# Adaptive Hierarchical Representation Learning for Long-Tailed Object Detection
# Supplementary Materials

Banghuai Li
MEGVII Technology
libanghuai@gmail.com

## A. Other Effective Attempts in Baseline++

### A.1. Data Augmentation

When facing scarce instances of tailed classes, data augmentation is the most straightforward method to battle against data hunger and alleviate overfitting. After trying several simple data augmentation methods, we found that zooming up or down images, randomly cropping the instance and color jitter positively affect the result.

### A.2. Regression Loss Function

Object detection is the combination of classification and regression, and the regression loss function always plays an vital role in better localization performance. Nowadays, IoU loss [13] is widely used in object detection for box regression. For a prediction box $A$ and the corresponding ground-truth box $B$, IoU loss is defined as follows:

$$\mathcal{L}_{IoU} = IoU = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

Compared with IoU loss function, GIoU [9] loss, another excellent box regression loss function, is feasible to optimize in the case of non overlapping bounding boxes and sensitive to the alignment of proposed boxes and ground-truth boxes. Thus, we adopt GIoU as our box regression loss function for better performance:

$$\mathcal{L}_{GIoU} = IoU - \frac{|C \backslash (A \cup B)|}{|C|} \tag{2}$$

where $C$ is the smallest enclosing convex box of A and B.

### A.3. Loss Balance

Loss balance technique is a common attempt towards long-tailed object detection problem. Among these approaches [1, 8, 10], we adopt EQL [10], a simple yet effective method, to address loss balance in long-tailed object detection. EQL [10] simply ignores the suppression to tailed classes when they act as the negative samples, aiming to make the network training fairer for each class. EQL is formulated as follows:

$$\hat{p}_j = \begin{cases} 1 - p_j, & y_j \neq 1 \\ p_j, & y_j = 1 \end{cases} \tag{3}$$

$$\mathcal{L}_{EQL} = -\sum_j (1 - E(r) T_{\lambda_r}(f_j)) log(\hat{p}_j) \tag{4}$$

where $r$ denotes a given region proposal, $E(r)$ outputs 1 when $r$ is a foreground region proposal, $f_j$ is the frequency of category $j$ in the dataset, $T_{\lambda_r}(f_j)$ will be 1 only if $f_j$ is larger than $\lambda_r$. Please refer to [10] for more details.

## B. Experiment Details

### B.1. Datasets

Large Vocabulary Instance Segmentation(LVIS) dataset, a large long-tailed vocabulary dataset in long-tailed detection, consists of 1230 categories in v0.5 and 1203 categories in v1.0. Since LVIS is a federated dataset [2], a few annotations are missing and few annotations are ambiguous. All categories are officially divided into three groups: frequent(more than 100 images), common(10 to 100 images), and rare(less than 10 images). Following the official guideline, we train our model on the train set and evaluate the result on the val set. Besides widely-used $AP$ across IoU threshold from 0.5 to 0.95, AP for frequent($AP_f$), common($AP_c$), rare($AP_r$) groups will be reported respectively for both object detection and instance segmentation results.

### B.2. Implementation Details

We use Mask R-CNN [3] as our base detector and ResNet-50 [4] with a Feature Pyramid Network [7] as the backbone. We use 8 GPUs with a batch size 16 for training. Our model is trained using stochastic gradient descent(SGD) with momentum 0.9 and weight decay 0.0001

