Instance-Aware Repeat Factor Sampling for Long-Tailed Object Detection

Burhaneddin Yaman
Bosch Research North America
Bosch Center for AI
burhaneddin.yaman@us.bosch.com

Tanvir Mahmud
University of Texas at Austin
tanvirmahmud@utexas.edu

Chun-Hao Liu
Amazon Prime Video
chunhaol@amazon.com

Abstract

We propose an embarrassingly simple method – instance-aware repeat factor sampling (IRFS) to address the problem of imbalanced data in long-tailed object detection. Imbalanced datasets in real-world object detection often suffer from a large disparity in the number of instances for each class. To improve the generalization performance of object detection models on rare classes, various data sampling techniques have been proposed. Repeat factor sampling (RFS) has shown promise due to its simplicity and effectiveness. Despite its efficiency, RFS completely neglects the instance counts and solely relies on the image count during re-sampling process. However, instance count may immensely vary for different classes with similar image counts. Such variation highlights the importance of both image and instance for addressing the long-tail distributions. Thus, we propose IRFS which unifies instance and image counts for the re-sampling process to be aware of different perspectives of the imbalance in long-tailed datasets. Our method shows promising results on the challenging LVIS v1.0 benchmark dataset over various architectures and backbones, demonstrating their effectiveness in improving the performance of object detection models on rare classes with a relative $+50\%$ average precision (AP) improvement over counterpart RFS. IRFS can serve as a strong baseline and be easily incorporated into existing long-tailed frameworks.

1 Introduction

Real-world datasets are often imbalanced and follow a long-tail distribution where few object classes dominates the distribution resulting in the heavy tail for rare classes Liu *et al.* (2019); Gupta *et al.* (2019); Zhao *et al.* (2022); Liu and Yaman (2022). This poses a significant challenge in many real-world applications, especially in computer vision tasks such as image recognition, object detection, and semantic segmentation Gupta *et al.* (2019). Without addressing this issue, models tend to be biased towards the majority classes, making it difficult to achieve good performance in the rare classes Tan *et al.* (2020); Chang *et al.* (2021); Alexandridis *et al.* (2022); Hu *et al.* (2022). However, performing well on rare classes is crucial for applications like autonomous driving, video surveillance, automated optical inspection, and medical imaging systems, where anomalies or unusual events can have significant implications. Therefore, addressing the long-tail distribution of real-world visual data is crucial for achieving reliable performance in many real-world applications.

The problem of data imbalance with a long-tail distribution has been approached in two major directions. The first direction involves data augmentation with class-balanced sampling Gupta *et al.*

NeurIPS 2023 Workshop on Heavy Tails in ML.

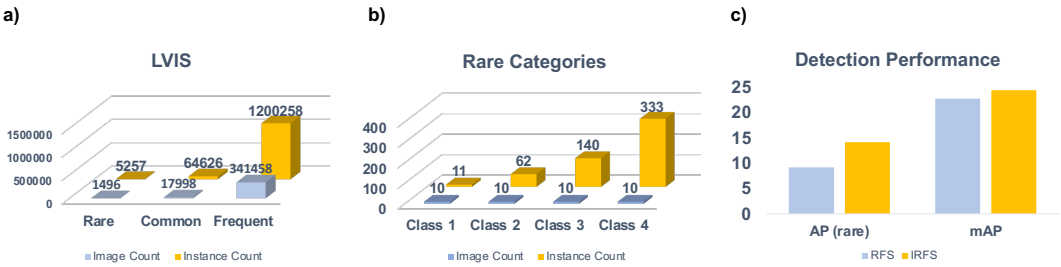

Figure 1: We illustrate the importance of instance count in addressing long-tail distributions. Figure a) shows image and instance counts for the LVIS v1.0 data distribution across three main categories, namely rare, common, and frequent. Figure b) demonstrates a representative case from the rare categories with the same image count but varying instance counts. It is observed that RFS, which ignores the instance count, assigns the same re-sampling factor to all classes with the same image count. In contrast, figure c) depicts the IRFS approach, which incorporates the instance count in the re-sampling process, resulting in a $+50\%$ relative improvement compared to RFS on rare categories and a better overall performance.

