Reconciling Object-Level and Global-Level Objectives for Long-Tail Detection

Shaoyu Zhang\textsuperscript{1,2}, Chen Chen\textsuperscript{1,2∗}, Silong Peng\textsuperscript{1,2,3∗}
\textsuperscript{1}Institute of Automation, Chinese Academy of Sciences, Beijing, China
\textsuperscript{2}School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
\textsuperscript{3}Beijing Visystem Co. Ltd, Beijing, China
\{zhangshaoyu2019, chen.chen, silong.peng\}@ia.ac.cn

Abstract

Large vocabulary object detectors are often faced with the long-tailed label distributions, seriously degrading their ability to detect rarely seen categories. On one hand, the rare objects are prone to be misclassified as frequent categories. On the other hand, due to the limitation on the total number of detections per image, detectors usually rank all the confidence scores globally and filter out the lower-ranking ones. This may result in missed detection during inference, especially for the rare categories that naturally come with lower scores. Existing methods mainly focus on the former problem and design various classification loss to enhance the object-level classification accuracy, but largely overlook the global-level ranking task. In this paper, we propose a novel framework that Reconciles Object-level and Global-level (ROG) objectives to address both problems. As a multi-task learning framework, ROG simultaneously trains the model with two tasks: classifying each object proposal individually and ranking all the confidence scores globally. Specifically, complementary to the object-level classification loss for model discrimination, we design a generalized average precision (GAP) loss to explicitly optimize the global-level score ranking across different objects. For each category, GAP loss generates balanced gradients to rectify the ranking errors. In experiments, we show that GAP loss is highly versatile to be plugged into various advanced methods and brings considerable benefits. Code is at https://github.com/EricZsy/ROG.

1. Introduction

The development of modern convolutional neural networks (CNNs) gives rise to great advances in object detection [14, 36, 29] and instance segmentation [49, 54]. So far, the state-of-the-art object detectors typically rely heavily on a huge amount of annotated data [27]. However, with the rapid growth of data scale, the long-tail effect has become a bottleneck in training object detectors for real-world use. In large vocabulary image datasets such as LVIS [15], there are only a few categories containing abundant instances, while most other rare categories seldom appear. Detectors naturally lean towards the frequent categories and usually fail in detecting those rarely seen objects.

In Figure 1, we analyze that the undesired phenomenon is mainly derived from two aspects. First, the significantly more frequent objects dominate the training process, and thereby suppress the rare categories by overwhelming discouraging gradients [42]. As a result, the latter is more likely to be predicted with lower confidence scores and mis-
classified, e.g., the *soccer* is classified as the frequent category *person* in Figure 1(b). Second, during inference, the rare objects are prone to be lost due to a seemingly trivial detail: the global ranking and filtering for all the scores in each image. Before non-maximum suppression (NMS), all the scores are collected together and the ones below a certain threshold will be filtered out. In addition, considering the memory limit, the object detectors usually have limitations on the maximum number of detections per image. Therefore, after NMS, all the detected objects in an image will be ranked by their confidence scores, and only the top-$K$ of them will be reserved for final results. Since the scores for rare categories are relatively low, they are easily squeezed out in the global-level and cross-object ranking competition (see Figure 1(c)). The two problems, respectively arising in training and inference phase, are relevant but of different emphasis. The former presents the challenge of classifying each object accurately, while the latter focuses on how to rank all the scores fairly.

Prior works in long-tail detection mainly focus on the former problem, i.e., misclassification from tail to head. To calibrate the classification bias, a variety of specialized classification loss functions are proposed. Some of them re-weight the classification loss based on class prior [42, 17, 1] or training status [41, 45, 48], while others [43, 37, 12, 19, 47] design class-wise margins for calibrated decision boundaries. Although modifying the classification loss could enhance model discrimination and improve object-level classification accuracy, it is not able to give a direct solution to the cross-object ranking task. In fact, even though a rare object is classified correctly, it is still possible to be missed due to the low confidence score. Recall that for object detection, the classification accuracy is not a comprehensive indicator, while average precision (AP) is a more widely used metric. Motivated by this, some work [4, 30, 31, 7] replace the classification task with a ranking task to learn to rank every positive sample above all the negative samples. Detectors trained by ranking-based loss may be good at ranking task but not discriminative enough. For example, the loss could be minimized to zero even if the scores of positive samples are only slightly higher than that of negative ones. More importantly, these ranking algorithms are category-agnostic that weights all ground-truth samples equally, no matter which categories they belong to. It is still inconsistent with the AP metric\(^1\) which is averaged over each category. Especially, under long-tail distribution, the gradients for ranking rare-category samples are weak, and will be easily deflected by frequent categories.