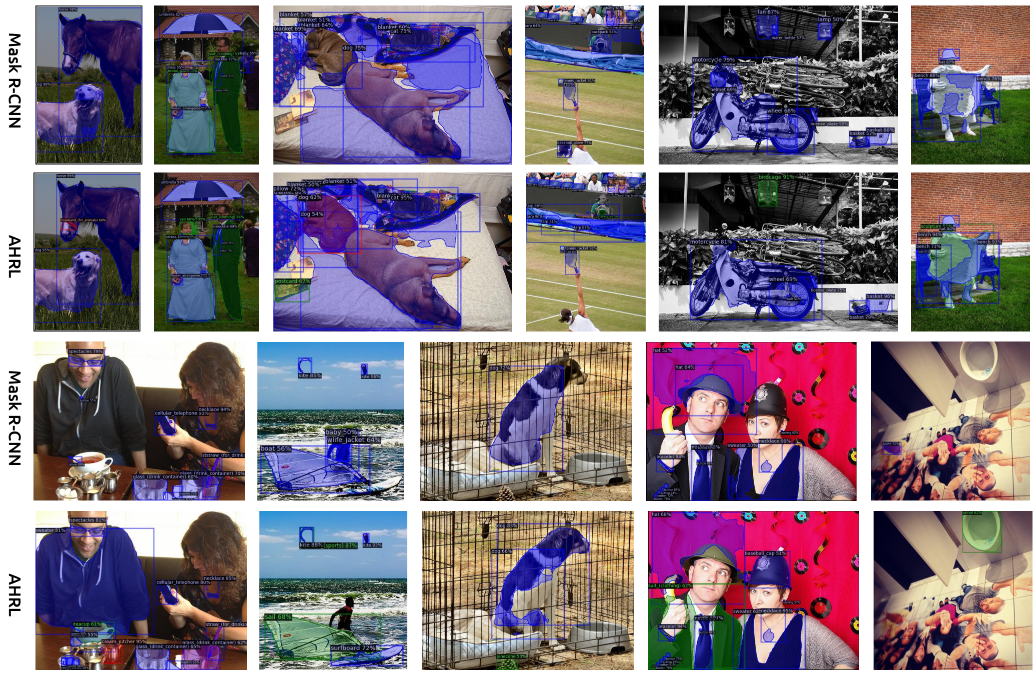


Figure 1. Mask R-CNN [3] *vs*. AHRL. Different colors stand for different frequency groups. Blue, green and red masks/boxes represent prediction objects from frequent, common and rare groups, respectively. Comparing with Mask R-CNN, AHRL can effectively eliminate the misclassification or missing detection problem, especially for scarce classes. For visualization, we apply the NMS with a threshold of 0.5 and filter out the predictions with a score lower than 0.05.

for 90k steps, with an initial learning rate of 0.02, which is decay to 0.002 and 0.0002 at 60k and 80k respectively. We adopt a class-specific branch for both mask and bounding box regression. The threshold of the prediction score is set to be 0.05. We follow [12] to set $\lambda_c$ and $\lambda_p$ as 20 in our experiments, respectively. We set $\lambda$ to 1 to balance the scale of the losses. Following [10], $\lambda_r$ is set to be $1.76 \times 10^{-3}$.

### B.3. Detailed Performance on LVIS v1.0

In this section, we report the performance on LVIS [2] v1.0 for each sub-category, i.e., $AP_f$, $AP_c$ and $AP_r$. We can find that AHRL outperforms all other SOTA methods for a large margin on each sub-category as well as the overall performance.

### B.4. Visualization on LVIS

As shown in Figure 1, we make an intuitive visualization comparison between our AHRL and Mask R-CNN on LVIS v0.5 dataset. It is obvious that AHRL can achieve a superior performance than Mask R-CNN.

### B.5. Visualization for Figure 2(b).

Detailed class information for those dots in Figure 2(b) in the paepr can be found in Figure 2.

## References


[1] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9268–9277, 2019. 1

[2] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5356–5364, 2019. 1, 2, 3

[3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1, 2, 3

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[5] Ting-I Hsieh, Esther Robb, Hwann-Tzong Chen, and Jia-Bin Huang. Droploss for long-tail instance segmentation. In *Proceedings of the Workshop on Artificial Intelligence Safety 2021 (SafeAI 2021) co-located with the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI 2021), Virtual, February 8, 2021*, 2021. 3

Table 1. Performance comparisons with the state-of-the-art methods on LVIS v1.0 [2].

| Method | Backbone | $AP^b$ | $AP^s$ | $AP_r$ | $AP_c$ | $AP_f$ |
|---|---|---|---|---|---|---|
| Mask R-CNN [3] | ResNet-50-FPN | 20.0 | 19.2 | 0 | 17.2 | 29.5 |
| EQL [10] | ResNet-50-FPN | 22.5 | 21.6 | 3.8 | 21.7 | 29.2 |
| BAGS [6] | ResNet-50-FPN | 23.7 | 23.1 | 13.1 | 22.5 | 28.2 |
| DropLoss [5] | ResNet-50-FPN | 22.9 | 22.3 | 12.4 | 22.3 | 26.5 |
| Mask R-CNN [3] | ResNet-101-FPN | 21.7 | 20.8 | 1.4 | 19.4 | 30.9 |
| EQL [10] | ResNet-101-FPN | 24.2 | 22.9 | 3.7 | 23.6 | 30.7 |
| BAGS [6] | ResNet-101-FPN | 26.5 | 25.8 | 16.5 | 25.7 | 30.1 |
| **AHRL(ours)** | ResNet-50-FPN | 26.4 | 25.7 | 16.6 | 25.4 | 29.7 |
| **AHRL(ours)** | ResNet-101-FPN | **28.7** | **27.6** | **19.3** | **27.6** | **31.4** |

