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

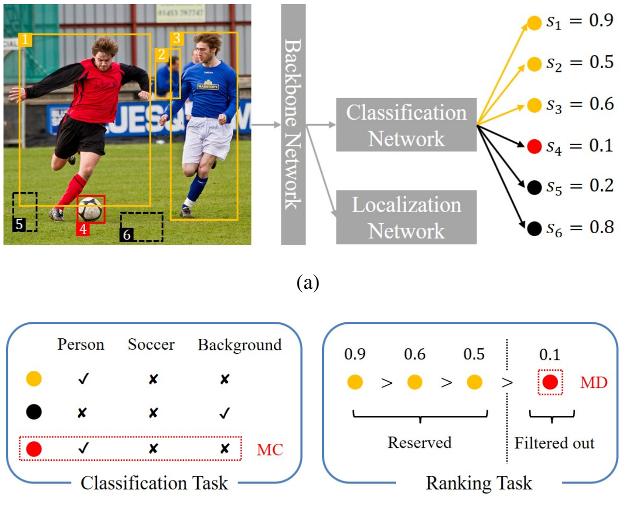
Shaoyu Zhang[1,2], Chen Chen[1,2*], Silong Peng[1,2,3*]
[1]Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3]Beijing Visystem Co. Ltd, Beijing, China
{zhangshaoyu2019, chen.chen, silong.peng}@ia.ac.cn

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.



(a)

(b)

(c)

Figure 1. The two major causes degrading detection performance for rare categories: misclassification (MC) and missed detection (MD). (a) The common pipeline of an object detector. The boxes with solid line and dashed line are foregrounds and backgrounds respectively. The dots denote the predictions for these boxes, followed by their confidence scores. (b) Rare object *soccer* is misclassified as *person* in the classification task. (c) Missed detection for *soccer* caused by global ranking and filtering.

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-

[*]Corresponding author.

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 global ranking, which may result in missed detection on rare objects. Very recently, Effective Class-Margin Loss [19] is proposed to implicitly bound the mean average precision by a margin-based classification loss. Orthogonal to designing the classification loss, we provide a new angle of view to explicitly optimize the global-level ranking task in long-tail detection.

Ranking for Object Detection. In object detection, a line of work [5, 30, 35, 28, 50] directly replaces the classification loss with ranking-based loss. AP Loss [5] explicitly models the category-agnostic ranking orders between all the positive-negative sample pairs, while RS Loss [31] extends this idea to additionally sort positive pairs w.r.t. their localization qualities. RankDetNet [28] systematically integrates the score-guided and IoU-guided ranking task to replace the classification task. However, the ranking loss itself is not discriminative enough. Moreover, these ranking-based losses are category-agnostic and mainly designed for general object detection. Under long-tailed distribution, the ranking objective will be dominated by frequent categories, while the incorrect ranking and missed detection for rare categories still remain unsolved. Unless prior works, our global-level ranking objective optimizes each category equally for ranking rectification.

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 $\boldsymbol{s} = [s_0, \dots, s_C]$, 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 $\mathcal{S}$ to calculate the precision and recall values. For a category m, $\mathcal{S}$ 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 $\mathcal{P}(m)$:

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

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

$$Prec(i) = \frac{R^{m+}(i)}{R(i)}. \tag{2}$$

$R^{m+}(i)$ 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:

$$AP = \frac{1}{C} \sum_{m=0}^{C-1} AP_m, \tag{3}$$

The AP metric has two desired characteristics. First, it effectively captures the cross-object ranking relation between 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

