Semi-Supervised Contrastive Learning with Orthonormal Prototypes

Huanran Li, Manh Nguyen, & Daniel Pimentel-Alarcón
Department of Electrical Engineering, Statistics, Biostatistics
Wisconsin Institute of Discovery
University of Wisconsin-Madison
{hli488}{mdnguyen4}{pimentelalar}@wisc.edu

Abstract

Contrastive learning has emerged as a powerful method in deep learning, excelling at learning effective representations through contrasting samples from different distributions. However, dimensional collapse, where embeddings converge into a lower-dimensional space, poses a significant challenge, especially in semi-supervised and self-supervised setups. In this paper, we first identify a critical learning-rate threshold, beyond which standard contrastive losses converge to collapsed solutions. Building on these insights, we propose CLOP, a novel semi-supervised loss function designed to prevent dimensional collapse by promoting the formation of orthogonal linear subspaces among class embeddings. Through extensive experiments on real and synthetic datasets, we demonstrate that CLOP improves performance in image classification and object detection tasks while also exhibiting greater stability across different learning rates and batch sizes.

1 Introduction

Recent advancements in deep learning have positioned Contrastive Learning as a leading paradigm, largely due to its effectiveness in learning representations by contrasting samples from different distributions while aligning those from the same distribution. Prominent models in this domain include SimCLR [4], Contrastive Multiview Coding (CMC) [40], VICReg [1], BarLowTwins [56], among others [16, 28, 46]. These models share a common two-stage framework: representation learning and fine-tuning. In the first stage, representation learning is performed in a self-supervised manner, where the model is trained to map inputs to embeddings using contrastive loss to separate samples from different labels. In the second stage, fine-tuning occurs under a supervised setup, where labeled data is used to classify embeddings correctly. For practical applicability, a small amount of labeled data is required in the fine-tuning stage to produce meaningful classifications, making the overall pipeline semi-supervised. Empirical evidence demonstrates that these models, even with limited labeled data (as low as 10%), can achieve performance comparable to fully-supervised approaches on moderate to large datasets [20].

Despite the effectiveness of contrastive learning on largely unlabeled datasets, a common issue encountered during the training process is dimensional collapse. As pointed out by [11, 13, 14, 21, 35, 39, 51], this phenomenon describes the collapse of output embeddings from the neural network into a lower-dimensional space, reducing their spatial utility and leading to indistinguishable classes. There are two main approaches to resolve this issue: augmentation modification [11, 21, 39, 51] and loss modification [11, 14, 35]. In this paper, we propose a semi-supervised loss function CLOP to address the issue of collapse. To address dimensional collapse, our approach selects prototypes similarly to [13, 60]. The key distinction is that our method aims to push the embeddings toward

Preprint. Under review.

distinct orthogonal linear subspaces, allowing them to occupy a higher-rank space. We demonstrate through experiments that CLOP is more effective for image classification and object detection tasks.

The main contributions of this paper can be viewed from two perspectives. First, we identify and analyze the detrimental effects of inappropriate learning rates on contrastive learning losses that rely exclusively on cosine similarity. Specifically, we show that when embeddings collapse into a rank-1 subspace, the optimizer converges to an uninformative stationary point. Through a gradient-descent analysis, we establish the existence of a critical learning-rate range outside of which training inevitably converges to this undesirable point. Building upon these insights, we propose a novel contrastive loss term, termed CLOP, designed to encourage embeddings of partial training datasets to cluster around a set of orthonormal prototypes. This loss term is applicable to both semi-supervised and fully-supervised contrastive learning scenarios, particularly when only a subset of the training data is labeled. We validate CLOP's effectiveness through extensive experiments on embedding visualization, image classification, and object detection tasks, using both balanced and imbalanced datasets. Our empirical results demonstrate that CLOP consistently outperforms baseline methods, exhibiting remarkable robustness and stability across a wide range of learning rates and reduced batch sizes.

Paper Organization In Section 2, we provide essential background information and discuss recent advancements in both self-supervised and supervised contrastive learning, as well as analyzing the dimensional collapse phenomenon associated with contrastive learning methods. Section 3 elaborates on the motivation behind our approach through simulations and detailed gradient analysis. Subsequently, in Section 4, we introduce our proposed model, **CLOP**. Finally, Section 5 presents extensive experimental evaluations conducted on image datasets.

