Equalized Focal Loss for Dense Long-Tailed Object Detection

Bo Li[1*]   Yongqiang Yao[2*]   Jingru Tan[1*]   Gang Zhang[3]

Fengwei Yu[2]   Jianwei Lu[1]   Ye Luo[1]

[1]Tongji University   [2]SenseTime Research   [3]Tsinghua University

{1911030,yeluo}@tongji.edu.cn, {soundbupt,tanjingru120}@gmail.com

yufengwei@sensetime.com, zhang-g19@mails.tsinghua.edu.cn, jwlu33@126.com

Abstract

Despite the recent success of long-tailed object detection, almost all long-tailed object detectors are developed based on the two-stage paradigm. In practice, one-stage detectors are more prevalent in the industry because they have a simple and fast pipeline that is easy to deploy. However, in the long-tailed scenario, this line of work has not been explored so far. In this paper, we investigate whether one-stage detectors can perform well in this case. We discover the primary obstacle that prevents one-stage detectors from achieving excellent performance is: categories suffer from different degrees of positive-negative imbalance problems under the long-tailed data distribution. The conventional focal loss balances the training process with the same modulating factor for all categories, thus failing to handle the long-tailed problem. To address this issue, we propose the Equalized Focal Loss (EFL) that rebalances the loss contribution of positive and negative samples of different categories independently according to their imbalance degrees. Specifically, EFL adopts a category-relevant modulating factor which can be adjusted dynamically by the training status of different categories. Extensive experiments conducted on the challenging LVIS v1 benchmark demonstrate the effectiveness of our proposed method. With an end-to-end training pipeline, EFL achieves 29.2% in terms of overall AP and obtains significant performance improvements on rare categories, surpassing all existing state-of-the-art methods. The code is available at https://github.com/ModelTC/EOD.

1. Introduction

Long-tailed object detection is a challenging task that has received growing attention recently. In the long-tailed scenario, data usually comes with a Zipfian distribution (*e.g.* LVIS [12]) in which a few head classes contain plenty of
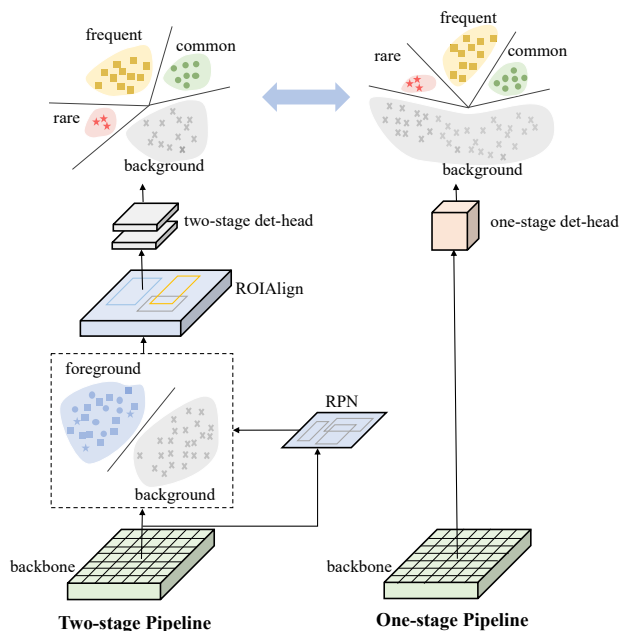
---

*Equal contribution.

Figure 1. Visualization of the training pipeline in one-stage detectors and two-stage detectors. Different shapes indicate different categories, and their corresponding colors indicate the labels of the classifier at different stages.

instances and dominate the training process. In contrast, a significant number of tail classes are instance-scarce thus perform poorly. Common solutions to long-tailed object detection are data resampling [6, 12, 34, 44, 46], decoupled training [18, 26, 44], and loss reweighting [38, 39, 43]. Despite those successes in alleviating the long-tailed imbalance problem, almost all long-tailed object detectors are developed based on two-stage approaches popularized by R-CNN [13, 35]. In practice, one-stage detectors are more appropriate for a realistic scenario than two-stage ones because they are computationally efficient and easy to deploy. However, there is no relevant work in this area.

In contrast to two-stage methods that incorporate a re-

gion proposal network (RPN [35]) to filter out most background samples before feeding proposals to the final classifier, one-stage detectors directly detect objects over a regular, dense set of candidate locations. As presented in Fig. 1, the extreme foreground-background imbalance problem is introduced in one-stage detectors due to the dense prediction schema. Combined with the foreground categories (*i.e.* foreground samples of categories) imbalance problem in the long-tailed scenario, it severely damages the performance of one-stage detectors.

Focal loss [28] is a conventional solution to the foreground-background imbalance problem. It focuses on the learning of hard foreground samples and reduces the impact of easy background samples with a specialized modulating factor. This loss redistribution technique works well under the category-balanced distribution but is inadequate to handle the imbalance problem among foreground categories in the long-tailed situation. To solve this issue, we start from the existing solutions (*e.g.* EQLv2 [38]) in the two-stage pipeline and adapt them to work with focal loss together in one-stage detectors. Unfortunately, we find that these solutions only bring marginal improvements compared with their applications to two-stage detectors (see Tab. 1). Then we argue that simply combining existing solutions with focal loss can not address the two types of imbalance problems simultaneously. By comparing the ratio of positive samples to negative samples in different data distributions (see Fig. 2), we further realize that the essence of these imbalance problems is that the positive-negative imbalance degrees are inconsistent among categories. Rare categories suffer from more severe positive-negative imbalance than frequent ones, thus requiring more emphasis.

In this paper, we propose Equalized Focal Loss (EFL) by introducing a category-relevant modulating factor into focal loss. The modulating factor with two decoupled dynamic factors (*i.e.* the focusing factor and the weighting factor) deals with the positive-negative imbalance of different categories independently. The focusing factor determines the concentration of learning on hard positive samples according to the imbalance degrees of their corresponding categories. The weighting factor increases the impact of rare categories, ensuring that the loss contribution of rare samples will not be overwhelmed by frequent ones. The synergy of these two factors enables EFL to overcome the foreground-background imbalance and the foreground categories imbalance uniformly when applying one-stage detectors to the long-tailed scenario.

We conduct extensive experiments on the challenging LVIS v1 [12] benchmark. With a simple and effective one-stage training pipeline, EFL achieves 29.2% AP, surpassing all existing long-tailed object detection methods. Experimental results on OpenImages [21] also demonstrate the generalization ability of our method.

