Self-supervised Learning is More Robust to Dataset Imbalance

Hong Liu[1], **Jeff Z. HaoChen**[1], **Adrien Gaidon**[2], **Tengyu Ma**[1]
[1]Stanford University, [2]Toyota Research Institute
{hliu99,jhaochen,tengyuma}@stanford.edu adrien.gaidon@tri.global

Abstract

Self-supervised learning (SSL) learns general visual representations without the need of labels. However, large-scale unlabeled datasets in the wild often have long-tailed label distributions, where we know little about the behavior of SSL. We investigate SSL under dataset imbalance, and find out that existing self-supervised representations are more robust to class imbalance than supervised representations. The performance gap between balanced and imbalanced pre-training with SSL is much smaller than the gap with supervised learning. Second, to understand the robustness of SSL, we hypothesize that SSL learns richer features from frequent data: it may learn label-irrelevant-but-transferable features that help classify the rare classes. In contrast, supervised learning has no incentive to learn features irrelevant to the labels of frequent examples. We validate the hypothesis with semi-synthetic experiments and theoretical analysis on a simplified setting.

1 Introduction

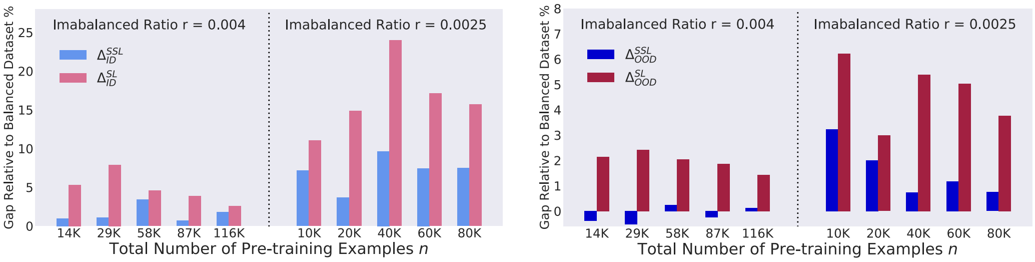
Self-supervised learning (SSL) is an important paradigm of machine learning, because it can leverage the availability of large-scale unlabeled datasets to learn representations for a wide range of downstream tasks and datasets [16, 8, 13, 7, 9]. Current SSL algorithms are mostly trained on curated, balanced datasets, but large-scale unlabeled datasets in the wild are inevitably imbalanced with a long-tailed label distribution [27, 24]. Curating a class-balanced unlabeled dataset requires the knowledge of labels, which defeats the purpose of leveraging unlabeled data by SSL.

The behavior of SSL algorithms under dataset imbalance remains largely underexplored in the literature, but extensive studies do not bode well for supervised learning (SL) with imbalanced datasets. The performance of vanilla supervised methods degrades significantly on class-imbalanced datasets [11, 5, 3], posing challenges to practical applications such as instance segmentation [29] and depth estimation [37]. Many recent works address this issue with various regularization and re-weighting/re-sampling techniques [1, 33, 18, 11, 5, 6, 30, 17, 32].

In this work, we systematically investigate the representation quality of SSL algorithms under class imbalance. Perhaps surprisingly, we find out that off-the-shelf SSL representations are already more robust to dataset imbalance than the representations learned by supervised pre-training. We evaluate the representation quality by linear probe on in-domain (ID) data and finetuning on out-of-domain (OOD) data. We compare the robustness of SL and SSL representations by computing the gap between the performance of the representations pre-trained on balanced and imbalanced datasets of the same sizes. We observe that the balance-imbalance gap for SSL is much smaller than SL, under a variety of configurations with varying dataset sizes and imbalance ratios and with both ID and OOD evaluations (see Figure 1 and Section 2 for more details). This robustness holds even with the same number of samples for SL and SSL, although SSL does not require labels and hence can be more easily applied to larger datasets than SL.

Why is SSL more robust to dataset imbalance? We identify the following underlying cause to answer this fundamental question: SSL learns richer features from the frequent classes than SL does. These features may help classify the rare classes under ID evaluation and are transferable to the downstream tasks under OOD evaluation. For simplicity, consider the situation where rare classes

Workshop on Distribution Shifts, 35th Conference on Neural Information Processing Systems (NeurIPS 2021).



(a) In Domain (ID).

(b) Out of Domain (OOD).

Figure 1: **Relative performance gap** (lower is better) between imbalanced and balanced representation learning. The gap is much smaller for self-supervised (MoCo v2) representations (Δ^{SSL} in blue) vs. supervised ones (Δ^{SL} in red) on long-tailed ImageNet with various number of examples n, across both ID (a) and OOD (b) evaluations. See Equation (1) for the precise definition of the relative performance gap and and Figure 2 for the absolute performance.

have so limited data that both SL and SSL models overfit to the rare data. In this case, it is important for the models to learn diverse features from the frequent classes which can help classify the rare classes. Supervised learning is only incentivized to learn those features relevant to predicting frequent classes and may ignore other features. In contrast, SSL may learn the structures within the frequent classes better—because it is not supervised or incentivized by any labels, it can learn not only the label-relevant features but also other interesting features capturing the intrinsic properties of the input distribution, which may generalize/transfer better to rare classes and downstream tasks.

We empirically validate this intuition by visualizing the features on a semi-synthetic dataset where the label-relevant features and label-irrelevant-but-transferable features are prominently seen by design (cf. Section 3.1). In addition, we construct a toy example where we can rigorously prove the difference between self-supervised and supervised features in Section 3.2.

We sum up our contributions as follows. (1) We are the first to systematically investigate the robustness of self-supervised representation learning to dataset imbalance. (2) We propose and validate an explanation of this robustness of SSL, empirically and theoretically.

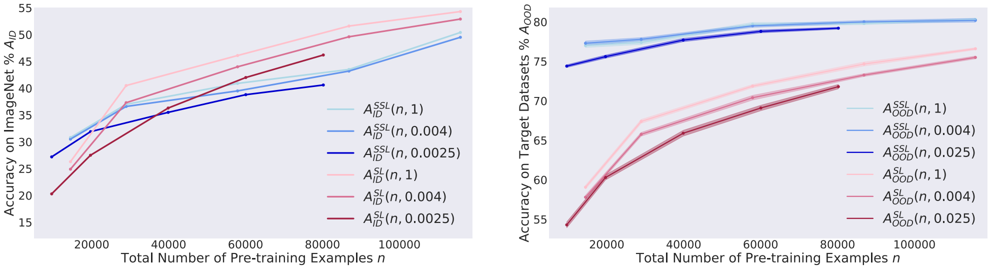
2 Exploring the Effect of Class Imbalance on SSL

2.1 Problem Formulation

Class-imbalanced pre-training datasets. x denotes the input and y denotes the label. Supervised pre-training has access to x and y, whereas self-supervised pre-training only observes x. Given a pre-training distribution $\mathcal{P}$, let r denote the ratio of class imbalance, i.e. the ratio between the probability of the rarest and the most frequent class: $r = \frac{\min_{j\in[C]} \mathcal{P}(y=j)}{\max_{j\in[C]} \mathcal{P}(y=j)} \leq 1$. We use $\mathcal{P}^r$ to denote the dataset with ratio r. We also use $\mathcal{P}^{\text{bal}}$ for the case where $r = 1$, i.e. the dataset is balanced. We assume that for any class $j \in [C]$, the class-conditional distribution $\mathcal{P}^r(x|y = j)$ is the same across datasets for all r. The pre-training dataset $\widehat{\mathcal{P}}^r_n$ consists of n i.i.d. samples from $\mathcal{P}^r$.