2 Related Work

Contrastive learning has gained prominence in deep learning for its ability to learn meaningful representations by pulling together similar (positive) pairs and pushing apart dissimilar (negative) pairs in the embedding space. Positive pairs are generated through techniques like data augmentation, while negative pairs come from unrelated samples, making contrastive learning particularly effective in self-supervised tasks like image classification. Pioneering models such as SimCLR [4], CMC [40], VICReg [1], and Barlow Twins [56] share the objective of minimizing distances between augmented versions of the same input (positive pairs) and maximizing distances between unrelated inputs (negative pairs). SimCLR maximizes agreement between augmentations using contrastive loss, while CMC extends this to multi-view learning [4, 40]. VICReg introduces variance-invariance-covariance regularization without relying on negative samples [1], and Barlow Twins reduce redundancy between different augmentations [56].

Recent innovations have improved contrastive learning across various domains. For instance, methods like structure-preserving quality enhancement in CBCT images [22] and false negative cancellation [18] have enhanced image quality and classification accuracy. In video representation, cross-video cycle-consistency and inter-intra contrastive frameworks [45, 38] have shown significant gains. Additionally, contrastive learning has advanced sentiment analysis [50], recommendation systems [52], and molecular learning with faulty negative mitigation [44]. [48] introduces GraphACL, a novel framework for contrastive learning on graphs that captures both homophilic and heterophilic structures without relying on augmentations.

2.1 Contrastive Loss

In unsupervised learning, [46] introduced InfoNCE, a loss function defined as:

$$\mathcal{L}_{\text{infoNCE}} = -\sum_{i \in I} \log \frac{\exp(\mathbf{z}_i^\top \mathbf{z}_{j(i)}/\tau)}{\sum_{a \neq i} \exp(\mathbf{z}_i^\top \mathbf{z}_a/\tau)} \tag{1}$$

where $\mathbf{z}_i$ is the embedding of sample i, $j(i)$ its positive pair, and τ controls the temperature.

Recent refinements focus on (1) component modifications, (2) similarity adjustments, and (3) novel approaches. [28] use EM with k-means to update centroids and reduce mutual information loss, while [43] add L2 distance to InfoNCE, though both underperform state-of-the-art (SOTA) techniques. [49] reduce noise with augmentations, and [55] improve gradient efficiency with Decoupled Contrastive

Learning, though neither surpasses SOTA. In similarity adjustments, [5] propose a debiased loss, and [12] use hyperbolic embeddings, but neither outperforms SOTA. Novel methods include min-max InfoNCE [41], Euclidean-based losses [1], and dimension-wise cosine similarity [56], achieving competitive performance without softmax-crossentropy.

2.2 Semi-Supervised Contrastive Learning

Semi-supervised contrastive learning effectively leverages both labeled and unlabeled data to learn meaningful representations. [57] introduced a framework with similarity co-calibration to mitigate noisy labels by adjusting the similarity between pairs. [19] proposed Generalized Contrastive Loss (GCL), which unifies supervised and unsupervised learning for speaker recognition, while [24] combined contrastive self-supervision with consistency regularization in SelfMatch. [37] introduced FixMatch, which combines pseudo-labeling with consistency regularization. In this approach, weakly augmented samples generate pseudo-labels that guide strongly augmented versions, ensuring robust semi-supervised learning (SSL) performance. [53] enhanced SSCL by enforcing class-wise consistency in learned representations, improving robustness to class imbalance and increasing generalization across datasets. [59] proposed SimMatch, a framework that unifies contrastive learning and consistency regularization by optimizing both instance-level alignment and class-level semantic consistency, leading to improved SSL feature representations. Building on this, [58] introduced SimMatch-V2, refining the balance between contrastive and consistency learning objectives, further enhancing transferability and performance in semi-supervised settings.