To address the aforementioned problems in long-tail detection, it is highly considerable to reconcile both the object-level discrimination objective and global-level ranking objective during training. Therefore, in this paper, we present ROG, a multi-task learning framework that simultaneously learns two tasks: classifying each object proposal individually and ranking all confidence scores globally. The object-level discrimination objective aims to train a discriminative classifier that could classify each object accurately. It also ensures a unified score distribution across categories. On this basis, the global-level ranking objective is proposed to optimize the cross-object ranking orders via a generalized average precision loss. For each specific category, the loss is calculated by ranking errors between category-specific positive-negative pairs. Then the generalized precision loss is averaged over all categories, and thus generates balanced gradients to re-rank samples for each category equally. Following [5], the error-driven update algorithm is adopted to optimize the non-differentiable ranking loss. As a whole, the classification task and the ranking task complement each other to jointly cater to the classification and ranking procedures in object detection. In addition, the two tasks could be trained harmoniously: the classification scores are provided for the ranking task, and the ranking task could improve scores of positive samples and reduce scores of negative samples which in turn promotes the classification task.

Extensive experiments are conducted on the challenging LVIS [15] and OpenImages [22] datasets. We show that the generalized average precision loss is highly versatile to cooperate with existing methods. As a whole framework, ROG consistently improves the performance of state-of-the-art methods, across various classification losses [41, 19], sampling strategies [15] and post-processing methods [33].

To sum up, the main contribution of this work is a ROG framework that considers both object-level classification task and global-level ranking task in long-tail detection. It is motivated by the overlooked ranking and filtering procedure in inference phase, and aims to learn to classify each object and rank all the confidence scores simultaneously. A generalized average precision loss is proposed to rectify the ranking errors for each category equally. The extensive experiments validate the effectiveness of ROG.

### 2. Related Works

**General Object Detection.** With deep learning, general object detection [53] is popularized by both two-stage and single-stage detectors. Two-stage detectors [14, 13, 38, 16] are equipped with a separate module to first generate region proposals, which are further refined for accurate classification and localization in the second stage. Based on this pipeline, Many follow-up improvements have been proposed from different concerns, including feature pyramid network (FPN) [25], mask branch for instance segmentation [16], and *etc.* Instead, single-stage detectors such as YOLO [36] and SSD [29] classify and localize semantic

\(^1\)Note that in COCO [27], there is no distinction between AP and mean average precision (mAP). They are both averaged over all categories. We follow their notations and use AP throughout the paper for consistency.
objects in a single shot using dense sampling. They are typically high in efficiency but have been lagging in accuracy until the introduction of RetinaNet [26]. In addition, there are other methods generating detection boxes by grouping key-points on objects [11, 23].

**Long-tail Object Classification and Detection.** Recently, long-tail object classification [51, 55] has drawn a lot of attention. Existing methods can be roughly categorized into class re-balancing and data augmentation. Class re-balancing, either by cost-sensitive learning [21, 2, 9, 40] or data re-sampling [18, 56, 3], aim to balance different classes during training. Kang et al. [20] further propose the decoupled training strategy which only re-balances the classifier learning. Alternatively, data augmentation methods attempt to increase the size and diversity of rare categories via data mixing [8] or head-to-tail knowledge transfer [52, 44].

