

## A Additional results beyond original paper

### A.1 Performance of OPeN on additional dataset and imbalance ratios

In addition to CIFAR-10-LT (IR=100) dataset, we compared the performance of OPeN [1] to DRS [2], RS, and ERM on CIFAR-10-LT (IR=50) and CIFAR-100-LT (IR=100,50) datasets. Consistent with claim 1, OPeN outperformed the baseline resampling schemes across all datasets. The original paper did not report the accuracy of deferred resampling (DRS) [2] for these additional datasets. Nonetheless, we compared OPeN with DRS because DRS provides a fair baseline, as OPeN uses the same deferred resampling schedule.

Dataset IR	CIFAR-10-LT		CIFAR-100-LT			
	50		100		50	
	Reported [1]	Ours	Reported [1]	Ours	Reported [1]	Ours
ERM	84.9	84.9	47.0	47.1	52.4	52.7
RS	82.2	80.9	42.5	41.6	48.0	46.5
DRS	-	86.9	-	50.8	-	55.8
OPeN	87.9	87.8	51.5	52.1	56.3	56.5

**Table 11.** Comparison of accuracy on CIFAR-10-LT (IR=50) and CIFAR-100-LT (IR=100,50).

### A.2 Hyperparameter Search

**Input normalization values** – The authors did not specify the mean and standard deviation used to normalize the dataset. We explored various prior works [2, 3, 9] and discovered that they differed from the values we computed (Tables 12 and 13). Surprisingly, we found that many prior works use mean values from the full CIFAR-10/100 datasets instead of the values from the long-tailed variants. This could result in an unfair evaluation, as the statistics from the full training dataset may resemble the validation dataset, as they are both balanced.

Source	Mean	Std
Zhong et al. [9]	(0.4914, 0.4822, 0.4465)	(0.2023, 0.1994, 0.2010)
Cao et al. [2]	(0.4914, 0.4822, 0.4465)	(0.2023, 0.1994, 0.2010)
Kim, Jeong, and Shin [3]	(0.4914, 0.4822, 0.4465)	(0.2023, 0.1994, 0.2010)
CIFAR-10 <sup>†</sup>	(0.4914, 0.4822, 0.4465)	(0.2470, 0.2435, 0.2616)
CIFAR-10-LT (IR=50) <sup>†</sup>	(0.4978, 0.5003, 0.4840)	(0.2505, 0.2477, 0.2722)
CIFAR-10-LT (IR=100) <sup>†</sup>	(0.4989, 0.5044, 0.4926)	(0.2513, 0.2485, 0.2734)

**Table 12.** Per-channel mean and standard deviations for experiments on CIFAR-10-LT. Values calculated in this paper are marked by <sup>†</sup>.

Source	Mean	Std
Zhong et al. [9]	(0.4914, 0.4822, 0.4465)	(0.2023, 0.1994, 0.2010)
Cao et al. [2]	(0.4914, 0.4822, 0.4465)	(0.2023, 0.1994, 0.2010)
Kim, Jeong, and Shin [3]	(0.5071, 0.4867, 0.4408)	(0.2675, 0.2565, 0.2761)
CIFAR-100 <sup>†</sup>	(0.5071, 0.4866, 0.4409)	(0.2673, 0.2564, 0.2762)
CIFAR-100-LT (IR=50) <sup>†</sup>	(0.5202, 0.4916, 0.4415)	(0.2676, 0.2609, 0.2778)
CIFAR-100-LT (IR=100) <sup>†</sup>	(0.5228, 0.4929, 0.4420)	(0.2677, 0.2617, 0.2780)

**Table 13.** Per-channel mean and standard deviations for experiments on CIFAR-100-LT. Values calculated in this paper are marked by <sup>†</sup>.

We train a WideResNet network with the default experiment setting in Section 3.3 to investigate the effect of these values. As shown in Table 14, using the calculated mean and standard deviation from the long-tailed dataset reduced performance. Regardless, we do find that OPeN still improves performance over ERM and DRS, supporting the central claim of Zada, Benou, and Irani [1].

Source	ERM	DRS	OPeN
Baseline [9, 2, 3]	81.18	83.22	85.04
CIFAR-10 <sup>†</sup>	80.50	82.39	84.72
CIFAR-10-LT (IR=100) <sup>†</sup>	79.28	81.03	84.12

**Table 14.** Performance on different input normalization values on CIFAR-10-LT (IR=100). Values calculated in this paper are marked by <sup>†</sup>.

**Batch size** – Before we communicated with the authors and confirmed that the batch size used was 128, we performed a hyperparameter search ourselves. In Table 15, we list the batch sizes and their respective performance. Experiments show that 128 is the best batch size for the set of hyperparameters.