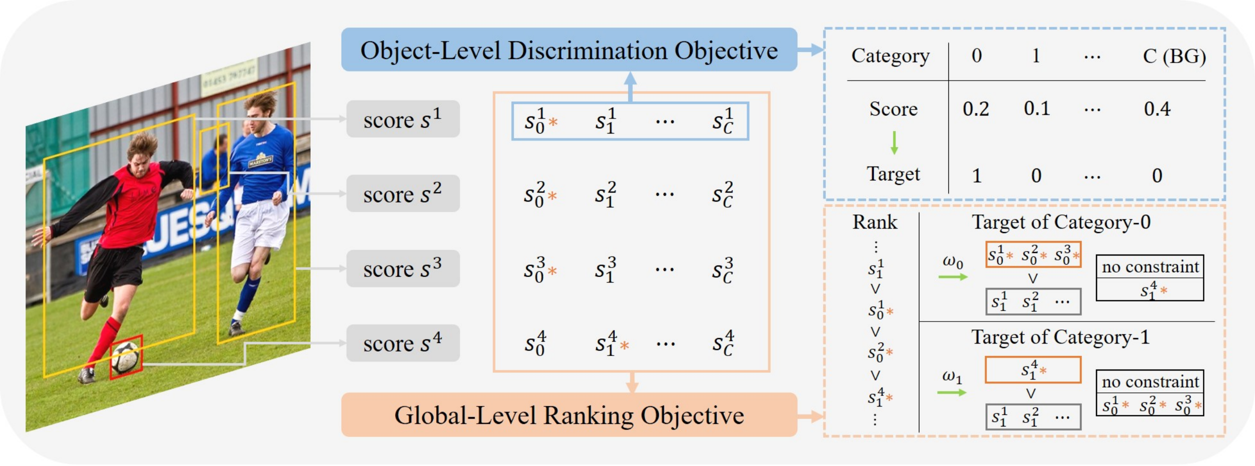

Object-Level Discrimination Objective
score s^1
score s^2
score s^3
score s^4
s_0^1* s_1^1 ··· s_C^1
s_0^2* s_1^2 ··· s_C^2
s_0^3* s_1^3 ··· s_C^3
s_0^4 s_1^4* ··· s_C^4
Global-Level Ranking Objective
Category 0 1 ··· C (BG)
Score 0.2 0.1 ··· 0.4
Target 1 0 ··· 0
Rank
s_1^1 > s_0^1* > s_0^2* > s_1^4*
Target of Category-0
ω_0
s_0^1* s_0^2* s_0^3* > s_1^1 s_1^2 ···
no constraint
s_1^4*
Target of Category-1
ω_1
s_1^4* > s_1^1 s_1^2 ···
no constraint
s_0^1* s_0^2* s_0^3*

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.

to adopt AP as a global-level ranking optimization objective on long-tailed data.

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 ω. 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:

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

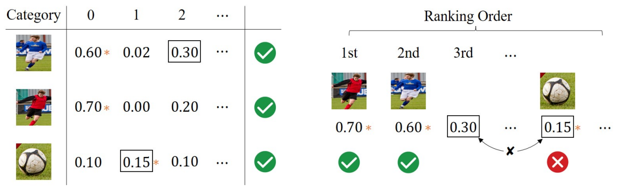
where $\boldsymbol{y} = [y_0, \ldots, y_C]$ is the one-hot label and $\boldsymbol{p} = [p_0, \ldots, p_C]$ is the probability vector for C foreground categories plus background.

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



(a) Object classification. (b) Score ranking.

Figure 3. Illustration of an example which succeeds in classification but fails in ranking. For the ranking task, the score comparisons across different objects, *e.g.*, 0.30 vs. 0.15 framed in boxes, are not considered in the object-level classification loss.

each training batch $\hat{\mathcal{B}}$. In $\hat{\mathcal{B}}$, all the positive samples of category m are collected in $\hat{\mathcal{P}}(m)$, while negative samples are collected in $\hat{\mathcal{N}}(m)$. Following [5], we define a pairwise ranking function $H(i,j)$:

$$H(i,j)=\begin{cases}0, & s_j-s_i<-\delta\\ \frac{(s_j-s_i)}{2\delta}+0.5, & -\delta\le s_j-s_i\le\delta\\ 1, & s_j-s_i>\delta\end{cases} \quad (5)$$

where δ is a parameter controlling the margin between s_i and s_j. Then the ranking position for each positive sample $i\in\hat{\mathcal{P}}(m)$ could be obtained via accumulative pairwise ranking:

$$R^{m+}(i)=1+\sum_{j\in\hat{\mathcal{P}}(m),j\neq i}H(i,j), \quad (6)$$

$$R(i)=1+\sum_{j\in\hat{\mathcal{P}}(m)\cup\hat{\mathcal{N}}(m),j\neq i}H(i,j). \quad (7)$$