(2019); Chang *et al.* (2021); Zhang *et al.* (2021). While up-sampling the minor classes through data augmentation and down-sampling the major classes may appear promising, such an approach may lead to performance bias towards the largely augmented minor classes. The utilization of object-centric sampling and class-balanced sampling can mitigate this issue by exploiting available data and achieving better performance on minor classes. However, significant imbalance between major and minor classes may require much larger up-sampling and down-sampling ratios that under-represent major classes and make training biased. The second direction involves modifying the loss function to handle the effect of large class imbalance Li *et al.* (2020); Tan *et al.* (2021); Alexandridis *et al.* (2022). Adaptive weighting of the classes in the loss function is one such solution that adds more weight to the minor class predictions. Grouping the predictions and loss functions according to the available samples of each class is another promising solution. Despite considerable improvement of performance with these approaches, the larger imbalance of the dataset can significantly under-represent the minor classes that suffer from similar challenges.

While class-based sampling approaches have focused on the imbalance in terms of the number of images in the dataset, instance level imbalance remains unaddressed. As a result, current class balanced sampling approaches assign the same oversampling ratio to all categories with the same number of images, despite the immense amount of imbalance in terms of instances (see Fig. 1). Thus, long-tailed datasets are not only imbalanced in terms of images but also instances. Therefore, disregarding instance count while addressing the long-tail distributions may be costly and lead to suboptimal results, especially for rare categories. To address this issue, we propose a novel instance-aware repeat factor sampling (IRFS) method that incorporates the instance count and fuses it with image count to determine the repeat factor for each category. Our main contributions are threefold.

- We propose an incredibly simple method to address long-tail distributions by incorporating the instance count in the re-sampling process.
- Our proposed approach can be used as a standalone replacement for current repeat factor sampling (RFS) approaches or combined with other re-weighting approaches.
- We conduct extensive experiments on challenging LVIS v1.0 benchmark. Our proposed instance-aware re-sampling approach provides significant performance improvement over RFS, and it achieves the state-of-the-art performance in both overall accuracy and accuracy of rare classes.

2 Related Works

Long-Tailed Challenges. In the realm of computer vision, long-tail distribution poses a persistent challenge that has yet to be fully resolved Yang *et al.* (2022). Such distribution patterns are common in real-world datasets, wherein a small subset of categories dominate the majority of instances, and each category's importance does not necessarily follow a power-law decayed distribution. Achieving

balanced performance for each category in this highly skewed distribution remains an open problem, particularly in the context of object detection, which is a more complex task than image recognition due to the presence of multiple classes in a single image. Various methods have been proposed to address long-tailed image classification, including assigning balanced weights for each class in the loss functions or adjusting the sampling rate for each class in the data loader Cui *et al.* (2019); Kim *et al.* (2020); Hong *et al.* (2021); Alshammari *et al.* (2022). However, applying these solutions to object detection is more challenging due to the nature of the task.

Re-sampling in Long-Tailed Detection. The state-of-the-art re-sampling technique in long-tailed detection – repeat factor sampling, was introduced in Gupta *et al.* (2019). This technique involves increasing the sampling rate of images that contain rare class objects, thereby improving the balance of the dataset. While RFS has proven to be effective, it fails to consider the object numbers factor jointly. In response to this limitation, the authors proposed re-sampling at object-level (OCS) Chang *et al.* (2021), which leverages a memory bank to replay the necessary objects in an image. The authors in Chang *et al.* (2021) further proposed using RFS together with OCS to achieve a balanced sampling in both image and object levels. We note that this approach does not modify or integrate instance count into the image re-sampling process. Instead, it utilizes the features of rare categories from the memory bank when computing the loss. Applying data augmentation strategies is another type of re-sampling technique. It has also shown promising results in addressing the long-tailed problem in object detection. For example, simple copy-and-paste techniques Ghiasi *et al.* (2021) have been used to randomly mix objects from different images, leading to improved performance in long-tailed instance segmentation tasks.

