SELF-SUPERVISED LEARNING IS MORE ROBUST TO DATASET IMBALANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

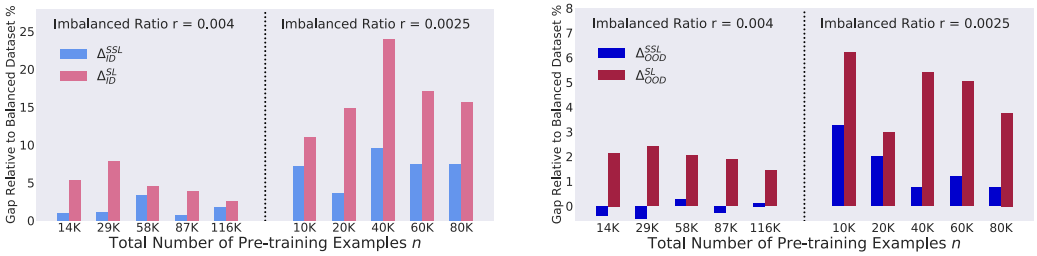
Self-supervised learning (SSL) is a scalable way to learn general visual representations since it learns without labels. However, large-scale unlabeled datasets in the wild often have long-tailed label distributions, where we know little about the behavior of SSL. In this work, we systematically investigate self-supervised learning under dataset imbalance. First, we find via extensive experiments that off-the-shelf self-supervised representations are already more robust to class imbalance than supervised representations. The performance gap between balanced and imbalanced pre-training with SSL is significantly smaller than the gap with supervised learning, across sample sizes, for both in-domain and, especially, out-of-domain evaluation. Second, towards understanding 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 and downstream tasks. In contrast, supervised learning has no incentive to learn features irrelevant to the labels from frequent examples. We validate this hypothesis with semi-synthetic experiments as well as rigorous mathematical analyses on a simplified setting. Third, inspired by the theoretical insights, we devise a re-weighted regularization technique that consistently improves the SSL representation quality on imbalanced datasets with several evaluation criteria, closing the small gap between balanced and imbalanced datasets with the same number of examples.

1 INTRODUCTION

Self-supervised learning (SSL) is an important paradigm for 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 (He et al., 2020; Chen et al., 2020; Grill et al., 2020; Caron et al., 2020; Chen & He, 2021). 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 (Reed, 2001). Curating a class-balanced unlabeled dataset requires the knowledge of labels, which defeats the purpose of leveraging unlabeled data by SSL.

The robustness of SSL algorithms to dataset imbalance remains largely underexplored in the literature, but extensive studies on supervised learning (SL) with imbalanced datasets do not bode well. The performance of vanilla supervised methods degrades significantly on imbalanced datasets (Cui et al., 2019; Cao et al., 2019; Buda et al., 2018), posing challenges to practical applications such as instance segmentation (Tang et al., 2020) and depth estimation (Yang et al., 2021). Many recent works address this issue with various regularization and re-weighting/re-sampling techniques (Ando & Huang, 2017; Wang et al., 2017; Jamal et al., 2020; Cui et al., 2019; Cao et al., 2019; 2021; Wang et al., 2020; Tian et al., 2020; Hong et al., 2021).

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 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



(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.

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 robust to dataset imbalance? We hypothesize 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 have so limited data that either SL or 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 3.2). In addition, we construct a toy example where we can rigorously prove the difference between self-supervised and supervised features in Section 3.1.

Finally, given our theoretical insights, we take a step towards further improving SSL algorithms, closing the small gap between balanced and imbalanced datasets. We identify the generalization gap between the empirical and population pre-training losses on rare data as the key to improvements.

To this end, we design a simple algorithm that first roughly estimates the density of examples with kernel density estimation and then applies a larger sharpness-based regularization (Foret et al., 2020) to the estimated rare examples. Our algorithm consistently improves the representation quality under several evaluation protocols.

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. (3) We propose a principled method to improve SSL under unknown dataset imbalance.

2 Exploring the Effect of Class Imbalance on SSL

Dataset class imbalance can pose challenge to self-supervised learning in the wild. Without access to labels, we cannot know in advance whether a large-scale unlabeled dataset is imbalanced. Hence, we need to study how SSL will behave under dataset imbalance to deploy SSL in the wild safely. In this section, we systematically investigate the effect of class imbalance on the self-supervised representations with experiments.

2.1 PROBLEM FORMULATION

Class-imbalanced pre-training datasets. We assume the datapoints / inputs are in $\mathbb{R}^d$ and come from C underlying classes. Let x denote the input and y denote the corresponding label. Supervised pre-training algorithms have access to the inputs and corresponding labels, whereas self-supervised pre-training only observes the inputs. Given a pre-training distribution $\mathcal{P}$ over over $\mathbb{R}^d \times [C]$, let r denote the ratio of class imbalance. That is, r is the ratio between the probability of the rarest class 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 will construct distributions with varying imbalance ratios and 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. Large scale data in the wild often follow heavily long-tailed label distributions, i.e., r is small. Throughout this paper we assume that for any class $j \in [C]$, the class-conditional distribution $\mathcal{P}^r(x|y = j)$ is the same across balanced and imbalanced datasets for all r. The pre-training dataset $\widehat{\mathcal{P}}_n^r$ consists of n i.i.d. samples from $\mathcal{P}^r$.

