Supplementary Materials for
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

1. Combined with YOLOX

YOLOX 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 EFL 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 YOLOX (we highly recommend readers to refer to https://github.com/Megvii-BaseDetection/YOLOX for more details).

As presented in Tab. 1, 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 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 (as mentioned in our papers). 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. 1, 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.

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 1. Results of the YOLOX detectors on the LVIS v1 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.

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

[*] Equal Contribution.
[†] Corresponding Author.

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 (the 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. As presented in Tab. 1, 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 categories 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) 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:

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

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 $y\prime \in [0, 1]$ here is the IoU score for a positive sample and 0 for a negative sample. 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.

2. Derivative of EFL

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 derivative for EFL of the j-th category is:

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

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. 1. 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.

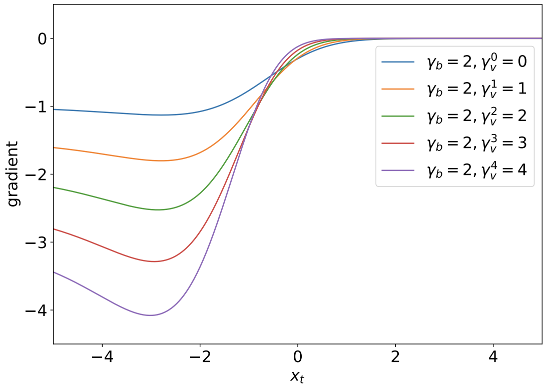

Figure 1. Derivative of our proposed EFL. $x_\mathrm{t} = (2y - 1)x$, where x is the output predicted logit and $y \in \{0, 1\}$ specific the ground-truth label of the binary classification. In this figure, different colors indicate different categories. We set $\gamma_b = 2$ and ignore the impact of α_t.

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

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 to one-stage detectors, we calculate the gradients of the EQLv2* (*i.e.* EQLv2&Focal, the combination of EQLv2 and focal loss). Here we also show the derivative for the EQLv2*:

$$\frac{\mathrm{d}EQLv2^*}{\mathrm{d}x} = w_\mathrm{t} \left(2y - 1\right) \left(1 - p_\mathrm{t}\right)^\gamma \left(\gamma p_\mathrm{t} \log\left(p_\mathrm{t}\right) + p_\mathrm{t} - 1\right) \quad (3)$$

where w_t indicates the gradient-guided weight similar to the weight in EQLv2. It is worth noting that the backward hook is the same applies in this situation.

3. 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,

method	loss	AP	AP$_r$	AP$_c$	AP$_f$
PAA	EQLv2*	24.1	16.5	22.1	29.8
	EFL	**25.6**	**19.8**	**23.8**	**30.2**
ATSS	EQLv2*	25.2	15.0	24.3	**30.8**
	EFL	**25.8**	**18.1**	**24.5**	30.6
Baseline†	EQLv2*	26.8	17.7	25.3	**32.6**
	EFL	**27.5**	**20.2**	**26.1**	32.4

Table 3. Comparison of EQLv2* and EFL, Baseline† is the improved baseline.

we conduct experiments on MS COCO dataset. COCO is a widely used object detection dataset that includes 80 categories with balanced data distribution. The ImageNet pretrained 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. 2, 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.

4. Compared with EQLv2*

We compare the performance of EFL and EQLv2* (*i.e.* EQLv2&Focal) in Tab. 3. Given stronger baselines under different high-performance one-stage detectors, EFL stably achieves a non-trivial improvement compared to the EQLv2*. Especially, for rare categories, EFL consistently outperforms EQLv2* by about 3% AP on these detectors. As mentioned in our paper, EFL focuses more on the learning of categories with extreme positive-negative imbalances. Such property enables EFL to perform well on rare categories. And the impressive improvements of EFL indicate that it has more advantages for addressing the one-stage long-tailed tasks than simply combining EQLv2 with the focal loss.

5. Compared with the NORCAL

We also compare the EFL with the NORCAL which is a recently proposed model calibration method. We adapt the NORCAL to the Faster R-CNN framework trained by a 2x schedule with the repeat factor sampler. As shown in Tab. 4, with the ResNet-101 backbone, EFL outperforms the NORCAL by 1.4% overall AP with 4.1% AP improvement on rare categories. One thing worth noting is that although the NORCAL provides the extension to multiple binary sigmoid classifiers, it performs poorly when combined with the one-stage detectors. We believe that the poor performance comes from the huge number of negative samples under the one-stage framework. All the results indicate that EFL could train the one-stage long-tailed detectors to achieve better representations rather than only adjust the logits of different categories.

backbone	method	AP	AP$_r$	AP$_c$	AP$_f$
ResNet-50	NORCAL	26.6	18.7	25.6	31.1
	EFL	**27.5**	**20.2**	**26.1**	**32.4**
ResNet-101	NORCAL	27.8	19.4	26.9	32.5
	EFL	**29.2**	**23.5**	**27.4**	**33.8**

Table 4. Comparison of NorCal (Faster R-CNN+RFS) and EFL.