Re-weighting in Long-Tailed Detection. Recently, the research community has shifted its focus towards inventing new re-weighting techniques to tackle the long-tailed detection problem. Equalization loss (EQL) Tan *et al.* (2020, 2021) was proposed to alleviate discouraging gradients for rare categories during parameter updating, thus improving the learning of better features for rare classes. Li et al. Li *et al.* (2020) proposed a balanced group softmax (BAGS) module that formulates a novel type of group-wise training. This module separates the tail and head classes, thus improving the learning of similar numbers of instances within each group. Observing that the classifier's weight norm can cause imbalance and make long-tailed detection challenging, the authors in Wang *et al.* (2022) proposed a C2AM loss, which enforces the margin between two categories to be proportional to the ratio of their classifiers' weight norms. Cho et al. Hyun Cho and Krähenbühl (2022) took a theoretical approach to derive a bound for object detection performance metric using classical margin-based binary classification theory. They proposed a surrogate objective named effective class-margin (ECM) loss. Specifically for instance segmentation, seesaw loss Wang *et al.* (2021) and Gumbel optimized loss Alexandridis *et al.* (2022) were proposed. The former dynamically re-balances gradients of positive and negative samples for each category, while the latter better aligns the distribution of the activation function with the long-tail distribution.

3 Methods

3.1 Repeat Factor Sampling

In this section, we introduce the Repeat Factor Sampling (RFS) method for training object detection models. RFS is a simple and effective approach that balances the class distribution by oversampling images containing rare classes Gupta *et al.* (2019). Specifically, RFS increases the occurrence rate of tail categories by repeating the images that contain them during training. This method has shown to be effective in addressing long-tail distribution challenges in object detection and has demonstrated promising results in improving the performance of models on rare categories.

For each category c, let f_c be the fraction of training images that contain at least one instance of c. Accordingly, a category-level repeat factor r_c is defined as

$$r_c = \max(1, \sqrt{t/f_c}), \tag{1}$$

where t is a hyper-parameter that intuitively controls the point at which oversampling begins. If f_c is greater than or equal to t, then there is no oversampling for that category. Additionally, as each image may contain different categories, an image-level repeat factor r_i is defined for each image i, which is determined as

$$r_i = \max_{c \in i} r_c, \tag{2}$$

where $c \in i$ denotes the categories labeled in the i-th image. During training, each image is repeated according to its repeat factor r_i.

3.2 Proposed Instance-Aware Repeat Factor Sampling

Class-balanced sampling through RFS has significantly improved performance, but exclusion of instances can lead to suboptimal results, especially for rare classes. Figure 1a) shows the image count and instance count for the three main categories. In addition to imbalance in image count, there exists a visibly significant further imbalance in terms of instance counts. Despite the varying instance count (Figure 1b)), classes with the same image count are assigned the same re-sampling factor which might lead to suboptimal results (Figure 1c)). Therefore, the f_c parameter in RFS may not be representative enough to handle long-tail distributions, especially if there is an additional imbalance layer between the number of images and the number of bounding boxes.

In order to address the challenges faced by RFS and to ensure that rare classes are properly prioritized during the sampling process, we propose an instance-aware repeat factor sampling (IRFS) method that takes both images and bounding box instances into account. To be more specific, we propose using both the image and bounding box instances to define f_c. Let $f_{(i,c)}$ and $f_{(b,c)}$ denote the fractions of images and bounding boxes in the training set that contain instances of category c. We define f_c as the mean of $f_{(i,c)}$ and $f_{(b,c)}$. For a geometric mean case, the Eq. 1 can be reformulated as

$$r_c = \max(1, \sqrt{t/\sqrt{f_{(i,c)} \times f_{(b,c)}}}). \tag{3}$$

Once instance-aware category-level repeat factor is calculated in Eq. 3, Eq. 2 is used to compute the image-level repeat factor.