To sum up, our key contributions can be summarized as follows: (1) We are the first to study one-stage long-tailed object detection. We hope it will inspire the community to rethink the power and the value of one-stage detectors in the long-tailed scenario. (2) We propose a novel Equalized Focal Loss (EFL) that extends the original focal loss with a category-relevant modulating factor. It is a generalized form of focal loss that can address the foreground-background imbalance and foreground categories imbalance simultaneously. (3) We conduct extensive experiments on LVIS v1 benchmark, and the results demonstrate the effectiveness of our approach. It establishes a new state-of-the-art and can be well applied to any one-stage detector.

2. Related Work

2.1. General Object Detection

In recent years, benefiting from the great success of convolutional neural networks (CNN) [14, 15, 17, 20, 31, 37], the computer vision community has witnessed a remarkable improvement in object detection. The modern object detection framework could be roughly divided into two-stage methods and one-stage methods.

Two-stage methods. With the emergence of Faster R-CNN [35], two-stage methods [2,3,10,11,13,35] occupy the dominant position in modern object detection. Two-stage detectors first generate object proposals by a region proposal mechanism (*e.g.* Selective Search [42], or RPN [35]) and then perform spatial extraction of feature maps based on those proposals for further predictions. Thanks to the proposal mechanism, plenty of background samples are filtered out. Following [35], the classifiers of most two-stage detectors are trained over a relatively balanced distribution of foreground and background samples, with a ratio of $1:3$.

One-stage methods. In general, one-stage detectors [5, 7, 22, 28, 30, 32, 33, 41] have a simple and fast training pipeline which are closer to real-world applications. In the one-stage scenario, detectors predict detection boxes from feature maps directly. The classifiers of one-stage detectors are trained over a dense set with about 10^4 to 10^5 candidates, but only a few candidates are foreground samples. Extensive studies [30, 36, 49] attempt to address the extreme foreground-background imbalance problem from the hard example mining view or more complex resampling/reweighing schemes [1]. Focal loss [28] and its derivatives [23–25] reshape the standard cross-entropy loss such that they down-weight the loss assigned to well-classified samples and focus training on hard samples. Benefit from the proposal of focal loss, one-stage detectors achieve very close performance to two-stage methods with a higher inference speed. Recently, there are also some attempts [8,19,47,50] to boost the performance from the label assignment perspective. Our proposed EFL could be ap-

plied well with those one-stage frameworks and bring significant performance gains in the long-tailed scenario.

2.2. Long-Tailed Object Detection

Compared with general object detection, long-tailed object detection is a more complex task since it suffers from an extreme imbalance among foreground categories. A straightforward solution to the imbalance is to perform data resampling during training. Repeat factor sampling (RFS) [12] over-samples the training data from tail classes while under-samples those from head classes from the image level. Wang *et al.* [44] train the detector in a decoupled way [18] and propose an extra classification branch with a class-balanced sampler from the instance level. Forest R-CNN [46] resamples proposals from RPN with different NMS thresholds. Other works [6, 34] implement data resampling through the meta-learning way or memory-augmented way. Loss reweighting is another widely used solution to solve the long-tailed distribution problem. Tan *et al.* [39] propose the Equalization Loss (EQL) that alleviates the gradient suppression of head classes to tail classes. EQLv2 [38] is an upgraded version of EQL which adopts a novel gradient-guided mechanism to reweight the loss contribution of each category. Wang *et al.* [45] interpret the long-tailed distribution problem from a statistic-free perspective and propose the adaptive class suppression loss (ACSL). DisAlign [48] proposes a generalized reweighting method that introduces a balanced class prior to the loss design. Besides data resampling and loss reweighting, many splendid works make attempts from different perspectives such as decoupled training [26, 44], margin modification [34, 43], incremental learning [16], and causal inference [40]. Nevertheless, all those approaches are developed with two-stage object detectors, and there has been no relevant work on one-stage long-tailed object detection so far. In this paper, we propose the first solution to the long-tailed object detection based on the one-stage pipeline. It surpasses all existing methods in a simple and effective way.

3. Methodology

3.1. Revisiting Focal Loss

In one-stage detectors, focal loss [28] is the widely-used solution to the foreground-background imbalance problem. It redistributes the loss contribution of easy samples and hard samples, which greatly weaken the influence of the majority of background samples. The formula of focal loss for binary classification is:

$$\mathrm{FL}\left(p_{\mathrm{t}}\right) = -\alpha_{\mathrm{t}}\left(1 - p_{\mathrm{t}}\right)^{\gamma}\log\left(p_{\mathrm{t}}\right) \qquad (1)$$

As declared in [28], the term $p_{\mathrm{t}} \in [0, 1]$ indicates the predicted confidence score of an object candidate and the term α_{t} is the parameter that balances the importance of
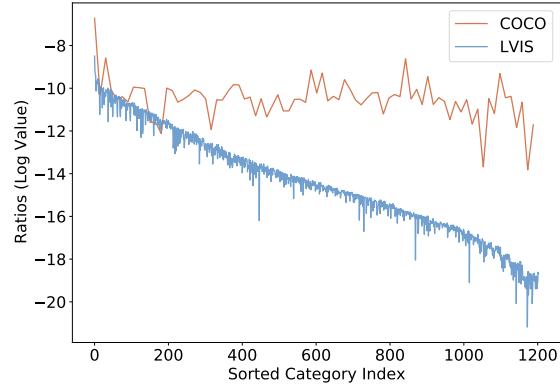

Figure 2. The ratio of the number of positive samples to negative samples on LVIS v1 [12] `train` split and COCO [29] `trainval35k` split. For better visualization, we demonstrate the log value of the ratio and align COCO 80 categories with LVIS 1203 categories. We adopt ATSS [47] as the sample selection strategy to distinguish between foreground samples and background samples.

positive samples and negative samples. The modulating factor $(1 - p_{\mathrm{t}})^{\gamma}$ is the key component of focal loss. It downweights the loss of easy samples and focuses on the learning of hard samples through the predicted score p_{t} and the focusing parameter γ. As mentioned in [23], a large number of negative samples are easy to classify while positive samples are usually hard. Hence, the imbalance between positive samples and negative samples could be roughly regarded as the imbalance between easy samples and hard samples. The focusing parameter γ determines the effect of focal loss. It could be concluded from Eq. (1) that a big γ will greatly reduce the loss contribution from most negative samples, thus improving the impact of positive samples. This conclusion indicates that the higher the imbalance degree between positive samples and negative samples, the bigger the expected value of γ.