2.3 Dimensional Collapse

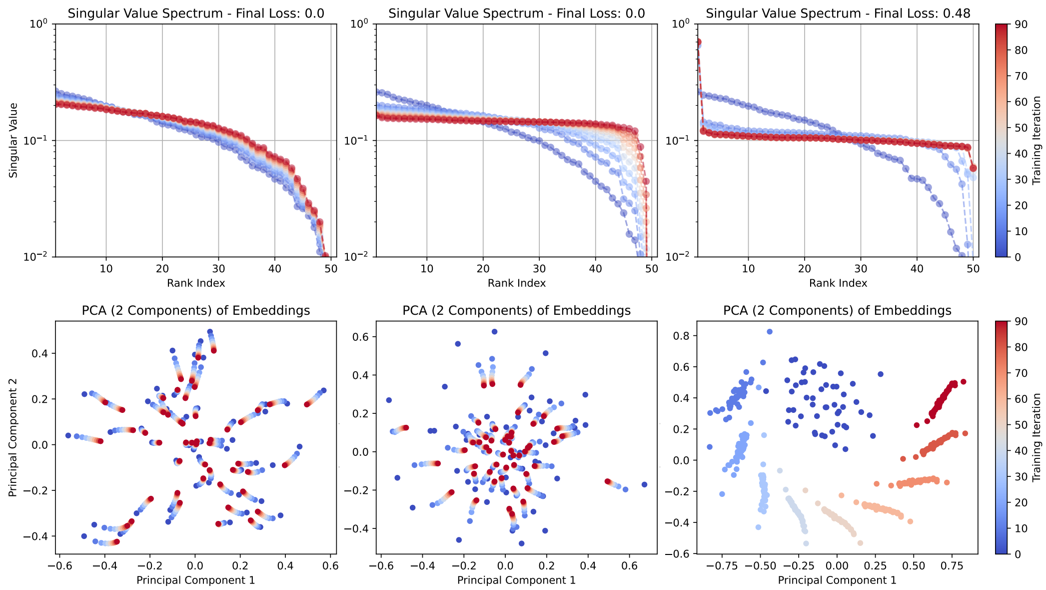
For dimensional collapse in contrastive learning, [21] examine dimensional collapse in self-supervised learning. They attribute this to strong augmentations distorting features and implicit regularization driving weights toward low-rank solutions. Similarly, [51] explore how simplicity bias leads to class collapse and feature suppression, with models favoring simpler patterns over complex ones. They suggest increasing embedding dimensionality and designing augmentation techniques that preserve class-relevant features to counter this bias and promote diverse feature learning. [11] emphasize the role of data augmentation and loss design in preventing class collapse, proposing a class-conditional InfoNCE loss term that uniformly pulls apart individual points within the same class to enhance class separation. In supervised contrastive learning, [13] propose loss function modifications to follow an ETF geometry by selecting prototypes that form this structure. In graph contrastive learning, [39] introduce a whitening transformation to decorrelate feature dimensions, avoiding collapse and enhancing representation capacity. Finally, [35] investigate the preference of contrastive learning for content over style features, leading to collapse. They propose to leverage adaptive temperature factors in the loss function to improve feature representation quality.

3 Motivation

This section investigates the cause of dimensional collapse in the perspective of learning rates. We first demonstrate that dimensional collapse represents a local stationary point of the InfoNCE loss by showing that linear embeddings result in a zero gradient (Lemma 1). While our demonstration utilizes the InfoNCE loss, this conclusion generalizes to most current loss functions relying solely on cosine similarity as their metric, encompassing unsupervised [16, 4, 7, 49, 55, 43, 27], semi-supervised [17, 36], and supervised contrastive learning [23, 7, 33, 29]. Subsequently, we discuss how a large learning rate induces a shift in the embedding mean through comparisons between pairs of negative samples, ultimately leading to dimensional collapse.

The InfoNCE loss (Eq. (1)) aims to encourage the embeddings to form distinguishable clusters in high-dimensional space, thereby facilitating classification for downstream models. However, in Lemma 1, we demonstrate that the worst-case scenario — where all embeddings become identical or co-linear — also constitutes a local optimum for the InfoNCE loss. This observation suggests that, from a theoretical perspective, InfoNCE exhibits instability, as both the best and worst solutions can lead to stationary points.

Lemma 1. *Let $\mathcal{F} : \mathbb{R}^m \to \mathbb{R}^{m'}$ be a family of Contrastive Learning structures, where m and m' denote the dimensions of the inputs and embeddings, respectively. If a function $f \in \mathcal{F}$ is trained*



(a) Learning rate of 0.01 **(b)** Learning rate of 0.1 **(c)** Learning rate of 1

Figure 1: Simulation with Repulsive Force on 50 simulated points in 50-dimensional space.

using the InfoNCE loss, then there exist infinitely many local stationary points where all embeddings produced by f are all equal or co-linear.

The proof of Lemma 1 relies on the observation that the embeddings are only compared against each other. If all embeddings are either identical or co-linear, the gradient vanishes due to the lack of angular differences, as well as the normalization process. The full proof is presented in Appendix A.

The remainder of this section examines the role of a large learning rate in causing dimensional collapse. To understand the dynamics of contrastive learning, it is crucial to consider two forces acting on each embedding: the *gravitational force* within the same pseudo class and the *repulsive force* between different pseudo classes. Contrary to the common belief that the gravitational force is responsible for inducing collapse, we observe that the overshooting of the repulsive force could be directly related to the dimensional collapse in contrastive learning.