4 Experiments

4.1 Dataset

In our experiments, we evaluate the proposed IRFS approach on the LVIS v1.0 dataset Gupta *et al.* (2019). LVIS contains 1.3 million object instances across 120K images and includes 1203 categories. This dataset is known for its heavy long-tail distribution and has been categorized into three groups: *frequent*, *common*, and *rare*, based on the frequency of occurrence. Rare categories are defined as those appearing in fewer than 10 images, common categories appear in more than 10 but less than 100 images, and frequent categories appear in over 100 images. The dataset is split into 100K training images and 20K validation images, and we perform our experiments on this split.

4.2 Implementation Details

We conducted all experiments using MMdetection Chen *et al.* (2019), a popular open-source library for object detection and instance segmentation. For our experiments, we utilized Mask R-CNN He *et al.* (2017), a state-of-the-art framework for object detection and instance segmentation. We trained Mask R-CNN with ResNet-50 and ResNet-101 backbones He *et al.* (2016) and Feature Pyramid Network (FPN) Lin *et al.* (2017). We also performed experiments on ATSS Zhang *et al.* (2020), a one-stage detector. We followed the standard LVIS setup and hyperparameters for the models. Specifically, we trained the models using SGD with 0.9 momentum and a batch size of 16 on 4 GPUs (NVIDIA Tesla V100 with 32 GB VRAM), with an initial learning rate of 0.02 and weight decay of 10^{-4}. The training and inference images were resized to a shorter and longer image edge of 800 and 1333 pixels, and we only used horizontal flipping for data augmentation unless stated otherwise. We trained all models for 12 epochs ($1\times$ schedule), with decay at the 8th and 11th epochs. We train the RFS models with the reported optimal value of $t = 10^{-3}$. We use IRFS with geometric mean and $t = 10^{-3}$ unless stated otherwise. In addition to RFS and IRFS, we report the performance of training without any sampling, i.e. $t = 0$, and denote it in the Tables as "w/o sampling".

4.3 Evaluation Metric

We evaluated our models using the LVIS metrics. The metrics include mean average precision (mAP), average precision (AP) with intersection over union (IoU) of 50% (AP_{50}), AP with IoU of

Table 1: LVIS v1.0 validation set results using Mask R-CNN and Cascade Mask R-CNN frameworks with ResNet-50/ResNet-101 backbones on $1\times$ schedule. RFS and IRFS are trained with $t = 10^{-3}$ and w/o sampling is trained with $t = 0$.

Framework	Backbone	**Detection**	mAP_{bbox}	AP_{50}	AP_{75}	AP_r	AP_c	AP_f
Mask R-CNN	ResNet-50	w/o sampling	16.9	28.1	17.7	0.0	12.3	29.6
		RFS	22.7	37.3	23.9	9.2	21.3	30.0
		IRFS	**24.4**	**39.8**	**25.8**	**14.1**	**22.8**	**30.7**
Mask R-CNN	ResNet-101	w/o sampling	18.6	30.0	19.6	0.0	14.2	31.6
		RFS	24.9	40.0	26.5	11.9	23.8	32.0
		IRFS	**26.4**	**41.8**	**28.3**	**16.9**	**24.7**	**32.5**
Cascade Mask R-CNN	ResNet-50	w/o sampling	20.1	29.5	21.3	0.3	16.1	33.3
		RFS	26.6	38.5	28.1	12.7	25.7	33.7
		IRFS	**28.8**	**41.4**	**30.6**	**19.1**	**27.4**	**34.4**
Cascade Mask R-CNN	ResNet-101	w/o sampling	21.8	31.5	23.1	0.2	18.4	35.0
		RFS	29.1	41.6	31.1	17.0	28.4	35.3
		IRFS	**30.0**	**42.8**	**32.1**	**19.1**	**29.0**	**35.9**
Framework	Backbone	**Segmentation**	mAP_{segm}	AP_{50}	AP_{75}	AP_r	AP_c	AP_f
Mask R-CNN	ResNet-50	w/o sampling	16.2	26.0	17.0	0.0	12.5	27.3
		RFS	21.9	34.8	23.2	10.4	21.1	27.8
		IRFS	**23.7**	**37.1**	**25.1**	**14.9**	**22.8**	**28.5**
Mask R-CNN	ResNet-101	w/o sampling	17.6	27.9	18.7	0.0	14.2	29.0
		RFS	23.8	37.0	25.3	12.3	23.3	29.3
		IRFS	**25.3**	**39.2**	**26.9**	**17.2**	**24.3**	**29.9**
Cascade Mask R-CNN	ResNet-50	w/o sampling	17.8	27.6	19.0	0.2	14.8	29.0
		RFS	23.8	36.3	25.4	12.3	23.3	29.4
		IRFS	**25.6**	**38.8**	**27.4**	**17.3**	**24.8**	**30.1**
Cascade Mask R-CNN	ResNet-101	w/o sampling	19.3	29.5	20.6	0.2	16.8	30.6
		RFS	26.0	39.2	27.8	15.9	25.6	30.8
		IRFS	**26.8**	**40.3**	**28.5**	**17.7**	**26.3**	**31.3**