When it comes to the multi-class case, focal loss is applied to C classifiers that act on the output logits transformed by the sigmoid function for each instance. C is the number of categories which means that a classifier is responsible for a specific category, *i.e.* a binary classification task. Since focal loss equally treats the learning of all categories with the same modulating factor, it fails to handle the long-tailed imbalance issue (see Tab. 2).

3.2. Equalized Focal Loss Formulation

In the long-tailed dataset (*i.e.* LVIS), in addition to the foreground-background imbalance, the classifier of a one-stage detector also suffers from the imbalance among foreground categories. As shown in Fig. 2, if we view from
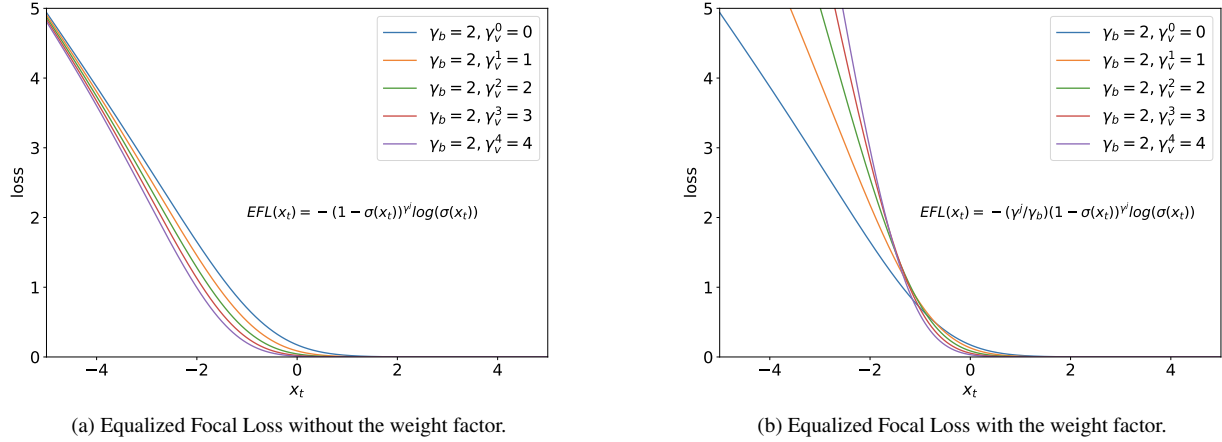
(a) Equalized Focal Loss without the weight factor.

(b) Equalized Focal Loss with the weight factor.

Figure 3. Comparison of the loss contribution between EFL with (b) and without (a) the weighting factor. $x_{\mathrm{t}} = (2y - 1)\,x$, where x is the output predicted logit and $y \in \{0, 1\}$ specific the ground-truth label of the binary classification. σ indicates the sigmoid function. We set the γ_b always to be equal to 2 and ignore the impact of α_{t} in EFL. Different colors indicate different categories.

the y-axis, the values of the ratio between positive samples to negative samples are much less than zero, which mainly reveal the imbalance between foreground and background samples. Here we refer to the value of the ratio as the positive-negative imbalance degree. From the view of the x-axis, we can see that the imbalance degrees across categories show great differences, which indicates the imbalance among foreground categories. Obviously, in the balanced data distribution (*i.e.* COCO), the imbalance degree of all categories are similar. So it is enough to use the same modulating factor for all categories in focal loss. In contrast, those imbalance degrees are different in the long-tailed data situation. Rare categories suffer from more severe positive-negative imbalance than frequent ones. As presented in Tab. 1, most one-stage detectors perform worse on rare categories than on frequent ones. It indicates that the same modulating factor is not suitable for all positive-negative imbalance problems with different degrees.

Focusing Factor. Based on the above analysis, we propose the Equalized Focal Loss (EFL) that adopts a category-relevant focusing factor to address the positive-negative imbalance of different categories separately. We formulate the loss of the j-th category as:

$$\mathrm{EFL}\,(p_{\mathrm{t}}) = -\alpha_{\mathrm{t}}\,(1 - p_{\mathrm{t}})^{\gamma^j}\,\log\,(p_{\mathrm{t}}) \qquad (2)$$

where α_{t} and p_{t} are the same as that in focal loss.

The parameter γ^j is the focusing factor for the j-th category, which plays a similar role as γ in focal loss. As mentioned in Sec. 3.1, different values of the γ correspond to different degrees of positive-negative imbalance issues. We adopt a big γ^j to alleviate the severe positive-negative imbalance problems in rare categories. And for frequent

categories with slight imbalance, a small γ^j is appropriate. The focusing factor γ^j is decoupled into two components, specifically a categories-agnostic parameter γ_b and a categories-specific parameter γ_v^j:

$$\begin{aligned} \gamma^j &= \gamma_b + \gamma_v^j \\ &= \gamma_b + s\left(1 - g^j\right) \end{aligned} \qquad (3)$$

where γ_b indicates the focusing factor in a balanced data scenario that controls the basic behavior of the classifier. And the parameter $\gamma_v^j \geq 0$ is a variable parameter associated with the imbalance degree of the j-th category. It determines how much the concentration of learning is on positive-negative imbalance problems. Inspired by EQLv2 [38], we adopt the gradient-guided mechanism to choose γ_v^j. The parameter g^j indicates the accumulated gradient ratio of positive samples to negative samples of the j-th category. As mentioned in [38], a big value of g^j indicates the j-th category (*e.g.* frequent) is trained balance while a small value indicates the category (*e.g.* rare) is trained imbalanced. To satisfy our requirement of γ^j, we clamp the value of g^j in the range $[0, 1]$ and adopt $1 - g^j$ to invert its distribution. The hyper-parameter s is a scaling factor that determines the upper limit of γ^j in EFL. Compared with focal loss, EFL handles the positive-negative imbalance problem of each category independently, which leads to the performance improvement (see Tab. 3).

Weighting Factor. Even with the focusing factor γ^j, there are still two obstacles damaging the performance: (1) For a binary classification task, a bigger γ is suitable to a more severe positive-negative imbalance problem. While in the multi-class case, as presented in Fig. 3a, for the same x_{t}[†],

---
[†] see the caption in Fig. 3.

the bigger the value of γ, the smaller the loss. It leads to the fact that when we want to increase the concentration on learning a category with severe positive-negative imbalance, we have to sacrifice part of its loss contribution in the overall training process. Such a dilemma prevents rare categories from achieving excellent performance. (2) When x_t is small, the losses of samples from different categories with distinct focusing factors will converge to a similar value. Actually, we expect rare hard samples to make more loss contribution than frequent hard ones since they are scarce and couldn't dominate the training process.