Pre-trained models. A feature extractor maps the input to the representations. A linear head classifier is composed with the feature extractor to produce the prediction. SSL algorithms learn the representations from unlabeled data. SL pre-training learns the feature extractor and the linear head from labeled data. We drop the head and only evaluate the quality of feature extractor. Following the standard protocol in prior works [16, 8], we measure the quality of representations on both *in-domain* and *out-of-domain* datasets with either *linear probe* or *fine-tuning*, as detailed below.

In-domain (ID) evaluation tests the performance of representations on the **balanced in-domain** distribution $\mathcal{P}^{\text{bal}}$ with linear probe. Given a feature extractor pre-trained on a pre-training dataset $\widehat{\mathcal{P}}^r_n$ with n data points and imbalance ratio r, we train a linear classifier on a *balanced* dataset and evaluate the representation quality with the accuracy of the learned linear head on $\mathcal{P}^{\text{bal}}$. We denote the ID accuracy of supervised pre-trained representations by $A^{\text{SL}}_{\text{ID}}(n, r)$. Note that $A^{\text{SL}}_{\text{ID}}(n, 1)$ stands for the result with balanced pre-training dataset. For SSL, we denote the accuracy by $A^{\text{SSL}}_{\text{ID}}(n, r)$.



(a) ImageNet, ID

(b) ImageNet, OOD

Figure 2: **Representation quality on balanced and imbalanced ImageNet.** The gap between balanced and imbalanced datasets with the same n is larger for SL. The accuracy of SL is better with reasonably large n in ID evaluation, while SSL representations perform better in OOD evaluation.

Out-of-domain (OOD) evaluation tests the performance of representations by fine-tuning the feature extractor and the head on the *downstream* target distribution $\mathcal{P}_t$. Starting from a feature extractor (pre-trained on a dataset of size n and imbalance ratio r) and a randomly initialized classifier, we fine-tune the whole model on the target dataset $\widehat{\mathcal{P}}_t$, and evaluate the representation quality by the expected top-1 accuracy on $\mathcal{P}_t$. We use $A_{\text{OOD}}^{\text{SL}}(n, r)$ and $A_{\text{OOD}}^{\text{SSL}}(n, r)$ to denote the resulting accuracies of supervised and self-supervised representations, respectively.

Summary of varying factors. We aim to study the effect of class imbalance to feature qualities on a diverse set of configurations with the following varying factors: (1) the number of examples in pre-training n, (2) the imbalance ratio of the pre-training dataset r, (3) ID or OOD evaluation, and (4) self-supervised learning algorithms: MoCo v2 [16], or SimSiam [9].

Experimental Setup. We pre-train the representations on ImageNet [28] with a wide range of numbers of examples and ratios of imbalance. Following Liu et al. [24], we consider the long-tailed Pareto label distributions with imbalance ratio in $\{1, 0.004, 0.0025\}$. For each imbalance ratio, we further downsample the dataset with a sampling ratio in $\{0.75, 0.5, 0.25, 0.125\}$ to form datasets with varying sizes. For ID evaluation, we use the original ImageNet training set for the training phase of linear probe and use the original validation set for the final evaluation. For OOD evaluation, we fine-tune the pre-trained feature extractors on the downstream tasks. For self-supervised pre-training, we consider MoCo v2 [16] and SimSiam [9]. Implementation details are deferred to Section A.1.

2.2 Results: SSL is More Robust than SL to Dataset Imbalance

In Figure 2, for both ID and OOD evaluations, the gap of SSL representations learned on balanced and imbalanced datasets with the same number of pre-training examples, i.e., $A^{\text{SSL}}(n, 1) - A^{\text{SSL}}(n, r)$, is smaller than the gap of SL representations, i.e., $A^{\text{SL}}(n, 1) - A^{\text{SL}}(n, r)$. Furthermore, we compute the relative accuracy gap to balanced dataset $\Delta^{\text{SSL}}(n, r) \triangleq (A^{\text{SSL}}(n, 1) - A^{\text{SSL}}(n, r))/A^{\text{SSL}}(n, 1)$ in Figure 1. With the same number of pre-training examples, the relative gap of SSL representations between balanced and imbalanced datasets is smaller than that of SL representations,

$$\Delta^{\text{SSL}}(n, r) \triangleq \frac{A^{\text{SSL}}(n, 1) - A^{\text{SSL}}(n, r)}{A^{\text{SSL}}(n, 1)} \ll \Delta^{\text{SL}}(n, r) \triangleq \frac{A^{\text{SL}}(n, 1) - A^{\text{SL}}(n, r)}{A^{\text{SL}}(n, 1)}. \tag{1}$$

We also note that comparing the robustness with the same number of data is actually in favor of SL, because SSL is more easily applied to larger datasets without the need of collecting labels.

ID vs. OOD. As shown in Figure 2, representations from SSL perform better than SL pre-training in ID evaluation with reasonably large n, while SSL pre-training is better in OOD evaluation. This phenomenon is orthogonal to the observation that SSL is more robust to dataset imbalance, and is consistent with recent works (e.g., Chen et al. [8], He et al. [16]) which also observed that SSL performs slightly worse than SL on balanced ID evaluation but better on OOD tasks.

3 Analysis

Where does the robustness stem from? In this section, we propose a possible reason and justify it with theoretical and empirical analyses. **SSL learns richer features from frequent data that are transferable to rare data.** The rare classes of the imbalanced dataset can contain only a few examples, making it hard to learn proper features from them. In this case, one may want to resort to

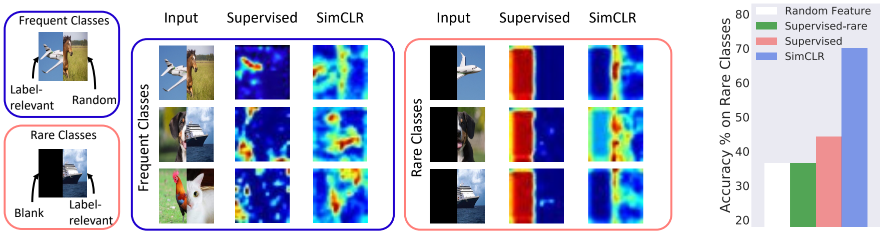


Figure 3: **Visualization of the semi-synthetic setting. Left**: The semi-synthetic dataset. **Middle**: Visualization of feature activations with Grad-CAM. **Right**: Accuracies evaluated on rare classes.

the features from the frequent classes. Due to the supervised nature of classification, SL mainly learns the features for classifying the frequent classes and may neglect features which can transfer to the rare classes. In contrast, in SSL, without the bias from the labels, the models learn richer features—both features for classifying the frequent classes and features transferable to the rare classes.

3.1 Illustrative Semi-synthetic Experiments

Dataset. We consider an imbalanced pre-training dataset with label-relevant and label-irrelevant-but-transferable features modified from CIFAR-10 as shown in Figure 3 (Left). Classes 1-5 (frequent) each contains 5000 examples. Classes 6-10 (rare) each has 10 examples. The left half of an image of classes 1-5 is from classes 1-5 of the original CIFAR-10 and corresponds to the label of that example. The right half is from a random label-irrelevant image of CIFAR-10. In contrast, the left half of an example from classes 6-10 is blank, whereas the right half is label-relevant and from classes 6-10 of the original CIFAR-10. In this setting, features from the left halves are correlated to the classification of the frequent classes, while features from the right halves are label-irrelevant for the frequent classes, but can help classify the rare classes. Note that features from the right halves cannot be directly learned from the rare examples with only 10 examples per class.