75% (AP_{75}), AP on rare classes (AP_r), AP on common classes (AP_c), and AP on frequent classes (AP_f). For Mask R-CNN, we reported both detection and segmentation metrics, which are denoted as mAP_{bbox} and mAP_{segm}, respectively.

4.4 Main Results

Table 1 illustrates the performance of IRFS with respect to counterpart RFS method.

Object Detection. In all frameworks, training without a balanced sampling results in the lowest AP. Due to the heavy-tailed distribution, the w/o sampling approach suffers greatly for rare class detection. While RFS provides a significant improvement over the w/o sampling approach, the proposed IRFS achieves the best performance with a significant improvement over RFS across all metrics. In particular, for Mask R-CNN with a ResNet-50 backbone, our proposed IRFS provides a relative $+50\%$ improvement over RFS for rare categories. Moreover, the superior performance of IRFS for rare classes does not result in degraded results for common and frequent classes. In fact, it relatively improves the performance of common and frequent categories by $+7\%$ and $+2\%$, respectively, compared to RFS. For Mask R-CNN with a ResNet-101 backbone, IRFS outperforms the RFS by 5.0 AP ($+40\%$) for rare categories and 1.5 mAP ($+5\%$) for overall classes. We achieve similar consistent results with the more advanced Cascade Mask R-CNN Cai and Vasconcelos (2018) framework across ResNet-50 and ResNet-101 backbones.

Instance Segmentation. Similar to object detection, IRFS exhibits significant improvement over RFS in instance segmentation. Particularly, for Mask R-CNN with ResNet-50 backbone, IRFS achieves a mAP of 23.7%, resulting in a $+8\%$ improvement over RFS. For rare categories, IRFS achieves a 14.9% AP, which is a $+43\%$ improvement over RFS. Consistently, IRFS outperforms RFS

Table 2: Comparison of the combination of RFS and IRFS with ECM loss, using Mask R-CNN framework with a ResNet-50 backbone, on the LVIS v1.0 validation set.

	Detection				Segmentation			
Method	mAP_{bbox}	AP_r	AP_c	AP_f	mAP_{segm}	AP_r	AP_c	AP_f
ECM	24.8	13.3	23.3	31.6	24.6	14.7	24.1	29.6
ECM+RFS	26.8	16.5	26.2	31.9	26.4	18.7	26.2	30.0
ECM+IRFS	**27.6**	**19.4**	**26.5**	**32.4**	**27.0**	**20.6**	**26.4**	**30.6**
RFS	22.7	9.2	21.3	30.0	21.9	10.4	21.1	27.8
IRFS	24.4	14.1	22.8	30.7	23.7	14.9	22.8	28.5

Table 3: Comparison of RFS and IRFS using a single-stage detector ATSS with different backbones on the LVIS v1.0 validation set.

Framework	Backbone	Method	mAP_{bbox}	AP_r	AP_c	AP_f
ATSS	ResNet-50	RFS	19.3	7.5	17.2	26.7
		IRFS	**19.7**	**9.5**	**17.2**	**26.9**
ATSS	ResNet-101	RFS	20.9	8.8	19.0	28.4
		IRFS	**22.0**	**10.7**	**20.4**	**28.6**

for common and frequent categories. With the ResNet-101 backbone and Cascade Mask R-CNN framework, IRFS maintains its superior performance compared to RFS across all cases.