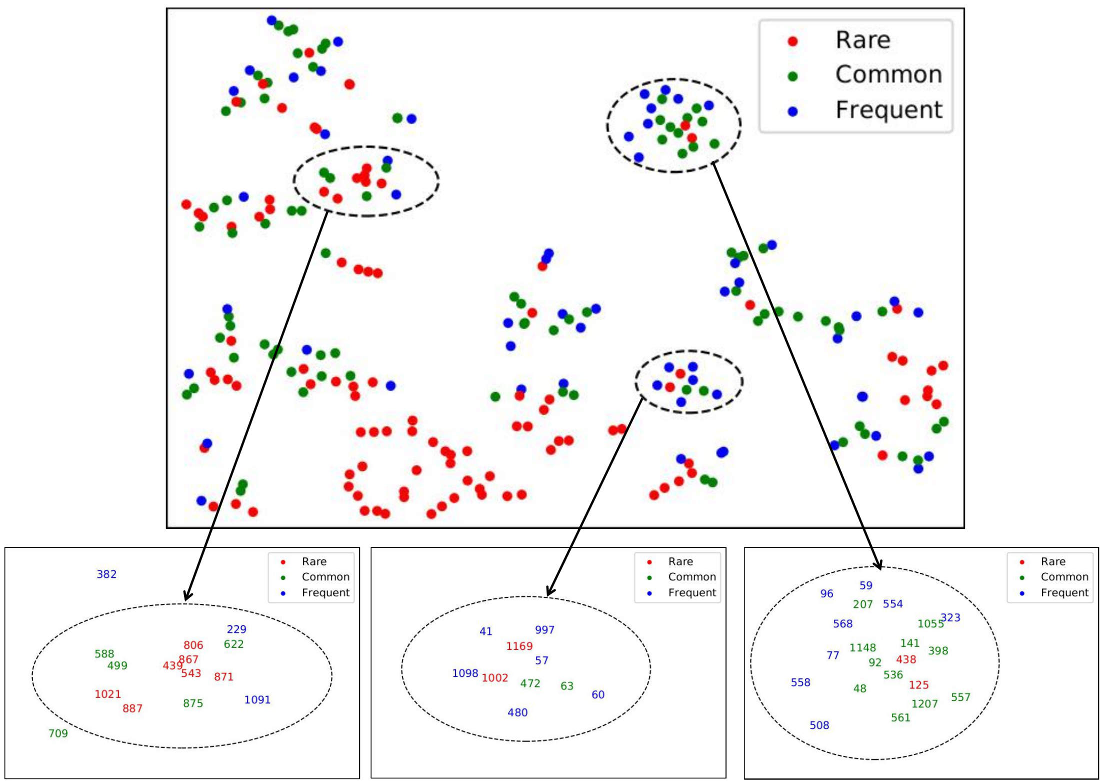


Figure 2. Details of t-SNE [11] visualization of class weights. Three magnified areas are shown at the bottom of the image. And in every magnified area, we show the concrete index of each class for a more detailed description.


[6] Yu Li, Tao Wang, Bingyi Kang, Sheng Tang, Chunfeng Wang, Jintao Li, and Jiashi Feng. Overcoming classifier imbalance for long-tail object detection with balanced group softmax. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10991–11000, 2020. 3

[7] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 1

[8] 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

[9] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[10] Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli

Ouyang, Changqing Yin, and Junjie Yan. Equalization loss for long-tailed object recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2, 3

[11] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 3

[12] Xin Wang, Thomas E Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. *arXiv preprint arXiv:2003.06957*, 2020. 2

[13] Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. Unitbox: An advanced object detection network. *MM '16: Proceedings of the 24th ACM international conference on Multimedia*, 08 2016. 1