Pre-trained models. A feature extractor is a function $f_\phi : \mathbb{R}^d \to \mathbb{R}^m$ parameterized by neural network parameters ϕ, which maps inputs to representations. A linear head is a linear function $g_\theta : \mathbb{R}^m \to \mathbb{R}^C$, which can be composed with f_ϕ to produce the label. SSL algorithms learn ϕ from unlabeled data. Supervised 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 ϕ.[1]

Following the standard evaluation protocol in prior works (He et al., 2020; Chen et al., 2020), we measure the quality of learned 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 f_ϕ pre-trained on a pre-training dataset $\widehat{\mathcal{P}}_n^r$ with n data points and imbalance ratio r, we train a C-way linear classifier θ on a *balanced* dataset[2] sampled i.i.d. from $\mathcal{P}^{\text{bal}}$. We evaluate the representation quality with the top-1 accuracy of the learned linear head on $\mathcal{P}^{\text{bal}}$. We denote the ID accuracy of supervised pre-trained representations by $A_{\text{ID}}^{\text{SL}}(n, r)$. Note that $A_{\text{ID}}^{\text{SL}}(n, 1)$ stands for the result with balanced pre-training dataset. For SSL representations, we denote the accuracy by $A_{\text{ID}}^{\text{SSL}}(n, r)$.

Out-of-domain (OOD) evaluation tests the performance of representations by fine-tuning the feature extractor and the head on a (or multiple) *downstream* target distribution $\mathcal{P}_t$. Starting from a feature extractor f_ϕ (pre-trained on a dataset of size n and imbalance ratio r) and a randomly initialized classifier θ, we fine-tune ϕ and θ 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.

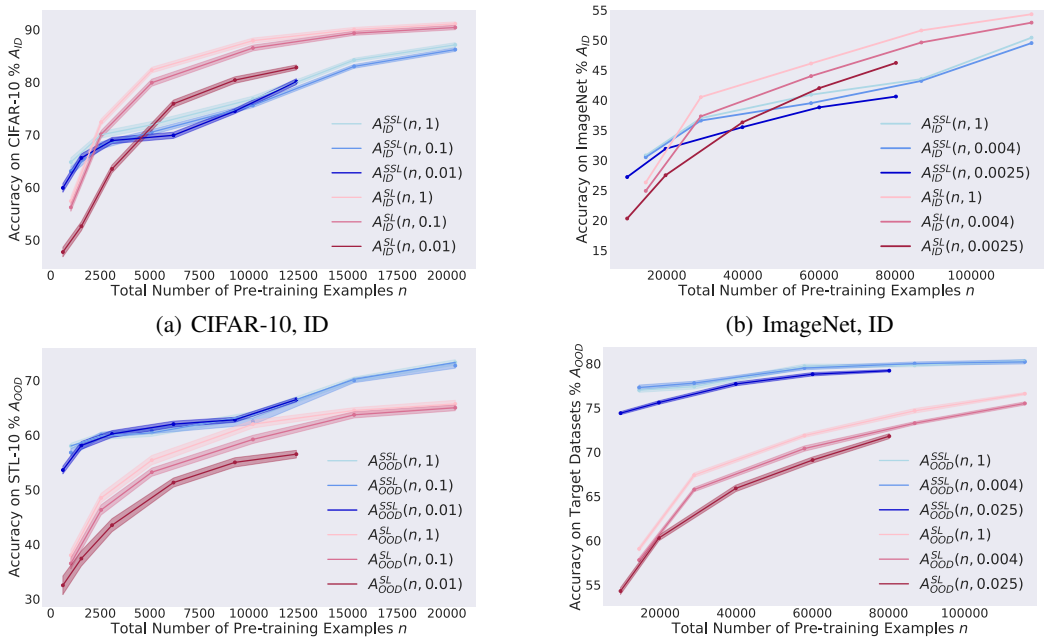
Summary of varying factors. We aim to study the effect of class imbalance to feature qualities on a diverse set of configurations with the following varying factors: (1) the number of examples in pre-training n, (2) the imbalance ratio of the pre-training dataset r, (3) ID or OOD evaluation, and (4) self-supervised learning algorithms: MoCo v2 (He et al., 2020), or SimSiam (Chen & He, 2021).

2.2 EXPERIMENTAL SETUP

Datasets. We pre-train the representations on variants of ImageNet (Russakovsky et al., 2015) or CIFAR-10 (Krizhevsky & Hinton, 2009) with a wide range of numbers of examples and ratios of imbalance. Following Liu et al. (2019), we consider exponential and Pareto distributions, which closely simulate the natural long-tailed distributions. We consider imbalance ratio in $\{1, 0.004, 0.0025\}$ for ImageNet and $\{1, 0.1, 0.01\}$ for CIFAR-10. 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. Note that we fix the variant of the dataset when comparing different algorithms. For ID evaluation, we use the original CIFAR-10 or ImageNet training set for the training phase of linear probe and use

[1] It is well-known that the composition of the head and features learned from supervised learning is more sensitive to imbalanced dataset than the quality of feature extractor ϕ (Cao et al., 2019; Kang et al., 2020). Please also see Table 4 in Appendix C for a comparison.

[2] We essentially use the largest balanced labeled ID dataset for this evaluation, which oftentimes means the entire curated training dataset, such as CIFAR-10 with 50,000 examples and ImageNet with 1,281,167 examples.



(a) CIFAR-10, ID (b) ImageNet, ID

(c) CIFAR-10, OOD (d) ImageNet, OOD

Figure 2: **Representation quality on balanced and imbalanced datasets.** Left: CIFAR-10, SL vs. SSL (SimSiam); Right: ImageNet, SL vs. SSL (MoCo v2). For both ID and OOD, the gap between balanced and imbalanced datasets with the same n is larger for supervised learning. The accuracy of supervised representations is better with reasonably large n in ID evaluation, while self-supervised representations perform better in OOD evaluation.[4]

the original validation set for the final evaluation. For OOD evaluation of representations learned on CIFAR-10, we use STL-10 (Coates et al., 2011) as the target /downstream dataset. For OOD evaluation of representations learned on ImageNet, we fine-tune the pre-trained feature extractors on CUB-200 (Wah et al., 2011), Stanford Cars (Krause et al., 2013), Oxford Pets (Parkhi et al., 2012), and Aircrafts (Maji et al., 2013), and measure the representation quality with average accuracy on the downstream tasks.