Table 2 presents a comparison of the impact of class and instance balanced sampling approaches on the state-of-the-art weighted loss function. It is observed that ECM without any class balanced sampling achieves a better overall bounding box and mask AP compared to RFS and IRFS. However, IRFS shows better results in terms of the target rare categories compared to standalone ECM. Class/Instance balanced sampling and weighted loss are complementary, and ECM with RFS and IRFS achieves significantly improved performance compared to standalone ECM or RFS/IRFS. Specifically, ECM with the IRFS achieves the best performance across all metrics. For target rare categories, it achieves a 19.4% and 20.6% AP for detection and segmentation tasks, respectively, which is a relative 17% and 10% improvement compared with ECM with RFS for detection and segmentation tasks.

Table 3 presents the performance comparison of RFS and IRFS on ATSS, a state-of-the-art single-stage detector, trained with ResNet-50 and ResNet-101 backbones. When trained with IRFS, ATSS exhibits a substantial relative improvement of 26% and 21% over its RFS-trained counterpart on rare classes, using ResNet-50 and ResNet-101 backbones, respectively. Furthermore, it consistently achieves better overall mAP compared to the counterpart RFS method.

Qualitative results and ablation studies are provided in the Supplementary Materials.

5 Discussion and Conclusions

In this work, we presented a novel approach named instance-aware repeat factor sampling (IRFS) for handling long-tail distributions in object detection. Extensive experimental results demonstrate that by incorporating instance count in addition to image count, IRFS consistently outperforms RFS by significantly improving detection performance of rare categories while maintaining the overall performance for all categories. These results emphasize the importance of instance count in improving the performance of rare categories.

IRFS is a simple method that can be used as a plug-in replacement for RFS, or combined with re-weighting methods to achieve state-of-the-art results in long-tailed object detection and instance segmentation. Due to its simplicity, easy integration with other methods, and effectiveness in improving performance, we believe IRFS can become a strong baseline for methods aiming to address long-tail distributions. For future work, we plan to investigate the application of IRFS in related fields in the computer vision domain, and explore new approaches to optimize the averaging method.

References

Konstantinos Panagiotis Alexandridis, Jiankang Deng, Anh Nguyen, and Shan Luo. Long-tailed instance segmentation using gumbel optimized loss. In *European Conference on Computer Vision*, pages 353–369. Springer, 2022.

Shaden Alshammari, Yu-Xiong Wang, Deva Ramanan, and Shu Kong. Long-tailed recognition via weight balancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6897–6907, June 2022.

Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Nadine Chang, Zhiding Yu, Yu-Xiong Wang, Anima Anandkumar, Sanja Fidler, and José Manuel Álvarez. Image-level or object-level? a tale of two resampling strategies for long-tailed detection. In *International Conference on Machine Learning*, 2021.

Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, 2019.

Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D. Cubuk, Quoc V. Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 2918–2928. Computer Vision Foundation / IEEE, 2021.

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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. Disentangling label distribution for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6626–6636, 2021.

Shu Hu, Chun-Hao Liu, Jayanta Dutta, Ming-Ching Chang, Siwei Lyu, and Naveen Ramakrishnan. Pseudoprop: Robust pseudo-label generation for semi-supervised object detection in autonomous driving systems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4390–4398, June 2022.

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

Jaehyung Kim, Jongheon Jeong, and Jinwoo Shin. M2m: Imbalanced classification via major-to-minor translation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13893–13902, 2020.

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.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2017.

Chun-Hao Liu and Burhaneddin Yaman. Object detection for autonomous dozers. *arXiv preprint arXiv:2208.08570*, 2022.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2537–2546, 2019.

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.

Jingru Tan, Xin Lu, Gang Zhang, Changqing Yin, and Quanquan Li. Equalization loss v2: A new gradient balance approach for long-tailed object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1685–1694, June 2021.

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 *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 9695–9704. Computer Vision Foundation / IEEE, 2021.

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

Lu Yang, He Jiang, Qing Song, and Jun Guo. A survey on long-tailed visual recognition. *Int. J. Comput. Vis.*, 130(7):1837–1872, 2022.

