Long-tail Detection with Effective Class-Margins Supplementary Materials

Jang Hyun Cho[1] and Philipp Krähenbühl[1]

The University of Texas at Austin, Austin TX 78712, USA
janghyuncho7@utexas.edu, philkr@cs.utexas.edu
https://github.com/janghyuncho/ECM-Loss

1 Pairwise Ranking Error

In this section, we will prove the second equality of Definition 2 of our main paper.

$$\begin{aligned} R_c &= P_{x' \sim D_{\neg c}, x \sim D_c}\left(s_x(x) < s_c(x')\right)) \\ &= 1 - E_{x' \sim D_{\neg c}}\left[P_{x \sim D_c}\left(s_x(x) > s_c(x')\right)\right)] \\ &= E_{x' \sim D_{\neg c}}\left[1 - r_c(s_c(x'))\right] \\ &= \int_0^1 \underbrace{P_{x' \sim D_{\neg c}}(r_c(s_c(x')) < \beta)}_{=g(\beta)} \mathrm{d}\beta = \tau. \end{aligned}$$

where the definition of g is

$$g(\beta) = P_{x' \sim D_{\neg c}}\left(s_c(x') > r_c^{-1}(\beta)\right) = P_{x' \sim D_{\neg c}}\left(r_c(s_c(x')) < \beta\right).$$

The above derivation connects the ranking error to g and the recall.

2 AP - Pairwise Ranking Error Bound

In this section, we will prove Theorem 2 of our main paper. We first derive the and lower bounds to the variational objective $\int_0^1 \frac{x}{x+\alpha g(x)}\mathrm{d}x$ under constraint $\int_0^1 g(x)\mathrm{d}x = \tau$ for a function $g(x) \geq 0$. The AP bounds then directly reduce to the variational objective.

Lemma 1. *Consider the following variational problem*

$$\begin{aligned} &\text{minimize}_g \int_0^1 \frac{x}{x + \alpha g(x)}\mathrm{d}x \\ &\text{subject to} \int_0^1 g(x)\mathrm{d}x = \tau \\ &\qquad g(x) \geq 0 \end{aligned}$$

The solution to this problem is

$$\max\left(1 - \sqrt{\frac{2}{3}\alpha\tau}, \frac{4}{9}\frac{1}{\frac{1}{2} + \alpha\tau}\right)$$

Proof. Consider the associated Euler-Lagrangian equation:

$$L(x, v(x), \lambda) = \int_0^1 \frac{x}{x + \alpha v(x)^2} \mathrm{d}x + \lambda \left(\int_0^1 v(x)^2 \mathrm{d}x - \tau \right)$$

where $g(x) = v(x)^2$ *for the non-negativity constraint. To solve for minima*

$$\begin{aligned}
\frac{d}{dv(x)} L(x, v(x), \lambda) &= -\frac{\alpha x v(x)}{(x + \alpha v(x)^2)^2} + \lambda v(x) = 0 \\
v(x)\alpha x &= \lambda v(x)(x + \alpha v(x)^2)^2 \\
\implies v(x) = 0 \quad &or \quad x + \alpha v(x)^2 = \sqrt{\frac{x\alpha}{\lambda}} = \sqrt{x}\sqrt{\frac{\alpha}{\lambda}} \\
\implies v(x)^2 &= \max\left(0, \sqrt{\frac{x}{\alpha\lambda}} - \frac{x}{\alpha}\right) = 1_{[x \leq \frac{\alpha}{\lambda}]} \left(\sqrt{\frac{x}{\alpha\lambda}} - \frac{x}{\alpha} \right) \\
\implies \int_0^1 \alpha v(x)^2 \mathrm{d}x = \alpha\tau &= \int_0^1 1_{[x \leq \frac{\alpha}{\lambda}]} \left(\sqrt{\frac{\alpha x}{\lambda}} - x \right) \mathrm{d}x \\
&= \int_0^\kappa \left(\sqrt{\frac{\alpha x}{\lambda}} - x \right) \mathrm{d}x = \frac{2}{3} \kappa^{\frac{3}{2}} \sqrt{\frac{\alpha}{\lambda}} - \frac{\kappa^2}{2}
\end{aligned}$$