Models. We use ResNet-18 on CIFAR-10 and ResNet-50 on ImageNet as backbones. For supervised pre-training, we follow the standard protocol of He et al. (2016) and Kang et al. (2020). For self-supervised pre-training, we consider MoCo v2 (He et al., 2020) and SimSiam (Chen & He, 2021). We run each evaluation experiment with 3 seeds and report the average and standard deviation in the figures. Further implementation details are deferred to Section A.1.

2.3 Results: Self-supervised Learning is More Robust than Supervised Learning to Dataset Imbalance

In Figure 2, we plot the results of ID and OOD evaluations, respectively. 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 supervised representations, i.e., $A^{\text{SL}}(n,1) - A^{\text{SL}}(n,r)$, consistently in all configurations. 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. We observe that 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 across the board,

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

[4]The maximum n is smaller for extreme imbalance. The standard deviation comes only from the randomness of evaluation. We do not include the stddev for ImageNet ID due to limitation of computation resources.

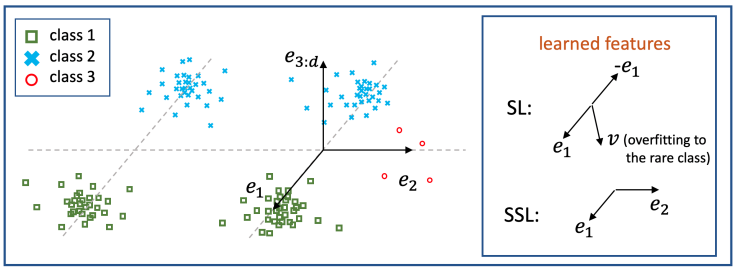


Figure 3: **Explaining SSL's robustness in a toy setting.** e_1 and e_2 are two orthogonal directions in the d-dimensional Euclidean space that decides the labels, and $e_{3:d}$ represents the other $d-2$ dimensions. Classes 1 and 2 are frequent classes and the third class is rare. To classify the three 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 interesting structure in 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 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, we observe that representations from supervised pre-training perform better than self-supervised pre-training in ID evaluation with reasonably large n, while self-supervised pre-training is better in OOD evaluation. This phenomenon is orthogonal to our observation that SSL is more robust to dataset imbalance, and is consistent with recent works (e.g., Chen et al. (2020); He et al. (2020)) which also observed that SSL performs slightly worse than supervised learning on balanced ID evaluation but better on OOD tasks.

3 Analysis

We have found out with extensive experiments that self-supervised representations are more robust to class imbalance than supervised representations. A natural and fundamental question arises: 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 for the rare classes. In this case, one may want to resort to the features learned from the frequent classes for help. However, due to the supervised nature of classification tasks, the supervised model mainly learns the features that help classify the frequent classes and may neglect other features which can transfer to the rare classes and potentially the downstream tasks. Partly because of this, Jamal et al. (2020) explicitly encourage the model to learn features transferable from the frequent to the rare classes with meta-learning. In contrast, in self-supervised learning, without the bias or incentive from the labels, the models can learn richer features that capture the intrinsic structures of the inputs—both features useful for classifying the frequent classes and features transferable to the rare classes simultaneously. In the following subsections, we study this reason in a setting where the two sets of features can be clearly separated. In Section 3.1, we proved SSL learns the with Gaussian input. In Section 3.2, we validate the hypothesis on a variant of CIFAR-10.

3.1 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 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 3 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 $f_{W_1,W_2}(x) \triangleq W_2W_1x$ 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_1x 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 $f_{W_1,W_2}(x)_y \geq f_{W_1,W_2}(x)_{y'} + 1$ for all data (x, y) in the dataset and $y' \neq y$.[5] Notice that when the norm is minimized, the row span of W_1 is exactly the same as the row span of matrix $W_{\text{SL}} \triangleq [w_1, w_2, w_3]^\top \in \mathbb{R}^{3\times d}$ which minimizes $\|w_1\|_2^2 + \|w_2\|_2^2 + \|w_3\|_2^2$ subject to the margin constraint $w_y^\top x \geq w_{y'}^\top x + 1$ for all empirical data (x, y) and $y' \neq y$. Therefore, feature W_1x is equivalent to $W_{\text{SL}}x$ for downstream linear probing, so we only need to analyze the feature quality of $W_{\text{SL}}x$ for supervised learning. For self-supervised learned, similar to SimSiam (Chen et al., 2020), 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 $W_{\text{SSL}} \in \mathbb{R}^{2\times d}$ 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 $W_{\text{SSL}}x$ 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 sampled from the data distribution $\mathcal{P}$ 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 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:*

- *Let $W_{\text{SL}} = [w_1, w_2, w_3]^\top$ be the feature learned by SL, then $|\langle e_2, w_i\rangle| \leq O(d^{-\frac{1}{20}})$ for $i \in [3]$.*
- *Let $W_{\text{SSL}} = [\tilde{w}_1, \tilde{w}_2]^\top$ 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 span of $\tilde{w}_1$ and $\tilde{w}_2$.*

Supervised learning results in features W_{SL} whose rows have small correlation with the transferable feature e_2, indicating that supervised learning only learns features for classifying the frequent classes and ignore the potential transferable features. In contrast, self-supervised learning recovers e_2 well, even though e_2 is not relevant to classifying the frequent classes.

3.2 Illustrative Semi-synthetic Experiments

In the previous subsection, we have shown that self-supervised learning provably learns label-irrelevant-but-transferable features from the frequent classes which can help classify the rare class in the toy case, while supervised learning mainly focuses on the label-relevant features. However, in real-world datasets, it is intractable to distinguish the two groups of features. To amplify this effect in a real-world dataset and highlight the insight of the theoretical analysis, we design a semi-synthetic experiment on SimCLR (Chen et al., 2020) to validate our conclusion.