To better illustrate the repulsive force, we conducted a simulation using 50 randomly generated embeddings in 50-d space, where positive pairs initially coincide at a single point, reflecting a scenario where the model has successfully merged augmented variants from the same input source into one embedding. We subsequently performed gradient descent on these embeddings using the InfoNCE loss with temperature of 0.1, recording the embedding trajectories throughout training. The simulations were executed across three distinct learning rates: 0.01, 0.1, and 1. The embedding singular value spectrum and the first two principal components of these embeddings are visualized in Figure 1. At learning rates of 0.01 and 0.1, the repulsive force inherent to the InfoNCE loss effectively redistributes the embeddings across a more uniform space, as indicated by the more evenly dispersed singular values across the embedding dimensions. Conversely, at a learning rate of 1, the embeddings fail to redistribute and instead collapse into a one-dimensional subspace, corroborating the stationary point described by Lemma 1. Furthermore, analysis of the first two principal components clearly illustrates a significant shift in the embedding mean and a reduction in variance throughout the training iterations. This observation indicates that larger learning rates induce a substantial shift in the embedding mean, consequently accelerating the collapse process by aligning gradients in similar directions. Thus, establishing an upper bound on the learning rates is crucial, as it limits the shift in embedding mean and would contribute to preventing dimensional collapse.

To better understand this phenomenon, we conducted a theoretical analysis of the gradient descent process using the InfoNCE loss, detailed in Appendix B. Briefly, our analysis reveals that, after performing a single gradient descent step, the upper bound on the norm of the embedding mean is scaled by the factor:

$$\|\boldsymbol{\mu}_1\|_2 \leq \left| \left(1 - \frac{2\eta}{\tau}\right) \frac{2}{\sigma^2} \frac{|\mathcal{N}|}{1+|\mathcal{N}|} \right| \|\boldsymbol{\mu}_0\|_2,$$

where σ represents the minimum embedding norm before normalization, and $|\mathcal{N}|$ is the number of negative samples within the batch. To prevent an increase in the embedding mean's norm, we constrain this scaling factor to be less than or equal to 1. This constraint yields an optimal learning rate range given by: $\eta \in \left[\frac{\tau}{2}\left(1 - \frac{\sigma^2(1+|\mathcal{N}|)}{2|\mathcal{N}|}\right), \frac{\tau}{2}\left(1 + \frac{\sigma^2(1+|\mathcal{N}|)}{2|\mathcal{N}|}\right)\right]$. This suggests that, under the assumption that positive sample pairs are successfully merged into the same embedding, the optimal learning rate that reduce mean shift is $\frac{\tau}{2}$.

4 CLOP: Populating Embedding Rank with Orthonormal Prototypes

To avoid the issue of embedding collapsing into a rank-1 linear subspace, we introduce a novel approach that promotes point isolation by adding an additional term to the loss function for contrastive learning. Specifically, we initialize a group of *orthonormal prototypes*. The number of orthonormal prototypes matches the total number of classes in the dataset. We then maximize the similarity between the orthonormal prototypes and the labeled samples in the training set.

Formally, let $\mathcal{S}$ be the labeled training set containing pairs of embeddings and labels, denoted as $\mathcal{S} = \{(\mathbf{z}_i, y_i) \mid i \in \{1, \dots, |\mathcal{S}|\}\}$. The set of prototypes, denoted as $\mathcal{C}$, is defined as $\mathcal{C} = \{\mathbf{c}_1, \dots, \mathbf{c}_k\}$, where k represents the number of classes in the dataset. To generate the prototypes $\mathcal{C}$, we randomly sample k i.i.d. vectors from an m'-dimensional space, where $|\mathbf{z}_i| = m'$. Subsequently, we apply singular value decomposition (SVD) to obtain the orthonormal basis, denoted as $\mathcal{C}$. This ensures that each prototype $\mathbf{c}_i$ is initialized as a unit vector, orthogonal to all other prototypes, at the beginning of the training process. The CLOP loss is formulated as follows:

$$\mathcal{L}_{\text{CLOP}} = \mathcal{L}_{\text{infoNCE}} + \lambda \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} (1 - s(\mathbf{z}_i, \mathbf{c}_{y_i})), \tag{2}$$

where $s(\cdot, \cdot)$ denotes the similarity metric, typically chosen to be the same as that used in $\mathcal{L}_{\text{infoNCE}}$, namely cosine similarity.