where $\kappa = \min(1, \frac{\alpha}{\lambda})$*. For* $\kappa = 1$*:*

$$\begin{aligned}
\alpha\tau = \frac{2}{3}\sqrt{\frac{\alpha}{\lambda}} - \frac{1}{2} &\implies \sqrt{\frac{\alpha}{\lambda}} = \frac{3}{2}\left(\alpha\tau + \frac{1}{2}\right) \\
\implies x + \alpha v(x)^2 &= \sqrt{x}\sqrt{\frac{\alpha}{\lambda}} = \sqrt{x}\frac{3}{2}\left(\alpha\tau + \frac{1}{2}\right) \\
\implies \int_0^1 \frac{x}{x + \alpha v(x)^2} \mathrm{d}x &= \int_0^1 \frac{x}{\sqrt{x}\frac{3}{2}(\alpha\tau + \frac{1}{2})} \mathrm{d}x = \int_0^1 \sqrt{x}\mathrm{d}x \frac{1}{\frac{3}{2}(\frac{1}{2} + \alpha\tau)} \\
&= \frac{4}{9} \frac{1}{\frac{1}{2} + \alpha\tau}
\end{aligned}$$

For $\kappa < 1$:

$$\begin{aligned}
\alpha\tau = \frac{2}{3}\left(\frac{\alpha}{\lambda}\right)^{\frac{3}{2}}\sqrt{\frac{\alpha}{\lambda}} - \frac{1}{2}\left(\frac{\alpha}{\lambda}\right)^{2} = \frac{1}{6}\left(\frac{\alpha}{\lambda}\right)^{2} \implies \lambda = \frac{1}{6}\sqrt{\frac{\alpha}{\tau}} \\
\implies \int_0^1 \frac{x}{x+\alpha v(x)^2}\mathrm{d}x = \int_0^\kappa \frac{x}{\sqrt{\frac{\alpha x}{\lambda}}}\mathrm{d}x + \int_\kappa^1 \frac{x}{x+0}\mathrm{d}x \\
= \sqrt{\frac{\lambda}{\alpha}}\int_0^\kappa \sqrt{x}\mathrm{d}x + \int_\kappa^1 \mathrm{d}x \\
= \frac{2}{3}\sqrt{\frac{\lambda}{\alpha}}\kappa^{\frac{3}{2}} + 1 - \kappa \\
= \frac{2}{3}\sqrt{\frac{\lambda}{\alpha}}\left(\frac{\alpha}{\lambda}\right)^{\frac{3}{2}} + 1 - \frac{\alpha}{\lambda} \\
= 1 - \frac{1}{3}\frac{\alpha}{\lambda} \\
= 1 - \sqrt{\frac{2}{3}\alpha\tau}
\end{aligned}$$

Each case yields on lower bound, hence the combined lower bound is

$$\max\left(1 - \sqrt{\frac{2}{3}\alpha\tau}, \frac{4}{9}\frac{1}{\frac{1}{2} + \alpha\tau}\right)$$

□

Bonus: The two bounds meet at $\frac{2}{3}$:

$$\frac{4}{9}\frac{1}{\frac{1}{2}+\alpha\tau} = 1 - \sqrt{\frac{2}{3}\alpha\tau} = \frac{2}{3} \quad \text{for} \quad \alpha\tau = \frac{1}{6}.$$

Lemma 2. *Consider the following variational problem*

$$\begin{aligned}
&\text{maximize}_g \int_0^1 \frac{x}{x+\alpha g(x)}\mathrm{d}x \\
&\text{subject to} \int_0^1 g(x)\mathrm{d}x = \tau \\
&g(x) \geq 0
\end{aligned}$$

Proof. First, let us re-formulate the problem as following

$$\begin{aligned}
&\text{minimize}_g \int_0^1 \frac{\alpha g(x)}{x+\alpha g(x)}\mathrm{d}x \\
&\text{subject to} \int_0^1 g(x)\mathrm{d}x = \tau \\
&g(x) \geq 0
\end{aligned}$$

