A Tale of Two Classes: Adapting Supervised Contrastive Learning to Binary Imbalanced Datasets

David Mildenberger[1,2,*], Paul Hager[1,*], Daniel Rueckert[1,2,3], Martin J. Menten[1,2,3]

[1]Technical University of Munich, [2]Munich Center for Machine Learning, [3]Imperial College London

{david.mildenberger, paul.hager, daniel.rueckert, martin.menten}@tum.de

[*]These authors contributed equally.

Abstract

Supervised contrastive learning (SupCon) has proven to be a powerful alternative to the standard cross-entropy loss for classification of multi-class balanced datasets. However, it struggles to learn well-conditioned representations of datasets with long-tailed class distributions. This problem is potentially exacerbated for binary imbalanced distributions, which are commonly encountered during many real-world problems such as medical diagnosis. In experiments on seven binary datasets of natural and medical images, we show that the performance of SupCon decreases with increasing class imbalance. To substantiate these findings, we introduce two novel metrics that evaluate the quality of the learned representation space. By measuring the class distribution in local neighborhoods, we are able to uncover structural deficiencies of the representation space that classical metrics cannot detect. Informed by these insights, we propose two new supervised contrastive learning strategies tailored to binary imbalanced datasets that improve the structure of the representation space and increase downstream classification accuracy over standard SupCon by up to 35%. We make our code available.[1]

1. Introduction

Supervised contrastive learning (SupCon) has emerged as a powerful alternative to the cross-entropy loss for supervised deep learning [8, 17, 20, 21]. SupCon combines full label information with a contrastive loss to cluster samples of the same class in similar regions of the representation space. Conversely, embeddings of different classes are pushed apart. This results in a well-conditioned representation space that preserves discriminative features of each sample while separating semantic classes [2]. SupCon has been used to achieve state-of-the-art results across diverse fields, including genetics [3], out-of-distribution detection [31], object detection [30], video action recognition [10], and neuroscience [28].

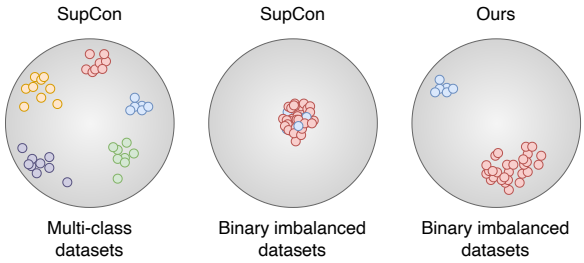

Figure 1. Supervised contrastive learning (SupCon) on multi-class balanced datasets returns a well-conditioned representation space, in which semantic classes are clearly separated. We show that for binary imbalanced datasets the prevalence of a dominant majority class causes the embeddings to collapse to a single point. Our proposed fixes restore the clear separation of semantic classes.

SupCon has been predominantly developed on and applied to multi-class balanced benchmark datasets like ImageNet, which consist of numerous equally prevalent classes. In contrast, real-world datasets often significantly deviate from these idealized conditions. There has been a growing focus on adapting supervised contrastive learning to long-tailed datasets, which are characterized by a few common "head" classes and many rare "tail" classes [7, 14, 21, 36, 42]. These works show that SupCon often yields representation spaces with dominating head classes when applied to long-tailed datasets, leading to reduced downstream utility.

SupCon's shortcomings on long-tailed datasets are potentially exacerbated on distributions with only two underlying classes: a common majority class and a rare minority class. Such binary imbalanced distributions are com-

[1]https://github.com/aiforvision/TTC

mon in real-world tasks, such as anomaly detection, fraud detection, and disease classification. For example, during medical screening, subjects are classified as either healthy or diseased, with the healthy cases typically outnumbering the pathological ones [13, 19]. A survey of representation learning for medical image classification found that 78 out of 114 studies focused on binary classification problems [13].

This work is the first to investigate the effectiveness of SupCon on binary imbalanced datasets, identifying limitations of existing supervised contrastive learning strategies and proposing algorithmic solutions to these issues. Our main contributions are:

- We empirically demonstrate that SupCon is ineffective on binary imbalanced datasets. In controlled experiments on seven natural and medical imaging datasets, we observe that the performance of SupCon degrades with increasing class imbalance, falling behind the standard cross-entropy loss even at moderate levels of class imbalance.
- To investigate these findings, we introduce two novel metrics for diagnosing the structural deficiencies of the representation space. We show that at high class imbalance all embeddings are closely clustered in a small region of the representation space, preventing separation of semantic classes (see Fig. 1). While canonical metrics fail to capture this problem, our proposed metrics detect this representation space collapse and diminished downstream utility. Furthermore, we theoretically substantiate our empirical observations in a proof.
- Informed by the insights gained through our metrics, we propose two new supervised contrastive learning strategies tailored to binary imbalanced datasets. These adjustments are easy to implement and incur minimal additional computational cost compared to the standard SupCon loss. We demonstrate that our fixes boost downstream classification performance by up to 35% over SupCon and outperform leading strategies for long-tailed data by up to 5%.

2. Related works

2.1. Analyzing representation spaces

To evaluate the quality of representation spaces, Wang and Isola have introduced the notions of representation space alignment and uniformity [37]. Alignment measures how closely semantically similar samples are located in the representation space. Uniformity quantifies the utilization of the representation space's capacity. Both metrics have been empirically validated as strong indicators of the representation space's quality and downstream utility. However, high uniformity and alignment alone do not guarantee separability of classes, as shown by Wang *et al*. [38]. In a subsequent study, Li *et al*. proposed analyzing alignment and uniformity at the level of semantic classes instead of individual samples [21]. While this improves upon sample-wise analysis, it still fails to properly compare representations *between* classes, which is crucial for downstream classification performance. We address this limitation by introducing two novel metrics, enabling the evaluation of sample and class consistency within representation neighborhoods.