We propose the weighting factor to alleviate the above problems by re-balancing the loss contribution of different categories. Similar to the focusing factor, we assign a big value of the weighting factor for rare categories to raise their loss contributions while keeping the weighting factor close to 1 for frequent categories. Specifically, we set the weighting factor of the j-th category to $\frac{\gamma_b + \gamma_v^j}{\gamma_b}$ to be consistent with the focusing factor. The final formula of EFL is:

$$\text{EFL}\,(p_\text{t}) = -\sum_{j=1}^{C} \alpha_\text{t} \left(\frac{\gamma_b + \gamma_v^j}{\gamma_b} \right) (1 - p_\text{t})^{\gamma_b + \gamma_v^j} \log(p_\text{t})$$

$$(4)$$

As shown in Fig. 3b, with the weighting factor, EFL significantly increases the loss contribution of rare categories. Meanwhile, it focuses more on the learning of rare hard samples compared with frequent hard ones.

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 according to its training status p_t and its corresponding category status γ^j. As presented in Sec. 4.3, both focusing factor and weighting factor play significant roles in EFL. Meanwhile, in the balanced data distribution, EFL with all $\gamma_v^j = 0$ is equivalent to focal loss. Such an appealing property makes EFL could be applied well with different data distributions and data samplers.

4. Experiments

4.1. Experimental Settings

Dataset. We perform our experiments on the challenging LVIS v1 [12] dataset. LVIS is a large vocabulary benchmark for long-tailed object recognition, which contains 1203 categories. Following the common practice [38, 43], all 100k images with about 1.3M instances in the `train` split are used for training, and all 20k images in the `val` split are used as validation for the analysis. The categories in the LVIS v1 dataset are divided into three groups according to the numbers of images that appear in the `train` split: rare categories (1-10 images), common categories (11-100 images), and frequent categories (>100 images).

method	AP	AP_r	AP_c	AP_f
RetinaNet [28]	18.5	9.6	16.1	25.0
+EQLv2&Focal	20.5 (+2.0)	12.6	19.2	25.4
FCOS* [41]	22.6	12.7	20.9	28.9
+EQLv2&Focal	23.0 (+0.4)	14.1	21.3	28.7
PAA [19]	23.7	14.2	21.6	30.2
+EQLv2&Focal	24.1 (+0.4)	16.5	22.1	29.8
ATSS [47]	24.7	13.7	23.4	31.1
+EQLv2&Focal	25.2 (+0.5)	15.0	24.3	30.8
Faster R-CNN [35]	24.1	14.7	22.2	30.5
Seasaw Loss [43]	**26.4 (+2.3)**	**17.5**	**25.3**	**31.5**

Table 1. Results of different one-stage detectors with the combination of focal loss and EQLv2 [38] on LVIS v1 benchmark. EQLv2&Focal indicates this combination. All detectors are trained with ResNet-50-FPN by a 2x schedule using a repeat factor sampler. FCOS* indicates that the reported FCOS result is trained with a center-sampling strategy [47].

Evaluation Metric. The results of object detection are evaluated with the widely-used metric AP of boxes prediction across IoU thresholds from 0.5 to 0.95. Additionally, we report AP_r, AP_c, and AP_f, which indicate the boxes AP for rare, common, and frequent categories, respectively.

Implementation Details. We adopt the ImageNet [4] pre-trained ResNet-50 [14] as the backbone with a Feature Pyramid Network (FPN) [27] as the neck. The network is trained using the SGD algorithm with a momentum of 0.9 and a weight decay of 0.0001. During the training phase, scale jitter and random horizontal flipping are adopted as the data augmentation. We train the model with a total batch size of 16 on 16 GPUs (1 image per GPU) and set the initial learning rate as 0.02. The prior probability of the last layer in the classification branch is initialized to 0.001 as mentioned in [28, 39]. During the inference phase, we resize the shorter edge of the input image to 800 pixels and keep the longer edge smaller than 1333 pixels without changing the aspect ratio. No test time augmentation is used. 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). Following [12], we select the top 300 confident boxes per image as the final detection results. Since most experimental results of two-stage approaches are based on the Mask R-CNN [13] framework, their released detection performance AP_b is affected by the segmentation performance. We report the detection results of those works by rerunning their code with the Faster R-CNN [35] framework for a fair comparison. All models are trained with the repeat factor sampler (RFS) by a 2x schedule. For our proposed EFL, we set the balanced factor $\alpha_\text{t} = 0.25$ and the base focusing parameter $\gamma_b = 2.0$ which

backbone	method	strategy	sampler	epoch	AP	AP_r	AP_c	AP_f
ResNet-50	*two-stage*							
	Faster R-CNN [35]	end-to-end	RFS	24	24.1	14.7	22.2	30.5
	EQL [39]	end-to-end	RFS	24	25.1	15.7	24.4	30.1
	EQLv2 [38]	end-to-end	RFS	24	25.5	16.4	23.9	31.2
	Seasaw Loss [43]	end-to-end	RFS	24	26.4	17.5	25.3	31.5
	cRT [18]	decoupled	RFS+CBS	24+12	24.8	15.9	23.6	30.1
	BAGS [26]	decoupled	RFS+CBS	24+12	26.0	17.2	24.9	31.1
	one-stage							
	RetinaNet [28]	end-to-end	RFS	24	18.5	9.6	16.1	25.0
	Baseline[†]	end-to-end	RFS	24	25.7	14.3	23.8	**32.7**
	EFL (Ours)	end-to-end	RFS	24	**27.5**	**20.2**	**26.1**	32.4
ResNet-101	*two-stage*							
	Faster R-CNN [35]	end-to-end	RFS	24	25.7	15.1	24.1	32.0
	EQLv2 [38]	end-to-end	RFS	24	26.9	18.2	25.4	32.4
	Seasaw Loss [43]	end-to-end	RFS	24	27.8	18.7	27.0	32.8
	BAGS [26]	decoupled	RFS+CBS	24+12	27.6	18.7	26.5	32.6
	one-stage							
	RetinaNet [28]	end-to-end	RFS	24	19.6	10.1	17.3	26.2
	Baseline[†]	end-to-end	RFS	24	27.0	14.4	25.7	**34.0**
	EFL (Ours)	end-to-end	RFS	24	**29.2**	**23.5**	**27.4**	33.8

Table 2. Main results of EFL compared with other methods on LVIS v1 `val` split. 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.

are consistent with focal loss [28]. The hyper-parameter s is set to 8 and more details about the impact of such a scale factor is showcased in Sec. 4.3.