Pre-training. We pre-train the representations on the aforementioned dataset. We use SimCLR with ResNet-50. After pre-training, we fix the representations evaluate with linear probe on the 5 rare classes to test if the model learns proper features for the rare classes.

Results. As a sanity check, we also include a supervised model trained on only the 50 rare examples (Supervised-rare in Figure 3 (Right)). As expected, the accuracy is 36.5%, which is almost the same as random representations, indicating the model cannot learn the features for the rare classes due to lack of data. We then compare supervised learning with self-supervised learning on the imbalanced dataset. SSL representations performs much better than supervised representations on the rare classes (70.1% vs 44.3%). In Figure 3 (Middle), SL mostly activate the left halves on frequent and rare examples, indicating it learns features on the left. In sharp contrast, SSL activates the whole image on the frequent examples and the right halves on the rare, indicating it learns features from both parts.

3.2 Rigorous Analysis on A Toy Setting

To justify the above conjecture, we instantiate supervised and self-supervised learning in a setting where the features helpful to classify the frequent classes and features transferable to the rare classes can be clearly separated. In this case, we prove that self-supervised learning learns better features than supervised learning.

Data distribution. Let e_1, e_2 be two orthogonal unit-norm vectors in the d-dimensional Euclidean space. Consider the following pre-training distribution $\mathcal{P}$ of a 3-way classification problem, where the class label $y \in [3]$. The input x is generated as follows. Let $\tau > 0$ and $\rho > 0$ be hyperparameters of the distribution. First sample q uniformly from $\{0, 1\}$ and $\xi \sim \mathcal{N}(0, I)$ from Gaussian distribution. For the first class ($y = 1$), set $x = e_1 - q\tau e_2 + \rho\xi$. For the second class ($y = 2$), set $x = -e_1 - q\tau e_2 + \rho\xi$. For the third class ($y = 3$), set $x = e_2 + \rho\xi$. Classes 1 and 2 are frequent classes, while class 3 is the rare class, i.e., $\frac{\mathcal{P}(y=3)}{\mathcal{P}(y=1)}, \frac{\mathcal{P}(y=3)}{\mathcal{P}(y=2)} = o(1)$. See Figure 8 for an illustration of this data distribution. In this case, both e_1 and e_2 are features from the frequent classes 1 and 2. However, only e_1 helps classify the frequent classes and only e_2 can be transferred to the rare classes.

Algorithm formulations. For supervised learning, we train a two-layer linear network $f_{W_1,W_2}(x) \triangleq W_2 W_1 x$ with weight matrices $W_1 \in \mathbb{R}^{m\times d}$ and $W_2 \in \mathbb{R}^{3\times m}$ for some $m \geq 3$, and then use the first layer $W_{\text{SL}} = W_1$ as the feature for downstream tasks. Given a linearly separable labeled dataset,

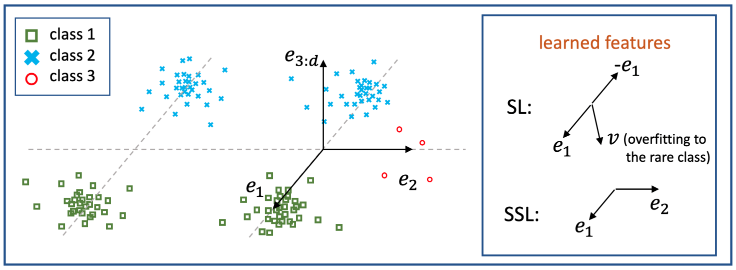


Figure 4: **Explaining SSL's robustness in a toy setting.** e_1 and e_2 are two orthogonal directions in the d-dimensional Euclidean space that decides the labels, and $e_{3:d}$ represents the other $d-2$ dimensions. Classes 1 and 2 are frequent classes and the third class is rare. To classify the three classes, the representations need to contain both e_1 and e_2 directions. Supervised learning learns direction e_1 from the frequent classes (which is necessary and sufficient to identify classes 1 and 2) and some overfitting direction v from the rare class which has insufficient data. Note that v might be mostly in the $e_{3:d}$ directions due to overfitting. In contrast, SSL learns both e_1 and e_2 directions from the frequent classes because they capture the intrinsic structures of the inputs (e.g., e_1 and e_2 are the directions with the largest variances), even though e_2 does not help distinguish the frequent classes. The direction e_2 learned from frequent data by SSL can help classify the rare class.

we learn such a network with minimal norm $\|W_1\|_F^2 + \|W_2\|_F^2$ subject to the margin constraint $f_{W_1,W_2}(x)_y \geq f_{W_1,W_2}(x)_{y'} + 1$ for all data (x, y) in the dataset and $y' \neq y$.[1] For self-supervised learned, similar to SimSiam [8], we construct positive pairs $(x+\xi, x+\xi')$ where x is from the empirical dataset, ξ and ξ' are independent random perturbations. We learn a matrix $W_{\text{SSL}} \in \mathbb{R}^{m\times d}$ which minimizes $-\hat{\mathbb{E}}[(W(x+\xi))^T(W(x+\xi'))] + \frac{1}{2}\|W^\top W\|_F^2$, where the expectation $\hat{\mathbb{E}}$ is over the empirical dataset and the randomness of ξ and ξ'. The regularization term $\frac{1}{2}\|W^\top W\|_F^2$ is introduced only to make the learned features more mathematically tractable. We use $W_{\text{SSL}}x$ as the feature of data x in the downstream task.

Main intuitions. We compare the features learned by SSL and supervised learning on an imbalanced dataset that contains an abundant (poly in d) number of data from the frequent classes but only a small (sublinear in d) number of data from the rare class. The key intuition behind our analysis is that supervised learning learns only the e_1 direction (which helps classify class 1 vs. class 2) and some random direction that overfits to the rare class. In contrast, self-supervised learning learns both e_1 and e_2 directions from the frequent classes. Since how well the feature helps classify the rare class (in ID evaluation) depends on how much it correlates with the e_2 direction, SSL provably learns features that help classify the rare class, while supervised learning fails. This intuition is formalized by the following theorem.

Theorem 3.1. *Let n_1, n_2, n_3 be the number of data from the three classes respectively. Let $\rho = d^{-\frac{1}{5}}$ and $\tau = d^{\frac{1}{5}}$ in the data generative model. For $n_1, n_2 = \Theta(\text{poly}(d))$ and $n_3 \leq d^{\frac{1}{5}}$, with probability at least $1 - O(e^{-d^{\frac{1}{10}}})$, the following statements hold for any feature dimension $m \geq 3$:*

- *Let $W_{\text{SL}} = [w_1, w_2, \cdots, w_m]^\top$ be the feature learned by SL, then $|\langle e_2, w_i\rangle| \leq O(d^{-\frac{1}{20}})$ for $i \in [m]$.*
- *Let $W_{\text{SSL}} = [\tilde{w}_1, \tilde{w}_2, \cdots, \tilde{w}_m]^\top$ be the feature learned by SSL, then $\|\Pi e_2\|_2 \geq 1 - O(d^{-\frac{1}{5}})$, where Π projects e_2 onto the row span of W_{SSL}.*

4 Conclusion

We discover that self-supervised representations are more robust to class imbalance than supervised representations and explore the underlying cause of this phenomenon. We hope our study can inspire analysis of self-supervised learning in broader environments in the wild such as domain shift.