Without the equality constraint $\int_0^1 g(x)\mathrm{d}x = \tau$*, the objective is minimized at* $g(x) = 0$ *for all* $x \in [0, 1]$*. The equality constraint assigns certain values* $g(x)$ *a positive mass. The optimal solution will assign* $g(x) = 0$ *for* $x < 1 - \tau$*, and* $g(x) = 1$ *for* $x \geq 1 - \tau$*. To see this, consider a value* $g(x_1) = \frac{\epsilon}{\alpha}$ *for* $x_1 < 1 - \tau$ *and one or move values* $g(x_2) \leq 1 - \frac{\epsilon}{\alpha}$ *for* $x_2 > 1 - \tau$*. Here, a solution* $\hat{g}(x_1) = 0$ *and* $\hat{g}(x_2) = g(x_2) + \frac{\epsilon}{\alpha}$ *has a lower objective*

$$\begin{aligned}
\Delta &= \left(\frac{\alpha g(x_1)}{x_1 + \alpha g(x_1)} + \frac{\alpha g(x_2)}{x_2 + \alpha g(x_2)}\right) - \left(\frac{\alpha \hat{g}(x_1)}{x_1 + \alpha \hat{g}(x_1)} + \frac{\alpha \hat{g}(x_2)}{x_2 + \alpha \hat{g}(x_2)}\right) \\
&= \left(\frac{\alpha g(x_1)}{x_1 + \alpha g(x_1)} + \frac{\alpha g(x_2)}{x_2 + \alpha g(x_2)}\right) - \frac{\alpha g(x_2) + \epsilon}{x_2 + \alpha g(x_2) + \epsilon} \\
&= \frac{\epsilon}{x_1 + \epsilon} - \frac{\epsilon x_2}{(x_2 + \alpha g(x_2))(x_2 + \alpha g(x_2) + \epsilon)} > 0
\end{aligned}$$

Here $\Delta > 0$ *and the new objective is lower since* $x_2 + \alpha g(x_2) > x_2$ *and* $x_2 > x_1$ *thus* $(x_2 + \alpha g(x_2))(x_2 + \alpha g(x_2) + \epsilon) > x_2(x_1 + \epsilon)$*.*

Thus the zero-mass region should be where x *is low as lower* x *increases the objective. Hence, the optimality will happen when* $g(x) = 0$ *for* $x \in [0, 1 - \tau]$*, and* $g(x) = 1$ *for* $x \in [1 - \tau, 1]$*. Thus:*

$$\begin{aligned}
\int_{1-\tau}^{1} \frac{\alpha}{\alpha + x}\mathrm{d}x &= \alpha \int_{1-\tau}^{1} \frac{1}{\alpha + x}\mathrm{d}x = \alpha(\log(1 + \alpha) - \log(1 - \tau + \alpha)) \\
&= -\alpha \log\left(1 - \frac{\tau}{1 + \alpha}\right) \\
\implies \max_g \int_0^1 \frac{x}{x + \alpha g(x)}\mathrm{d}x &= 1 + \alpha \log\left(1 - \frac{\tau}{1 + \alpha}\right)
\end{aligned}$$

which concludes the proof. □

Lemma 1 and Lemma 2 for the bounds to the AP.

Theorem 1. *Average Precision can be bounded from above and below as following*

$$1 + \alpha_c \log\left(1 - \frac{R_c}{1 + \alpha_c}\right) \geq AP_c \geq \max\left(1 - \sqrt{\frac{2}{3}\alpha_c R_c}, \frac{8}{9}\frac{1}{1 + 2\alpha_c R_c}\right) \tag{1}$$

Proof. Let us recap the definitions of AP and R:

$$\begin{aligned}
AP_c &= \int_0^1 \frac{\beta}{\beta + \alpha_c P_{x \sim D_{\neg c}}(s_c(x) > r_c^{-1}(\beta))}\mathrm{d}\beta \\
&= \int_0^1 \frac{\beta}{\beta + \alpha_c \underbrace{P_{x \sim D_{\neg c}}(r_c(s_c(x)) < \beta)}_{=g(\beta)}}\mathrm{d}\beta \\
R_c &= \int_0^1 \underbrace{P_{x' \sim D_{\neg c}}(r_c(s_c(x')) < \beta)}_{=g(\beta)}\mathrm{d}\beta = \tau
\end{aligned}$$

where the second line of AP_c is because r_c is strictly monotonously decreasing. With $x = \beta$ and $g(x) = P_{x\sim D_{\neg c}}(r_c(s_c(x)) < \beta)$, Lemma 1 and Lemma 2 are directly applicable for a function $0 \leq g(x) \leq 1$ with a fixed $R_c = \tau$. The corresponding upper and lower bounds of $\mathcal{L}_c^{Det}$ in the main paper is a direct consequence of this theorem since $\mathcal{L}_c^{Det} = 1 - AP_c$. □