2.2. Supervised contrastive learning for long-tailed datasets

When applying SupCon to long-tailed datasets, samples from the majority class often occupy a disproportionate amount of the representation space [7, 14, 15, 21, 42]. In the most severe cases the representation space collapses completely, losing all utility [4, 8, 9, 39]. Many works thus aim to enhance latent space uniformity by spreading features evenly, irrespective of data imbalance. Zhu *et al*. balance gradient contributions of classes to achieve a regular simplex latent structure [42]. Hou *et al*. split the majority classes into smaller sub-classes according to their latent features [14]. Cui *et al*. leverage parametric class prototypes to adaptively balance the learning signal [7]. Kang *et al*. and Li *et al*. limit the number of positives that contribute to the loss with Li *et al*. also using fixed class prototypes [15, 21]. Although these methods outperform SupCon on long-tailed distributions, they remain untested on binary imbalanced distributions whose unique characteristics are not explicitly addressed.

3. Metrics to diagnose representation spaces of binary data distributions

Alignment and uniformity are established metrics for evaluating the quality and structure of representation spaces obtained via unsupervised contrastive learning [37]. Alignment measures the average distance between positive pairs in the feature space. Low average distance, or high alignment, indicates consistent embeddings that are robust to noise. Uniformity measures the average Gaussian potential between embeddings on the unit hypersphere. Low average potential, or high uniformity, corresponds to expressive embeddings that fully utilize the entire representation space. Effective representation spaces exhibit both high alignment and high uniformity and should in theory yield good linear separability of the semantic classes.

However, because these metrics were originally developed for unsupervised contrastive learning, they operate on a per-sample level and ignore latent class information. To address this limitation, Li *et al*. have extended alignment and uniformity to SupCon, by analyzing the alignment of classes instead of samples [21]. Although this class-level alignment better reflects semantic separability, it still does not capture the relationships *between* classes.

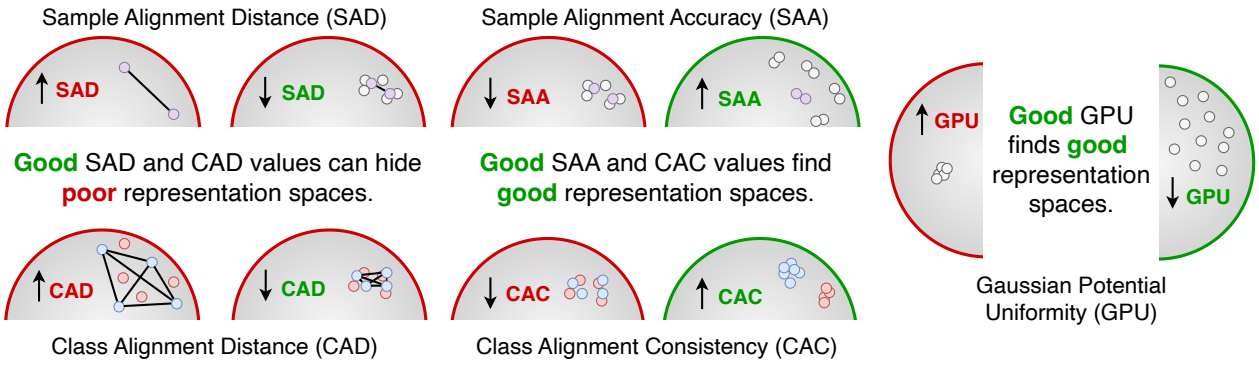

Figure 2. Our novel sample alignment accuracy (SAA) and class alignment consistency (CAC) metrics capture the relationships *between* embeddings of different classes instead of just within one class. By more directly measuring the separability of latent classes, it is a stronger indicator of a representation space's downstream utility.

We therefore propose two new alignment metrics: sample alignment accuracy (SAA) and class alignment consistency (CAC). Our metrics compare the alignment both within and between samples and classes. This is in contrast to the canonical sample alignment distance (SAD) [37] and class alignment distance (CAD) [21] which only measure the alignment within one sample or class, as shown in Fig. 2.

3.1. Definitions

We define a dataset $\mathcal{X}$ containing N images, $\mathcal{X} = \{x_k\}_{k=1}^{N}$. Given that the images in $\mathcal{X}$ are labeled with a binary class distribution, we denote the samples of class $i \in \{0, 1\}$ as $x \in \mathcal{X}_i$ with the function $g : \mathcal{X} \to \{0, 1\}$ mapping each image to its label. Let $\tilde{x}_k$ and $\tilde{x}_k^+$ denote two randomly augmented instances of x_k that form a positive pair. Conversely, any view that is not generated from x_k, denoted by $\tilde{x}_k^-$ denotes, forms a negative pair with $\tilde{x}_k$. The set of all views is called $\mathcal{W}$ with $|\mathcal{W}| = 2N$. The set of all views of class i is $\mathcal{W}_i$. The function $f : \mathcal{W} \to \mathcal{S}^{d-1}$ maps a view $\tilde{x}$ onto a d-dimensional representation on the unit sphere $\mathcal{S}^{d-1}$.

3.1.1. Sample alignment distance (SAD)

Sample alignment distance (SAD), as defined by Wang *et al*. [37], measures the average distance between representations of two augmented views of the same sample. Formally, SAD is computed as the average pairwise ℓ_2 distance between sample-wise positive pairs:

$$\text{SAD} = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \|(f(\tilde{x}) - f(\tilde{x}^+))\|_2 \tag{1}$$

A low SAD, or high alignment, implies that two different views of the same image produce similar embeddings. Obtaining consistent embeddings despite perturbations from augmentations typically indicates that generalized class-level semantic features are being captured, which is ultimately beneficial for downstream applications.