Stabilized Setting and Improved Baseline. As presented in Tab. 1, there is a large performance gap between the two-stage baseline Faster R-CNN [35] and the widely-used one-stage detector RetinanNet [28]. To bridge this gap, we investigate plenty of one-stage frameworks to establish an improved baseline that is more appropriate for the long-tailed task. ATSS [47] stands out among those methods with its simplicity and high performance. From experiments, we discover that the training processes of most one-stage detectors are pretty unstable with result fluctuations and sometimes encountering NaN problems. Intuitively, the primary culprit is the abnormal gradients in the early training stage caused by severe imbalance problems. To stabilize the training processes, we adopt the stabilized settings that extend the warm-up iterations from 1000 to 6000 and utilize the gradient clipping with a maximum normalized value of 35. Meanwhile, in ATSS, we adopt an IoU branch to replace the centerness branch and set the anchor scale from {8} to {6, 8} with the hyper-parameter $k = 18$ to cover more potential candidates. The combination of stabilized and improved settings is adopted as our improved baseline. Unless otherwise stated, EFL is trained with the improved baseline.

4.2. Benchmark Results

To show the effectiveness of our proposed method, we compare our approach with other works that report state-of-the-art performance. As demonstrated in Tab. 2, with ResNet-50-FPN backbone, our proposed method achieves an overall 27.5% AP, which improves the proposed improved baseline by 1.8% AP, and even achieves 5.9 points improvement on rare categories. The result shows that EFL could handle the extreme positive-negative imbalance problem of rare categories well. Compared with other end-to-end methods like EQL [39], EQLv2 [38], and Seesaw Loss [43], our proposed method outperforms them by 2.4% AP, 2.0% AP, and 1.1% AP, respectively. And compared with decoupled training approaches like cRT [18] and BAGS [26], our approach surpasses them with an elegant end-to-end training strategy (by 2.7% AP and 1.5% AP). Besides the high performance, we keep the advantages of one-stage detectors like simplicity, rapidity, and ease of deployment.

With the larger ResNet-101-FPN backbone, our approach still performs well on the improved 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 estab-

WF	FF	AP	AP_r	AP_c	AP_f
		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

Table 3. Ablation study of each component in Equalized Focal Loss. WF, FF indicate the weighting factor and the focusing factor, respectively.

s	AP	AP_r	AP_c	AP_f
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

Table 4. Ablation study of the hyper-parameter s. $s = 8$ is adopted as the default setting in other experiments.

STS	IB	IA	AP	AP_r	AP_c	AP_f
			25.8	18.1	24.5	30.6
✓			26.1	18.6	24.8	30.8
✓	✓		26.8	18.7	25.4	31.9
✓		✓	26.5	20.0	24.8	31.1
✓	✓	✓	**27.5**	**20.2**	**26.1**	**32.4**

Table 5. Ablation study of each component in our proposed improved baseline. STS, IB, IA indicate the stabilized settings, IoU branch, and increased anchor scale, respectively.

lishes a new state-of-the-art. Notably, the performance of rare categories on the improved 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.

4.3. Ablation Studies

Influence of components in EFL. There are two components in EFL, which are the focusing factor and the weighting factor. To demonstrate the effect of each component, we train the model with our proposed improved baseline by a 2x schedule and repeat factor sampler. As shown in Tab. 3, 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 on 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 γ_b among all categories in EFL. Thus the function of the weighting factor could also be regarded as a reweighting approach combined with focal loss. As expected, the weighting factor outperforms the improved baseline by 0.4% AP. By the synergy of these two components, EFL dramatically improves the performance of improved 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 that is also one of the advantages of our work. We study the influence of s with different values and find that $s = 8$ achieves the best performance. As presented in Tab. 4, almost all $s \geq 0$ could increases the performance of the improved baseline. What's more, s keeps relatively high performance in a wide range. It indicates that our EFL is hyper-parameter insensitive.

Influence of components in the improved baseline. We investigate the influence of three components in our proposed improved baseline: the stabilized settings, the IoU branch, and the increased anchor scale. The models are all trained with our proposed EFL. As presented in Tab. 5, the stabilized settings stabilize the training process and bring some performance improvement (+0.3% AP). Combined with the stabilized settings, the IoU branch and the increased anchor scale improve the original EFL by 1.0% AP and 0.7% AP, respectively. Notably, compared with the ATSS in Tab. 1, even without the stabilized and improved settings, our EFL still brings a significant performance improvement (from 24.7% AP to 25.8% AP, +1.1% AP). And combined with those settings, the performance improvement is further increased. Our approaches outperform the improved baseline by 1.8% AP as demonstrated in Tab. 2. The results indicate that the improved baseline is beneficial to the improvement of EFL but not necessary. Meanwhile, the stabilized settings are detector-agnostic that could be applied well with other one-stage detectors. Actually, all RetinaNet and FCOS* experiments are trained with the stabilized settings to avoid the NaN problem.

4.4. Model Analysis

Combined with other one-stage detectors. To demonstrate the generalization ability of EFL across different one-stage detectors, we combine it with RetinaNet, FCOS*, PAA, and our improved baseline, separately. As presented in Tab. 6, EFL performs well with all those one-stage detectors. Compared with the combination of EQLv2 and focal loss in Tab. 1, our proposed EFL maintains a stable big performance improvement (about +2% AP) to the original detector. What's more, EFL dramatically improves the performance of those detectors on rare categories, which shows

method	EFL	AP	AP_r	AP_c	AP_f
RetinaNet [28]		18.5	9.6	16.1	**25.0**
	✓	**20.5 (+2.0)**	**15.2**	**18.6**	24.8
FCOS* [41]		22.6	12.7	20.9	**28.9**
	✓	**23.4 (+0.8)**	**14.9**	**21.9**	28.7
PAA [19]		23.7	14.2	21.6	30.2
	✓	**25.6 +(1.9)**	**19.8**	**23.8**	30.2
Baseline†		25.7	14.3	23.8	**32.7**
	✓	**27.5 +(1.8)**	**20.2**	**26.1**	32.4