Batch size	Accuracy		
	ERM	DRS	OPeN
32	74.38	80.15	80.50
64	78.69	82.89	83.89
128	<b>81.18</b>	<b>83.22</b>	<b>85.04</b>
256	79.11	75.28	82.36
512	75.19	71.39	79.22

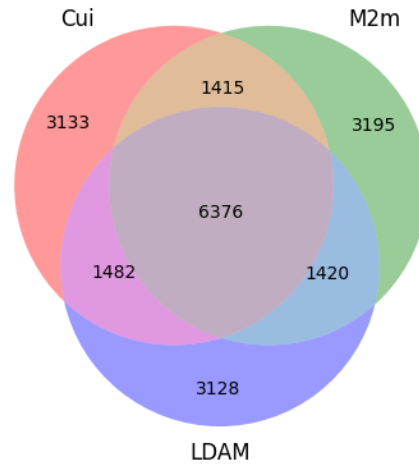
**Table 15.** Hyperparameter search for batch size. Experiment was done on CIFAR-10-LT (IR=100) with the default setting in Table 2 except for the batch size. Results with highest accuracy for each method are boldfaced.

**Noise ratio** – The noise ratio is the hyperparameter that defines the probability of replacing an oversampled image with pure noise. The authors used a noise ratio of  $\frac{1}{3}$  across all datasets. We compared the performance of OPeN across increasing levels of noise ratios. Surprisingly, replacing up to  $\frac{2}{3}$  of the oversampled images with pure noise continued to provide higher validation accuracy than DRS, while the train accuracy dropped with increasing noise ratio, as expected.

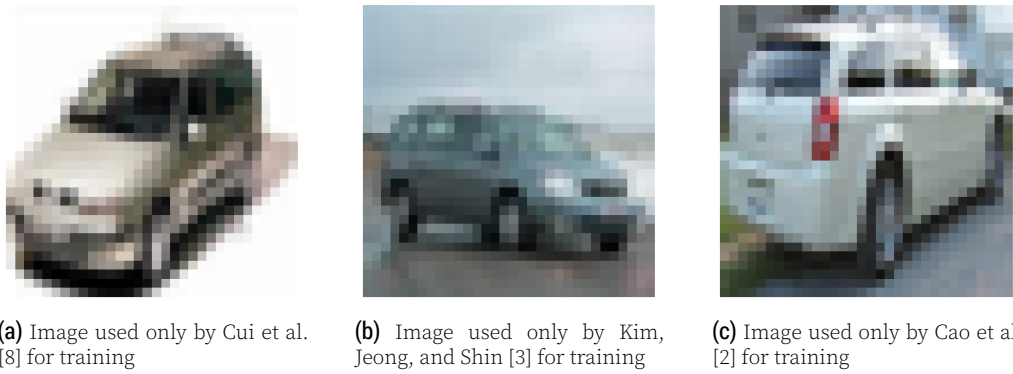
Noise ratio	1/6	2/6	3/6	4/6	5/6	6/6
Accuracy	83.23	<b>85.04</b>	84.34	84.23	83.31	80.01

**Table 16.** Hyperparameter search for noise ratio. Experiment was done on CIFAR-10-LT (IR=100) with the default setting in Table 2 except for the noise ratio. Result with highest accuracy is boldfaced.

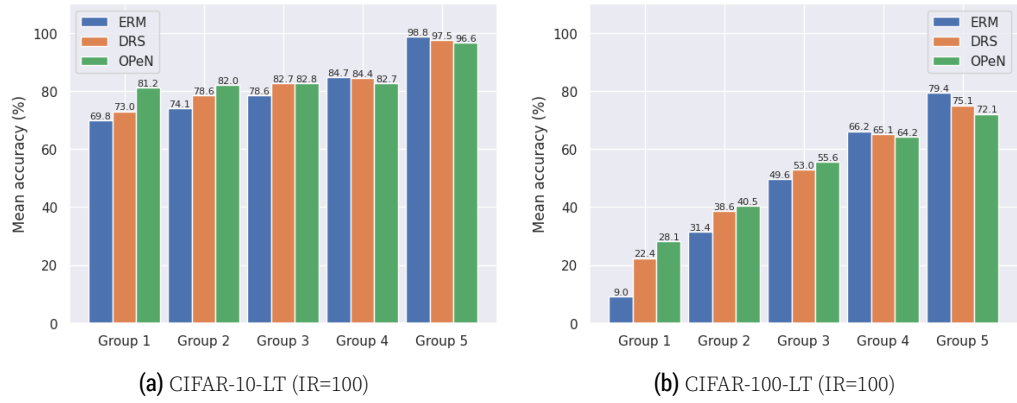
## B Additional Figures



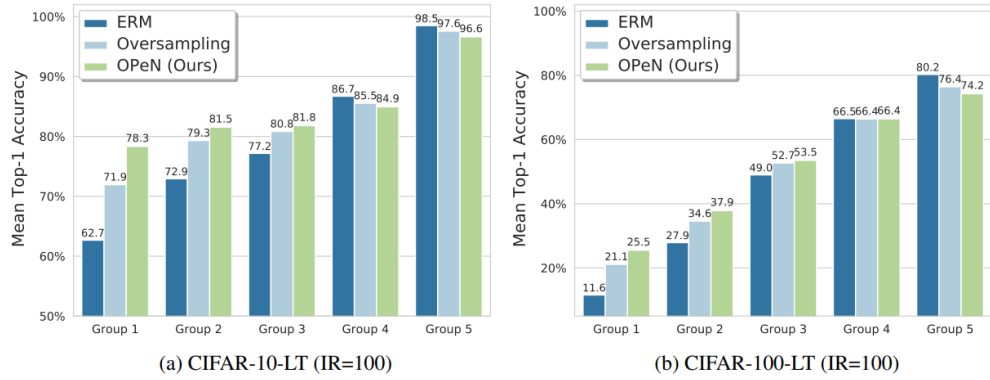
**Figure 3.** Venn diagram showing the intersections of training dataset used by Cui et al. [8] (denoted Cui), Cao et al. [2] (LDAM), and Kim, Jeong, and Shin [3] (M2m) for CIFAR-10-LT (IR=100).