3.1.2. Sample alignment accuracy (SAA)

We introduce the concept of sample alignment accuracy (SAA) to determine if the embeddings of positive pairs are more closely aligned with each other compared to other samples. SAA is the proportion of all sample-wise positive pairs for which the ℓ_2 distance between their embeddings is smaller than that to all negative pairs:

$$\text{SAA} = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \mathbb{1}\Big(\|(f(\tilde{x}) - f(\tilde{x}^+))\|_2 < \min_{\tilde{x}^- \in \mathcal{W} \setminus \{\tilde{x}, \tilde{x}^+\}} \|(f(\tilde{x}) - f(\tilde{x}^-))\|_2 \Big) \tag{2}$$

Here, $\mathbb{1}(\cdot)$ is the indicator function, which outputs 1 if the condition $\cdot$ holds, and 0 otherwise.

Compared to SAD, SAA is more insightful in cases in which many samples, both positives and negatives, are placed in close proximity to each other. While SAD would indicate high alignment despite low separability of semantic classes, SAA would correctly diagnose a partially degenerate representation space.

3.1.3. Class alignment distance (CAD)

Class alignment distance (CAD), introduced by Li *et al*. [21], calculates the average distance between all representations within a class to evaluate how well the learned representation space clusters samples according to their semantic labels [21]. Let C be the number of classes and $\tilde{x}, \tilde{x}' \in \mathcal{W}_i$ all unique pairs of samples in $\mathcal{W}_i$:

$$\text{CAD} = \frac{1}{C}\sum_{i=1}^{C}\frac{1}{\binom{|\mathcal{X}_i|+1}{2}}\sum_{\tilde{x},\tilde{x}'\in\mathcal{W}_i}\|(f(\tilde{x})-f(\tilde{x}'))\|_2 \quad (3)$$

Compared to SAD, CAD captures alignment across an entire class, indicating how well a class clusters on the representation space's hypersphere.

3.1.4. Class alignment consistency (CAC)

To measure how pure embedding neighborhoods are with respect to the latent class, we introduce class alignment consistency (CAC). We define class alignment within a local neighborhood of the closest r views to $\tilde{x}$, which we call $\mathcal{W}_{\tilde{x}}$. For our analysis we set r to 5% of all views. Let $\mathcal{D}$ denote the set of sets containing each $\tilde{x}$ and its local neighborhood $\mathcal{W}_{\tilde{x}}$, with $(\tilde{x}, \mathcal{W}_{\tilde{x}}) \in \mathcal{D}$.

$$\text{CAC} = \frac{1}{|\mathcal{D}|}\sum_{(\tilde{x},\mathcal{W}_{\tilde{x}})\in\mathcal{D}}\frac{1}{|\mathcal{W}_{\tilde{x}}|}\sum_{\tilde{x}'\in\mathcal{W}_{\tilde{x}}}\mathbb{1}(g(\tilde{x})=g(\tilde{x}')) \quad (4)$$

Unlike CAD, CAC also measures the distance of embeddings to those of the opposite class. This provides a more direct signal of the separability of classes that better correlates with downstream classification performance.

3.1.5. Gaussian potential uniformity (GPU)

Uniformity measures how evenly representations are distributed across the unit hypersphere. Wang *et al*. [37] define uniformity through the logarithm of the average pairwise Gaussian potential:

$$\text{GPU} = \log\left(\frac{1}{\binom{|N|+1}{2}}\sum_{k=1}^{N}\sum_{j=1}^{N}e^{-\|f(\tilde{x}_k)-f(\tilde{x}_j)\|_2^2}\right) \quad (5)$$

A lower GPU indicates that the embeddings are more evenly spread across the hypersphere. Utilizing large portions of the hypersphere for embeddings suggests a broader range of features learned and thus increased generalizability to unseen data.

4. Methods

4.1. SupCon

SupCon is best understood as an extension of the original unsupervised contrastive NT-Xent loss. In essence, the unsupervised loss maximizes the cosine similarity between embeddings of positive pairs while minimizing the similarity between embeddings of negative pairs. Using the notation introduced in Sec. 3.1, the loss is:

$$\mathcal{L}_{\text{NT-Xent}} = -\sum_{x\in\mathcal{X}}\log\frac{e^{f(\tilde{x})\cdot f(\tilde{x}^+)/\tau}}{\sum_{\tilde{x}^-\in\mathcal{W}\setminus\{\tilde{x}\}}e^{f(\tilde{x})\cdot f(\tilde{x}^-)/\tau}} \quad (6)$$

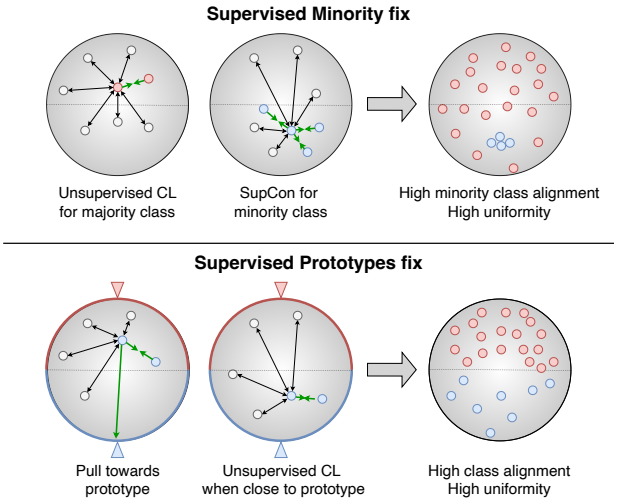

Figure 3. We introduce two fixes for supervised contrastive learning. Supervised Minority applies supervision exclusively to the minority class, preventing class collapse and enhancing alignment of the minority class. Supervised Prototypes attracts samples to fixed class prototypes, improving both class alignment and uniformity.

Here, $\cdot$ denotes the dot product and $\tau \in \mathbb{R}^+$ is the temperature parameter. In practice, $\mathcal{X}$ and $\mathcal{W}$ are restricted to the elements and views of a single batch.