The primary objective of the CLOP loss is to align all embeddings corresponding to the same class towards a common target prototype, $\mathbf{c}_{y_i}$. Beyond the “gravitational force" and “repulsive force" provided by the main contrastive loss, the CLOP loss introduces a supervised “pulling force" that prevents collapse by pulling labeled embeddings into class-specific orthogonal subspaces. It is important to note that, without additional constraints, samples outside of set $\mathcal{S}$ may still converge to other unspecified embeddings, potentially collapsing into a rank-1 subspace. However, a fundamental assumption in contrastive learning is that augmented samples are treated as being drawn from the same distribution as the original input data from the same class. Thus, the “gravitational force” between embeddings of the same class should pull unsupervised embeddings toward the target prototypes.

To assess the supervisory efficacy of CLOP, we visualized the CIFAR100's representations learned by both CLOP and SupCon under two label regimes (10% and 100%) in Table 1. Embeddings were projected to two dimensions using t-distributed Stochastic Neighbor Embedding (t-SNE) [42], and we also examined each model's singular value spectrum. Both methods employ a ResNet-50 backbone with linear projection heads four times wider than standard (denoted as *ResNet-50 (4x)*), trained for 500 epochs with a batch size of 256 and an initial learning rate of 1.0. CLOP yields a higher effective rank in its embedding space and produces more compact, well-separated clusters under both limited (10%) and full (100%) supervision. Quantitatively, linear fine-tuning on ImageNet confirms this advantage: at 10% label usage, CLOP achieves a top-1 accuracy of 59.57%, outperforming SupCon's 55.01% by 4.56%; at 100% label usage, CLOP reaches 82.06% versus SupCon's 74.75%, a gain of 7.31%. This experiment is conducted on a single NVIDIA A100 GPU, with each run completing within 4 hours of training.

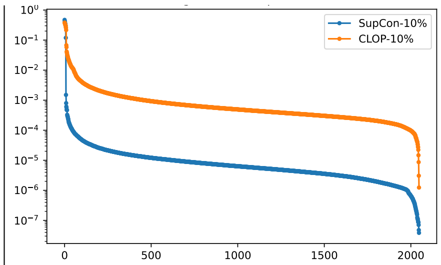
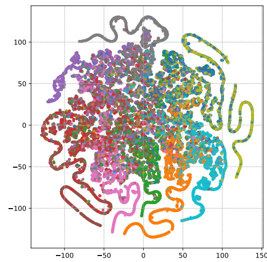
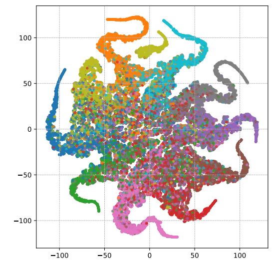
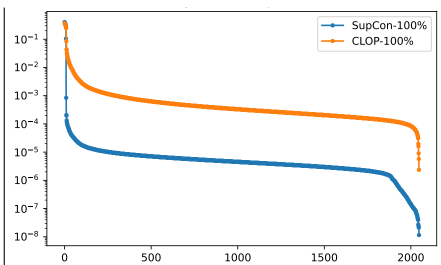
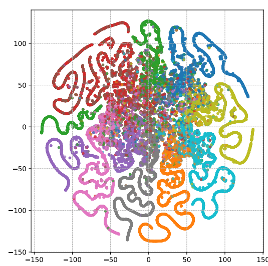
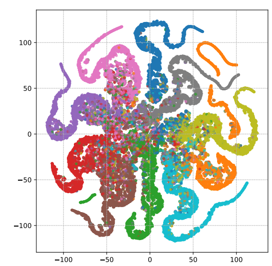
<table>
<tr><th></th><th>Singular Value Spectrum</th><th>SupCon</th><th>CLOP</th></tr>
<tr><td>10% Label</td><td>
</td><td>
Top-1 Acc: 55.01%</td><td>
Top-1 Acc: 59.57%</td></tr>
<tr><td>100% Label</td><td>
</td><td>
Top-1 Acc: 74.75%</td><td>
Top-1 Acc: 82.06%</td></tr>
</table>

Table 1: SupCon vs. CLOP under different label usage

Methods	10% Labels		50% Labels	
	Top-1	Top-5	Top-1	Top-5
SupCon	0.595	0.865	0.625	0.884
FixMatch	0.588	0.883	0.611	0.887
SsCL	0.715	0.829	0.746	0.887
CCSSL	0.735	0.841	0.757	0.894
SimMatch	0.719	0.829	0.724	0.866
SimMatch-V2	0.729	0.840	0.729	0.877
CLOP	**0.743**	**0.904**	**0.76**	**0.92**

Table 2: Top-1 and Top-5 accuracy on CIFAR-100.