[5] Previous work shows that deep linear networks trained with gradient descent using logistic loss converge to this min norm solution in direction (Ji & Telgarsky, 2018).

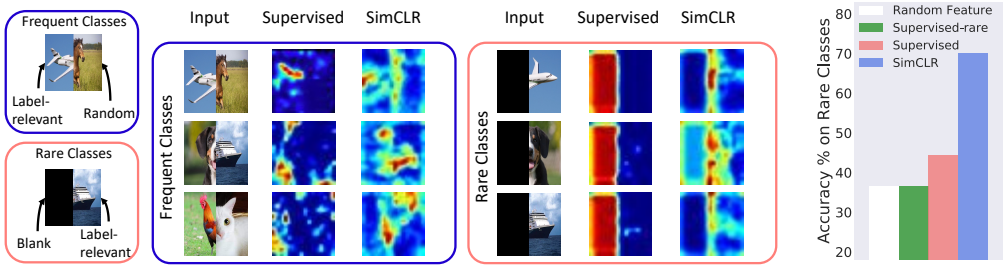


Figure 4: **Visualization of SSL's features in semi-synthetic settings. Left**: The right halves of the rare examples decide the labels, while the left are blank. The left halves of the frequent examples decide the labels, while the right halves are random half images, which contain label-irrelevant-but-transferable features. **Middle**: Visualization of feature activations with Grad-CAM (Selvaraju et al., 2017). SimCLR learns features from both left and right sides, whereas SL mainly learns label-relevant features from the left side of frequent data and ignore label-irrelevant features on the right side. **Right**: Accuracies evaluated on rare classes. The head linear classifiers are trained on 25000 examples from the 5 rare classes. Indeed, SimCLR learns much better features for rare classes than SL. Random Feature (feature extractor with randomly weights) and supervised-rare (features trained with only the rare examples) are included for references.

Dataset. In the theoretical analysis above, the frequent classes contain both features related to the classification of frequent classes and features transferable to the the rare classes. Similarly, we consider an imbalanced pre-training dataset with two groups of features modified from CIFAR-10 as shown in Figure 4 (Left). We construct classes 1-5 as the frequent classes, where each class contains 5000 examples. Classes 6-10 are the rare classes, where each class has 10 examples. In this case, the ratio of imbalance $r = 0.002$. Each image from classes 1-5 consists of a left half and a right half. The left half of an example 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 image of CIFAR-10, which is label-irrelevant. 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 of the images 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 classes since they have only 10 examples per class. This is consistent with the setting of Theorem 3.1.

Pre-training. We pre-train the representations on the semi-synthetic imbalanced dataset. For supervised learning, we use ResNet-50 on this 10-way classification task. For self-supervised learning, we use SimCLR with ResNet-50. To avoid confusing the left and right parts, we disable the random horizontal flip in the data augmentation. After pre-training, we fix the representations and train a linear classifier on top of the representations with balanced data from the 5 rare classes (25000 examples in total) to test if the model learns proper features for the rare classes during pre-training. In Figure 4 (Right), we test the classifier on the rare classes. In Figure 4 (Middle), we further visualize the Grad-CAM (Selvaraju et al., 2017) of the representations on the held-out set[6].

Results. As a sanity check, we first pre-train a supervised model with only the 50 rare examples and train the linear head classifier with 25000 examples from the rare classes (5-way classification) to see if the model can learn proper features for the rare classes with only rare examples (Supervised-rare in Figure 4 (Right)). As expected, the accuracy is 36.5%, which is almost the same as randomly initialized representations with trained head classifier, indicating that the model cannot learn the features for the rare classes with only rare examples due to the limited number of examples. We then compare supervised learning with self-supervised learning on the whole semi-synthetic dataset. In Figure 4 (Right), self-supervised representations performs much better than supervised representations on the rare classes (70.1% vs 44.3%). We further visualize the activation maps of representations with Grad-CAM. Supervised learning mostly activate the left halves of the examples for both frequent and rare classes, indicating that it mainly learn features on the left. In sharp contrast, self-supervised

[6]CIFAR images are of low resolution. For visualization, we use high resolution version of the CIFAR-10 images in Figure 4 (Middle). We also provide the visualization on original CIFAR-10 images in Figure 7.

learning activates the whole image on the frequent examples and the right part on the rare examples, indicating that it learns features from both parts.

4 IMPROVING SSL ON IMBALANCED DATASETS WITH REGULARIZATION

In this section, we aim to further improve the performance of SSL to close the gap between imbalanced and balanced datasets. Many prior works on imbalanced supervised learning regularize the rare classes more strongly, motivated by the observation that the rare classes suffer from more overfitting (Cao et al., 2019; 2021). Inspired by these works, we compute the generalization gaps (i.e., the differences between empirical and validation pre-training losses) on frequent and rare classes for the step-imbalance CIFAR-10 datasets (where 5 classes are frequent class with 5000 examples per class and the rest are rare with 50 examples per class). Indeed, as shown in Table 1, we still observe a similar phenomenon—the frequent classes have much smaller pre-training generalization gap than the rare classes (0.035 vs. 0.081), which indicates the necessity of more regularization on the rare classes.

We need a data-dependent regularizer that can have different effects on rare and frequent examples. Thus, weight decay or dropout (Srivastava et al., 2014) are not suitable. The prior work of Cao et al. (2019) regularizes the rare classes more strongly with larger margin, but it does not apply to SSL where no labels are available. Inspired by Cao et al. (2021), we adapt sharpness-aware minimization (SAM) (Foret et al., 2021) to imbalanced SSL.