[1] Previous work shows that deep linear networks trained with gradient descent using logistic loss converge to this min norm solution in direction [19].

References

[1] Shin Ando and Chun Yuan Huang. Deep over-sampling framework for classifying imbalanced data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 770–785, 2017.

[2] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning*, 2019.

[3] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

[4] Jonathon Byrd and Zachary Lipton. What is the effect of importance weighting in deep learning? In *International Conference on Machine Learning*, pp. 872–881. PMLR, 2019.

[5] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, volume 32, pp. 1565–1576. Curran Associates, Inc., June 2019.

[6] Kaidi Cao, Yining Chen, Junwei Lu, Nikos Arechiga, Adrien Gaidon, and Tengyu Ma. Heteroskedastic and imbalanced deep learning with adaptive regularization. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=mEdwVCRJuX4`.

[7] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 33:9912–9924, 2020.

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, PMLR, 13–18 Jul 2020.

[9] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15750–15758, June 2021.

[10] Elijah Cole, Xuan Yang, Kimberly Wilber, Oisin Mac Aodha, and Serge Belongie. When does contrastive visual representation learning work? *arXiv preprint arXiv:2105.05837*, 2021.

[11] 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*, pp. 9268–9277, 2019.

[12] Priya Goyal, Mathilde Caron, Benjamin Lefaudeux, Min Xu, Pengchao Wang, Vivek Pai, Mannat Singh, Vitaliy Liptchinsky, Ishan Misra, Armand Joulin, et al. Self-supervised pretraining of visual features in the wild. *arXiv preprint arXiv:2103.01988*, 2021.

[13] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 33:21271–21284, 2020.

[14] Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *arXiv preprint arXiv:2106.04156*, 2021.

[15] 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*, pp. 770–778, 2016.

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, June 2020.

[17] 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*, pp. 6626–6636, 2021.

[18] Muhammad Abdullah Jamal, Matthew Brown, Ming-Hsuan Yang, Liqiang Wang, and Boqing Gong. Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[19] Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. *arXiv preprint arXiv:1810.02032*, 2018.

[20] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020.

[21] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.

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

[23] Jason D Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. Predicting what you already know helps: Provable self-supervised learning. *arXiv preprint arXiv:2008.01064*, 2020.

[24] 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 (CVPR)*, June 2019.

[25] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[26] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505, 2012.

[27] William J Reed. The pareto, zipf and other power laws. *Economics letters*, 74(1):15–19, 2001.

[28] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[29] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1513–1524. Curran Associates, Inc., 2020.

[30] Junjiao Tian, Yen-Cheng Liu, Nathan Glaser, Yen-Chang Hsu, and Zsolt Kira. Posterior re-calibration for imbalanced datasets. *arXiv preprint arXiv:2010.11820*, 2020.

[31] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[32] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *International Conference on Learning Representations*, 2021.

[33] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 7032–7042, 2017.

[34] Colin Wei, Sang Michael Xie, and Tengyu Ma. Why do pretrained language models help in downstream tasks? an analysis of head and prompt tuning. *arXiv preprint arXiv:2106.09226*, 2021.

[35] Da Xu, Yuting Ye, and Chuanwei Ruan. Understanding the role of importance weighting for deep learning. In *International Conference on Learning Representations*, 2021.

[36] Yuzhe Yang and Zhi Xu. Rethinking the value of labels for improving class-imbalanced learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19290–19301. Curran Associates, Inc., 2020.

[37] Yuzhe Yang, Kaiwen Zha, Yingcong Chen, Hao Wang, and Dina Katabi. Delving into deep imbalanced regression. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11842–11851, 18–24 Jul 2021.

[38] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. 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.

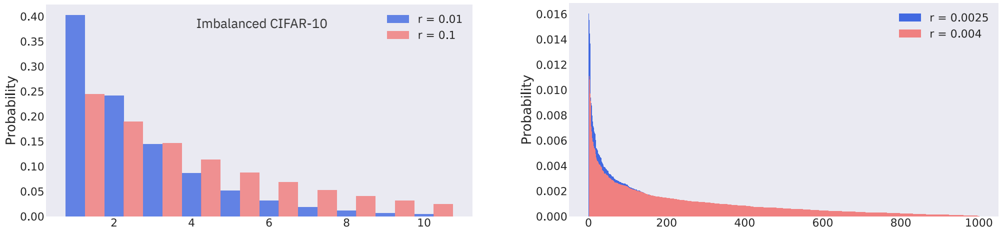


Figure 5: **Visualization of the label distributions.** We visualize the label distributions of the imbalanced CIFAR-10 and ImageNet. We consider two imbalance ratios r for each dataset.

A Details of Section 2

A.1 Implementation Details

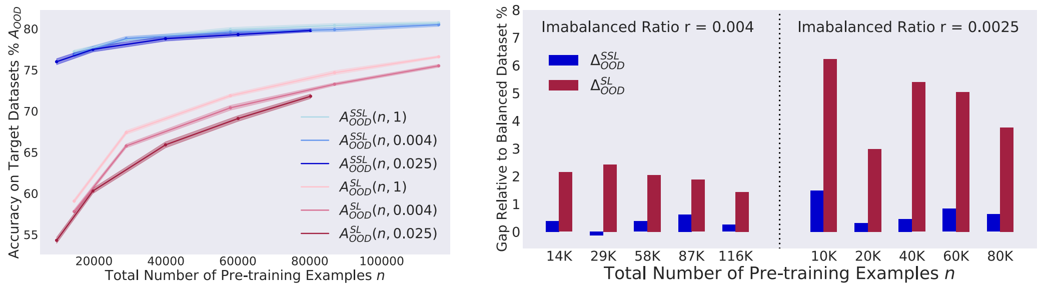
Generating Pre-training Datasets. CIFAR-10 [22] contains 10 classes with 5000 examples per class. We use exponential imbalance, i.e. for class c, the number of examples is $5000 \times e^{\beta(c-1)}$. we consider imbalance ratio $r \in \{0.1, 0.01\}$, i.e. the number of examples belonging to the rarest class is 500 or 50. The total n_s is therefore 20431 or 12406. ImageNet-LT is constructed by Liu et al. [24], which follows the Pareto distribution with the power value 6. The number of examples from the rarest class is 5. We construct a long tailed ImageNet following the Pareto distribution with more imbalance, where the number of examples from the rarest class is 3. The total number of examples n_s is 115846 and 80218 respectively. For each ratio of imbalance, we further downsample the dataset with the sampling ratio in $\{0.75, 0.5, 0.25, 0.125\}$ to formulate different number of examples. To compare with the balanced setting fairly, we also sample balanced versions of datasets with the same number of examples. Note that each variant of the dataset is fixed after construction for all algorithms. See the visualization of label distributions of dataset variants in Figure 5.

Training Procedure. We use ResNet-50 as backbones. For supervised pre-training, we follow the standard protocol of He et al. [15] and Kang et al. [20]. On the standard ImageNet-LT, we train the models for 90 epochs with step learning rate decay. For down-sampled variants, the training epochs are selected with cross validation. Fo self-supervised learning, the initial learning rate on the standard ImageNet-LT is set to 0.025 with batch-size 256. We train the model for 300 epochs on the standard ImageNet-LT and adopt cosine learning rate decay following [16, 9]. We train the models for more epochs on the down sampled variants to ensure the same number of total iterations. The code on CIFAR-10 LT is adapted from `https://github.com/Reza-Safdari/SimSiam-91.9-top1-acc-on-CIFAR10`. We run each evaluation experiment with 3 seeds and report the average and standard deviation in the figures.