Methods	10% Labels		50% Labels	
	Top-1	Top-5	Top-1	Top-5
SupCon	0.704	0.874	0.734	0.830
FixMatch	0.720	0.886	0.774	0.911
SsCL	0.721	0.909	0.786	0.899
CCSSL	0.751	0.923	0.771	0.907
SimMatch	0.740	**0.930**	0.776	0.904
SimMatch-V2	0.748	0.917	0.788	0.916
CLOP	**0.791**	0.927	**0.829**	**0.949**

Table 3: Top-1 and Top-5 accuracy on ImageNet.

5 Experiment

In this section, we evaluate the performance of CLOP in various learning settings. First, we compare CLOP against existing semi-supervised contrastive learning methods on image classification tasks. Furthermore, we show that CLOP consistently outperforms competitors under both balanced and imbalanced class distributions. Next, we highlight the generalizability of CLOP by presenting its outstanding transfer learning results on image classification and object detection tasks. Finally, we conduct extensive ablation studies on key hyperparameters, including learning rate, batch size, λ, similarity metrics, and augmentation strategies, highlighting CLOP's robustness to varying learning rates and its effectiveness in small-batch learning.

5.1 Semi-Supervised Image Classification

For balanced-class training, we utilize the full CIFAR-100 [26] and ImageNet [8] datasets, considering scenarios where either 10% or 50% of labels are available for contrastive learning. CLOP is implemented with a ResNet-50 (4x) backbone using the SimCLR loss function. We benchmark CLOP against several state-of-the-art semi-supervised contrastive learning methods, including SupCon [23], FixMatch [37], SsCL [57], CCSSL [53], SimMatch [59], and SimMatch-V2 [58]. The classification

Methods	CIFAR-100		ImageNet-200		ImageNet	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
SupCon	0.585	0.858	0.505	0.749	0.672	0.780
FixMatch	0.576	0.840	0.621	0.759	0.708	0.860
SsCL	0.719	0.860	0.585	0.733	0.730	0.862
CCSSL	0.744	0.874	0.631	0.889	0.710	0.865
SimMatch	0.685	0.829	0.527	0.768	0.705	0.855
SimMatchV2	0.689	0.862	0.659	0.846	0.733	0.845
CLOP	**0.763**	**0.918**	**0.689**	**0.898**	**0.799**	**0.932**

Table 4: Top-1 and Top-5 accuracy on CIFAR-100, ImageNet-200, and ImageNet under imbalanced-class training.

Method	Food	CIFAR10	CIFAR100	SUN397	DTD	Caltech-101	Flowers
SimCLR	0.8820	0.9770	0.8590	**0.6350**	0.7320	0.9210	0.9700
SupCon	0.8723	0.9742	0.8427	0.5804	0.7460	0.9104	0.9600
FixMatch	0.8824	0.9639	0.8553	0.5774	0.7269	0.9123	0.9669
SsCL	0.8546	0.9866	0.8481	0.5800	0.7300	0.9115	0.9574
CCSSL	0.8663	0.9637	0.8352	0.5818	0.7270	0.9029	0.9456
SimMatch	**0.8881**	0.9759	0.8435	0.6004	0.7306	0.9013	0.9646
SimMatch-V2	0.8568	0.9638	0.8270	0.5886	**0.7526**	0.9185	0.9613
CLOP (this paper)	0.8792	**0.9989**	**0.8809**	0.6267	0.7385	**0.9331**	**0.9718**

Table 5: Transfer learning results for classification tasks (pretrained on ImageNet). Numbers are mean-per-class accuracy for Caltech and Flowers; and top-1 accuracy for all other datasets.

results for CIFAR-100 and ImageNet are presented in Table 2 and Table 3, respectively. Our findings indicate that CLOP consistently outperforms all competing methods across all experimental settings. Notably, on CIFAR-100, CLOP achieves the highest Top-1 accuracy of 0.743 with only 10% of labels and maintains a strong lead at 0.760 with 50% labels, while also significantly improving Top-5 performance. On ImageNet, CLOP achieved a Top-1 accuracy of 0.791 at 10% label availability and 0.829 at 50%, outperforming the next best methods by margins of over 4% in some cases.

For imbalanced-class training, we generate class-wise sample ratios by sampling from a uniform distribution in the range of $[0.001, 1]$, while maintaining all other experimental settings identical to the balanced-class training setup. The results are presented in Table 4, demonstrating that CLOP significantly outperforms all benchmark methods.