Reweighted SAM (rwSAM). SAM improves model generalization by penalizing loss sharpness. Suppose the training loss of the representation f_ϕ is $\widehat{L}(\phi)$, i.e. $\widehat{L}(\phi) = \frac{1}{n}\sum_{j=1}^{n} \ell(x_j, \phi)$. SAM seeks parameters where the loss is uniformly low in the neighboring area,

$$\min_{\phi} \widehat{L}(\phi + \epsilon(\phi)), \quad \text{where} \quad \epsilon(\phi) = \operatorname*{arg\,max}_{\|\epsilon\|<\rho} \epsilon^\top \nabla_\phi \widehat{L}(\phi). \tag{2}$$

To take the weight of different examples into account, we add reweighting to the inner maximization step of SAM. Intuitively, we wish the optimization landscape to be flatter for rare examples, which is in effect regularizing the model more on rare examples. Concretely, consider the reweighted training loss associated with weight vector $w \in \mathbb{R}^n$, $\widehat{L}_w(\phi) = \frac{1}{n}\sum_{j=1}^{n} w_j \ell(x_j, \phi)$. The reweighted SAM objective re-weights the regularization-related terms (e.g., ϵ_w) but not the training loss $\widehat{L}$:

$$\min_{\phi} \widehat{L}(\phi + \epsilon_w(\phi)), \quad \text{where} \quad \epsilon_w(\phi) = \operatorname*{arg\,max}_{\|\epsilon\|<\rho} \epsilon^\top \nabla_\phi \widehat{L}_w(\phi). \tag{3}$$

Assigning Weight with Kernel Density Estimation. The weight w_j of an example x_j should be inversely correlated with the frequency of the corresponding class y_j. However, we have no access to the labels. In order to approximate the frequency of examples, we use kernel density estimation on top of the representations f_ϕ. Concretely, denote by $K(\cdot, h)$ the Gaussian density with bandwidth h. We assign w_i to be inversely correlated with the estimated density, i.e., $w_i = \left(\frac{1}{n}\sum_{j=1}^{n} K(f_\phi(x_i) - f_\phi(x_j), h)\right)^{-\alpha}$ where h and $\alpha > 0$ are hyperparameters selected by cross validation.

4.1 EXPERIMENTS

We test the proposed rwSAM on CIFAR-10 with step or exponential imbalance and ImageNet-LT (Liu et al., 2019). After self-supervised pre-training on the long-tailed dataset, we evaluate the representations by (1) linear probing on the balanced in-domain dataset and (2) fine-tuning on downstream target datasets. For (1) and (2), we compare with SSL, SSL+SAM (w/o reweighting), and SSL balanced, which learns the representations on the balanced dataset *with the same number of examples*. Implementation details and additional results are deferred to Section C.

Results. Table 1 summarizes results on long tailed CIFAR-10. With both step and exponential imbalance, rwSAM improves the performance of SimSiam over 1%, and even surpasses the performance of SimSiam on balanced CIFAR-10 with the same number of examples. Gap Freq. refers to the gap between training and test SimSiam loss on the 5 frequent classes, while Gap Rare refers to that on the 5 rare classes. Note that compared to SimSiam, rwSAM closes the generalization gap of pre-training loss on rare examples from 0.081 to 0.066, which verifies the effect of re-weighted regularization. In Table 2, we provide the result of fine-tuning on downstream tasks with representations pre-trained on ImageNet-LT. The proposed method improves the transferability of representations to downstream tasks consistently.

Table 1: Results on CIFAR-10-LT with linear probe.

	$r = 0.01$, step			$r = 0.01$, exp
Method	Acc. (%)	Gap Freq.	Gap Rare	Acc. (%)
SimSiam	84.3 ± 0.2	0.035	0.081	81.4 ± 0.3
SimSiam+SAM	84.7 ± 0.3	0.044	0.075	82.1 ± 0.4
SimSiam, balanced	85.8 ± 0.2	0.038	0.037	82.0 ± 0.4
SimSiam+rwSAM	85.6 ± 0.4	0.037	0.066	82.7 ± 0.5

Table 2: Results on ImageNet-LT with fine-tuning.

	ImageNet-LT				
Method	CUB	Cars	Aircrafts	Pets	Avg.
MoCo v2	69.9 ± 0.7	88.4 ± 0.4	82.9 ± 0.6	80.1 ± 0.6	80.3
MoCo v2+SAM	69.9 ± 0.5	88.8 ± 0.5	83.4 ± 0.4	81.5 ± 0.8	80.9
MoCo v2, balanced	69.8 ± 0.5	88.6 ± 0.4	82.7 ± 0.5	80.0 ± 0.4	80.2
MoCo v2+rwSAM	70.3 ± 0.7	88.7 ± 0.3	84.9 ± 0.6	81.7 ± 0.4	81.4
SimSiam	70.0 ± 0.3	87.0 ± 0.6	81.5 ± 0.7	83.8 ± 0.5	80.6
SimSiam, balanced	70.5 ± 0.8	87.9 ± 0.7	81.8 ± 0.7	82.7 ± 0.4	80.7
SimSiam+rwSAM	70.7 ± 0.8	88.4 ± 0.6	82.6 ± 0.6	84.0 ± 0.4	81.4

5 Related Work

Supervised Learning with Dataset Imbalance. There exists a line of works studying supervised imbalanced classification. Ando & Huang (2017); Buda et al. (2018) 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 (Cui et al., 2019; Tang et al., 2020; Wang et al., 2017). Byrd & Lipton (2019); Xu et al. (2021) studied the effect of importance weighting and found out that importance weighting does not change the solution without regularization. Cao et al. (2019) studied reweighted regularization based on classifier margin, but these techniques are limited to supervised imbalanced recognition. Cao et al. (2021) 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 (Jamal et al., 2020; Hong et al., 2021; Wang et al., 2021; Zhang et al., 2021). Kang et al. (2020) found out that the representations of supervised learning perform better than the classifier itself with class imbalance. Yang & Xu (2020) 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 (He et al., 2020; Chen et al., 2020). Siamese networks predict the output of the other branch, and use stop-gradient to avoid collapsing (Grill et al., 2020; Chen & He, 2021). Clustering methods learn representations by performing clustering on the representations and improve the representations with cluster index (Caron et al., 2020). Cole et al. (2021) investigated the effect of data quantity and task granularity on self-supervised representations. Goyal et al. (2021) 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 (Arora et al., 2019; Lee et al., 2020; HaoChen et al., 2021; Wei et al., 2021).