Evavluation. For in-domain evaluation (ID), we first train the the representations on the aforementioned dataset variants, and then train the linear head classifier on the *full balanced* CIFAR10 or ImageNet. We set the initial learning rate to 30 when training the linear head with batch-size 4096 and train for 100 epochs in total. For in-domain out-of-domain evaluation (OOD) on ImageNet, we first train the the representations on the aforementioned dataset variants, and then fine-tune the model to CUB-200 [31], Stanford Cars [21], Oxford Pets [26], and Aircrafts [25]. The number of examples of these target datasets ranges from 2k to 10k, which is a reasonable scale as the number of examples of the pre-training dataset variants ranges from 10k to 110k. The representation quality is evaluated with the average performance on the four tasks. We set the initial learning rate to 0.1 in fine-tuning train for 150 epochs in total. For in-domain out-of-domain evaluation (OOD) on CIFAR-10, we use STL-10 as the downstream target tasks and perform linear probe.

A.2 Additional Results

To validate the phenomenon observed in Section 2 is consistent for different self-supervised learning algorithms, we provide the OOD evaluation results of SimSiam trained on ImageNet variants in Figure 6. SimSiam representations are also less sensitive to class imbalance than supervised representations.



(a) OOD Accuracy.

(b) Relative Gap.

Figure 6: **OOD Results of SimSiam on ImageNet.** Simsiam also demonstrates robustness to class imbalance compared to supervised learning

Table 1: Numbers in Figure 2 and Figure 6.

Imbalanced Ratio r	$r = 1$, balanced					$r = 0.004$					$r = 0.0025$				
Data Quantity n	116K	87K	58K	29K	14K	116K	87K	58K	29K	14K	80K	60K	40K	20K	10K
MoCo V2, ID	50.4	43.5	40.9	37.0	30.8	49.5	43.2	39.5	36.6	30.5	40.6	38.8	35.5	31.9	27.2
MoCo V2, OOD	80.3	79.8	79.7	77.4	77.0	80.2	80.1	79.5	77.8	77.3	79.2	78.8	77.7	75.6	74.4
Supervised, ID	54.3	51.6	46.1	40.5	26.3	52.9	49.6	44.0	37.3	24.9	46.1	42.0	36.3	27.5	20.3
Supervised, OOD	76.6	74.7	71.9	67.4	59.1	75.5	73.3	70.4	65.8	57.8	71.8	69.1	65.9	60.3	54.3
SimSiam, OOD	80.7	80.4	79.9	78.7	77.2	80.6	79.9	79.6	78.8	76.9	79.8	79.3	78.8	77.5	76.0

We also provide the numbers of Figure 2 and Figure 6 in Table 1.

B Details of Section 3.1

We first generate the balanced semi-synthetic dataset with 5000 examples per class. The left halves of images from classes 1-5 correspond to the labels, while the right halves are random. The left halves of images from class 6-10 are blank, whereas the right halves correspond to the labels. We then generate the imbalanced dataset, which consists of the 5000 examples per class from classes 1-5 (frequent classes), and 10 examples per class from classes 6-10 (rare classes). We use Grad-CAM implementation based on `https://github.com/meliketoy/gradcam.pytorch` and SimCLR implementation from `https://github.com/leftthomas/SimCLR`. We provide examples and Grad-CAM of the semi-synthetic datasets in Figure 7. To avoid confusing the left and right parts, we disable the random horizontal flip in data augmentation.

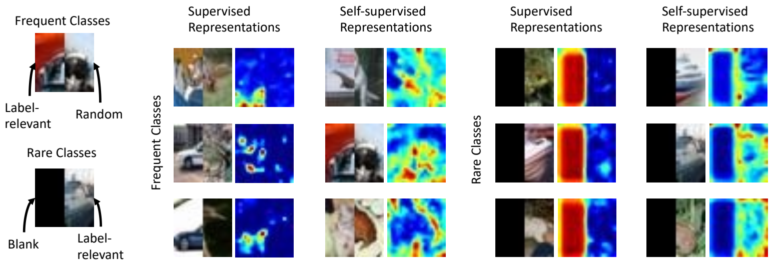

Frequent Classes
Label-relevant
Random
Rare Classes
Blank
Label-relevant
Supervised Representations
Self-supervised Representations
Frequent Classes
Supervised Representations
Self-supervised Representations
Rare Classes

Figure 7: **Examples of the semi-synthetic Datasets and Grad-CAM visualizations.**

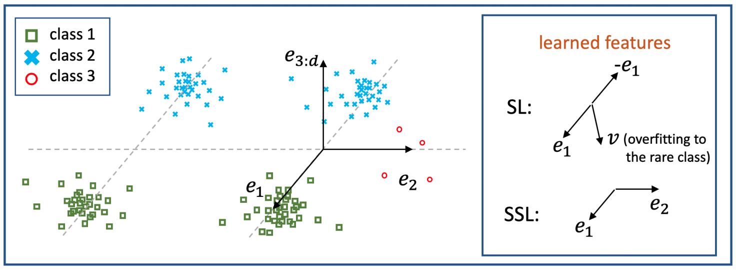


Figure 8: **Explaining SSL's robustness in a toy setting.** e_1 and e_2 are two orthogonal directions in the d-dimensional space that decide the labels, and $e_{3:d}$ contains the other $d-2$ dimensions. Classes 1 and 2 are frequent classes and class 3 is rare. To classify them, the representations need both e_1 and e_2. SL learns e_1 from the frequent classes to identify classes 1 and 2, and an overfitting direction v from class 3 with limited data. In contrast, SSL learns both e_1 and e_2 from the frequent classes, even though e_2 does not help distinguish the frequent classes. e_2 can help classify the rare class.

C Proof of Theorem 3.1

Data distribution. Let e_1, e_2 be two orthogonal unit-norm vectors in the d-dimensional Euclidean space. Consider the following pre-training distribution $\mathcal{P}$ of a 3-way classification problem, where the class label $y \in [3]$. The input x is generated as follows. Let $\tau > 0$ and $\rho > 0$ be hyperparameters of the distribution. First sample q uniformly from $\{0, 1\}$ and $\xi \sim \mathcal{N}(0, I)$ from Gaussian distribution. For the first class ($y = 1$), set $x = e_1 - q\tau e_2 + \rho\xi$. For the second class ($y = 2$), set $x = -e_1 - q\tau e_2 + \rho\xi$. For the third class ($y = 3$), set $x = e_2 + \rho\xi$. Classes 1 and 2 are frequent classes, while class 3 is the rare class, i.e., $\frac{\mathcal{P}(y=3)}{\mathcal{P}(y=1)}, \frac{\mathcal{P}(y=3)}{\mathcal{P}(y=2)} = o(1)$. See Figure 8 for an illustration of this data distribution. In this case, both e_1 and e_2 are features from the frequent classes 1 and 2. However, only e_1 helps classify the frequent classes and only e_2 can be transferred to the rare classes.