Following above definitions, the generalized average precision loss is proposed as $\langle 1-\mathrm{AP}\rangle_{\hat{\mathcal{B}}}$, with special designs to adapt to the mini-batch training paradigm. Intuitively, the AP metric should reflect the ranking performance for each category equally over the entire dataset. However, due to mini-batch training, the statistical information in each training batch is usually biased, which could not reflect the global statistics. Typically, only a small subset of foreground categories $\hat{C}$ occurs in a mini-batch and the frequency of each category may also vary sharply in different $\hat{\mathcal{B}}$. Therefore, we need to inject the global statistics of the entire dataset into the batch-wise ranking loss. Let $L_{ij}=H(i,j)/R(i)$, then the generalized average precision loss on the training batch is:

$$L_{gap}=\frac{1}{|\hat{C}|}\sum_{m\in\hat{C}}\frac{\omega_m}{|\hat{\mathcal{P}}(m)|}\sum_{i\in\hat{\mathcal{P}}(m)}\sum_{j\in\hat{\mathcal{N}}}L_{ij}. \quad (8)$$

The global statistic for category m is introduced by ω_m:

$$\omega_m=\frac{(\frac{1}{|\mathcal{P}(m)|})^\gamma}{\sum_{k=0}^{C-1}(\frac{1}{|\mathcal{P}(k)|})^\gamma}\times C, \quad (9)$$

where γ is a hyper-parameter controlling the strength of global information. The total size $|\mathcal{P}(\cdot)|$ for each category is empirically accumulated by $|\hat{\mathcal{P}}(\cdot)|$ from each training batch. Note that in Eq. 8, the negative samples are defined in $\hat{\mathcal{N}}$ 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 $\left\{(i,j)|i\in\hat{\mathcal{P}}(m),j\in\hat{\mathcal{P}}(n),m\neq n\right\}$ 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:

$$L_{rog}=L_{cls}+\lambda_{gap}L_{gap}, \quad (10)$$

where λ_{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_{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_{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_{gap} is:

$$g^k_{gap}=\begin{cases}-\frac{1}{|\hat{C}|}\frac{\omega_m}{|\hat{\mathcal{P}}(m)|}\sum_{j\in\hat{\mathcal{N}}}L_{kj}, & k\in\hat{\mathcal{P}}(m)\\ \frac{1}{|\hat{C}|}\sum_{m\in\hat{C}}\frac{\omega_m}{|\hat{\mathcal{P}}(m)|}\sum_{i\in\hat{\mathcal{P}}(m)}L_{ik}, & k\in\hat{\mathcal{N}}.\end{cases} \quad (11)$$

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.

Sampler	L_{cls}	L_{gap}	AP^{bbox}	AP	AP$_r$	AP$_c$	AP$_f$
Random	Sigmoid CE	✗	17.3	16.7	0.8	13.4	27.4
		✓	**24.7 (+7.4)**	**23.9 (+7.2)**	**14.1 (+13.3)**	**23.7**	**28.0**
	Softmax CE	✗	16.7	16.1	0.0	12.0	**27.4**
		✓	**22.0 (+5.3)**	**22.3 (+6.2)**	**12.1 (+12.1)**	**21.9**	27.3
	CE + NorCal [33] †	✗	20.2	19.6	2.7	18.4	**28.3**
		✓	**24.2 (+4.0)**	**24.1 (+4.5)**	**16.7 (+14.0)**	**24.2**	27.2
	EQLv2 [41] †	✗	24.2	23.6	15.0	22.5	28.5
		✓	**25.0 (+0.8)**	**24.4 (+0.8)**	**17.2 (+2.2)**	**23.5**	**28.7**
	ECM [19] †	✗	22.4	21.3	5.1	20.9	**28.9**
		✓	**25.0 (+2.6)**	**24.7 (+3.4)**	**16.7 (+11.6)**	**24.2**	28.7
RFS	Sigmoid CE	✗	22.9	22.2	11.8	21.4	27.6
		✓	**25.9 (+3.0)**	**25.1 (+2.9)**	**18.2 (+6.4)**	**24.6**	**28.7**
	GOL [1] †	✗	25.8	26.0	19.1	26.2	28.8
		✓	**26.1 (+0.3)**	**26.4 (+0.4)**	**20.3 (+1.2)**	**26.6**	**28.9**
	ECM [19] †	✗	26.7	26.3	19.5	26.0	29.8
		✓	**27.2 (+0.5)**	**26.9 (+0.6)**	**20.1 (+0.6)**	**26.8**	**30.0**