Table 6. Results of EFL combined with other one-stage object detectors. FCOS* indicates that the reported FCOS result is trained with a center-sampling strategy [47]. Baseline† indicates our proposed improved baseline.

method	AP	AP1	AP2	AP3	AP4	AP5
R-50 w/ FPN						
Faster R-CNN [35]	43.1	26.3	42.5	45.2	48.2	**52.6**
RetinaNet [28]	32.1	21.0	34.0	35.4	35.6	34.2
Baseline† [47]	43.3	19.4	44.3	49.5	**50.6**	52.2
EFL (Ours)	**51.5**	**52.8**	**52.9**	**50.8**	50.2	50.9
R-101 w/ FPN						
Faster R-CNN [35]	46.0	29.2	45.5	49.3	50.9	**54.7**
RetinaNet [28]	35.8	26.4	38.9	38.3	38.1	36.8
Baseline† [47]	44.7	19.6	46.6	51.1	**52.3**	53.4
EFL (Ours)	**52.6**	**53.4**	**53.8**	**51.4**	51.8	52.3

Table 7. Results on OpenImages Challenge 2019 `val` split. Baseline† indicates our proposed improved baseline.

our strength in settling the long-tailed distribution problem.

A more distinct decision boundary. We investigate whether EFL has more distinct decision boundaries than the focal loss on rare categories classifiers. Since the decision boundaries cannot be displayed intuitively, the empirical margins between positive samples and negative samples among categories are adopted to reflect those boundaries. The margin is calculated by subtracting the mean predicted score of negative samples from the mean predicted score of positive samples for each category. As EFL greatly focuses learning on the positive-negative imbalance problem of rare categories, the margins between positive samples and negative samples of those categories in EFL should be prominent. As presented in Fig. 4, the baseline focal loss keeps small margins between positive samples and negative samples on rare categories. In contrast, our proposed EFL increases the margins of all categories especially for rare categories, resulting in more distinct decision boundaries.
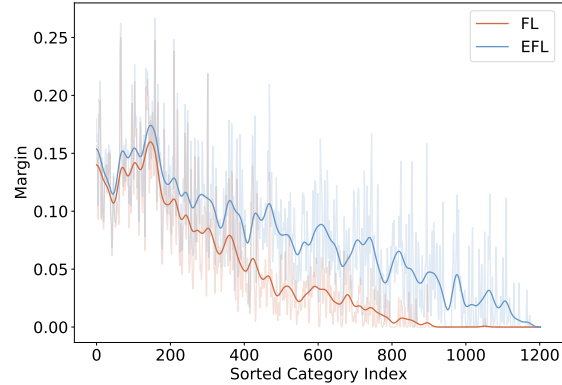

Figure 4. Comparisons of the margins among categories of the focal loss and EFL on LVIS v1 `val` split. The margin is calculated by subtracting the mean predicted score of negative samples from the mean predicted score of positive samples for each category. A larger margin indicates that the classifier has a more distinct decision boundary.

4.5. Results on Open Images Detection

To verify the generalization ability to other datasets, we conduct a series of experiments on the challenging Open-Image dataset [21]. We adopt the Challenge 2019 split as the benchmark. Challenge 2019 is a subset of Open-Images V5, and it contains 500 categories that also suffer from the long-tailed distribution problem. The 1.7M images in the `train` split are used for training while the 41k `val` split is used for validation. It is worth noting that all experimental settings are consistent with them searched in the LVIS v1 benchmark without any tuning. To better understand the improvement on rare categories, following the settings of [38, 39], the categories in Challenge 2019 are divided into five groups (100 categories per group) according to the instance numbers. We adopt the widely-used mAP@IoU=0.5 metric for evaluation and indicate the mAP of the above five groups as AP1 to AP5. Where AP1 is the AP of the first group with the rarest categories and AP5 is the AP of the last group with the most frequent categories. All models are trained by a schedule of 120k/160k/180k with the random sampler. As presented in Tab. 7, our proposed improved baseline greatly improves the performance of one-stage detectors. It brings about 11% AP improvement compared with the widely used RetinaNet. Combined with the improved 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 improved 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 AP1 split. With the larger ResNet-101 backbone, our pro-

model	loss	YOLOX*	AP	AP$_r$	AP$_c$	AP$_f$
small	Sigmoid		15.2	2.9	11.6	24.7
	FL	✓	18.5	3.6	15.7	**28.2**
	EFL(Ours)	✓	**23.3**	**18.1**	**21.2**	28.0
	QFL	✓	22.5	11.0	20.6	**29.7**
	EQFL(Ours)	✓	**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**	**28.2**	**34.7**
	QFL	✓	28.9	16.8	27.2	36.1
	EQFL(Ours)	✓	**31.0**	**24.0**	**29.1**	**36.2**

Table 8. Results of the YOLOX [9] detectors on the LVIS v1 [12] dataset. All experiments are trained from scratch by 300 epochs with the repeat factor sampler (RFS). The YOLOX* indicates the enhanced YOLOX detector that is trained with our proposed improved settings. FL and QFL indicate the focal loss and the quality focal loss, respectively. EFL and EQFL are the methods proposed in this paper that indicate the equalized version of FL and QFL. The scale factor s in EFL and EQFL are all set to 4.

posed 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.

5. Conclusion

In this work, we study how to build a high-performance one-stage object detector in the long-tailed case. We identify the positive-negative imbalance degree inconsistency among categories as the primary difficulty. A novel Equalized Focal Loss (EFL) is proposed to protect the learning of one-stage detectors from extreme imbalance problems. Our proposed EFL is the first solution to the one-stage long-tailed object detection. Combined with some improved techniques and stabilized settings, a strong one-stage detector with EFL beats all existing state-of-the-art methods on the challenging LVIS v1 benchmark.

A. Combined with YOLOX

YOLOX [9] is a recently proposed one-stage detector based on the YOLO series. Its remarkable performance and extremely fast inference speed have won the favor of researchers and developers. In this paper, we investigate whether the advanced YOLOX detector works well under the long-tailed data distribution. Then we introduce our proposed method into the YOLOX detector to help it achieve excellent performance. The experiments are conducted on the small and medium models of YOLOX (YOLOX-S and

YOLOX-M). The challenging LVIS v1 dataset is adopted as the benchmark. All networks are trained from scratch by 300 epochs with the repeat factor sampler (RFS). Unless otherwise stated, our experimental settings are aligned with the original settings in [9][†].