Algorithm formulations. For supervised learning, we train a two-layer linear network $f_{W_1,W_2}(x) \triangleq W_2 W_1 x$ with weight matrices $W_1 \in \mathbb{R}^{m\times d}$ and $W_2 \in \mathbb{R}^{3\times m}$ for some $m \ge 3$, and then use the first layer $W_1 x$ as the feature for downstream tasks. Given a linearly separable labeled dataset, we learn such a network with minimal norm $\|W_1\|_F^2 + \|W_2\|_F^2$ subject to the margin constraint $f_{W_1,W_2}(x)_y \ge f_{W_1,W_2}(x)_{y'} + 1$ for all data (x, y) in the dataset and $y' \ne y$.[2] Notice that when the norm is minimized, the row span of W_1 is exactly the same as the row span of matrix $W_{\mathrm{SL}} \triangleq [w_1, w_2, w_3]^\top \in \mathbb{R}^{3\times d}$ which minimizes $\|w_1\|_2^2 + \|w_2\|_2^2 + \|w_3\|_2^2$ subject to the margin constraint $w_y^\top x \ge w_{y'}^\top x + 1$ for all empirical data (x, y) and $y' \ne y$. Therefore, feature $W_1 x$ is equivalent to $W_{\mathrm{SL}} x$ for downstream linear probing, so we only need to analyze the feature quality of $W_{\mathrm{SL}} x$ for supervised learning. For self-supervised learned, similar to SimSiam [8], we construct positive pairs $(x + \xi, x + \xi')$ where x is from the empirical dataset, ξ and ξ' are independent random perturbations. We learn a matrix $W_{\mathrm{SSL}} \in \mathbb{R}^{2\times d}$ which minimizes $-\hat{\mathbb{E}}[(W(x + \xi))^T (W(x + \xi'))] + \frac{1}{2}\|W^\top W\|_F^2$, where the expectation $\hat{\mathbb{E}}$ is over the empirical dataset and the randomness of ξ and ξ'. The regularization term $\frac{1}{2}\|W^\top W\|_F^2$ is introduced only to make the learned features more mathematically tractable. We use $W_{\mathrm{SSL}} x$ as the feature of data x in the downstream task.

We notate data from the first class as $x_i^{(1)} = e_1 - q_i^{(1)}\tau e_2 + \rho\xi_i^{(1)}$ where $i \in [n_1]$ and $q_i^{(1)} \in \{0, 1\}$. Similarly, we notate data from the second class as $x_i^{(2)} = -e_1 - q_i^{(2)}\tau e_2 + \rho\xi_i^{(2)}$ where $i \in [n_2]$ and $q_i^{(1)} \in \{0, 1\}$. We notate data from the third class as $x_i^{(3)} = e_2 + \rho\xi_i^{(3)}$ where $i \in [n_3]$. Notice that all $\xi_i^{(k)}$ are independently sampled from $\mathcal{N}(0, I)$.

We first introduce the following lemma, which gives some high probability properties of independent Gaussian random variables.

Lemma C.1. *Let $\xi_i \sim \mathcal{N}(0, I)$ for $i \in [n]$. Then, for any $n \le \mathrm{poly}(d)$, with probability at least $1 - e^{-d^{\frac{1}{10}}}$ and large enough d, we have:*

- $|\langle \xi_i, e_1\rangle| \le d^{\frac{1}{10}}$, $|\langle \xi_i, e_2\rangle| \le d^{\frac{1}{10}}$ *and* $|\|\xi_i\|_2^2 - d| \le 4d^{\frac{3}{4}}$ *for all* $i \in [n]$.

[2] Previous work shows that deep linear networks trained with gradient descent using logistic loss converge to this min norm solution in direction [19].

- $|\langle \xi_i, \xi_j \rangle| \leq 3d^{\frac{3}{5}}$ *for all* $i \neq j$.

Proof of Lemma C.1. Let $\xi, \xi' \sim \mathcal{N}(0, I)$ be two independent random variables. By the tail bound of normal distribution, we have

$$\Pr\left(|\langle \xi, e_1 \rangle| \geq d^{\frac{1}{10}}\right) \leq d^{-\frac{1}{10}} \cdot e^{-\frac{d^{\frac{1}{5}}}{2}}. \tag{2}$$

By the tail bound of χ^2_d distribution, we have

$$\Pr\left(|\|\xi\|_2^2 - d| \geq 4d^{\frac{3}{4}}\right) \leq 2e^{-\sqrt{d}}. \tag{3}$$

Since the directions of ξ and ξ' are independent, we can bound their correlation with the norm of ξ times the projection of ξ' onto ξ:

$$\Pr\left(|\langle \xi, \xi' \rangle| \geq 3d^{\frac{3}{5}}\right) \leq \Pr\left(\|\xi\|_2 \geq \sqrt{d} + 2d^{\frac{3}{8}}\right) + \Pr\left(|\langle \xi', \frac{\xi}{\|\xi\|} \rangle| \geq d^{\frac{1}{10}}\right) \leq \frac{e^{-\frac{d^{\frac{1}{5}}}{2}}}{d^{\frac{1}{10}}} + 2e^{-\sqrt{d}}. \tag{4}$$

Since every ξ_i and ξ_j are independent when $i \neq j$, by the union bound, we know that with probability at least $1 - (n^2 + 2n)(\frac{e^{-\frac{d^{\frac{1}{5}}}{2}}}{d^{\frac{1}{10}}} + 2e^{-\sqrt{d}})$, we have $|\langle \xi_i, e_1 \rangle| \leq d^{\frac{1}{10}}$, $|\langle \xi_i, e_2 \rangle| \leq d^{\frac{1}{10}}$ and $|\|\xi_i\|_2^2 - d| \leq 4d^{\frac{3}{4}}$ for all $i \in [n]$, and also $|\langle \xi_i, \xi_j \rangle| \leq 3d^{\frac{3}{5}}$ for all $i \neq j$. Since the error probability is exponential in d, for large enough d, the error probability is smaller than $e^{-d^{\frac{1}{10}}}$, which finishes the proof. □

Using the above lemma, we can prove the following lemma which constructs a linear classifier of the empirical dataset with relatively large margin and small norm.

Lemma C.2. *In the setting of Theorem 3.1, let* $w_1^* = e_1$, $w_2^* = -e_1$, $w_3^* = \frac{1}{\rho d}\sum_{i=1}^{n_3} \xi_i^{(3)}$. *Apply Lemma C.1 to the set of all* $\xi_i^{(k)}$ *where* $k \in [3]$ *and* $i \in [n_k]$. *When the high probability outcome of Lemma C.1 happens, the margin of classifier* $\{w_1^*, w_2^*, w_3^*\}$ *is at least* $1 - O(d^{-\frac{1}{10}})$. *Furthermore, we have* $\|w_3^*\|_2^2 \leq O(d^{-\frac{1}{5}})$.

Proof of Lemma C.2. When the high probability outcome of Lemma C.1 happens, we give a lower bound on the margin for all data in the dataset. For data $x = x_i^{(1)}$ in class 1, we have

$$w_1^{*\top} x = 1 + \langle \xi_i^{(1)}, e_1 \rangle \rho \geq 1 - \rho d^{\frac{1}{10}}, \tag{5}$$

$$w_2^{*\top} x = -1 + \langle \xi_i^{(1)}, e_1 \rangle \rho \leq -1 + \rho d^{\frac{1}{10}}, \tag{6}$$

$$w_3^{*\top} x = \frac{1}{\rho d}\left(e_1 - q_i^{(1)} \tau e_2 + \rho \xi_i^{(1)}\right)^\top \left(\sum_{j=1}^{n_3} \xi_j^{(3)}\right) \leq \frac{n_3(\tau + 1)}{\rho d} d^{\frac{1}{10}} + \frac{3n_3}{d} d^{\frac{3}{5}}. \tag{7}$$

So the margin on data $(x_i^{(1)}, 1)$ is

$$w_1^{*\top} x - w_3^{*\top} x \geq 1 - \rho d^{\frac{1}{10}} - \frac{n_3(\tau + 1)}{\rho d} d^{\frac{1}{10}} - \frac{3n_3}{d} d^{\frac{3}{5}} \geq 1 - O(d^{-\frac{1}{10}}). \tag{8}$$

Similarly, for data $x_i^{(2)}$ in class 2, the margin is at least $1 - O(d^{-\frac{1}{10}})$.