The long-tailed data distribution also degrades the performance of object detectors [32]. Wang et al. [46] show that the performance drop mainly arises in the classification sub-network. Therefore, with the precedents in long-tail classification, earlier attempts in long-tail detection utilize decoupled training [46, 24] or cost-sensitive loss [37, 42, 17]. To alleviate the suppression on rare categories, Equalization Loss (EQL) [42] is proposed with class-wise weights to reduce discouraging gradients from frequent instances, and it is further improved by directly re-weighting the classification loss based on gradient statistics [41]. Similarly, Seesaw Loss [45] dynamically balances gradients for each category with the mitigation factor and the compensation factor. Although these methods could alleviate the object-level misclassification, they overlook the cross-object ranking relation between all the ent foreground categories. Then, all the scores are ranked globally and the lower-ranking ones are filtered out. Since the ranking task is performed on the scores, here we use the term sample to refer to each score $s_i$. The reserved samples are collected in $S$ to calculate the precision and recall values. For a category $m$, $S$ is divided into a set of positive samples $\mathcal{P}(m)$ and a set of negative samples $\mathcal{N}(m)$. The average precision on $m$ could be written as the mean of precision over $\mathcal{P}(m)$:

$$AP_m = \frac{1}{|\mathcal{P}(m)|} \sum_{i \in \mathcal{P}(m)} Prec(i),$$

where $|\mathcal{P}(m)|$ is the size of $\mathcal{P}(m)$, and the precision for $i$-th positive sample is

$$Prec(i) = \frac{R^{m+}(i)}{R(i)}.$$  

$R^{m+}(i)$ and $R(i)$ stands for the ranking position of sample $i$ in $\mathcal{P}(m)$ and $\mathcal{P}(m) \cup \mathcal{N}(m)$. Then the average precision over all foreground categories is formulated as:

$$AP = \frac{1}{C} \sum_{m=0}^{C-1} AP_m,$$

The AP metric has two desired characteristics. First, it effectively captures the cross-object ranking relation between the ranking confidence scores, which is in concert with the ranking task in inference phase. Second, it measures each category equally so that the influence of data imbalance on overall assessment could be eliminated. Therefore, it is promising.

### 3. Our Method

We propose to reconcile object-level and global-level (ROG) objectives to improve long-tail detection. In Sec. 3.1, we first revisit the evaluation of object detection as preliminary, and present an overview of our method. Then we introduce the two components in ROG: the object-level discrimination objective (see Sec. 3.2) and the global-level ranking objective (see Sec. 3.3). Lastly, the multi-task objective and its optimization are introduced in Sec. 3.4.

#### 3.1. Preliminary and Overview

Given an image, the aim of object detection is to detect semantic objects with their locations and categories. For better insight into our method, we begin by revisiting how to measure the performance of an object detector.

When evaluated on a dataset, the detector collects all the object proposals, each with a score vector $s = [s_0, \ldots, s_C]$, where $s_C$ is the score for background and others for different foreground categories. Then, all the scores are ranked globally and the lower-ranking ones are filtered out. Since the ranking task is performed on the scores, here we use the term sample to refer to each score $s_i$. The reserved samples are collected in $S$ to calculate the precision and recall values. For a category $m$, $S$ is divided into a set of positive samples $\mathcal{P}(m)$ and a set of negative samples $\mathcal{N}(m)$. The average precision on $m$ could be written as the mean of precision over $\mathcal{P}(m)$:

$$AP_m = \frac{1}{|\mathcal{P}(m)|} \sum_{i \in \mathcal{P}(m)} Prec(i),$$

where $|\mathcal{P}(m)|$ is the size of $\mathcal{P}(m)$, and the precision for $i$-th positive sample is

$$Prec(i) = \frac{R^{m+}(i)}{R(i)}.$$  

$R^{m+}(i)$ and $R(i)$ stands for the ranking position of sample $i$ in $\mathcal{P}(m)$ and $\mathcal{P}(m) \cup \mathcal{N}(m)$. Then the average precision over all foreground categories is formulated as:

$$AP = \frac{1}{C} \sum_{m=0}^{C-1} AP_m,$$
Figure 2. The framework of ROG. Detector predicts score vectors for all the detected boxes, each of which contains the scores for \( C \) foreground categories and a background (BG) category. The overall objective includes an object-level discrimination objective and a global-level ranking objective. The former aims for accurate classification for each object, while the latter rectifies the cross-object score ranking for each specific category. * denotes the ground-truth, and the green arrows denote the optimization directions in each objective.