SupCon extends the NT-Xent loss to include full label information. By considering all samples of the same class in the numerator, it aims to maximize the similarity between all projections of a class. For each class i, the SupCon loss is defined as:

$$\mathcal{L}^i_{\text{SupCon}} \triangleq -\sum_{x\in\mathcal{X}_i}\frac{1}{|\mathcal{W}_i\setminus\{\tilde{x}\}|}\sum_{p\in\mathcal{W}_i\setminus\{\tilde{x}\}}\log\frac{e^{f(\tilde{x})\cdot f(p)/\tau}}{\sum_{a\in\mathcal{X}\setminus\{\tilde{x}\}}e^{f(\tilde{x})\cdot f(a)/\tau}} \quad (7)$$

The total loss in the binary case is then:

$$\mathcal{L}_{\text{SupCon}} = \mathcal{L}^0_{\text{SupCon}} + \mathcal{L}^1_{\text{SupCon}} \quad (8)$$

4.2. Supervised Minority

We introduce Supervised Minority, a novel supervised contrastive learning strategy specifically for binary imbalanced datasets. Supervised Minority applies supervision exclusively to the minority class (see Fig. 3). Formally, we combine SupCon in the minority (min) class with the NT-Xent [5] loss in the majority (maj) class:

$$\mathcal{L}_{\text{SupMin}} = \mathcal{L}^{\text{min}}_{\text{SupCon}} + \mathcal{L}^{\text{maj}}_{\text{NT-Xent}} \quad (9)$$

By using the NT-Xent loss for the majority class, we aim to guard against class collapse and increase uniformity. Additionally, by using SupCon in the minority class we enhance its alignment.

4.3. Supervised Prototypes

Our second approach, Supervised Prototypes, builds upon the concept of fixed prototypes [16, 21, 26, 41]. We initialize two fixed class prototypes at opposite ends of the representation space's hypersphere. Each prototype attracts samples of its respective class (see Fig. 3). While prototypes improve class alignment, they can reduce latent space uniformity [21]. To mitigate this, we attract samples towards the prototype only if their cosine similarity with that prototype is less than 0.5. When a sample's representation already has high similarity to its prototype, it is influenced only by the NT-Xent loss.

We place the majority class prototype p_{maj} on the $\mathcal{S}^{d-1}$ unit sphere so that it minimizes the average distance to all encodings of unaugmented training samples. This position is determined through gradient descent. Let $p_{min} = -p_{maj}$ represent the minority class prototype, ensuring maximal separation on the hypersphere from p_{maj}.

The loss of a sample and its class prototype is given by:

$$\mathcal{L}^{i}_{p_{\tilde{x}}} = -\log \frac{e^{f(\tilde{x})\cdot p_i/\tau}}{\sum_{\tilde{x}^- \in \mathcal{W}\setminus\{\tilde{x}\}} e^{f(\tilde{x})\cdot f(\tilde{x}^-)/\tau}} \tag{10}$$

The complete contrastive loss with prototype alignment for class i, $\mathcal{L}^{i}_{\text{SupConProto}}$, and overall binary supervised contrastive loss with prototype alignment, $\mathcal{L}_{\text{SupConProto}}$, are defined as:

$$\mathcal{L}^{i}_{\text{SupConProto}} \triangleq \sum_{x\in\mathcal{X}_i} \begin{cases} \left[\mathcal{L}_{\text{NT-Xent}}(\tilde{x}) + \mathcal{L}^{(i)}_{p_{\tilde{x}}}\right] & \text{if } f(\tilde{x})\cdot p_i \leq 0.5 \\ \mathcal{L}_{\text{NT-Xent}}(\tilde{x}) & \text{otherwise} \end{cases} \tag{11}$$

$$\mathcal{L}_{\text{SupConProto}} = \mathcal{L}^{0}_{\text{SupConProto}} + \mathcal{L}^{1}_{\text{SupConProto}} \tag{12}$$

5. Experimental setup

5.1. Datasets

We utilize a total of seven datasets with binary class distributions that can be grouped into two categories: subsets of the iNaturalist21 (iNat21) dataset [35], where we artificially control class imbalance, and real-world medical datasets that naturally exhibit binary distributions and class imbalances. Our three subsets of iNat21 comprise plants (oaks and flowering plants), insects (bees and wasps), and mammals (hoved animals and carnivores). For each subset, we fix the class ratio to 50%-50%, 95%-5%, and 99%-1%, while keeping the total number of samples constant. Our real-world medical datasets include a cardiac datasets curated from the UK Biobank population study [29], two datasets from the medical MNIST collection, BreastMNIST and PneumoniaMNIST [40], and the FracAtlas dataset [1]. Additional details about dataset characteristics and preprocessing are provided in supplementary Sec. S1.

5.2. Network architecture and training

In line with prior work and baselines, we use a ResNet-50 image encoder [11], and follow established pre-training protocols[5, 6, 17]. After pre-training, we fine-tune a linear layer using a balanced subset comprising 1% of the training data and report accuracy on a balanced test set. As the medical datasets contain far fewer samples, we do not subsample them for fine-tuning and report the receiver operating characteristic area under the curve (AUC). Further information on the training protocols can be found in supplementary Sec. S2.

5.3. Baselines

In addition to standard SupCon [17] and weighted cross-entropy, we have included the five leading supervised contrastive learning methods for long-tailed datasets as baselines: parametric contrastive learning (PaCo) [7], k-positive contrastive learning (KCL) [15], targeted supervised contrastive learning (TSC) [21], subclass-balancing contrastive learning (SBC) [14], and balanced contrastive learning (BCL) [42]. We include results for KCL and TSC with 3 and 6 positives to fairly adapt them to the heavily imbalanced binary case. A brief description of each method can be found in Sec. 2.2. Further details about the setup and tuning of the baselines is included in supplementary Sec. S3. Additional baselines using classical, non-contrastive strategies to handle class imbalance, such as focal loss, oversampling, and undersampling, can be found in supplementary Sec. S4.

6. Results

First, we measure the performance of SupCon on binary imbalanced distributions, showing that it inversely correlates with dataset imbalance. Next, we show how the newly introduced SAA and CAC can diagnose representation space collapse, an issue that canonical alignment metrics fail to detect. We substantiate these findings by introducing a proof that provides a mathematical explanation for the observed behavior. Finally, we show the benefit of our proposed supervised contrastive learning strategies for binary imbalanced datasets compared to existing baselines for long-tailed distributions.

