A. Appendix

A.1. Details on learning rate schedule

Following FixMatch [10], we use cosine learning rate decay [9] which sets the learning rate to $\eta \cos\left(\frac{l\pi k}{16K}\right)$ where l is a hyper-parameter controlling the decay rate, η is the initial learning rate, k is the current training step, and K is the total number of training steps per generation. We use $l = 1$ for CReST and CReST+ on a FixMatch [10] base, on $l = 5$ for CReST models on a MixMatch [1] base.

A.2. Details on datasets

Details of DARP datasets. We directly compare with DARP [5], the most recent state-of-the-art SSL algorithm specifically designed for class-imbalanced data. We apply our method on exactly the same datasets used in DARP, which are class-imbalanced datasets constructed from CIFAR10 [6]. Similar to long-tailed CIFAR10 (CIFAR10-LT), the training images are randomly selected per class. However, instead of maintaining different label fractions β, DAPR keeps the number of training samples in the most majority class $N_1 = 4500$, including 3000 unlabeled samples and 1500 labeled samples. That is, the label fraction is set to be $\beta = \frac{1500}{4500} = 33.3\%$ This setting is applied to all evaluated imbalance ratios $\gamma = 50, 100, 150$. Please refer to DAPR [5] for more details.

Class distribution of ImageNet127. We apply our method on ImageNet127 [4] to test the efficacy of CReST and CReST+ on a large-scale dataset. Both the training set and the validation set of ImageNet127 are built by grouping the 1000 classes of ImageNet [3] into 127 classes based on their top-down hierarchy in WordNet. It is originally introduced to study the relationship between coarse classes and their fine-grained classes. And in [4] and [12], only the instance-wise accuracy on the imbalanced validation set is considered which is a class-imbalanced metric. In our work, however, we focus on its naturally class-imbalanced property, and our metric is class-balanced, *i.e.*, averaged recall over all classes.

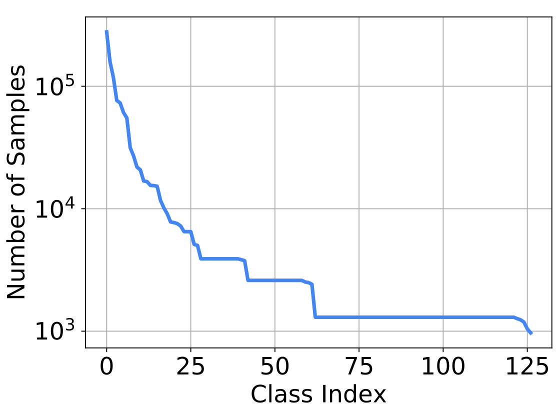
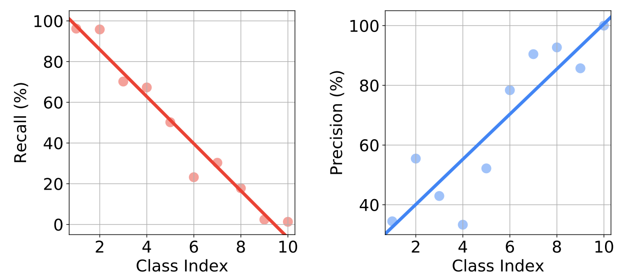
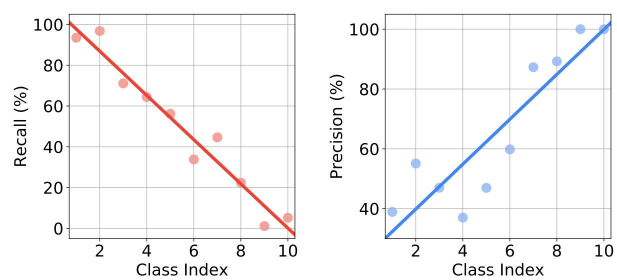


Figure 1. The imbalanced class distribution of the training set of ImageNet127 [4].

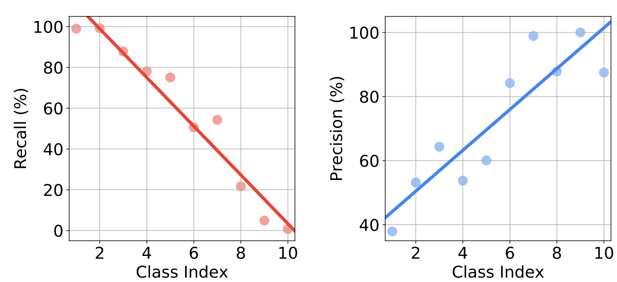
In Fig. 1, we show the number of samples of each merged class on the training set. The class index is sorted by the size of the class in descending order. Note that the y-axis is log-scaled. As shown in Fig. 1, the class-distribution is highly skewed, while the most minority class still contains 969 training examples, which is adequate to be split to labeled and unlabeled subsets to form an SSL task. This is unlike common benchmark datasets designed for fully-supervised learning (*e.g.*, [8, 2]) where the minority classes have too few examples to bulld an valid SSL task. These two prop-



(c) Pseudo-Labeling [7]



(f) Mean Teacher [11]



(i) MixMatch [1]

Figure 2. Per-class recall and precision on CIFAR10-LT ($\gamma = 100$, $\beta = 10\%$) with three different SSL algorithms. The class index is sorted by the number of examples in descending order. In line with our observation on FixMatch [10], these models obtain high recall but low precision on majority classes, while obtaining low recall but high precision on minority classes.

erties, *i.e*. the skewed class-distribution and the adequate training samples in minority classes, make ImageNet127 a good test bed for class-imblanced semi-supervised learning.

A.3. Precision and recall of other SSL algorithms

In this section, we provide recall and precision of each class with three different SSL algorithms, including Pseudo-Labeling [7], Mean Teacher [11] and MixMatch [1]. We directly apply these three SSL algorithms on CIFAR10-LT with imbalanced ratio 100 and label fraction 10%. The results are presented in Fig. 2. All three algorithms behave similarly, where majority classes obtain high recall but low precision, and minority classes suffer from low recall but achieve surprisingly high precision. The opposite bias of recall and precision of these models is in line with our observation made on FixMatch [10]. This common phenomenon shared by different SSL algorithms motivates our to exploit the high precision of minority classes to alleviate their recall degradation, re-balancing the model during the process of self-training.

References

[1] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. MixMatch: A holistic approach to semi-supervised learning. In *NeurIPS*, 2019. 1, 2

[2] Yin Cui, Yang Song, Chen Sun, Andrew Howard, and Serge Belongie. Large scale fine-grained categorization and domain-specific transfer learning. In *CVPR*, 2018. 1

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 1

[4] Minyoung Huh, Pulkit Agrawal, and Alexei A Efros. What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*, 2016. 1

[5] Jaehyung Kim, Youngbum Hur, Sejun Park, Eunho Yang, Sung Ju Hwang, and Jinwoo Shin. Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning. In *NeurIPS*, 2020. 1

[6] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009. 1

[7] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop*, 2013. 1, 2

[8] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world. In *CVPR*, 2019. 1

[9] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *ICLR*, 2017. 1

[10] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, 2020. 1, 2

[11] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017. 1, 2

[12] Zhirong Wu, Alexei A Efros, and Stella X Yu. Improving generalization via scalable neighborhood component analysis. In *ECCV*, 2018. 1