3 Optimal Margins

Similar to Cao et al. [2], we aim to find optimal binary margins γ_+ and γ_- under separability condition. This reduces the problem into following:

$$\text{minimize}_{\gamma_+,\gamma_-} \quad \frac{1}{\gamma_+}\sqrt{\frac{1}{n_+}} + \frac{1}{\gamma_-}\sqrt{\frac{1}{n_-}} \tag{2}$$

$$\text{subject to} \quad \gamma_+ + \gamma_- = 1 \tag{3}$$

Here the constraint is due to the fact that $s_-(x) = 1 - s_+(x)$ in binary case and thus $\gamma_- = 1 - \gamma_+$. Solving the constrained optimization problem

$$L(\gamma_+, \gamma_-, \lambda) = \frac{1}{\gamma_+}\sqrt{\frac{1}{n_+}} + \frac{1}{\gamma_-}\sqrt{\frac{1}{n_-}} + \lambda(\gamma_+ + \gamma_- - 1) \tag{4}$$

$$\implies \frac{\partial}{\partial\gamma_+}L(\gamma_+, \gamma_-, \lambda) = -\frac{1}{{\gamma_+}^2}\sqrt{\frac{1}{n_+}} + \lambda = 0 \tag{5}$$

$$\implies \gamma_+ = \sqrt{\frac{{n_+}^{-\frac{1}{2}}}{\lambda}}, \quad \gamma_- = \sqrt{\frac{{n_-}^{-\frac{1}{2}}}{\lambda}} \tag{6}$$

$$\implies \frac{\partial}{\partial\lambda}L(\gamma_+, \gamma_-, \lambda) = \gamma_+ + \gamma_- - 1 = 0 \tag{7}$$

$$\implies \gamma_+ + \gamma_- = \sqrt{\frac{{n_+}^{-\frac{1}{2}}}{\lambda}} + \sqrt{\frac{{n_-}^{-\frac{1}{2}}}{\lambda}} = 1 \tag{8}$$

$$\implies \sqrt{\lambda} = \frac{\sqrt{{n_+}^{-\frac{1}{2}}} + \sqrt{{n_-}^{-\frac{1}{2}}}}{1} \tag{9}$$

$$\implies \gamma_+ = \frac{{n_+}^{-\frac{1}{4}}}{{n_+}^{-\frac{1}{4}} + {n_-}^{-\frac{1}{4}}} = \frac{{n_-}^{\frac{1}{4}}}{{n_+}^{\frac{1}{4}} + {n_-}^{\frac{1}{4}}} \tag{10}$$

$$\gamma_- = \frac{{n_-}^{-\frac{1}{4}}}{{n_+}^{-\frac{1}{4}} + {n_-}^{-\frac{1}{4}}} = \frac{{n_+}^{\frac{1}{4}}}{{n_+}^{\frac{1}{4}} + {n_-}^{\frac{1}{4}}} \tag{11}$$

which are as desired. The exact same process can be repeated for each class $c \in C$ and we will have our Effective Class-Margins. □

4 Surrogate Scoring Function

In this section, we will justify the choice of our surrogate scoring function.

$$\hat{s}_c(x) = \frac{w_c^+ e^{f(x)_c}}{w_c^+ e^{f(x)_c} + w_c^- e^{-f(x)_c}} \tag{12}$$