6.1. SupCon performance on binary datasets degrades with increasing class imbalance

We first evaluate SupCon on the three binary natural image datasets while varying the degree of class imbalance (see Tab. 1). We observe a sharp drop in linear probing accuracy as class imbalance increases. Specifically, models trained with 1% and 5% minority class representation achieve downstream accuracies between 50% and 60%, compared to over 90% accuracy in the balanced case. While

Table 1. Balanced accuracy of all evaluated methods on three binary natural imaging datasets at varying degrees of class imbalance. We compare standard weighted cross-entropy loss and supervised contrastive learning (top rows) to five baselines for supervised contrastive learning on long-tailed distributions (middle rows) and our two proposed fixes (bottom rows). Supervised Minority strategy does not apply to balanced settings and thus it is not reported there.

Method	Plants			Insects			Animals		
	50%	5%	1%	50%	5%	1%	50%	5%	1%
Weighted CE	81.1	61.4	60.1	82.4	63.4	62.8	70.7	61.9	57.3
SupCon [17]	93.7 ± 0.6	56.2 ± 1.6	54.4 ± 2.0	93.3 ± 0.1	62.6 ± 0.9	56.4 ± 0.1	80.8 ± 1.2	54.4 ± 1.6	56.9 ± 1.8
PaCo [7]	91.5 ± 0.9	59.2 ± 1.4	55.9 ± 2.2	92.4 ± 2.2	66.4 ± 0.6	53.7 ± 1.2	79.2 ± 1.5	65.3 ± 1.1	55.4 ± 1.6
KCL (K=3) [15]	90.6 ± 0.7	87.6 ± 0.4	81.1 ± 0.3	89.8 ± 0.3	81.1 ± 1.0	73.3 ± 1.4	81.8 ± 0.3	76.6 ± 0.6	71.2 ± 0.8
KCL (K=6) [15]	94.2 ± 0.4	86.6 ± 1.4	78.6 ± 0.5	91.5 ± 0.6	79.8 ± 1.2	69.2 ± 2.1	82.6 ± 0.6	75.3 ± 0.7	70.1 ± 2.0
TSC (K=3) [21]	93.4 ± 0.8	88.0 ± 0.5	79.4 ± 1.1	88.2 ± 0.6	81.2 ± 1.0	73.3 ± 1.6	82.4 ± 2.5	76.1 ± 0.9	71.2 ± 0.8
TSC (K=6) [21]	94.6 ± 0.4	87.5 ± 0.7	80.1 ± 0.4	91.1 ± 0.5	79.8 ± 1.3	71.2 ± 2.0	83.0 ± 1.3	75.2 ± 2.1	72.2 ± 0.9
SBC [14]	75.4 ± 0.9	57.2 ± 2.6	55.6 ± 1.9	77.6 ± 1.6	51.8 ± 2.3	54.0 ± 3.8	72.4 ± 1.1	55.1 ± 1.3	56.1 ± 2.1
BCL [42]	94.1 ± 0.3	85.0 ± 4.4	71.3 ± 2.9	**94.5 ± 0.1**	80.0 ± 2.5	74.0 ± 0.1	**86.2 ± 0.2**	76.5 ± 0.5	60.3 ± 0.7
Sup Minority	–	**89.8 ± 0.6**	**85.4 ± 0.5**	–	**82.8 ± 1.1**	**78.8 ± 0.8**	–	77.9 ± 0.9	**75.3 ± 0.5**
Sup Prototypes	**95.1 ± 0.2**	88.7 ± 0.7	83.4 ± 1.8	93.0 ± 0.3	81.2 ± 0.7	73.7 ± 1.3	82.9 ± 0.7	**79.2 ± 1.3**	73.0 ± 1.3

SupCon outperforms the weighted cross-entropy baseline on balanced datasets by over 10%, it underperforms this simple baseline by over 5% on binary imbalanced distributions. We find that the performance of SupCon drops below that of weighted cross-entropy around 20% imbalance, before completely collapsing between 5% and 1% (see supplementary Sec. S5).

6.2. Analyzing representation spaces using canonical metrics and novel metrics

To understand the underlying causes of the observed degradation, we analyze the learned representation spaces using both the canonical and new metrics (see Fig. 4). Both the canonical SAD and CAD are close to zero across all imbalances, suggesting high alignment of the learned representation space. However, they fail to put the distance in context to samples from other instances or classes. In comparison, our novel SAA and CAA metrics indicate that embeddings do not form distinct class-wise clusters. An SAA of 0 shows that the learned embeddings cannot differentiate between samples by input semantics. A CAC close to 50% suggests that both minority and majority class samples are almost equally mixed in local neighborhoods. Together, the new metrics confirm that SupCon's representation space collapses under heavy imbalance.

To further substantiate this empirically observed behavior, we present a proof in supplementary Sec. S7. The proof shows that gradients in the final network layer are upper-bounded by the inverse of the number of positives for a given sample. When the majority class dominates training batches, the gradient quickly saturates, preventing meaningful updates for that class and causing collapse towards a single point in the representation space.

6.3. Performance of fixes on natural and medical imaging datasets

Next, we compare our two fixes, Supervised Minority and Supervised Prototypes, against five established baselines for long-tailed supervised contrastive learning. Our Supervised Minority fix achieves the best linear probing performance across all iNat21 datasets (see Tab. 1), outperforming SupCon by 20% to 35%.

Our Supervised Minority fix also surpasses the performance of the five baselines developed for long-tailed datasets. Compared to these, its effectiveness increases at higher class imbalance. At 5% class imbalance it outperforms all baselines by at least 1%, and at 1% imbalance by a margin of 3%. Supervised Prototypes performed second best on all natural imaging datasets.

Supervised Prototypes performed best on three of the four medical datasets (see Tab. 2). Both of our fixes generally match or outperform all five baselines developed for long-tailed data distributions. On the infarction dataset which has the strongest imbalance (4%) we see the largest gain of +2% AUC over the best performing baseline.