Method	Birdsnap	Cars	Aircraft	VOC2007	Pets
SimCLR	0.7590	0.9130	0.8785	0.8410	0.8920
SupCon	0.7515	0.9169	0.8409	0.8517	0.9347
FixMatch	0.7545	0.9004	0.8462	0.8372	**0.9515**
SsCL	0.7343	0.9089	0.8341	0.8504	0.9288
CCSSL	0.7613	**0.9247**	0.8528	0.8580	0.9471
SimMatch	0.7562	0.9127	0.8269	0.8393	0.9200
SimMatchV2	0.7516	0.9186	0.8453	0.8546	0.9494
CLOP	**0.7794**	0.9171	**0.8810**	**0.8646**	0.8982

Table 6: Transfer learning results for objects detection task (pretrained on ImageNet). Numbers are mAP for VOC2007; mean-per-class accuracy for Aircraft and Pets; and top-1 accuracy for Birdsnap.

Statistic	Non-Orthonormal Prototypes	Orthonormal Prototypes
Mean	0.768	**0.780**
Median	0.762	**0.781**
Std	0.015	**0.009**
Lower Quantile	0.760	**0.775**
Upper Quantile	0.775	**0.786**

Table 7: Ablation study on the effect of prototype initialization. We compare non-orthonormal and orthonormal initialization over 10 independent runs on CIFAR-100 under the full label setting. Results report the distribution statistics of Top-1 classification accuracy.

5.2 Transfer Learning on Image Classification and Object Detection

To assess CLOP's generalization capability on unseen datasets, we conduct transfer learning experiments on both image classification and object detection tasks. Specifically, we first pretrain the ResNet-50 (4x) backbone using various contrastive loss functions on ImageNet with all labels. Subsequently, we replace the projection head with either a one-layer prediction head or a two-layer object detection head with ReLU activation and fine-tune the network on the nature image datasets, including Food [3], CIFAR-10 [25], CIFAR-100 [26], SUN397 [47], DTD [34], Caltech-101 [10], Flowers [31], Birdsnap [2], Cars [54], Aircraft [30], VOC2007 [9], Pets [32]. For image classification tasks, we report the accuracy results in Table 5. Following the standard evaluation metrics in [4, 23], we present mean-per-class accuracy for Caltech and Flowers, while reporting top-1 accuracy for all other datasets. The results indicate that CLOP generally outperforms competing methods. Although SimMatch and SimMatch-V2 achieve slightly higher accuracy on the Food and DTD datasets, the performance gain is less than 2%. In contrast, CLOP surpasses these methods on the remaining four datasets, with the most significant improvement exceeding 3.5% on CIFAR-100. For object detection tasks, we report mean average precision (mAP) for VOC2007, mean-per-class accuracy for Aircraft and Pets, and top-1 accuracy for Birdsnap. The results, presented in Table 6, demonstrate that CLOP outperforms competing methods on three out of five datasets.

5.3 Ablation Studies

In this section, we present the experimental results for image classification, conducted with various batch sizes and learning rates on the CIFAR-100 and ImageNet datasets. For baseline methods, we implement the InfoNCE [46] with a supervised linear classifier for semi-supervised learning and the SupCon [23] for fully-supervised learning. All experiments are performed using the SimCLR [4] framework with ResNet-50 (4x) [15].

Effect of Prototype Initialization. To assess the impact of prototype initialization in our framework, we conduct an ablation study comparing non-orthonormal prototypes with orthonormal prototypes on CIFAR-100 under full supervision. As shown in Table 7, initializing prototypes with orthonormal vectors consistently yields better performance across all statistical measures. Specifically, orthonormal initialization improves the mean top-1 accuracy from 0.768 to 0.780 and reduces the standard deviation from 0.015 to 0.009, indicating both improved performance and training stability.

Orthonormal (CLOP) vs ETF prototypes. To ensure a fair comparison with ETF, we also evaluate performance using ETF prototypes as an alternative to the orthonormal prototypes. For fully-supervised learning, we utilize all labels in the training datasets for both SupCon and CLOP. In the semi-supervised setting, we employ 10% of the labeled data for both linear classifier and CLOP training. We report both top-1 classification accuracy on ImageNet in Figure 2 and accuracies on CIFAR100 and Tiny-ImageNet in Appendix C using the supervised linear classifier.