The decision boundary is then

$$\hat{s}_c(x) = \hat{s}_{\neg c}(x) = 1 - \hat{s}_c(x) \tag{13}$$

$$\implies \frac{w_c^+ e^{f(x)_c}}{w_c^+ e^{f(x)_c} + w_c^- e^{-f(x)_c}} = \frac{w_c^- e^{-f(x)_c}}{w_c^+ e^{f(x)_c} + w_c^- e^{-f(x)_c}} \tag{14}$$

$$\implies \log w_c^+ + f(x)_c = \log w_c^- - f(x)_c \tag{15}$$

$$\implies f(x)_c = \frac{1}{2}(\log w_c^- - \log w_c^+) \tag{16}$$

In the unweighted sigmoid function, this point will lie at

$$s_c(x) = \frac{e^{f(x)_c}}{e^{f(x)_c} + e^{-f(x)_c}} \tag{17}$$

$$= \frac{e^{\frac{1}{2}(\log w_c^- - \log w_c^+)}}{e^{\frac{1}{2}(\log w_c^- - \log w_c^+)} + e^{\frac{1}{2}(\log w_c^+ - \log w_c^-)}} \tag{18}$$

$$= \frac{\sqrt{\frac{w_c^-}{w_c^+}}}{\sqrt{\frac{w_c^-}{w_c^+}} + \sqrt{\frac{w_c^+}{w_c^-}}} = \frac{\sqrt{\frac{\gamma_c^+}{\gamma_c^-}}}{\sqrt{\frac{\gamma_c^+}{\gamma_c^-}} + \sqrt{\frac{\gamma_c^-}{\gamma_c^+}}} = \frac{\gamma_c^+}{\gamma_c^+ + \gamma_c^-} \tag{19}$$

$$= \gamma_c^+ \tag{20}$$

$$\implies s_{\neg c}(x) = \gamma_c^- \tag{21}$$

since $\gamma_c^+ + \gamma_c^- = 1$. Hence, we have shown that our surrogate scoring function with effective class-margins shifts the decision boundary of sigmoid function to γ_c^+ and γ_c^- as desired. □

5 Margins vs Weights vs Gradients

In this section, we provide more intuitions about the relationship between margins, weights, and gradients. The gradient of positive and negative samples with ECM Loss is the following:

$$\frac{\partial}{\partial f(x)_c} \ell_{\text{ECM}}(x, y) = \frac{2 w_{\neg c} e^{f(x)_{\neg c}}}{w_c e^{f(x)_c} + w_{\neg c} e^{f(x)_{\neg c}}} \propto w_{\neg c} = \frac{1}{\gamma_{\neg c}} \propto n_{\neg c} \tag{22}$$

$$\frac{\partial}{\partial f(x)_{\neg c}} \ell_{\text{ECM}}(x, y) = \frac{2 w_c e^{f(x)_c}}{w_c e^{f(x)_c} + w_{\neg c} e^{f(x)_{\neg c}}} \propto w_c \ = \frac{1}{\gamma_c} \ \propto n_c \tag{23}$$

where we omit detection weight m_c for simplicity. The positive gradient is greater for rare classes (higher $n_{\neg c}$) compared to frequent classes, whereas the negative

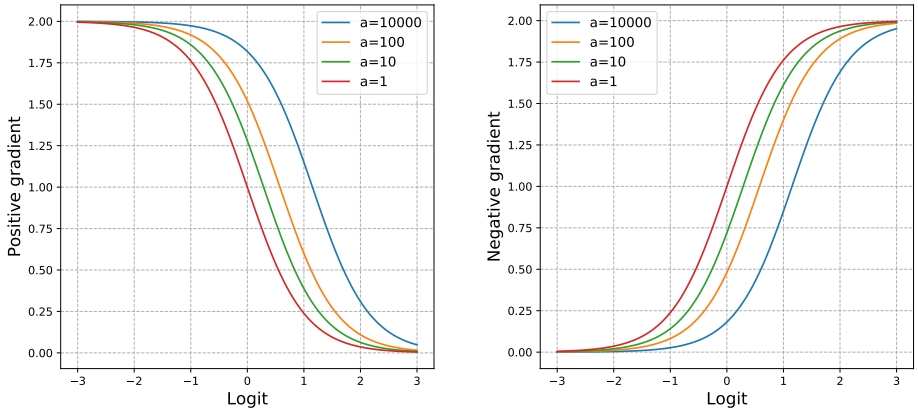


Fig. 1: Visualization of the positive and negative gradients from ECM Loss with different positive and negative samples ratios.