(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 λ_{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

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

Object-level	Global-level	AP^{bbox}	AP	AP$_r$	AP$_c$	AP$_f$
N/A	AP loss [5]	17.2	17.9	3.4	15.9	26.4
	RS loss [31]	20.5	19.8	3.6	17.6	**29.4**
	GAP loss (ours)	**21.1**	**21.1**	**12.0**	**21.0**	25.2
Softmax CE loss	AP loss [5]	16.7	17.8	4.5	16.5	25.2
	GAP loss (ours)	**22.0**	**22.3**	**12.1**	**21.9**	**27.3**
Sigmoid CE loss	AP loss [5]	14.6	15.8	2.9	13.8	23.7
	GAP loss (ours)	**24.7**	**23.9**	**14.8**	**23.0**	**28.8**

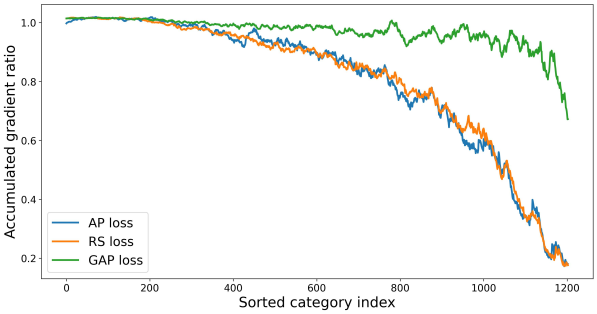


Figure 4. Ratios of accumulated positive gradients to negative gradients from different ranking loss.

that ROG largely compensates missed detection for rare categories caused by global ranking and filtering.

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

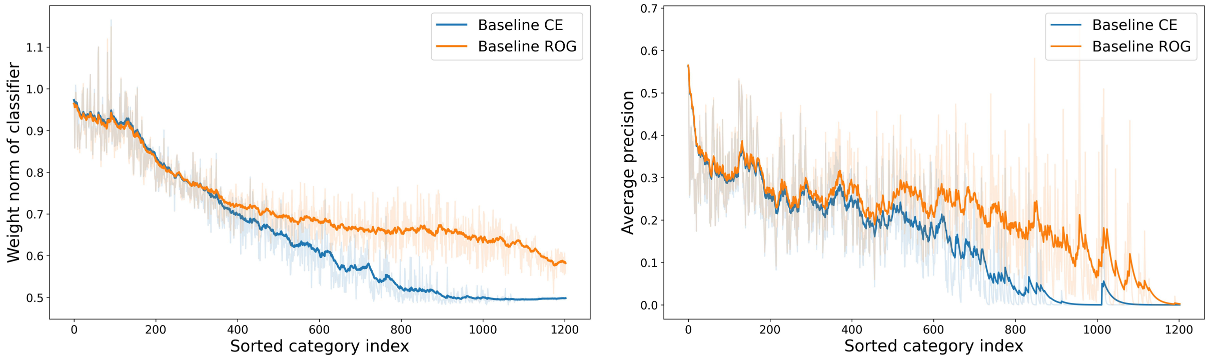
Method	K	AP$_r$	AP$_c$	AP$_f$
RFS	50	3.9	12.4	22.6
	100	6.7 **(+71.8%)**	16.1 **(+29.8%)**	25.1 **(+11.0%)**
ROG	50	11.4	16.5	23.7
	100	13.4 **(+17.5%)**	19.9 **(+20.6%)**	26.1 **(+10.1%)**

Table 4. Ablation study on designs in GAP loss. The global statistic ω 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.