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*, 2020.

Cheng Zhang, Tai-Yu Pan, Yandong Li, Hexiang Hu, Dong Xuan, Soravit Changpinyo, Boqing Gong, and Wei-Lun Chao. MosaicOS: A simple and effective use of object-centric images for long-tailed object detection. In *ICCV*, 2021.

Ruiyang Zhao, Burhaneddin Yaman, Yuxin Zhang, Russell Stewart, Austin Dixon, Florian Knoll, Zhengnan Huang, Yvonne W Lui, Michael S Hansen, and Matthew P Lungren. fastmri+, clinical pathology annotations for knee and brain fully sampled magnetic resonance imaging data. *Scientific Data*, 9(1):152, 2022.

6 Supplementary Materials

6.1 Qualitative Results

Fig. 2 illustrates the detection and segmentation results for w/o sampling, RFS and IRFS models trained with Mask R-CNN framework using ResNet-50 backbone on $1\times$ schedule. Note that only targeted rare categories are shown for visibility. As expected, w/o sampling fails in detecting the rare categories. IRFS detects and segment with a higher accuracy for rare categories detected by RFS as well (Fig. 2a). IRFS further shows its strength by detecting rare categories that are not detected by RFS (Fig. 2b).

Fig. 3 illustrates the detection and segmentation results for ECM, ECM + RFS and ECM + IRFS models trained with Mask R-CNN framework using ResNet-50 backbone on $1\times$ schedule. Note that only targeted rare categories are shown for visibility. In this challenging representative scenario, ECM without any re-sampling as well as RFS and IRFS without ECM fails to detect the rare categories. While ECM combined with RFS also fails detecting any rare category, ECM with IRFS achieves a better performance and successfully detects and segments a rare category.

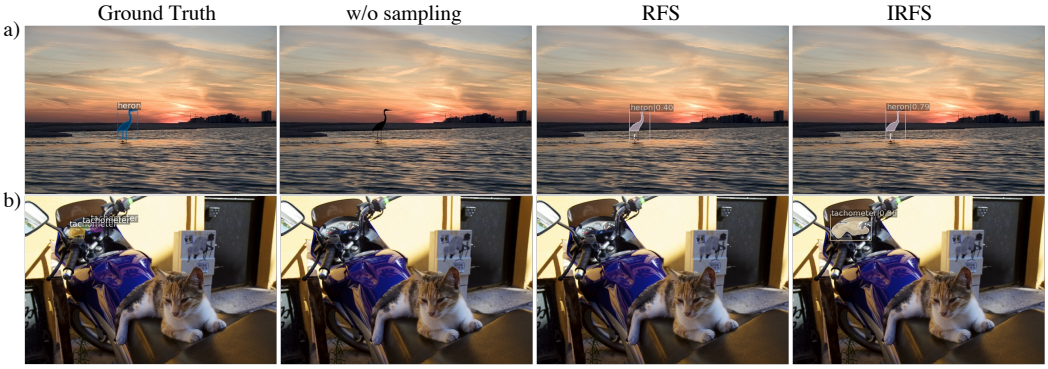

Figure 2: Representative detection and instance segmentation visualizations on LVIS v1.0 dataset. For convenience, we present the results specifically focused on the rare categories.

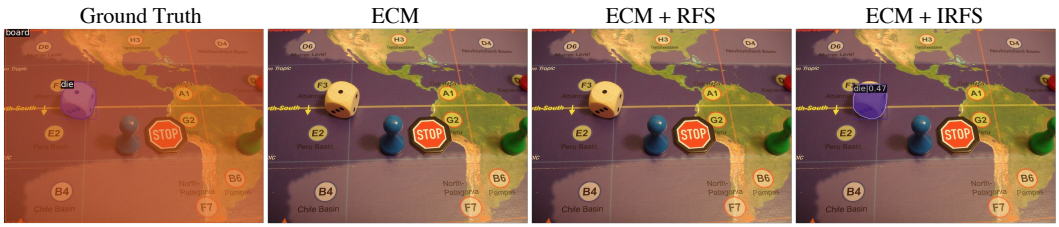

Figure 3: Illustration of detection and instance segmentation results for ECM re-weighting method and its combination with re-sampling approaches. For convenience, we exclusively display the results for the targeted rare categories.