For data $x = x_i^{(3)}$ in class 3, we have

$$w_3^{*\top} x = \frac{1}{\rho d}\left(\sum_{j=1}^{n_3} \xi_j^{(3)}\right)^\top \left(e_2 + \rho \xi_i^{(3)}\right) \geq \frac{1}{d}\|\xi_i^{(3)}\|_2^2 - \frac{3n_3}{d} d^{\frac{3}{5}} - \frac{n_3 d^{\frac{1}{10}}}{\rho d} \geq 1 - O(d^{-\frac{1}{5}}). \tag{9}$$

On the other hand,

$$w_1^{*\top} x = \langle \rho \xi_i^{(3)}, e_1 \rangle \leq \rho d^{\frac{1}{10}}, \tag{10}$$

$$w_2^{*\top} x = \langle \rho \xi_i^{(3)}, -e_1 \rangle \leq \rho d^{\frac{1}{10}}. \tag{11}$$

So the margin is

$$w_3^{*\top} x - \max\{w_1^{*\top} x, w_2^{*\top} x\} \geq w_3^{*\top} x - \rho d^{\frac{1}{10}} \geq 1 - O(d^{-\frac{1}{10}}). \tag{12}$$

Finally, noticing that $\|w_3^*\|_2 \leq \frac{2n_3\sqrt{d}}{\rho d} \leq 2d^{-\frac{1}{10}}$ finishes the proof. □

Now we are ready to prove the supervised learning part of Theorem 3.1 using our construction:

Proof of Theorem 3.1 (supervised learning part). We frist apply Lemma C.1 to the set of all $\xi_i^{(k)}$ where $k \in [3]$ and $i \in [n_k]$. We consider the situation when the high probability outcome of Lemma C.1 holds (which happens with probability at least $1 - e^{-d^{\frac{1}{10}}}$). By Lemma C.2, the constructed classifier $\{w_1^*, w_2^*, w_3^*\}$ has margin $\alpha \geq 1 - O(d^{-\frac{1}{10}})$ in this case. As a result, $\{\frac{1}{\alpha}w_1^*, \frac{1}{\alpha}w_2^*, \frac{1}{\alpha}w_3^*\}$ is a classifier with margin 1 and norm bounded by

$$\|\frac{1}{\alpha}w_1^*\|_2^2 + \|\frac{1}{\alpha}w_2^*\|_2^2 + \|\frac{1}{\alpha}w_3^*\|_2^2 = \frac{2 + \|w_3^*\|_2^2}{\alpha^2} \leq 2 + O(d^{-\frac{1}{10}}). \tag{13}$$

Let $\{w_1, w_2, w_3\}$ be the solution of supervised learning. Since its norm cannot be larger than the constructed one, we have $\|w_1\|_2^2 + \|w_2\|_2^2 + \|w_3\|_2^2 \leq 2 + O(d^{-\frac{1}{10}})$. By standard concentration inequality, when $n_1 \geq \text{poly}(d)$, with probability at least $1 - e^{d^{-\frac{1}{10}}}$, we have

$$\left|\mathbb{E}_{i\in[n_1], q_i^{(1)}=0}[x_i^{(1)}] - e_1\right| \leq d^{-\frac{1}{10}}, \tag{14}$$

where the expectation is over all the data from class 1 that satisfies $q_i^{(1)} = 0$. By the definition of $\{w_1, w_2, w_3\}$ we know $(w_1 - w_3)^\top x_i^{(1)} \geq 1$ for all $i \in [n_1]$, hence averaging over all the class 1 data with $q_i^{(1)} = 0$ and using the above inequality gives us

$$(w_1 - w_3)^\top e_1 \geq 1 - \|w_1 - w_3\|_2 \cdot d^{-\frac{1}{10}} \geq 1 - O(d^{-\frac{1}{10}}). \tag{15}$$

A similar analysis for class 2 data gives us

$$(w_2 - w_3)^\top(-e_1) \geq 1 - O(d^{-\frac{1}{10}}). \tag{16}$$

Now we prove that w_1, w_2, w_3 all have small correlation with e_2. Without loss of generality, we assume $w_3^\top e_1 \triangleq t \geq 0$. If $t \geq \frac{1}{2}$, we have

$$\langle w_1, e_1\rangle^2 + \langle w_2, e_1\rangle^2 + \langle w_3, e_1\rangle^2 \geq \left(t + 1 - O(d^{-\frac{1}{10}})\right)^2 > 2.25 - O(d^{-\frac{1}{10}}), \tag{17}$$

which contradicts with $\|w_1\|_2^2 + \|w_2\|_2^2 + \|w_3\|_2^2 \leq 2 + O(d^{-\frac{1}{10}})$. Therefore, there must be $t \leq \frac{1}{2}$, hence

$$\begin{aligned}
&\langle w_1, e_1\rangle^2 + \langle w_2, e_1\rangle^2 + \langle w_3, e_1\rangle^2 && (18)\\
&\geq \left(1 + t - O(d^{-\frac{1}{10}})\right)^2 + \left(1 - t - O(d^{-\frac{1}{10}})\right)^2 + t^2 && (19)\\
&\geq 2 + 3t^2 - O(d^{-\frac{1}{10}}) && (20)\\
&\geq 2 - O(d^{-\frac{1}{10}}). && (21)
\end{aligned}$$

As a result,

$$\begin{aligned}
&\langle w_1, e_2\rangle^2 + \langle w_2, e_2\rangle^2 + \langle w_3, e_2\rangle^2 && (22)\\
&\leq \|w_1\|_2^2 + \|w_2\|_2^2 + \|w_3\|_2^2 - \langle w_1, e_1\rangle^2 - \langle w_2, e_1\rangle^2 - \langle w_3, e_1\rangle^2 && (23)\\
&\leq \left(2 + O(d^{-\frac{1}{10}})\right) - \left(2 - O(d^{-\frac{1}{10}})\right) && (24)\\
&\leq O(d^{-\frac{1}{10}}), && (25)
\end{aligned}$$

which finishes the proof. □

To prove the self-supervised learning part of Theorem 3.1, we first introduce the following lemma which gives some helpful properties of the empirical data matrix.