Extensive ablations across temperatures, batch sizes and varying degrees of supervision in both the minority and majority class can be found in supplementary Sec. S8. Visualizations of the learned embeddings via UMAP in supplementary Sec. S9 also corroborate that Supervised Minority and Supervised Prototypes avoid the representation collapse observed in standard SupCon.

6.4. Our proposed metrics correlate with downstream classification performance

Classical uniformity (GPU) and alignment (SAD) metrics focus on pairwise sample-level relationships without considering class-level context. While class alignment dis-

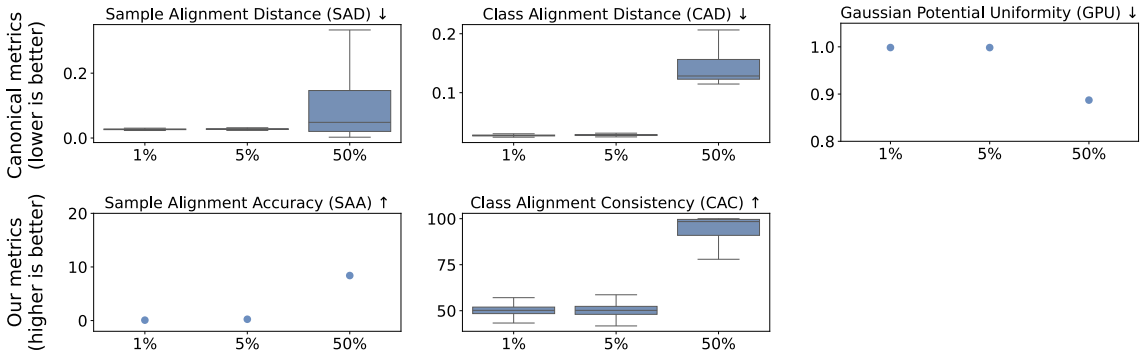

Figure 4. Boxplots of metrics analysing SupCon's representation space learned from the plants dataset. As class imbalance grows the representation space collapses despite the canonical SAD and CAD metrics being low. In contrast, SAA and CAC correctly identify the collapse. Similar results are observed on the insects and animals datasets (see supplementary Sec. S6).

Table 2. Area under the curve (AUC) of all evaluated methods on four medical imaging datasets. We compare standard weighted cross-entropy loss and supervised contrastive learning (top rows) to five baselines for supervised contrastive learning for long-tailed distributions (middle rows) and our two proposed fixes (bottom rows).

Method	UKBB	MedMNIST		FracAtlas
	Infarction (4%)	BreastMNIST (37%)	PneumoniaMNIST (35%)	Fractures (21%)
Weighted CE	72.4	75.1	98.8	79.8
SupCon [17]	61.9 ± 1.9	75.1 ± 0.7	99.5 ± 0.1	84.8 ±0.1
PaCo [7]	66.6 ± 1.3	66.0 ± 1.9	98.7 ± 0.2	83.7 ± 0.6
KCL (K=3) [15]	75.3 ± 0.4	89.9 ± 0.8	99.6 ± 0.1	**88.2 ± 0.1**
KCL (K=6) [15]	73.6 ± 0.2	89.5 ± 0.4	98.9 ± 0.1	86.5 ± 0.1
TSC (K=3) [21]	75.7 ± 0.1	89.2 ± 0.1	99.5 ± 0.1	87.1 ± 0.1
TSC (K=6) [21]	75.0 ± 0.1	88.5 ± 0.1	99.5 ± 0.1	86.3 ± 0.1
SBC [14]	70.0 ± 0.3	80.8 ± 0.7	99.3 ± 1.2	80.9 ± 5.3
BCL [42]	74.0 ± 0.1	90.5 ± 0.1	99.6 ± 0.1	84.9 ± 0.1
Sup Minority	77.7 ± 1.1	86.4 ± 0.2	99.6 ± 0.1	82.3 ±0.7
Sup Prototypes	**77.9 ± 0.4**	**90.7 ± 0.5**	**99.8 ± 0.1**	86.0 ± 0.1

tance (CAD) incorporates intra-class similarity for supervised contrastive learning, it does not account for the critical inter-class relationships that influence downstream separability.

As shown in Fig. 5, the canonical metrics exhibit near-zero correlation with linear probing accuracy when plotted globally across all datasets. In contrast, our class alignment consistency achieves an R^2 value of 0.69, indicating a strong global linear relationship with downstream classification accuracy.

When considering each dataset individually and averaging the correlations over all datasets, we find weak correlations between 0.15 and 0.3 for the classical metrics. In contrast, our novel metrics achieve a mean R^2 value of 0.45 and 0.65 due to the fact that they consider inter-sample and inter-class statistics. This shows the suitability of our metrics for evaluating representation space quality for downstream utility both on a global and local scale.

7. Discussion and conclusion

Although imbalanced binary distributions are commonly found in real-world machine learning problems and especially in medical applications, they have received little attention in the context of supervised contrastive learning. In extensive experiments on seven natural and medical imaging datasets, we have shown that SupCon on binary imbalanced distributions often results in collapsed representations, leading to poor downstream performance.