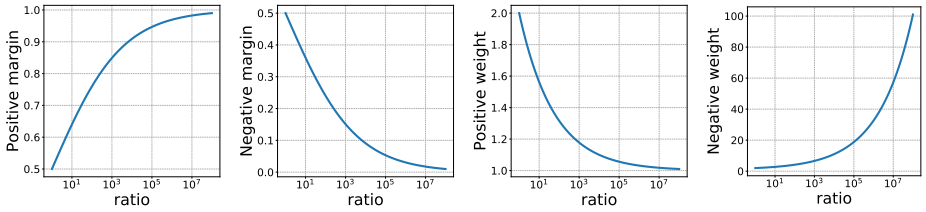
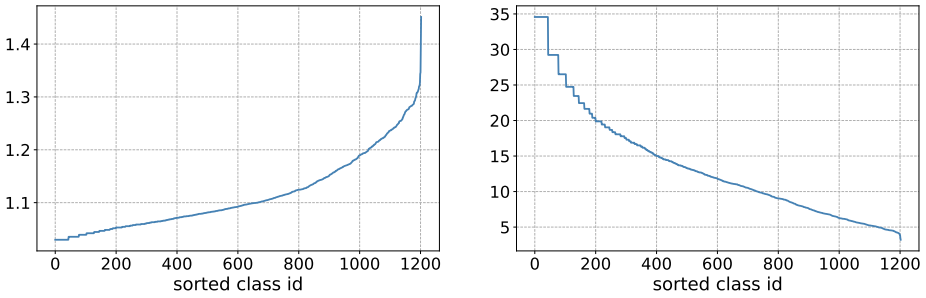


Fig. 2: Visualization of the positive and negative margins and weights as a function of the sample ratio α_c.

gradient is lower. This coincides with the intuitions from prior works based on heuristics or indirect measure of a model. Below, we show visualization of positive and negative gradients as a function of logit with different positive and negative ratios $a_c = \frac{n_{\neg c}}{n_c}$. In Figure 1, low a means frequent classes whereas high a means rare classes. Our surrogate scoring function $\hat{s}_c$ balances the gradient values based on the frequency of each class. Frequent classes get lower positive gradient and higher negative gradient (red in 1) whereas rare classes get higher positive gradient and lower negative gradient (blue in 1). In Figure 2, we further visualize the relationship between the positive and negative margins and weights as a function of the ratio α_c.

In Figure 3, we visualize the computed weights for our scoring function $\hat{s}_c$ in Equation (11) of the main paper for w_c^+ (left) and w_c^- (right).



(a) w_c^+ (b) w_c^-

Fig. 3: Visualization of the computed the positive and negative weights used in our scoring function $\hat{s}_c$ for each class $c \in C$. The prior distribution is from LVIS v1 training annotations over 1203 classes. The weights are sorted in ascending order and background probability is measured with Mask R-CNN with ResNet-50 backbone.

6 Implementation Details

In this section, we will discuss details of the experiments and implementation.

Background count. We empirically measure the frequency of background samples as a ratio of foreground and background samples within a batch, r, in the *classification layer* of a detector during the first few iterations. This ratio will then be used to derive *dataset-level* count of background samples as

$$n_{bg} = r \sum_{c \in C} n_c \tag{24}$$

where n_c is the number of positive samples of class c in the training dataset. Then, we compute the sample count of each class *with background* as

$$n_c^+ = n_c \tag{25}$$

$$n_c^- = \Big(\sum_{c' \in C \cup \{bg\}} n_{c'} \Big) - n_c \tag{26}$$

and use it to compute the effective class-margins. For one-stage detectors, we only count for foreground classes as foreground and background imbalance is managed from focal weight [9].

Two-stage detectors. We train two-stage instance segmentation models based on Mask R-CNN [6] and Cascade Mask R-CNN [1] with various backbones, ResNet-50 and ResNet-101 [7] with Feature Pyramid Network [8] pretrained on ImageNet-1K [4], Swin Transformer [10] pretrained on ImageNet-21K with