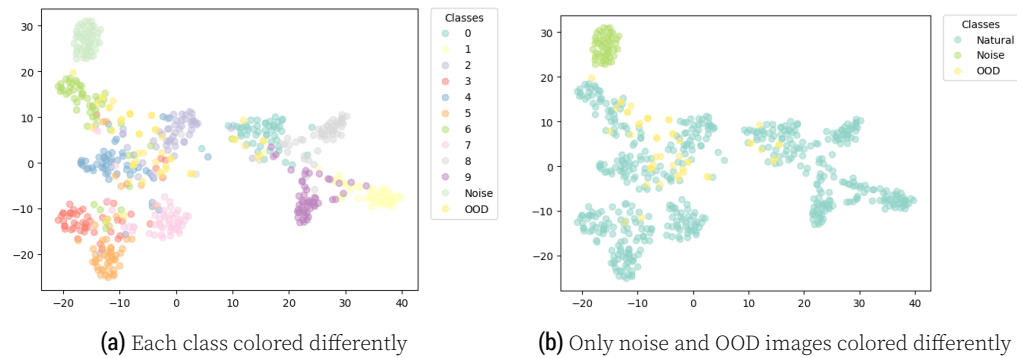
**Figure 4.** Examples of automobile images only found in one but not the other 2 CIFAR-10-LT (IR=100) datasets.



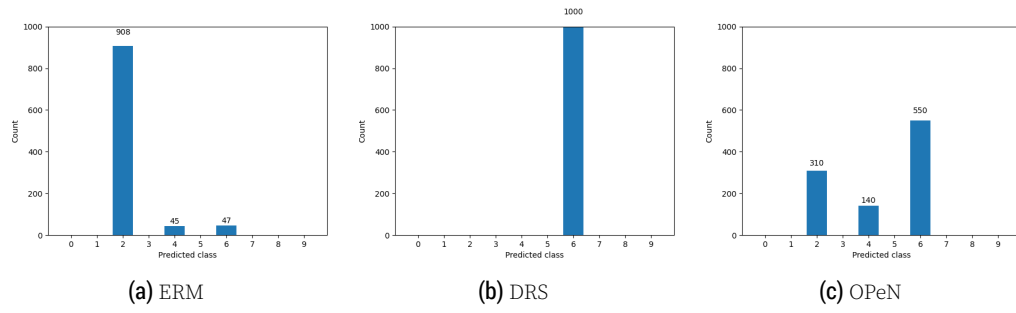
**Figure 5.** Validation accuracy for CIFAR-10-LT and CIFAR-100-LT with IR=100. Classes partitioned into 5 groups, where Group 1 is the least frequent and Group 5 is the most frequent. Reproduction of Figure 6.



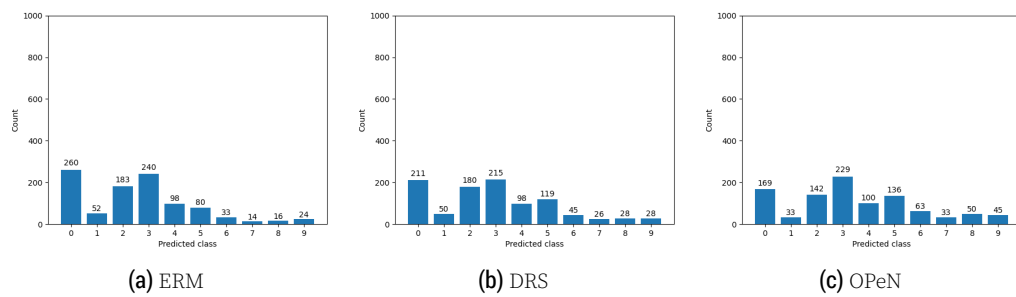
**Figure 6.** Figure from Zada, Benou, and Irani [1] reporting validation accuracy for CIFAR-10-LT and CIFAR-100-LT with IR=100 where classes are partitioned into 5 groups. Group 1 is the least frequent and Group 5 is the most frequent.



**Figure 7.** t-SNE of 50 examples from each class in the CIFAR-10 validation dataset, 50 out-of-distribution images from CIFAR-100 validation dataset, and 50 random pure noise. Images are embedded with model trained with DRS.



**Figure 8.** Histogram of predicted classes for 1000 noise images.



**Figure 9.** Histogram of predicted classes for 1000 out-of-distribution images.