As presented in Tab. 8, both YOLOX-S and YOLOX-M perform poorly in the long-tailed scenario. We argue that the poor performance mainly comes from two aspects. On the one hand, the supervisor (especially the OTA [8] label assignment strategy) in the YOLOX detector is influenced by the long-tailed data distribution which results in low-quality supervision during the training phase. On the other hand, the classification loss in the YOLOX detectors is the sigmoid loss which is incapable to handle the severe positive-negative imbalance degree inconsistency problem (see Sec. 3.2 for more details). Based on these analyses, we make some modifications to the YOLOX detector and apply our proposed method to it. With the following settings, the medium model of YOLOX even achieves an overall AP of 31.0% which indicates the effectiveness of our method:

Enhancements on the YOLOX. Firstly, we replace the supervisor and the predictor of the YOLOX with the settings in our improved baseline (including the ATSS label assignment strategy, IoU branch, and increased anchor scale). The IoU loss combined with the L1 loss is adopted as the localization loss (YOLOX has the same behavior during the last 15 training epochs). We denote the YOLOX detector combined with our improved settings as the YOLOX* series. As shown in Tab. 8, with these enhancements, the YOLOX* outperforms the YOLOX by a large margin (from 20.9% AP to 25.0% AP on the medium model) which indicates that the supervision in the YOLOX* is more reliable than the YOLOX.

Adapt EFL to the YOLOX*. Although the YOLOX* performs better than the YOLOX, its training process is still highly biased towards the frequent categories (AP$_r$ is only 7.1% on the medium model). Thus we adapt our proposed EFL to the YOLOX* to address the long-tailed imbalance issues. It is worth noting that we empirically set the weight decays of bias parameters in the last layer of the classification head to 0.0001 (original setting in the YOLOX is 0), because we discover from experiments that this setting is of vital importance on the performance of EFL. Without this slight modification, the gradient collection mechanism in EFL will malfunction. The hyper-parameter s in our EFL is set to 4 when applied to the YOLOX*. As presented in Tab. 8, combined with the YOLOX* series, EFL achieves excellent performance in the long-tailed situation. On the medium model, it reaches an overall AP of 30.0% that outperforms the YOLOX-M* detector by 5.0% AP. What's more, it greatly improves the performance of the rare cat-

---

[†]We highly recommend readers to refer to https://github.com/Megvii-BaseDetection/YOLOX for more details.
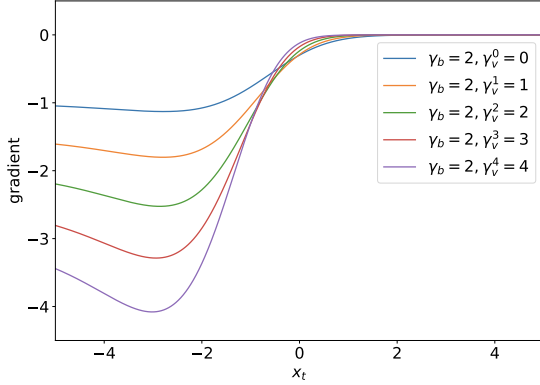
Figure 5. Derivative of our proposed EFL. $x_{\mathrm{t}} = (2y-1)x$, where x is the output predicted logit and $y \in \{0,1\}$ specific the ground-truth label of the binary classification. In this figure, different colors indicate different categories. We set $\gamma_b = 2$ and ignore the impact of α_{t}.

egories with +16.7% AP. The results demonstrate that our proposed EFL is a very practical approach that could greatly alleviate the long-tailed imbalance problem for almost all one-stage detectors.

Equalized Quality Focal Loss. Meanwhile, we also investigate the performance of the quality focal loss (QFL) [25] combined with the YOLOX* series. It could be concluded from experiments that QFL achieves more competitive results compared with the focal loss. We wonder whether the performance of QFL could be further improved by drawing ideas from the EFL. Then the class-relevant modulating factor is designed for the QFL and we denote the novel loss as the equalized quality focal loss (EQFL). The EQFL of the j-th category is formulated as:

$$\mathrm{EQFL}\,(p) = -m_f^j \left(y' \log (p) + (1-y') \log (1-p)\right) \quad (5)$$

where $m_f^j = w_f^j \left(|y'-p|\right)^{f_f^j}$ is the specific form of the modulating factor in EQFL. The weighting factor and the focusing factor (*e.g.* the w_f^j and the f_f^j) are the same as them in EFL. It should be noticed that $y' \in [0,1]$ here is the IoU score for a positive sample and 0 for a negative sample as declared in [25]. Our proposed EQFL achieves 31.0% AP on the medium model. We hope that the impressive performance and powerful generalization ability of our proposed method could inspire the community to raise more attention on one-stage detectors in the long-tailed case.

B. Derivative

The derivative is a crucial part of the parameter g^j in EFL. It could be used to calculate the accumulated gradient of positive samples and negative samples. For reference, the

loss	AP	AP$_{50}$	AP$_{75}$	AP$_s$	AP$_m$	AP$_l$
FL	42.3	61.0	45.7	26.7	46.2	53.0
EFL(s=2)	42.4	61.0	46.1	26.5	46.2	53.2
EFL(s=4)	42.3	61.2	45.6	26.6	46.1	52.9
EFL(s=8)	42.2	60.8	45.6	26.4	46.1	52.7

Table 9. Results in the COCO dataset. All results are came from the improved baseline with the ResNet-50 backbone. The models are trained by a 2x schedule with the random sampler.

derivative for EFL of the j-th category is:

$$\frac{\mathrm{dEFL}}{\mathrm{d}x} = \frac{\gamma^j \,(2y-1)}{\gamma_b} \,(1-p_{\mathrm{t}})^{\gamma^j} \left(\gamma^j p_{\mathrm{t}} \log (p_{\mathrm{t}}) + p_{\mathrm{t}} - 1\right) \quad (6)$$

where $y \in \{0,1\}$ specific the ground-truth label of the binary classification. Plots for the derivatives of different categories are shown in Fig. 5. For all categories, EFL has small derivatives for easy samples ($x_{\mathrm{t}} > 0$). As γ_v^j increases (*e.g.* the category becomes rare), EFL gradually improves the gradient contribution of hard samples, resulting in more concentration on learning them.