6. Qualitative Analysis

The qualitative analysis and the failure cases of the EFL are showcased in Fig. 2. The EFL is compared with the improved baseline (FL) to demonstrate its strength. We only show the detected boxes of rare categories with confidence scores greater than 0.2. It could be concluded from Fig. 2 (top) that the EFL detects the rare categories more accurately with higher confidence scores than the improved baseline. However, there are also some failure cases. For example, many pennants are missed detected by the EFL. And the three tachometers are detected into one. These phenomena suggest that the performance of EFL needs to be further improved in some dense and hard situations. Meanwhile, since the EFL only cares about the classification problems, the localization of some objects are not so precise (*e.g.* the armor and the chap).

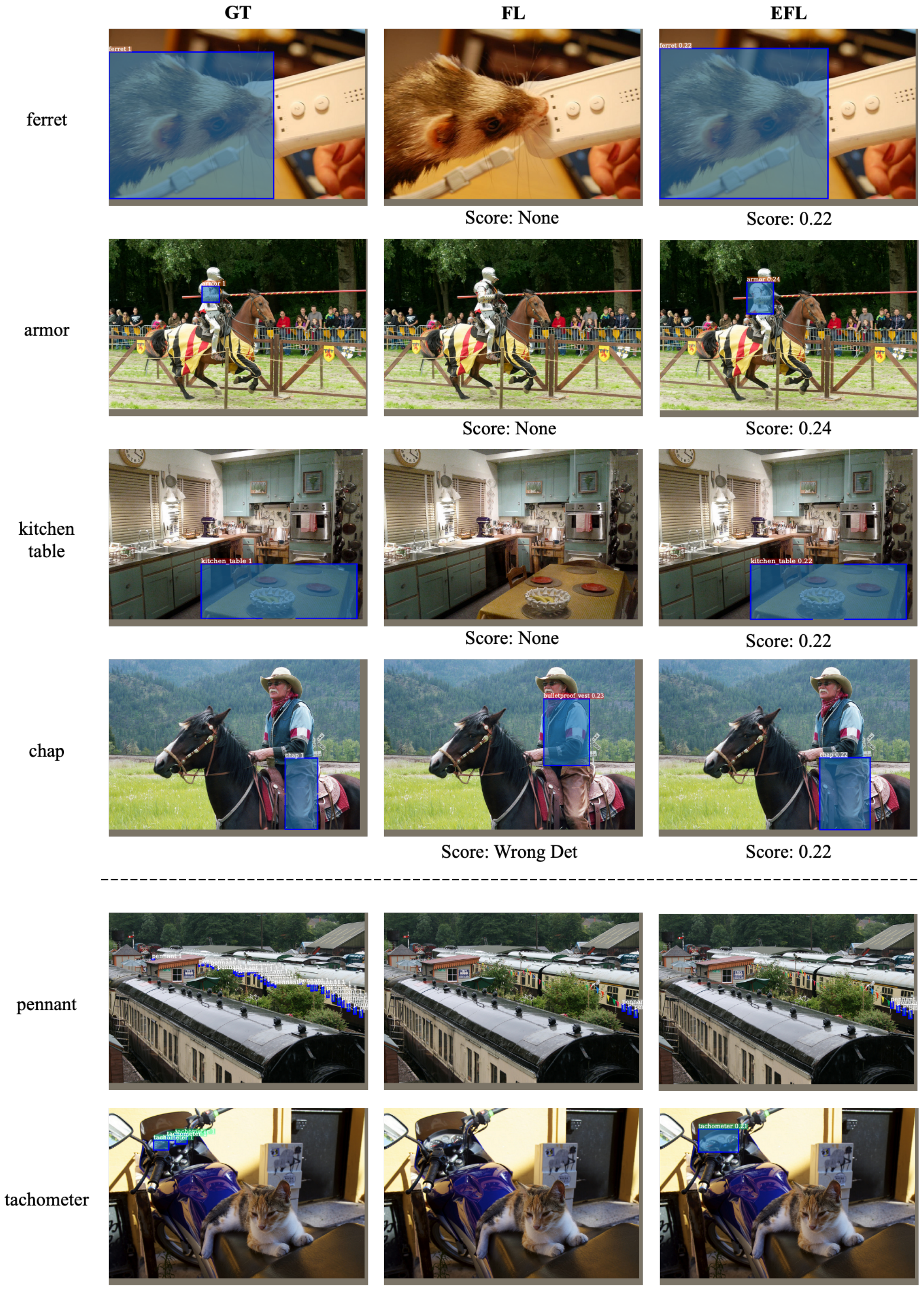

Figure 2. Qualitative Analysis (above the dotted line) and Failure Cases (below the dotted line) of EFL.