Our goal is to improve long-tail detection by reducing both classification errors and ranking errors, especially for rare categories. To this end, we propose to reconcile object-level objective and global-level objective during training. The overall framework of ROG is illustrated in Figure 2. ROG jointly optimizes an object-level discrimination objective and a global-level ranking objective on the confidence scores. The object-level discrimination objective aims to ensure a discriminative model to classify each object accurately. Complementarily, the global-level ranking objective caters to the ranking procedure of inference phase and guides the detector to rank all scores across objects. It is achieved by a generalized average precision loss, which generates balanced gradients for each category via global statistics \( \omega \). The generalized average precision loss is particularly suitable for long-tailed data as it aims to rectify ranking errors for each specific category equally and pays enough attention to rare categories. Under the ROG framework, the two objectives complement with each other to jointly promote the detection performance.

### 3.2. Object-Level Discrimination Objective

In object detection, the classification sub-network plays a role in classifying each object proposal. It is usually trained with the cross-entropy (CE) classification loss to ensure the discrimination. For each proposal, the CE classification loss could be written as:

\[
L_{cls} = - \sum_{j=0}^{C} y_j \log p_j, \tag{4}
\]

where \( y = [y_0, \ldots, y_C] \) is the one-hot label and \( p = [p_0, \ldots, p_C] \) is the probability vector for \( C \) foreground categories plus background.

In view of the long-tailed characteristic, there are many novel classification losses proposed for balancing the training among categories, such as Seesaw Loss [45] and Effective Class-Margin Loss [19]. These methods effectively promote the rare categories and can be freely adopted as the classification loss in our framework. Overall though, the object-level classification loss only ensures the discriminative ability of the model, which means that the model could discriminate the ground-truth from other categories for each proposal individually. However, it is not able to manage the cross-object ranking task. We present an example in Figure 3. Owing to the classification loss, all the three objects are classified accurately with the highest confidence scores for their ground-truth categories. In terms of the ranking task, however, the ranking order could still be incorrect. The ground-truth of rare category, i.e. soccer, tends to be predicted with a low confidence score, and thus may rank behind some negative samples. Although the classification loss could improve the score of ground-truth category, it fails to provide explicit comparisons among different objects to rectify the incorrect ranking orders.

### 3.3. Global-Level Ranking Objective

The classification loss aims to enhance the discrimination of the model, but it could not optimize the cross-object ranking task directly. To overcome this limitation, we further propose a generalized average precision (GAP) loss to explicitly adjust the ranking orders for each category.

The generalized average precision loss is calculated on
where $\gamma$ is a hyper-parameter controlling the strength of global information. The total size $|\mathcal{P}(m)|$ for each category is empirically accumulated by $|\mathcal{P}(\cdot)|$ from each training batch. Note that in Eq. 8, the negative samples are defined in $\mathcal{N}$ rather than $\hat{\mathcal{N}}(m)$, which means they are not positive samples for any other category. This is because that positive samples for other categories would be also included in $\hat{\mathcal{N}}(m)$, leading to conflicting gradients in optimization. Therefore, the ranking orders between positive sample pairs from different categories $\{(i, j) | i \in \mathcal{P}(m), j \in \hat{\mathcal{P}}(n), m \neq n\}$ are not constrained in the loss as illustrated in Figure 2.

3.4. Multi-Task Objective and Optimization

Combining both object-level discrimination objective and global-level ranking objective, the total loss of ROG is:

$$L_{\text{rog}} = L_{\text{cls}} + \lambda_{\text{gap}} L_{\text{gap}},$$

(10)

where $\lambda_{\text{gap}}$ is the parameter controlling the weight of two tasks. Since we mainly focus on the classification sub-network, the bounding box regression loss in the localization sub-network is omitted here.

The $L_{\text{cls}}$ could be optimized easily via backpropagation and automatic differentiation [34]. However, note that the ranking function $H(i, j)$ is non-differentiable, the optimization of $L_{\text{gap}}$ is not trivial. Inspired by [5], we adopt the error-driven update algorithm [39] to efficiently calculate the gradients. Given the input $(s_j - s_i)$ and the resulting pairwise loss $L_{ij}$, the update for input is directly set as the difference between target loss value and current loss value, i.e., $(0 - L_{ij})$. Therefore, $\frac{\partial L_{ij}}{\partial s_i}$ and $\frac{\partial L_{ij}}{\partial s_j}$ could be replaced by $-L_{ij}$ and $L_{ij}$, respectively. For a score $s_k$, the gradient propagated from $L_{\text{gap}}$ is:

$$g_{\text{gap}}^k = \begin{cases} 0, & \text{if } s_k \neq 0, \\ \frac{1}{|\mathcal{C}|} \frac{\sum_{m \in \mathcal{C}} \omega_m}{|\mathcal{P}(m)|} \sum_{j \in \mathcal{N}} L_{kj}, & k \in \hat{\mathcal{P}}(m), \end{cases}$$

(11)

The manually set gradients could be further back propagated by the chain rule to train the whole network. Please refer to supplementary material for the detailed derivations.