6 Conclusion

In this paper, we study self-supervised learning under class imbalance. 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, and provide insights for the design of future unsupervised learning methods.

REPRODUCIBILITY STATEMENT

To ensure reproducibility, we describe the implementation details of the algorithms and the construction of the datasets in Section A.1 and Section C. The code of the experiments in Section 2, Section 3.2 and Section 4 is provided in the supplementary material. We describe the setting and data assumptions of the toy case in Section 3.1 and provide the proof in Section D.

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.

REFERENCES

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.

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.

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.

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.

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.

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.

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.

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.

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.

Adam Coates, Andrew Y Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 215–223, 2011.

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.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 702–703, 2020.

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.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

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.

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.

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.

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.

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.

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.

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.

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

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.

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.

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

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.

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.

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

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.

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

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.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

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.

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.

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

Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X Yu. Long-tailed recognition by routing diverse distribution-aware experts. *arXiv preprint arXiv:2010.01809*, 2020.

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.

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.

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.

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

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.

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.

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.

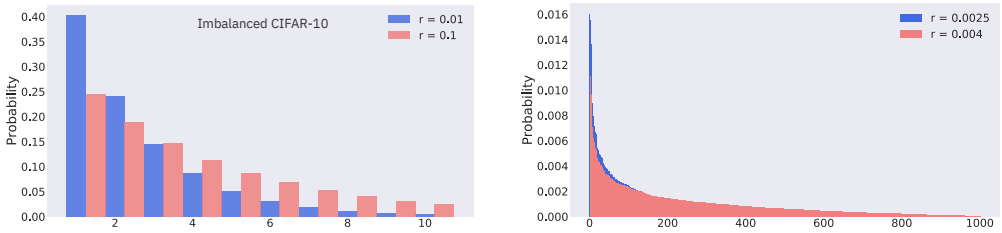


Figure 5: **Visualization of the label distributions.** We visualize the label distributions of the imbalanced CIFAR-10 and ImageNet. We consider two imbalance ratios r for each dataset.

A Details of Section 2

A.1 Implementation Details

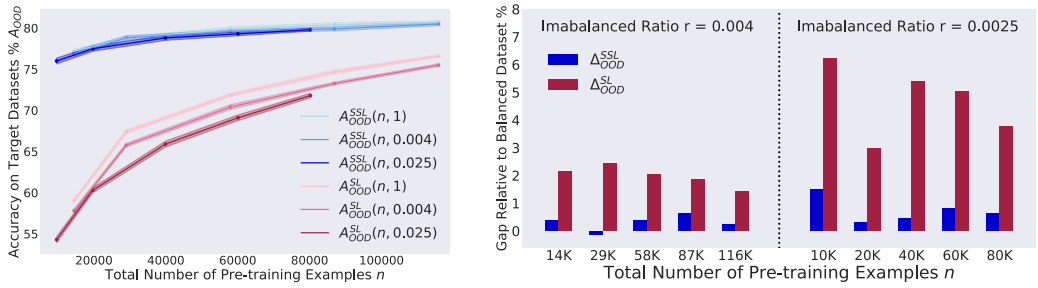
Generating Pre-training Datasets. CIFAR-10 (Krizhevsky & Hinton, 2009) contains 10 classes with 5000 examples per class. We use exponential imbalance, i.e. for class c, the number of examples is $5000 \times e^{\beta(c-1)}$. we consider imbalance ratio $r \in \{0.1, 0.01\}$, i.e. the number of examples 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. (2019), 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. For supervised pre-training, we follow the standard protocol of He et al. (2016) and Kang et al. (2020). 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 (He et al., 2020; Chen & He, 2021). 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`.

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 (Wah et al., 2011), Stanford Cars (Krause et al., 2013), Oxford Pets (Parkhi et al., 2012), and Aircrafts (Maji et al., 2013). 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



(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 3: 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

in Figure 6. SimSiam representations are also less sensitive to class imbalance than supervised representations.

We also provide the numbers of Figure 2 and Figure 6 in Table 3.

B Details of Section 3.2

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.

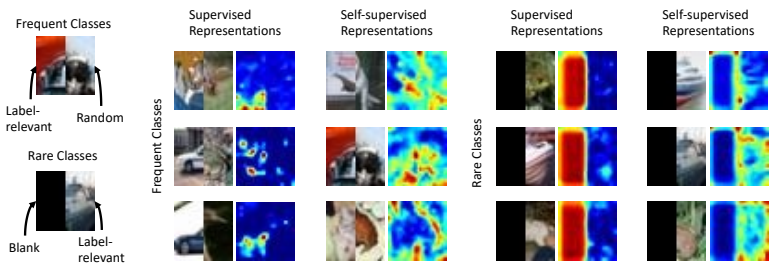


Figure 7: **Examples of the semi-synthetic Datasets and Grad-CAM visualizations.**