Lemma C.3. *In the setting of Theorem 3.1, let $M \triangleq \mathbb{E}_x[xx^\top]$ where the expectation is over empirical data. Then, when $n_1, n_2 \geq \text{poly}(d)$, with probability at least $1 - e^{-d^{\frac{1}{10}}}$, we have: (1) $e_2^\top M e_2 \geq \Omega(d^{\frac{2}{5}})$, and (2) $u^\top M u \leq O(1)$ for all $u \in \mathbb{R}^d$ such that $u^\top e_2 = 0$ and $\|u\|_2 = 1$.*

Proof of Lemma C.3. Let $n = n_1 + n_2 + n_3$. We abuse notation and let ξ_i ($i \in [n]$) be the set of all $\xi_i^{(k)}$ that appears in the empirical data. Let matrix $M' = \frac{1}{n}\sum_{i=1}^n \xi_i \xi_i^\top$. By standard concentration inequalities and union bound, for $n \geq \text{poly}(d)$, with probability at least $1 - \frac{1}{2}e^{-d^{\frac{1}{10}}}$, we have that $|M'_{i,j}| \leq \frac{1}{d}$ for all $i \neq j$ and $|M'_{i,i} - 1| \leq \frac{1}{d}$ for all $i \in [d]$. In this case, for any vector $u \in \mathbb{R}^d$ such that $\|u\|_2 = 1$ and $u^\top e_2 = 0$, we have

$$u^\top M u \leq 2\|u\|_2^2 + 2u^\top(\rho^2 M')u \leq 2 + 2\rho^2 + 2\rho^2\|M' - I\|_F \leq O(1). \tag{26}$$

On the other hand, by the definition of data distirbution and standard concentration inequalities, for $n \geq \text{poly}(d)$, with probability at least $1 - \frac{1}{2}e^{-d^{\frac{1}{10}}}$ we have that: at least $\frac{1}{3}$ of all data either is class 1 with $q_i^{(1)} = 1$ or class 2 with $q_i^{(2)} = 1$, and $\|\frac{1}{n}\sum_{i=1}^n \xi_i\|_2 \leq O(\frac{1}{d})$. In this case,

$$e_2^\top M e_2 = \mathbb{E}_x[(e_2^\top x)^2] \geq (\mathbb{E}_x[e_2^\top x])^2 \geq \left(\frac{1}{3}\tau - e_2^\top\left(\frac{1}{n}\sum_{i=1}^n \xi_i\right)\right)^2 \geq \Omega(\tau^2) = \Omega(d^{\frac{2}{5}}). \tag{27}$$

□

Using the above lemma, we can prove the self-supervised learning part of Theorem 3.1.

Proof of Theorem3.1(self-supervised learning part). Let $M = \mathbb{E}_x[xx^\top]$ be the empirical data matrix, where the expectation is over the dataset. Notice that self-supervised learning objective has the same minimizer as the matrix factorization objective $\|M - W^\top W\|_F^2$, by Eckart–Young–Mirsky theorem we know that the span of $\tilde{w}_1, \tilde{w}_2$ is exactly the span of the top 2 eigenvectors of matrix M. Let $M = \sum_{i=1}^d \lambda_i v_i v_i^\top$ where λ_i is the i-th largest eigenvalue of M with the corresponding eigenvector v_i. We decompose e_2 in the eigenvector basis as $e_2 = \sum_{i=1}^d \zeta_i v_i$.

By Lemma C.3, we know that with probability at least $1 - e^{-d^{\frac{1}{10}}}$, we have $e_2^\top M e_2 \geq \Omega(d^{\frac{2}{5}})$ and $\frac{u^\top M u}{\|u\|_2^2} \leq O(1)$ for all u orthogonal to e_2. To prove the result regarding self-supervised learning, we only need to prove that $\zeta_1^2 \geq 1 - O(d^{-\frac{1}{5}})$ in this case.

We first show that $\zeta_1^2 \geq \frac{1}{2}$. For contradiction, first assume $\zeta_1^2 \leq \frac{1}{2}$. Define vector

$$u = \zeta_1 v_1 - \frac{\sum_{i=2}^d \zeta_i v_i}{(1 - \zeta_1^2)}. \tag{28}$$

which satisfies $u^\top e_2 = 0$ and $\|u\|_2^2 \leq 4$. Notice that

$$\frac{u^\top M u}{\|u\|_2^2} \geq \frac{e_2^\top M e_2}{4} \geq \Omega(d^{\frac{2}{5}}), \tag{29}$$

which contradicts to $\frac{u^\top M u}{\|u\|_2^2} \leq O(1)$. Therefore, we have $\zeta_1^2 \geq \frac{1}{2}$.

To prove that ζ_1^2 is close to 1, we let scalar $t = \frac{1}{\zeta_1^2} - 1$ and define vector

$$u = -t\zeta_1 v_1 + \sum_{i=2}^d \zeta_i v_i, \tag{30}$$

which satisfies $u^\top e_2 = 0$. Since $\zeta_1^2 \geq \frac{1}{2}$, we have $t \leq 1$ and $\|u\|_2^2 \leq 1$. As a result, we have

$$\frac{u^\top M u}{\|u\|_2^2} \geq t^2 \lambda_1 \zeta_1^2 + \sum_{i=2}^{d} \lambda_i \zeta_i^2 \geq t^2 e_2^\top M e_2 \geq \Omega(t^2 d^{\frac{2}{5}}). \tag{31}$$

On the other hand, we know that $\frac{u^\top M u}{\|u\|_2^2} \leq O(1)$. Comparing these two bounds gives us $t^2 \leq O(d^{-\frac{1}{5}})$, which means $\zeta_1^2 \geq 1 - O(d^{-\frac{1}{5}})$. □

D Related Work

Supervised Learning with Dataset Imbalance. There exists a line of works studying supervised imbalanced classification. Ando & Huang [1], Buda et al. [3] proposed to re-sample the data to make the frequent and rare classes appear with equal frequency in training. Re-weighting assigns different weights for head and tail classes and eases the optimization difficulty under class imbalance [11, 29, 33]. Byrd & Lipton [4], Xu et al. [35] studied the effect of importance weighting and found out that importance weighting does not change the solution without regularization. Cao et al. [5] studied reweighted regularization based on classifier margin, but these techniques are limited to supervised imbalanced recognition. Cao et al. [6] proposed to regularize the local curvature of loss on imbalanced and noisy datasets. Recent works also designed specific losses or training pipelines for imbalanced recognition [18, 17, 32, 38]. Kang et al. [20] found out that the representations of supervised learning perform better than the classifier itself with class imbalance. Yang & Xu [36] studied the effect of self-training and self-supervised pre-training on supervised imbalanced recognition classifiers. In contrast, the focus of our paper is the effect of class imbalance on *self-supervised representations*, which is orthogonal to the self-supervised learning literature.

Self-supervised Learning. Recent works on self-supervised learning successfully learn representations that approach the supervised baseline on ImageNet and various downstream tasks. Contrastive learning methods attract positive pairs and drive apart negative pairs [16, 8]. Siamese networks predict the output of the other branch, and use stop-gradient to avoid collapsing [13, 9]. Clustering methods learn representations by performing clustering on the representations and improve the representations with cluster index [7]. Cole et al. [10] investigated the effect of data quantity and task granularity on self-supervised representations. Goyal et al. [12] studied self-supervised methods on large scale datasets in the wild, but they do not consider dataset imbalance explicitly. Several works have also theoretically studied the success of self-supervised learning [2, 23, 14, 34].

Ethics Statement

Our paper is the first to study the problem of robustness to imbalanced training of self-supervised representations. This setting is of critical importance to AI Ethics (especially fairness or bias concerns), as large real-world datasets tend to be imbalanced in practice, for instance including less examples from under-represented minorities. Furthermore, pre-training is a standard practice in deep learning, especially for quickly adapting models to new domains, which corresponds to our OOD evaluation scenario.

Our experiments and theoretical analysis show that SSL is more robust than supervised pre-training, especially in the OOD scenario. As supervised learning is still the de facto standard for pre-training, our work should have a wide impact, encouraging practitioners to use SSL for pre-training instead, or at least consider evaluating the impact of imbalanced pre-training on their downstream task.

We also remark that the paper does not imply at all that the algorithms proposed or studied can guarantee any form of fairness, and they in fact should still suffer from biases. The paper should be considered as a step towards studying the important technical issue of dataset imbalance, which is closely related to the fairness or biases questions.