ω	Online	Excl. FG	AP^{bbox}	AP
✗	✗	✗	19.2	19.6
✓	✗	✗	23.8	23.5
✓	✓	✗	24.1	23.4
✓	✗	✓	24.1	23.9
✓	✓	✓	**24.7**	**23.9**

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.

Designs in GAP loss. In Eq. 8, our GAP loss differs from previous ranking loss [5] mainly in the category-specific ranking manner and the introduction of global statistic ω. In Table 4, we start by omitting the ω in calculating the loss. As analyzed in Sec. 3.3, the statistic in a single batch could not represent the statistic of the entire training set. It fails to optimize the AP over all categories in the face of long-tail distribution, and thus leads to a sub-optimal result. When introducing ω by the pre-defined category distribution from dataset, the AP gains by a large margin (4.6 points on AP^{bbox} and 3.9 points on AP). Furthermore, we replace the pre-defined distribution by the online statistic distribution, which slightly improves the AP for object detection by 0.3 points. In addition, when calculating GAP loss between positive and negative samples for category m,



(a) Weight norms of classifier.

(b) Average precision of each category.

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.

Method	AP^{bbox}	AP	AP_r	AP_c	AP_f
RFS [15]	26.6	25.5	16.6	24.5	30.6
EQLv2 [41]	27.9	27.2	20.6	25.9	31.4
LOCE [12]	29.0	28.0	19.5	27.8	**32.0**
Seesaw [45]	28.9	28.1	20.0	28.0	31.9
ROG (ours)	**29.3**	**28.8**	**21.1**	**29.1**	31.8

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

Method	AP^{bbox}	AP_1	AP_2	AP_3	AP_4	AP_5
Faster-R101	46.0	29.2	45.5	49.3	50.9	54.7
EQL [42]	48.0	36.1	47.2	50.5	51.0	55.0
Seesaw [45]	47.5	37.2	46.0	48.7	50.2	55.1
EQLv2 [41]	55.1	51.0	55.2	56.6	55.6	57.5
ROG (ours)	**58.2**	**54.8**	**59.2**	**60.6**	**58.0**	**58.5**

we consider the negative set $\hat{\mathcal{N}}$ instead of $\hat{\mathcal{N}}(m)$. 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

[1] Konstantinos Panagiotis Alexandridis, Jiankang Deng, Anh Nguyen, and Shan Luo. Long-tailed instance segmentation using gumbel optimized loss. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*, pages 353–369. Springer, 2022.

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

[3] Nadine Chang, Zhiding Yu, Yu-Xiong Wang, Animashree Anandkumar, Sanja Fidler, and Jose M Alvarez. Image-level or object-level? a tale of two resampling strategies for long-tailed detection. In *International conference on machine learning*, pages 1463–1472. PMLR, 2021.

[4] Kean Chen, Jianguo Li, Weiyao Lin, John See, Ji Wang, Lingyu Duan, Zhibo Chen, Changwei He, and Junni Zou. Towards accurate one-stage object detection with ap-loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5119–5127, 2019.

[5] Kean Chen, Weiyao Lin, Jianguo Li, John See, Ji Wang, and Junni Zou. Ap-loss for accurate one-stage object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3782–3798, 2020.

[6] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[7] Tao Chenxin, Zizhang Li, Xizhou Zhu, Gao Huang, Yong Liu, et al. Searching parameterized ap loss for object detection. *Advances in Neural Information Processing Systems*, 34:22021–22033, 2021.

[8] Hsin-Ping Chou, Shih-Chieh Chang, Jia-Yu Pan, Wei Wei, and Da-Cheng Juan. Remix: rebalanced mixup. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 95–110. Springer, 2020.

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

[10] Achal Dave, Piotr Dollár, Deva Ramanan, Alexander Kirillov, and Ross Girshick. Evaluating large-vocabulary object detectors: The devil is in the details. *arXiv preprint arXiv:2102.01066*, 2021.

