 2018, arXiv:1811.00982.

[19] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 801–818.
Jingru Tan received the bachelor’s degree from Tongji University, in 2017, and the PhD degree from Tongji University, in 2022. He was a postdoctoral fellow with Shanghai Jiao Tong University. He is currently working with Central South University. His research interests include deep learning and computer vision.

Bo Li received the BS degree in software engineering from Tongji University, Shanghai, China, in 2019. He is currently working toward the PhD degree in software engineering with the School of Software Engineering, Tongji University. His research interests include computer vision and long-tailed recognition.

Xin Lu received the master’s degree in computer science from Zhejiang University. He is a computer vision researcher with SenseTime Research in Shenzhen. His research interests include computer vision application, AutoML, semi-supervised learning, and unsupervised learning.

Yongqiang Yao received the master’s degree from the Beijing University of Posts and Telecommunications, in 2020. He is currently a researcher with SenseTime. His main research interests include computer vision, training framework research, and machine learning.

Fengwei Yu received the master’s degree in computer science from Beihang University. He previously served as the deputy director of R&D with SenseTime Research Institute, responsible for deep learning infrastructure. His research interests include model training systems, deep learning compilers, model quantization, compression, and more. He has published numerous academic papers in conferences such as CVPR, ICCV, ECCV, ICLR, NeurIPS, ICPP, MLsys, and IPDPS.

Tong He received the PhD degree in computer science from the University of Adelaide, Australia, in 2020. He is currently a researcher with Shanghai AI Laboratory. His research interests include computer vision and machine learning.

Wanli Ouyang (Senior Member, IEEE) received the PhD degree from the Department of Electronic Engineering, Chinese University of Hong Kong. He is now a professor with Shanghai AI Lab. His research interests include pattern recognition, machine learning, and AI for science. Before that, he was an associate professor with the University of Sydney. He served as the associate editor of IEEE Transactions on Pattern Analysis and Machine Intelligence, International Journal of Computer Vision, and PR, the senior area chair of CVPR.