224x224 image resolution. We train with for 12 or 24 epochs with Repeated Factor Sampler (RFS) [5] on a 1× or 2× schedule respectively. For CNN backbones, we use the SGD optimizer with 0.9 momentum, initial learning rate of 0.02, weight decay of 0.0001, with step-wise scheduler decaying learning rate by 0.1 after 8 and 11 epochs for 1x, and 20 and 22 epochs for 2x, and batch size of 16 on 8 GPUs. Please note that the baseline methods are trained with their optimal learning rate schedule with decaying schedule of 16 and 22 epochs. For example, Mask R-CNN with ResNet-50 trained with Seesaw Loss [12] on decaying schedule of 20 and 22 epochs result with 26.7 mAP_{segm} and 26.9 mAP_{bbox}, whereas decaying schedule of 16 and 22 (default) result with 26.7 mAP_{segm} and 27.3 mAP_{bbox}. For Transformer backbones, we use the AdamW optimizer with an initial learning rate of 0.00005, beta set to $(0.9, 0.999)$, weight decay of 0.05, with Cosine-annealing scheduler. For all our models, we follow the standard data augmentation during training: random horizontal flipping and multi-scale image resizing to fit the shorter side of image to $(640, 672, 704, 736, 768, 800)$ at random, and the longer side kept smaller than 1333. For Swin Transformer, we use a larger range of scale of the short side of image from 480 to 800. For two-stage detectors, we normalize the classification layers with temperature $\tau = 20$ for both box and mask classifications, and apply foreground calibration as post-process following prior practices [3,12,15]. LVIS has more instances than COCO. We thus increase the per-image detection limit to 300 from 100 and set the confidence threshold to 0.0001. This is common practice in LVIS [5]. For OpenImages, we train Cascade R-CNN with ResNet-50 backbone following the baseline and experimental setup of Zhou *et al.* [16]. All models in this experiment were trained for 180k iterations with a class-aware Sampler.

One-stage detectors. For one-stage detectors, Focal Loss [9] is the standard choice of the loss function. It effectively diminishes loss values for "easy" samples such as the background. In this experiment, we test the compatibility of our method with Focal weights. Specifically, we apply the computed focal weights to our ECM Loss. Instead of a binary cross-entropy on the surrogate scoring function $\hat{s}$, we use the focal loss. We use a number of popular one-stage detectors: FCOS [11], ATSS [14] and VarifocalNet [13]. Each method uses either Focal Loss or a variant [13]. We use the default hyperparameter for all types of focal weights: $\gamma = 2, \alpha = 0.25$ for Focal Loss, $\gamma = 2$ and $\alpha = 0.75$ for Varifocal Loss [13]. In LVIS v1.0 models expect to see more instances. We thus double the per-pyramid level number of anchor candidates from 9 to 18 for ATSS and VarifocalNet. Similar to 2-stage detectors, we increased per-image detection to 300 and set the confidence threshold to 0.0001. We train on ResNet-50 and ResNet-101 backbones for 12 epochs for 1x and 24 epochs for 2x, batch size of 16 on 8 GPUs. We set the learning rate to be 0.01 which was the optimal learning rate for the baselines. For all other settings, we follow the two-stage experiments.

References

1. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6154–6162 (2018) 8
2. Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss. arXiv preprint arXiv:1906.07413 (2019) 5
3. Dave, A., Dollár, P., Ramanan, D., Kirillov, A., Girshick, R.: Evaluating large-vocabulary object detectors: The devil is in the details. arXiv preprint arXiv:2102.01066 (2021) 9
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848 8
5. Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5356–5364 (2019) 9
6. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017) 8
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 8
8. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017) 8
9. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017) 8, 9
10. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. International Conference on Computer Vision (ICCV) (2021) 8
11. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9627–9636 (2019) 9
12. Wang, J., Zhang, W., Zang, Y., Cao, Y., Pang, J., Gong, T., Chen, K., Liu, Z., Loy, C.C., Lin, D.: Seesaw loss for long-tailed instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9695–9704 (2021) 9
13. Zhang, H., Wang, Y., Dayoub, F., Sunderhauf, N.: Varifocalnet: An iou-aware dense object detector. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8514–8523 (2021) 9
14. Zhang, S., Chi, C., Yao, Y., Lei, Z., Li, S.Z.: Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9759–9768 (2020) 9
15. Zhang, S., Li, Z., Yan, S., He, X., Sun, J.: Distribution alignment: A unified framework for long-tail visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2361–2370 (2021) 9
16. Zhou, X., Koltun, V., Krähenbühl, P.: Simple multi-dataset detection. In: arXiv preprint arXiv:2102.13086 (2021) 9