4. Experiments

4.1. Experimental Setup

**Dataset.** We evaluate ROG on LVIS v1 [15] and OpenImages [22] dataset. LVIS v1 is a large vocabulary object detection and instance segmentation dataset, which contains 100k images for training and 19.8k images for validation. There are totally 1,203 object categories with long-tailed distribution. According to the number of training images per category, they are divided into three groups: rare (with 1-10 images), common (with 11-100 images), and frequent
Table 1. Results on the validation set of LVIS v1 with ResNet-50 backbone. The proposed generalized average precision loss brings consistent improvements on existing methods under different sampling strategies. † indicates reproduced results from their released codes.

<table>
<thead>
<tr>
<th>Sampler</th>
<th>$L_{cls}$</th>
<th>$L_{gap}$</th>
<th>AP$^{bbox}$</th>
<th>AP</th>
<th>AP$_r$</th>
<th>AP$_c$</th>
<th>AP$_f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sigmoid CE</td>
<td>✓</td>
<td></td>
<td>17.3</td>
<td>16.7</td>
<td>0.8</td>
<td>13.4</td>
<td>27.4</td>
</tr>
<tr>
<td>Softmax CE</td>
<td>✓</td>
<td></td>
<td>24.7 (+7.4)</td>
<td>23.9</td>
<td>(+7.2)</td>
<td>14.1 (+13.3)</td>
<td>23.7</td>
</tr>
<tr>
<td>CE + NorCal [33] †</td>
<td>X</td>
<td>✓</td>
<td>20.2</td>
<td>19.6</td>
<td>2.7</td>
<td>18.4</td>
<td>28.3</td>
</tr>
<tr>
<td>EQLv2 [41] †</td>
<td>X</td>
<td>✓</td>
<td>24.2 (+4.0)</td>
<td>24.1</td>
<td>(+4.5)</td>
<td>16.7 (+14.0)</td>
<td>24.2</td>
</tr>
<tr>
<td>ECM [19] †</td>
<td>X</td>
<td>✓</td>
<td>25.0 (+0.8)</td>
<td>24.4</td>
<td>(+0.8)</td>
<td>17.2 (+2.2)</td>
<td>23.5</td>
</tr>
<tr>
<td>RFS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sigmoid CE</td>
<td>✓</td>
<td></td>
<td>22.9</td>
<td>22.2</td>
<td>11.8</td>
<td>21.4</td>
<td>27.6</td>
</tr>
<tr>
<td>Softmax CE</td>
<td>✓</td>
<td></td>
<td>25.9 (+3.0)</td>
<td>25.1</td>
<td>(+2.9)</td>
<td>18.2 (+6.4)</td>
<td>24.6</td>
</tr>
<tr>
<td>ECM [19] †</td>
<td>X</td>
<td>✓</td>
<td>26.7 (+0.5)</td>
<td>26.9</td>
<td>(+0.6)</td>
<td>20.1 (+0.6)</td>
<td>26.8</td>
</tr>
</tbody>
</table>

(x with over 100 images). OpenImages is another long-tailed object detection dataset with 500 categories, which are further divided into five groups following [41].

Evaluation Metrics. The evaluation metric is the mean average precision across IOU threshold from 0.5 to 0.95. We use AP$^{bbox}$ to assess the detection performance, and AP to assess the segmentation performance. In addition to the average precision over all categories, we also report AP$_r$, AP$_c$, and AP$_f$ on LVIS to measure the performance for the rare, common and frequent groups respectively.