6.2 Ablation Study

In this section we conduct ablation studies on two key parameters of the proposed IRFS method, namely threshold value and averaging method.

Table 4 presents the ablation study for IRFS using thresholds, $t \in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$. The best overall mAP is achieved by IRFS with $t = 10^{-3}$. For the target rare category, $t = 10^{-3}$ significantly outperforms other thresholds for both detection and segmentation. It is noteworthy that IRFS's best performing threshold value, $t = 10^{-3}$ matches RFS's best matching threshold value Gupta *et al.* (2019).

Table 4: An ablation study on IRFS with different thresholds t using Mask R-CNN with a ResNet-50 backbone on a $1\times$ schedule on LVIS v1.0 validation set.

	Detection				Segmentation			
t	mAP_{bbox}	AP_r	AP_c	AP_f	mAP_{segm}	AP_r	AP_c	AP_f
w/o sampling ($t = 0$)	16.9	0.0	12.3	29.6	16.2	0.0	12.5	27.3
$t = 0.1$	19.8	6.0	17.8	28.0	18.9	6.0	17.5	26.1
$t = 0.01$	23.7	11.0	22.2	**30.9**	22.9	11.3	22.2	**28.7**
$t = 0.001$	**24.4**	**14.1**	**22.8**	30.7	**23.7**	**14.9**	**22.8**	28.5
$t = 0.0001$	19.5	3.7	16.5	29.7	18.7	3.9	16.8	27.4
RFS ($t = 0.001$)	22.7	9.2	21.3	30.0	21.9	10.4	21.1	27.8

Table 5 shows the results of different common averaging options, including geometric, harmonic, arithmetic, and quadratic averaging.[1] In addition to averaging, we also propose and include the instance only repeat factor sampling. Instance only disregards the image count in the re-sampling process. We present the results for $t = 10^{-3}$. Both harmonic and geometric averaging show improved performance over arithmetic and quadratic averaging, with the former providing the best overall results. The results indicate that all averaging options provide significant improvement over RFS. This further emphasizes the importance of including instances when addressing long-tail distributions. More interestingly, instance only approach achieves the best overall mAP in detection and it achieves more than $+50\%$ improvement over RFS, the counterpart image count only based re-sampling method. However, disregarding image count can lead to suboptimal results when dealing with real-world long-tailed datasets that exhibit variations in image instance count distributions. Thus, we assert that IRFS is better suited for addressing long-tailed datasets with diverse distributions.

Table 5: An ablation study on different averaging options for IRFS using Mask R-CNN with a ResNet-50 backbone on a $1\times$ schedule on LVIS v1.0 validation set. We also propose instance only approach for comparison purposes.

	Detection				Segmentation			
Mean	mAP_{bbox}	AP_r	AP_c	AP_f	mAP_{segm}	AP_r	AP_c	AP_f
w/o sampling ($t = 0$)	16.9	0.0	12.3	29.6	16.2	0.0	12.5	27.3
Geometric	24.4	14.1	22.8	30.7	23.7	14.9	22.8	28.5
Harmonic	24.5	**15.0**	22.7	30.8	**23.8**	**15.5**	22.7	28.6
Arithmetic	24.1	12.2	**23.2**	30.5	23.3	12.9	**23.0**	28.3
Quadratic	23.6	11.7	22.2	30.4	22.7	12.0	22.0	28.2
Instance Only	**24.6**	14.4	22.9	**31.0**	23.7	14.0	22.9	**28.7**
RFS	22.7	9.2	21.3	30.0	21.9	10.4	21.1	27.8

[1] Note the mathematical definition for geometric, harmonic, arithmetic, and quadratic average for two numbers x and y is defined as $\sqrt{xy}$, $\frac{2xy}{x+y}$, $\frac{x+y}{2}$, and $\sqrt{\frac{x^2+y^2}{2}}$, respectively.