In addition to manually calculating the gradients, an alternate approach is to register a backward hook on the classification loss function to get the gradients of positive and negative samples. Meanwhile, when adapting EQLv2 [38] to one-stage detectors, we calculate the gradients of the EQLv2&Focal. EQLv2&Focal is the combination of EQLv2 and focal loss. It directly applies the gradient-guided reweighting mechanism to the focal loss. Here we also show the derivative for EQLv2&Focal:

$$\frac{\mathrm{d}f\,(x)}{\mathrm{d}x} = w_{\mathrm{t}}\,(2y-1)\,(1-p_{\mathrm{t}})^{\gamma} \left(\gamma p_{\mathrm{t}} \log (p_{\mathrm{t}}) + p_{\mathrm{t}} - 1\right) \quad (7)$$

where w_{t} indicates the gradient-guided weight to the focal loss. It is worth noting that the backward hook is the same applies in this situation.

C. Performance on COCO Dataset

As we claim in this paper, EFL is equivalent to the focal loss in the balanced data scenario. To verify this analysis, we conduct experiments on MS COCO dataset [29]. COCO is a widely used object detection dataset that includes 80 categories with a balanced data distribution. The ImageNet pre-trained ResNet-50 is adopted as the backbone, and the networks are trained with our proposed improved baseline. We train the focal loss and our proposed EFL by a 2x schedule with the random sampler. All other settings are consistent with those in LVIS.

As presented in Tab. 9, 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. EFL does not rely on pre-computing the distribution of training data and could operate well with any data sampler. This distribution-agnostic property enables EFL to work well with real-world applications in different data distributions.

References

[1] Samuel Rota Bulo, Gerhard Neuhold, and Peter Kontschieder. Loss max-pooling for semantic image segmentation. In *CVPR*, pages 7082–7091. IEEE, 2017. 2

[2] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *CVPR*, pages 6154–6162, 2018. 2

[3] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *CVPR*, pages 4974–4983, 2019. 2

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 5

[5] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *ICCV*, pages 6569–6578, 2019. 2

[6] Chengjian Feng, Yujie Zhong, and Weilin Huang. Exploring classification equilibrium in long-tailed object detection. *arXiv preprint arXiv:2108.07507*, 2021. 1, 3

[7] Cheng-Yang Fu, Wei Liu, Ananth Ranga, Ambrish Tyagi, and Alexander C Berg. Dssd: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*, 2017. 2

[8] Zheng Ge, Songtao Liu, Zeming Li, Osamu Yoshie, and Jian Sun. Ota: Optimal transport assignment for object detection. In *CVPR*, pages 303–312, 2021. 2, 9

[9] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 9

[10] Ross Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015. 2

[11] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587, 2014. 2

[12] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, pages 5356–5364, 2019. 1, 2, 3, 5, 9

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 1, 2, 5

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2, 5

[15] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018. 2

[16] Xinting Hu, Yi Jiang, Kaihua Tang, Jingyuan Chen, Chunyan Miao, and Hanwang Zhang. Learning to segment the tail. In *CVPR*, pages 14045–14054, 2020. 3

[17] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, pages 4700–4708, 2017. 2

[18] 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. 1, 3, 6

[19] Kang Kim and Hee Seok Lee. Probabilistic anchor assignment with iou prediction for object detection. In *ECCV*, pages 355–371. Springer, 2020. 2, 5, 8

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *NIPS*, 25:1097–1105, 2012. 2

[21] 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. *IJCV*, 128(7):1956–1981, 2020. 2, 8

[22] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *ECCV*, pages 734–750, 2018. 2

[23] Buyu Li, Yu Liu, and Xiaogang Wang. Gradient harmonized single-stage detector. In *AAAI*, volume 33, pages 8577–8584, 2019. 2, 3

[24] Xiang Li, Wenhai Wang, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss v2: Learning reliable localization quality estimation for dense object detection. In *CVPR*, pages 11632–11641, 2021. 2

[25] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *arXiv preprint arXiv:2006.04388*, 2020. 2, 10

[26] 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 *CVPR*, pages 10991–11000, 2020. 1, 3, 6

[27] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 5

[28] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017. 2, 3, 5, 6, 8

[29] 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 *ECCV*, pages 740–755. Springer, 2014. 3, 10

[30] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, pages 21–37. Springer, 2016. 2

[31] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *CVPR*, pages 10428–10436, 2020. 2

[32] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016. 2

[33] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *CVPR*, pages 7263–7271, 2017. 2

[34] Jiawei Ren, Cunjun Yu, Shunan Sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. *arXiv preprint arXiv:2007.10740*, 2020. 1, 3

[35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NIPS*, 28:91–99, 2015. 1, 2, 5, 6, 8

[36] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *CVPR*, pages 761–769, 2016. 2

[37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[38] 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 *CVPR*, pages 1685–1694, 2021. 1, 2, 3, 4, 5, 6, 8, 10

[39] Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan. Equalization loss for long-tailed object recognition. In *CVPR*, pages 11662–11671, 2020. 1, 3, 5, 6, 8

[40] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *arXiv preprint arXiv:2009.12991*, 2020. 3

[41] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *ICCV*, pages 9627–9636, 2019. 2, 5, 8

[42] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *IJCV*, 104(2):154–171, 2013. 2

[43] 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 *CVPR*, pages 9695–9704, 2021. 1, 3, 5, 6

[44] 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 *ECCV*, pages 728–744. Springer, 2020. 1, 3

[45] Tong Wang, Yousong Zhu, Chaoyang Zhao, Wei Zeng, Jinqiao Wang, and Ming Tang. Adaptive class suppression loss for long-tail object detection. In *CVPR*, pages 3103–3112, 2021. 3

[46] Jialian Wu, Liangchen Song, Tiancai Wang, Qian Zhang, and Junsong Yuan. Forest r-cnn: Large-vocabulary long-tailed object detection and instance segmentation. In *ACMMM*, pages 1570–1578, 2020. 1, 3

[47] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *CVPR*, pages 9759–9768, 2020. 2, 3, 5, 6, 8

[48] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *CVPR*, pages 2361–2370, 2021. 3

[49] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z Li. Single-shot refinement neural network for object detection. In *CVPR*, pages 4203–4212, 2018. 2

[50] Benjin Zhu, Jianfeng Wang, Zhengkai Jiang, Fuhang Zong, Songtao Liu, Zeming Li, and Jian Sun. Autoassign: Differentiable label assignment for dense object detection. *arXiv preprint arXiv:2007.03496*, 2020. 2