C Details of Section 4

C.1 Implementation Details

We use the same implementation as Section 2 for supervised and self-supervised learning baselines. We implement sharpness-aware minimization following (Foret et al., 2021). In each step of update, we first compute the reweighted loss $\widehat{L}_w(\phi) = \frac{1}{n_s}\sum_{j=1}^{n_s} w_j \ell(x_j^s, \phi)$ and compute its gradient w.r.t. ϕ, $\nabla_\phi \widehat{L}_w(\phi)$. Then we can compute ϵ_ϕ as $\epsilon_\phi = \rho \text{sgn}(\nabla_\phi \widehat{L}_w(\phi))|\nabla_\phi \widehat{L}_w(\phi)|^{q-1} / \left(\|\nabla_\phi \widehat{L}_w(\phi)\|_q^q\right)^{1/p}$, where $\frac{1}{p} + \frac{1}{q} = 1$. Finally, we update the model on the loss $\widehat{L}(\phi)$ by $\phi = \phi - \eta \nabla_\phi \widehat{L}(\phi + \epsilon(\phi))$. We select the hyperparameters ρ and α with cross validation. On ImageNet-LT and iNaturalist, $\rho = 2$ and $\alpha = 0.5$. On CIFAR-10-LT, $\rho = 5$ and $\alpha = 1.2$.

C.2 Additional Results

We further introduce another evaluation protocol of the representations: following the protocol of Kang et al. (2020); Yang & Xu (2020), fine-tuning on imbalanced dataset with supervision, and then re-training the linear classifier with class-aware resampling, to compare with supervised imbalanced recognition methods. In MoCo V2 pre-training, we use the standard data augmentation following He et al. (2020). In fine-tuning, we use RandAugment (Cubuk et al., 2020). For (3), we further compare with CRT (Kang et al., 2020), LADE (Hong et al., 2021), and RIDE (Wang et al., 2021), which are strong methods tailored to supervised imbalanced recognition. Results are provided in Table 4. Supervised here refers to training the feature extractor and linear classifier with supervision on the imbalanced dataset directly. CRT first trains the feature extractor with supervision, and then *re-trains the classifier with resampled loss*. Note that CRT is performing better than Supervised, indicating that the composition of the head and features learned from supervised learning is more sensitive to imbalanced dataset than the quality of feature extractor itself.

Table 4: ImageNet-LT with Supervision.

Method	Backbone	Acc.
Supervised	ResNet-50	49.3
CRT (Kang et al., 2020)	ResNet-50	52.0
LADE (Hong et al., 2021)	ResNeXt-50	53.0
RIDE (Wang et al., 2021)	ResNet-50	54.9
RIDE (Wang et al., 2021)	ResNeXt-50	56.4
MoCo V2	ResNet-50	55.0
MoCo V2+rwSAM	ResNet-50	55.5

Even with a simple pre-training and fine-tuning pipeline, MoCo V2 representations can be comparable with much more complicated state-of-the-arts tailored to supervised imbalanced recognition, further corroborating the power of SSL under class imbalance. With rwSAM, we can further improve the result of MoCo V2.

D PROOF OF THEOREM 3.1

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 D.1. *Let $\xi_i \sim \mathcal{N}(0, I)$ for $i \in [n]$. Then, for any $n \leq \text{poly}(d)$, with probability at least $1 - e^{-d^{\frac{1}{10}}}$ and large enough 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* $i \in [n]$.
- $|\langle \xi_i, \xi_j\rangle| \leq 3d^{\frac{3}{5}}$ *for all* $i \neq j$.

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

$$\Pr\left(|\langle \xi, e_1\rangle| \geq d^{\frac{1}{10}}\right) \leq d^{-\frac{1}{10}} \cdot e^{-\frac{d^{\frac{1}{5}}}{2}}. \tag{4}$$

By the tail bound of χ_d^2 distribution, we have

$$\Pr\left(|\|\xi\|_2^2 - d| \geq 4d^{\frac{3}{4}}\right) \leq 2e^{-\sqrt{d}}. \tag{5}$$

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:

$$\Pr\left(|\langle \xi, \xi'\rangle| \geq 3d^{\frac{3}{5}}\right) \leq \Pr\left(\|\xi\|_2 \geq \sqrt{d} + 2d^{\frac{3}{8}}\right) + \Pr\left(|\langle \xi', \frac{\xi}{\|\xi\|}\rangle| \geq d^{\frac{1}{10}}\right) \leq \frac{e^{-\frac{d^{\frac{1}{5}}}{2}}}{d^{\frac{1}{10}}} + 2e^{-\sqrt{d}}. \tag{6}$$

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 D.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 D.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 D.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 D.2. When the high probability outcome of Lemma D.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

$$w_1^{*\top} x = 1 + \langle \xi_i^{(1)}, e_1\rangle\rho \geq 1 - \rho d^{\frac{1}{10}}, \tag{7}$$

$$w_2^{*\top} x = -1 + \langle \xi_i^{(1)}, e_1\rangle\rho \leq -1 + \rho d^{\frac{1}{10}}, \tag{8}$$

$$w_3^{*\top} x = \frac{1}{\rho d}\left(e_1 - q_i^{(1)}\tau e_2 + \rho\xi_i^{(1)}\right)^\top \left(\sum_{j=1}^{n_3} \xi_j^{(3)}\right) \leq \frac{n_3(\tau + 1)}{\rho d} d^{\frac{1}{10}} + \frac{3n_3}{d} d^{\frac{3}{5}}. \tag{9}$$

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

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

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

$$w_3^{*\top} x = \frac{1}{\rho d} \left(\sum_{j=1}^{n_3} \xi_j^{(3)} \right)^\top \left(e_2 + \rho \xi_i^{(3)} \right) \geq \frac{1}{d} \|\xi_i^{(3)}\|_2^2 - \frac{3n_3}{d} d^{\frac{3}{5}} - \frac{n_3 d^{\frac{1}{10}}}{\rho d} \geq 1 - O(d^{-\frac{1}{5}}). \tag{11}$$