Implementation Details. Our implementation is based on MMDetection toolbox[6]. We adopt the Mask R-CNN [16] and Faster R-CNN [38] with Feature Pyramid Networks (FPN) [25] as baseline models for LVIS and OpenImages, respectively. Models are trained by SGD with a momentum of 0.9 and a weight decay of 0.0001. For 1x schedule with 12 training epochs, the learning rate is initialized as 0.02, and then decays by 0.1 at the end of epoch 8 and 11. For 2x schedule, models are trained with 24 epochs, and the learning rate decays at the end of epoch 16 and 22. During training, the default data augmentations such as random horizontal flipping and scale jitter are used. During inference, the score threshold is set to 0.0001 and the maximum number of detections per image is set to 300 following the convention. For ROG, we set $\delta = 0.5$ and $\gamma = 1$ unless specified. The $\lambda_{gap}$ is set to 0.1 for experiments with repeat factor sampling (RFS) [15] and 1.0 for other experiments.

4.2. Ablation Study

We conduct the ablation studies on LVIS, with ResNet-50 backbone network and 1x training schedule.

Effectiveness of ROG. We firstly evaluate the effectiveness of the generalized average precision loss $L_{gap}$. Since GAP loss is based on ranking and complementary to the classification task, it can be seamlessly plugged into existing classification-based methods. As shown in Table 1, GAP loss consistently enhances the classification-based baselines and state-of-the-arts. For the baseline model trained with sigmoid CE loss and random sampler, GAP loss could improve the AP of object detection and instance segmentation by 7.4 and 7.2, respectively. It is noted that the AP for rare categories rises from 0.8 to 14.1, which clearly demonstrates the effectiveness of GAP loss for promoting rare categories. Furthermore, we apply our GAP loss with the recently proposed ECM loss [19]. We find that GAP loss still brings solid improvements (e.g., +11.6 AP$_r$). Considering that ECM loss is already the state-of-the-art classification loss, we attribute the additional improvements to the global-level ranking objective. In addition, our method is also compatible with the RFS sampler that repeatedly samples images containing rare objects. This is attributed to the manner of online statistics in each batch, which precisely calculates the total number of training instances from each category, even with different samplers.

We further verify the effectiveness of ROG that reduces missed detection caused by global ranking. To quantitatively analyze missed detection, an intuitive idea is to increase the maximum number of detections per image (denoted by $K$) to include more missed samples, and observe how much the AP will be improved [10]. In Table 3, by including more samples into the final detection (from $K = 50$ to 100), the improvements for rare categories are far more than frequent ones (71.8% vs. 11.0%), which indicates that rare samples are more likely to be missed. In contrast, for our ROG, the improvements brought by increasing $K$ are relatively balanced among all the categories. This validates
that ROG largely compensates missed detection for rare categories caused by global ranking and filtering.

**Analysis of different ranking objectives.** The global-level ranking objective in ROG trains the model to rank confidence scores across different objects. We investigate how the performance is affected by different choices of the ranking objectives. In Table 2, our GAP loss is compared with AP loss [5] and RS loss [31], which are both category-agnostic. When using AP loss or RS loss as the global-level ranking objective alone, we find that the results are comparable with the CE loss baseline. However, the performance for rare categories are still significantly lower than common and frequent categories. Due to the category-agnostic nature, these losses weight each positive sample equally without considering their labels, leading to insufficient training for the ranking task on rare categories. In contrast, our GAP loss is category-specific and thus induces balanced training over all categories. As a result, the performance on rare categories are improved over 8 AP via GAP loss, which validates our analysis. Furthermore, we also find that training AP loss together with the object-level CE loss even leads to worse performance. Our GAP loss works more harmoniously with CE loss and consistently improves the AP on all the rare, common and frequent categories. Note that [31] designs a special way to balance the RS loss and bounding box regression loss, so we did not conduct experiments on the combination of RS loss and classification loss. To further explore the balanced performance from GAP loss, we follow [41] to accumulate the gradients from positive samples and negative samples for each category and calculates their ratios. The ratios from different ranking loss are presented in Figure 4, which shows that GAP loss produces more balanced gradients for ranking on each category.