[11] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6569–6578, 2019.

[12] Chengjian Feng, Yujie Zhong, and Weilin Huang. Exploring classification equilibrium in long-tailed object detection. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 3417–3426, 2021.

[13] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[14] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

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

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[17] Ting-I Hsieh, Esther Robb, Hwann-Tzong Chen, and Jia-Bin Huang. Droploss for long-tail instance segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 1549–1557, 2021.

[18] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5375–5384, 2016.

[19] Jang Hyun Cho and Philipp Krähenbühl. Long-tail detection with effective class-margins. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, pages 698–714. Springer, 2022.

[20] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019.

[21] Salman H Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous A Sohel, and Roberto Togneri. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE transactions on neural networks and learning systems*, 29(8):3573–3587, 2017.

[22] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.

[23] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 734–750, 2018.

[24] Yu Li, Tao Wang, Bingyi Kang, Sheng Tang, Chunfeng Wang, Jintao Li, and Jiashi Feng. Overcoming classifier imbalance for long-tail object detection with balanced group softmax. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10991–11000, 2020.

[25] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

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

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[28] Ji Liu, Dong Li, Rongzhang Zheng, Lu Tian, and Yi Shan. Rankdetnet: Delving into ranking constraints for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 264–273, 2021.

[29] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016.

[30] Kemal Oksuz, Baris Can Cam, Emre Akbas, and Sinan Kalkan. A ranking-based, balanced loss function unifying classification and localisation in object detection. *Advances in Neural Information Processing Systems*, 33:15534–15545, 2020.

[31] Kemal Oksuz, Baris Can Cam, Emre Akbas, and Sinan Kalkan. Rank & sort loss for object detection and instance segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3009–3018, 2021.

[32] Kemal Oksuz, Baris Can Cam, Sinan Kalkan, and Emre Akbas. Imbalance problems in object detection: A review. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3388–3415, 2020.

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

[34] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[35] Qi Qian, Lei Chen, Hao Li, and Rong Jin. Dr loss: Improving object detection by distributional ranking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12164–12172, 2020.

[36] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

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

[38] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[39] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[40] Saptarshi Sinha, Hiroki Ohashi, and Katsuyuki Nakamura. Class-difficulty based methods for long-tailed visual recognition. *International Journal of Computer Vision*, 130(10):2517–2531, 2022.

[41] Jingru Tan, Xin Lu, Gang Zhang, Changqing Yin, and Quanquan Li. Equalization loss v2: A new gradient balance approach for long-tailed object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1685–1694, 2021.

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

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

[44] Jianfeng Wang, Thomas Lukasiewicz, Xiaolin Hu, Jianfei Cai, and Zhenghua Xu. Rsg: A simple but effective module for learning imbalanced datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3784–3793, 2021.

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

[46] Tao Wang, Yu Li, Bingyi Kang, Junnan Li, Junhao Liew, Sheng Tang, Steven Hoi, and Jiashi Feng. The devil is in classification: A simple framework for long-tail instance segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 728–744. Springer, 2020.

[47] Tong Wang, Yousong Zhu, Yingying Chen, Chaoyang Zhao, Bin Yu, Jinqiao Wang, and Ming Tang. C2am loss: Chasing a better decision boundary for long-tail object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6980–6989, 2022.

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

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

tation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8587–8601, 2021.

[50] Dongli Xu, Jinhong Deng, and Wen Li. Revisiting ap loss for dense object detection: Adaptive ranking pair selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14187–14196, 2022.

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

[52] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature transfer learning for face recognition with under-represented data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5704–5713, 2019.

[53] Syed Sahil Abbas Zaidi, Mohammad Samar Ansari, Asra Aslam, Nadia Kanwal, Mamoona Asghar, and Brian Lee. A survey of modern deep learning based object detection models. *Digital Signal Processing*, page 103514, 2022.

[54] Yuhang Zang, Chen Huang, and Chen Change Loy. Fasa: Feature augmentation and sampling adaptation for long-tailed instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3457–3466, 2021.

[55] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *arXiv preprint arXiv:2110.04596*, 2021.

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