On the other hand,

$$w_1^{*\top} x = \langle \rho \xi_i^{(3)}, e_1 \rangle \leq \rho d^{\frac{1}{10}}, \tag{12}$$

$$w_2^{*\top} x = \langle \rho \xi_i^{(3)}, -e_1 \rangle \leq \rho d^{\frac{1}{10}}. \tag{13}$$

So the margin is

$$w_3^{*\top} x - \max\{w_1^{*\top} x, w_2^{*\top} x\} \geq w_3^{*\top} x - \rho d^{\frac{1}{10}} \geq 1 - O(d^{-\frac{1}{10}}). \tag{14}$$

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 D.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 D.1 holds (which happens with probability at least $1 - e^{-d^{\frac{1}{10}}}$). By Lemma D.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

$$\|\frac{1}{\alpha} w_1^*\|_2^2 + \|\frac{1}{\alpha} w_2^*\|_2^2 + \|\frac{1}{\alpha} w_3^*\|_2^2 = \frac{2 + \|w_3^*\|_2^2}{\alpha^2} \leq 2 + O(d^{-\frac{1}{10}}). \tag{15}$$

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

$$\left| \mathbb{E}_{i \in [n_1], q_i^{(1)} = 0} [x_i^{(1)}] - e_1 \right| \leq d^{-\frac{1}{10}}, \tag{16}$$

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

$$(w_1 - w_3)^\top e_1 \geq 1 - \|w_1 - w_3\|_2 \cdot d^{-\frac{1}{10}} \geq 1 - O(d^{-\frac{1}{10}}). \tag{17}$$

A similar analysis for class 2 data gives us

$$(w_2 - w_3)^\top (-e_1) \geq 1 - O(d^{-\frac{1}{10}}). \tag{18}$$

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

$$\langle w_1, e_1 \rangle^2 + \langle w_2, e_1 \rangle^2 + \langle w_3, e_1 \rangle^2 \geq \left(t + 1 - O(d^{-\frac{1}{10}}) \right)^2 > 2.25 - O(d^{-\frac{1}{10}}), \tag{19}$$

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

$$\langle w_1, e_1\rangle^2 + \langle w_2, e_1\rangle^2 + \langle w_3, e_1\rangle^2 \tag{20}$$

$$\ge \left(1 + t - O(d^{-\frac{1}{10}})\right)^2 + \left(1 - t - O(d^{-\frac{1}{10}})\right)^2 + t^2 \tag{21}$$

$$\ge 2 + 3t^2 - O(d^{-\frac{1}{10}}) \tag{22}$$

$$\ge 2 - O(d^{-\frac{1}{10}}). \tag{23}$$

As a result,

$$\langle w_1, e_2\rangle^2 + \langle w_2, e_2\rangle^2 + \langle w_3, e_2\rangle^2 \tag{24}$$

$$\le \|w_1\|_2^2 + \|w_2\|_2^2 + \|w_3\|_2^2 - \langle w_1, e_1\rangle^2 - \langle w_2, e_1\rangle^2 - \langle w_3, e_1\rangle^2 \tag{25}$$

$$\le \left(2 + O(d^{-\frac{1}{10}})\right) - \left(2 - O(d^{-\frac{1}{10}})\right) \tag{26}$$

$$\le O(d^{-\frac{1}{10}}), \tag{27}$$

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 D.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 $n_1, n_2 \ge \mathrm{poly}(d)$, with probability at least $1 - e^{-d^{\frac{1}{10}}}$, we have: (1) $e_2^\top M e_2 \ge \Omega(d^{\frac{2}{5}})$, and (2) $u^\top M u \le O(1)$ for all $u \in \mathbb{R}^d$ such that $u^\top e_2 = 0$ and $\|u\|_2 = 1$.*

Proof of Lemma D.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 $n \ge \mathrm{poly}(d)$, with probability at least $1 - \frac{1}{2}e^{-d^{\frac{1}{10}}}$, we have that $|M'_{i,j}| \le \frac{1}{d}$ for all $i \ne j$ and $|M'_{i,i} - 1| \le \frac{1}{d}$ 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

$$u^\top M u \le 2\|u\|_2^2 + 2u^\top(\rho^2 M')u \le 2 + 2\rho^2 + 2\rho^2\|M' - I\|_F \le O(1). \tag{28}$$

On the other hand, by the definition of data distirbution and standard concentration inequalities, for $n \ge \mathrm{poly}(d)$, 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 $\|\frac{1}{n}\sum_{i=1}^n \xi_i\|_2 \le O(\frac{1}{d})$. In this case,

$$e_2^\top M e_2 = \mathbb{E}_x[(e_2^\top x)^2] \ge (\mathbb{E}_x[e_2^\top x])^2 \ge \left(\frac{1}{3}\tau - e_2^\top\left(\frac{1}{n}\sum_{i=1}^n \xi_i\right)\right)^2 \ge \Omega(\tau^2) = \Omega(d^{\frac{2}{5}}). \tag{29}$$

□

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

$$u = \zeta_1 v_1 - \frac{\sum_{i=2}^d \zeta_i v_i}{(1-\zeta_1^2)}. \tag{30}$$

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

$$\frac{u^\top M u}{\|u\|_2^2} \geq \frac{e_2^\top M e_2}{4} \geq \Omega(d^{\frac{2}{5}}), \tag{31}$$

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

$$u = -t\zeta_1 v_1 + \sum_{i=2}^d \zeta_i v_i, \tag{32}$$

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

$$\frac{u^\top M u}{\|u\|_2^2} \geq t^2 \lambda_1 \zeta_1^2 + \sum_{i=2}^d \lambda_i \zeta_i^2 \geq t^2 e_2^\top M e_2 \geq \Omega(t^2 d^{\frac{2}{5}}). \tag{33}$$

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 $\zeta_1^2 \geq 1 - O(d^{-\frac{1}{5}})$. $\square$