To diagnose these failures, we introduced two new metrics: Sample Alignment Accuracy (SAA) and Class Alignment Consistency (CAC) which extend the notion of alignment to measure how well samples and classes are distinguished from each other in the learned space. These metrics uncovered shortcomings that canonical measures overlooked, showing that SupCon fails to form meaningful embeddings under high imbalance. Crucially, CAC correlates well with linear probing accuracy and is thus a suitable metric for measuring representation space quality for down-

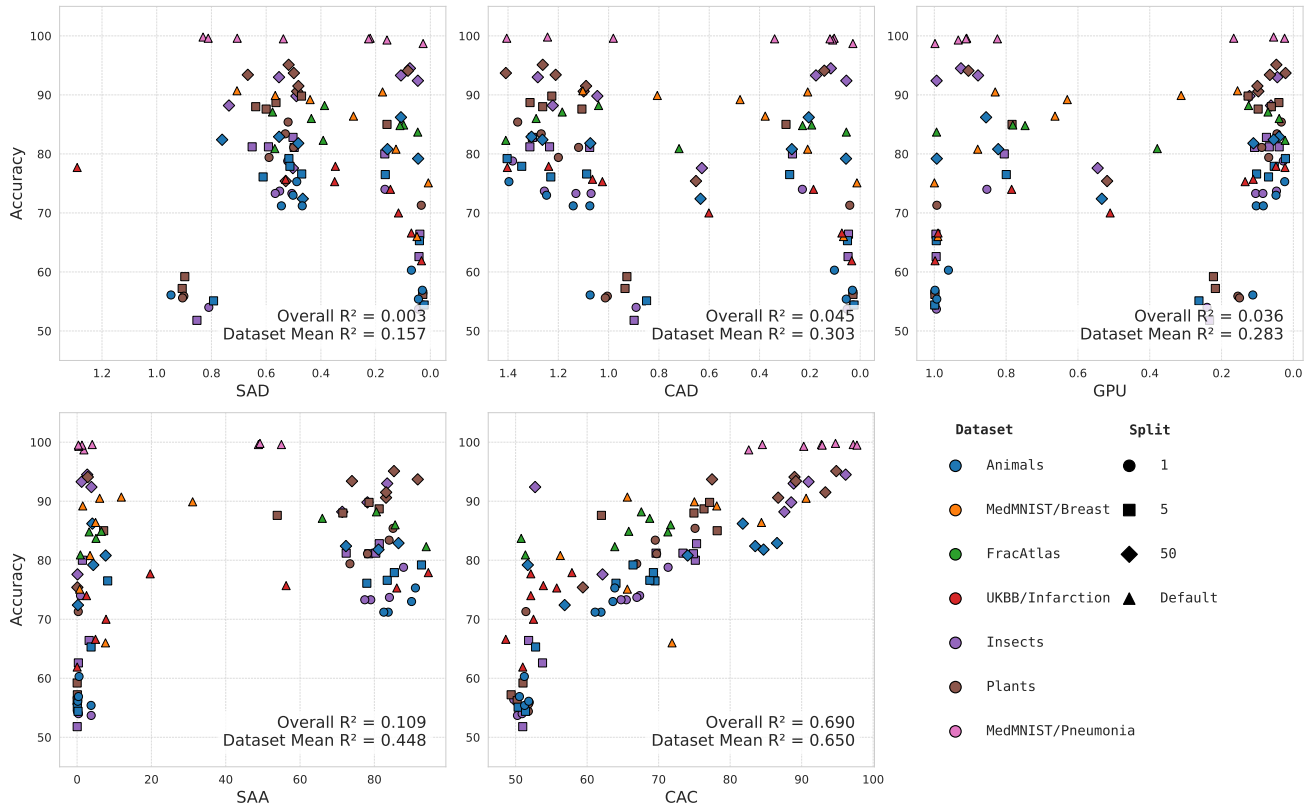

Figure 5. Correlations between five representation-space metrics and linear probing performance across all datasets and all considered methods. The overall R^2 is calculated globally over all points while the dataset mean R^2 is calculated per dataset and then averaged. As SAA and CAC are the only metrics that account for relationships between samples and classes instead of simply within them, they correlate much stronger with downstream performance.

stream applications.

Finally, we proposed two fixes, *Supervised Minority* and *Supervised Prototypes*, specifically tailored to address binary imbalance. Both solutions boost accuracy by up to 35% over standard SupCon and surpass existing methods for long-tailed distributions by up to 5%. With minimal additions to the standard SupCon loss and negligible computational overhead, these fixes offer a straightforward path to improved performance on binary imbalanced classification problems.

Limitations A limitation of our methods is that the Supervised Minority fix cannot be used when data is balanced (as there is no minority class) and there does not seem to be a particularly clear pattern when Supervised Minority performs better than Supervised Prototypes or vice versa. We observed that Supervised Prototypes always achieved the best performance on medical datasets, while Supervised Minority usually performs better on natural image datasets. We hypothesize that this could either be due to the domain-specific data characteristics or a dependence of both methods on the degree of class imbalance that slightly differed in our experiments. When choosing an approach for a new dataset, it would be prudent to test both methods.

Conclusion Our study complements and extends a series of previous works that have aimed to explore the theoretical foundations of contrastive learning and researched its application to long-tailed datasets. By focusing on the particularly challenging case of binary imbalanced datasets, we have improved the understanding of the dynamics of contrastive learning and developed tools to diagnose and enhance methods dealing with such datasets, which are very common in real-world applications, such as medicine.

Acknowledgments This research has been conducted using the UK Biobank Resource under Application Number 87802. This work was supported in part by the European Research Council grant Deep4MI (Grant Agreement no. 884622). Martin J. Menten is funded by the German Research Foundation under project 532139938.

References

[1] Iftekharul Abedeen, Md Ashiqur Rahman, Fatema Zohra Prottyasha, Tasnim Ahmed, Tareque Mohmud Chowdhury, and Swakkhar Shatabda. Fracatlas: A dataset for fracture classification, localization and segmentation of musculoskeletal radiographs. *Scientific Data*, 10(1):521, 2023. 5, 11, 13

[2] Mido Assran, Randall Balestriero, Quentin Duval, Florian Bordes, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, and Nicolas Ballas. The hidden uniform cluster prior in self-supervised learning. In *The Eleventh International Conference on Learning Representations*, 2023. 1

[3] Antonio Pedro Camargo, Simon Roux, Frederik Schulz, Michal Babinski, Yan Xu, Bin Hu, Patrick SG Chain, Stephen Nayfach, and Nikos C Kyrpides. Identification of mobile genetic elements with genomad. *Nature Biotechnology*, 42(8):1303–1312, 2024. 1