**Table 2.** Comparisons on different choices of global-level ranking objectives.

<table>
<thead>
<tr>
<th>Object-level Global-level</th>
<th>AP\textsuperscript{bbox}</th>
<th>AP</th>
<th>AP\textsubscript{r}</th>
<th>AP\textsubscript{c}</th>
<th>AP\textsubscript{f}</th>
</tr>
</thead>
<tbody>
<tr>
<td>N/A</td>
<td>AP loss [5]</td>
<td>17.2</td>
<td>17.9</td>
<td>3.4</td>
<td>15.9</td>
</tr>
<tr>
<td></td>
<td>RS loss [31]</td>
<td>20.5</td>
<td>19.8</td>
<td>3.6</td>
<td>17.6</td>
</tr>
<tr>
<td></td>
<td>GAP loss (ours)</td>
<td>21.1</td>
<td>21.1</td>
<td>12.0</td>
<td>21.0</td>
</tr>
<tr>
<td>Softmax CE loss</td>
<td>AP loss [5]</td>
<td>16.7</td>
<td>17.8</td>
<td>4.5</td>
<td>16.5</td>
</tr>
<tr>
<td></td>
<td>GAP loss (ours)</td>
<td>22.0</td>
<td>22.3</td>
<td>12.1</td>
<td>21.9</td>
</tr>
<tr>
<td>Sigmoid CE loss</td>
<td>AP loss [5]</td>
<td>14.6</td>
<td>15.8</td>
<td>2.9</td>
<td>13.8</td>
</tr>
<tr>
<td></td>
<td>GAP loss (ours)</td>
<td>24.7</td>
<td>23.9</td>
<td>14.8</td>
<td>23.0</td>
</tr>
</tbody>
</table>

**Table 3.** Relative improvements by increasing the maximum number of detections per image (denoted by \(K\)) on LVIS v1.

<table>
<thead>
<tr>
<th>Method</th>
<th>(K)</th>
<th>AP\textsuperscript{bbox}</th>
<th>AP</th>
<th>AP\textsubscript{r}</th>
<th>AP\textsubscript{c}</th>
<th>AP\textsubscript{f}</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>50</td>
<td>6.7 (+71.8%)</td>
<td>16.1 (+29.8%)</td>
<td>25.1 (+11.0%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>13.4 (+17.5%)</td>
<td>19.9 (+20.6%)</td>
<td>26.1 (+10.1%)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 4.** Ablation study on designs in GAP loss. The global statistic \(\omega\) is defined in Eq. 9. Online means the online statistic manner for calculating category size. Excl. FG means excluding other foreground positive samples from the negative set of category \(m\).

<table>
<thead>
<tr>
<th>(\omega)</th>
<th>Online</th>
<th>Excl. FG</th>
<th>AP\textsuperscript{bbox}</th>
<th>AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>✗ ✗ ✗</td>
<td></td>
<td></td>
<td>19.2</td>
<td>19.6</td>
</tr>
<tr>
<td>✓ ✗ ✗</td>
<td></td>
<td></td>
<td>23.8</td>
<td>23.5</td>
</tr>
<tr>
<td>✓ ✓ ✗</td>
<td></td>
<td></td>
<td>24.1</td>
<td>23.4</td>
</tr>
<tr>
<td>✓ ✗ ✓</td>
<td></td>
<td></td>
<td>24.1</td>
<td>23.9</td>
</tr>
<tr>
<td>✓ ✓ ✓</td>
<td>✓</td>
<td></td>
<td>24.7</td>
<td>23.9</td>
</tr>
</tbody>
</table>
Figure 5. Visualization on weight norms of classifier and AP of each category.

Table 5. Comparisons with state-of-the-art methods on LVIS v1.

<table>
<thead>
<tr>
<th>Method</th>
<th>AP$^{bbox}$</th>
<th>AP</th>
<th>AP$_r$</th>
<th>AP$_c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>RFS [15]</td>
<td>26.6</td>
<td>25.5</td>
<td>16.6</td>
<td>24.5</td>
</tr>
<tr>
<td>EQLv2 [41]</td>
<td>27.9</td>
<td>27.2</td>
<td>20.6</td>
<td>25.9</td>
</tr>
<tr>
<td>LOCE [12]</td>
<td>29.0</td>
<td>28.0</td>
<td>19.5</td>
<td>27.8</td>
</tr>
<tr>
<td>Seesaw [45]</td>
<td>28.9</td>
<td>28.1</td>
<td>20.0</td>
<td>28.0</td>
</tr>
<tr>
<td>ROG (ours)</td>
<td>29.3</td>
<td>28.8</td>
<td>21.1</td>
<td>29.1</td>
</tr>
</tbody>
</table>

Table 6. Comparisons with state-of-the-art methods on OpenImages.