[4] Mayee Chen, Daniel Y Fu, Avanika Narayan, Michael Zhang, Zhao Song, Kayvon Fatahalian, and Christopher Ré. Perfectly balanced: Improving transfer and robustness of supervised contrastive learning. In *International Conference on Machine Learning*, pages 3090–3122. PMLR, 2022. 2

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020. 4, 5, 14

[6] Tsai Shien Chen, Wei Chih Hung, Hung Yu Tseng, Shao Yi Chien, and Ming Hsuan Yang. Incremental false negative detection for contrastive learning. In *10th International Conference on Learning Representations, ICLR 2022*, 2022. 5, 14

[7] Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 715–724, 2021. 1, 2, 5, 6, 7, 15

[8] Florian Graf, Christoph Hofer, Marc Niethammer, and Roland Kwitt. Dissecting supervised contrastive learning. In *International Conference on Machine Learning*, pages 3821–3830. PMLR, 2021. 1, 2

[9] Paul Hager, Martin J. Menten, and Daniel Rueckert. Best of both worlds: Multimodal contrastive learning with tabular and imaging data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23924–23935, 2023. 2, 12

[10] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *Advances in neural information processing systems*, 33:5679–5690, 2020. 1

[11] 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*, pages 770–778, 2016. 5, 13, 17

[12] 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*, pages 9729–9738, 2020. 15

[13] Kim Hee et al. Transfer learning for medical image classification: a literature review. *BMC Medical Imaging*, 2022. 2

[14] Chengkai Hou, Jieyu Zhang, Haonan Wang, and Tianyi Zhou. Subclass-balancing contrastive learning for long-tailed recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5395–5407, 2023. 1, 2, 5, 6, 7, 15

[15] Bingyi Kang, Yu Li, Sa Xie, Zehuan Yuan, and Jiashi Feng. Exploring balanced feature spaces for representation learning. In *International Conference on Learning Representations*, 2021. 2, 5, 6, 7, 14, 21

[16] Tejaswi Kasarla, Gertjan Burghouts, Max Van Spengler, Elise Van Der Pol, Rita Cucchiara, and Pascal Mettes. Maximum class separation as inductive bias in one matrix. *Advances in neural information processing systems*, 35:19553–19566, 2022. 5

[17] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. 1, 5, 6, 7, 14, 17

[18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. 14

[19] Vinod Kumar, Gotam Singh Lalotra, Ponnusamy Sasikala, Dharmendra Singh Rajput, Rajesh Kaluri, Kuruva Lakshmanna, Mohammad Shorfuzzaman, Abdulmajeed Alsufyani, and Mueen Uddin. Addressing binary classification over class imbalanced clinical datasets using computationally intelligent techniques. *Healthcare*, 10(7), 2022. 2

[20] Shikun Li, Xiaobo Xia, Shiming Ge, and Tongliang Liu. Selective-supervised contrastive learning with noisy labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 316–325, 2022. 1

[21] Tianhong Li, Peng Cao, Yuan Yuan, Lijie Fan, Yuzhe Yang, Rogerio S Feris, Piotr Indyk, and Dina Katabi. Targeted supervised contrastive learning for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6918–6928, 2022. 1, 2, 3, 5, 6, 7

[22] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022. 17

[23] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 15

[24] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2022. 14

[25] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and

projection. *Journal of Open Source Software*, 3(29):861, 2018. 22

[26] Pascal Mettes, Elise Van der Pol, and Cees Snoek. Hyperspherical prototype networks. *Advances in neural information processing systems*, 32, 2019. 5

[27] Richard W Nesto. Screening for asymptomatic coronary artery disease in diabetes. *Diabetes Care*, 22(9):1393, 1999. 12

[28] Steffen Schneider, Jin Hwa Lee, and Mackenzie Weygandt Mathis. Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, 617(7960):360–368, 2023. 1

[29] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3): e1001779, 2015. 5, 11, 12

[30] Bo Sun, Banghuai Li, Shengcai Cai, Ye Yuan, and Chi Zhang. Fsce: Few-shot object detection via contrastive proposal encoding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7352–7362, 2021. 1

[31] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pages 20827–20840. PMLR, 2022. 1

[32] Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive learning, multi-view redundancy, and linear models. In *Algorithmic Learning Theory*, pages 1179–1206. PMLR, 2021. 11

[33] Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Demystifying self-supervised learning: An information-theoretical framework. *arXiv preprint arXiv:2006.05576*, 2020. 11

[34] Paul Valensi, Luc Lorgis, and Yves Cottin. Prevalence, incidence, predictive factors and prognosis of silent myocardial infarction: a review of the literature. *Archives of cardiovascular diseases*, 104(3):178–188, 2011. 12

[35] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12884–12893, 2021. 5, 11

[36] Peng Wang, Kai Han, Xiu-Shen Wei, Lei Zhang, and Lei Wang. Contrastive learning based hybrid networks for long-tailed image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 943–952, 2021. 1

[37] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pages 9929–9939. PMLR, 2020. 2, 3, 4

[38] Yifei Wang, Qi Zhang, Yisen Wang, Jiansheng Yang, and Zhouchen Lin. Chaos is a ladder: A new theoretical understanding of contrastive learning via augmentation overlap. In *International Conference on Learning Representations*, 2022. 2, 11

[39] Yihao Xue, Siddharth Joshi, Eric Gan, Pin-Yu Chen, and Baharan Mirzasoleiman. Which features are learnt by contrastive learning? On the role of simplicity bias in class collapse and feature suppression. In *Proceedings of the 40th International Conference on Machine Learning*, pages 38938–38970. PMLR, 2023. 2

[40] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023. 5, 11, 13

[41] Yibo Yang, Shixiang Chen, Xiangtai Li, Liang Xie, Zhouchen Lin, and Dacheng Tao. Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network? *Advances in neural information processing systems*, 35:37991–38002, 2022. 5

[42] Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. Balanced contrastive learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6908–6917, 2022. 1, 2, 5, 6, 7