<table>
<thead>
<tr>
<th>Method</th>
<th>AP$^{bbox}$</th>
<th>AP$_1$</th>
<th>AP$_2$</th>
<th>AP$_3$</th>
<th>AP$_4$</th>
<th>AP$_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster-R101</td>
<td>46.0</td>
<td>29.2</td>
<td>45.5</td>
<td>49.3</td>
<td>50.9</td>
<td>54.7</td>
</tr>
<tr>
<td>EQL [42]</td>
<td>48.0</td>
<td>36.1</td>
<td>47.2</td>
<td>50.5</td>
<td>51.0</td>
<td>55.0</td>
</tr>
<tr>
<td>Seesaw [45]</td>
<td>47.5</td>
<td>37.2</td>
<td>46.0</td>
<td>48.7</td>
<td>50.2</td>
<td>55.1</td>
</tr>
<tr>
<td>EQLv2 [41]</td>
<td>55.1</td>
<td>51.0</td>
<td>55.2</td>
<td>56.6</td>
<td>55.6</td>
<td>57.5</td>
</tr>
<tr>
<td>ROG (ours)</td>
<td>58.2</td>
<td>54.8</td>
<td>59.2</td>
<td>60.6</td>
<td>58.0</td>
<td>58.5</td>
</tr>
</tbody>
</table>

we consider the negative set $\hat{N}$ instead of $\hat{N}(m)$. This will exclude positive samples of other foreground categories and avoids conflicting gradients among categories in optimization. The experimental results verify the effectiveness of our designs, which achieve AP$^{bbox}$ of 24.7 and AP of 23.9.

**Weight norm of classifier and AP of each category.** Previous work [20] has shown that classifier trained on long-tail data exhibits imbalanced weight norms across categories. In Figure 5(a), we visualize the weight norms of classifier in the Mask R-CNN. We train a model by the baseline sigmoid CE loss and another by the baseline ROG. By applying the GAP loss on the CE loss, we observe that the baseline ROG leads to more balanced weight norms. In addition, we represent the AP of each category in Figure 5(b). We find that ROG significantly lifts up the performance on rare categories, meanwhile without obvious performance drop on other categories.

Please refer to supplementary material for more results and analysis.

4.3. Main Results

In this section, we compare the proposed ROG with several state-of-the-art methods on LVIS v1 and OpenImages. We adopt ResNet-101 with FPN as backbone network, and use 2x training schedule. For LVIS, Seesaw loss [45] is adopted as the classification loss in ROG. The results listed in Table 5 show that ROG improves the Seesaw loss by 0.7 points on AP and 1.1 points on AP$_r$. For OpenImages, we choose the EQLv2 [41] as the classification loss. In Table 6, on the basis of EQLv2, ROG brings 3.1 AP gains for object detection. It improves the performance on all the groups, especially the rarest one (3.8 points for AP$_1$).

5. Conclusion and Limitation

In this paper, we propose ROG to reconcile object-level and global-level objectives in long-tail detection. Different from previous works focusing mainly on classification, we analyze that the ranking and filtering procedure during inference may cause the missed detection for rare categories. Motivated by this, complementary to the object-level discrimination objective, a generalized average precision loss is proposed as a global-level ranking objective, with a balanced view to rectify ranking errors for each specific category. Since the two objectives aim for two separate tasks, existing classification losses could be plugged into ROG without any interference. Meanwhile, the two objectives works harmoniously with each other, and jointly promote the performance of object detection and instance segmentation especially for the rare categories. Experimental results show that ROG consistently improves the performance of state-of-the-art methods.

**Limitation.** In ROG, the proposed generalized average precision loss is based on pairwise ranking of all positive samples. As a result, it takes slightly longer for training than the classification-based baseline. We plan to address these limitations in the future work.
Acknowledgements. This work is supported by the National Key R&D Program of China under Grant 2021YFF0602101.

References


[49] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Solo: A simple framework for instance segmen-