CLOP Prevents Collapse with Large Learning Rates. We trained models with learning rates ranging from 0.1 to 10 on CIFAR-100 and ImageNet for 200 epochs and Tiny-ImageNet for 100 epochs, using a batch size of 1024. The corresponding classification accuracies are presented in Figure 2 and 3. Across both datasets, CLOP consistently outperforms the baseline methods. Moreover, as demonstrated by Section 3, excessively large learning rates can lead to complete collapse, as clearly observed in the baseline methods at a learning rate of 10 on both datasets. However, with

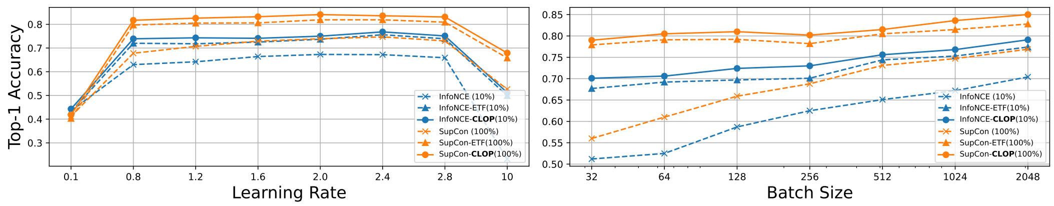


Figure 2: Top-1 classification accuracy on ImageNet across different learning rates and batch sizes. The percentage of labels used for supervised training is indicated in the legend.

Table 8: Ablation studies on λ, similarity metric for CLOP, and augmentation strategies.

(a) Acc of different λ.

λ	Top-1	Top-5
0.1	0.745	0.935
0.5	0.740	0.931
1.0	0.754	**0.938**
1.5	**0.760**	0.937

(b) Acc of different similarity metric.

Similarity Metric	Top-1	Top-5
Cosine	**0.754**	**0.938**
Euclidean	0.749	0.933
Manhattan	0.715	0.899

(c) Acc of augmentation strategies.

Augmentation	Top-1	Top-5
AutoAugment	0.625	0.847
SimCLR	**0.754**	**0.938**
RandAugment	0.726	0.899

the incorporation of CLOP into the loss function, we observe a significantly smaller performance degradation on both datasets.

CLOP Enables Smaller Batch Sizes. We trained models with batch sizes of 32, 64, 128, 256, 512, 1024, and 2048 on CIFAR-100 and ImageNet for 200 epochs and on Tiny-ImageNet for 100 epochs. The learning rate was fixed at $(0.3 \times \text{batch size}/256)$ for optimal performance. The corresponding classification accuracies are presented in Figure 2 and 4. CLOP consistently outperformed the baseline methods across all batch sizes. As reported in the original papers [4, 23], contrastive learning performs optimally when the batch size exceeds 1024, a finding corroborated by our experiments. However, with the addition of CLOP, we observe significantly less performance degradation at smaller batch sizes. Remarkably, CLOP achieved similar accuracy with a batch size of 32 compared to the baseline SupCon with a batch size of 2048 for CIFAR-100.

Tuning λ. To evaluate the sensitivity of the tuning parameter λ in CLOP, we trained the model with SupCon loss across different λ values, keeping the batch size fixed at 1024. The classification accuracy on both CIFAR-100 is reported in Table 8(a). We observe that the performance remains stable for λ values ranging from 0.1 to 1.5, with $\lambda = 1.0$ and $\lambda = 1.5$ yielding the best overall performance.

Choice of Similarity Metric. To evaluate the impact of different similarity functions on Eq. (2), we trained the same ResNet-50 (4x) architecture on CIFAR-100 using cosine similarity, Euclidean similarity, and Manhattan similarity. The results, presented in Table 8(b), indicate that cosine similarity, which aligns with $\mathcal{L}_{CL}$ in Eq. (2), achieves the highest performance.

Augmentation Strategies. To evaluate the impact of augmentation strategies on CLOP, we trained the same ResNet-50 (4x) model on CIFAR-100 with a batch size of 1024. We selected three commonly used augmentation methods: 1) RandAugment: Augmentation with three operations randomly chosen from all image processing functions in PyTorch (e.g., padding, resizing, cropping, rotation, color jitter, Gaussian blur, inversion, contrast adjustment, equalization); 2) AutoAugment using the ImageNet policy proposed in [6]; 3) SimCLR Augmentation Policy from [4]. The results are shown in Table 8(c), indicate that SimCLR augmentation works best with CLOP.

6 Conclusion and Limitations

In this paper, we addressed the issue of dimensional collapse in contrastive learning by analyzing the impact of learning rates on representation stability. We identified a critical threshold beyond which standard contrastive objectives, particularly those based on cosine similarity, converge to degenerate low-rank solutions. Motivated by this observation, we proposed **CLOP**, a novel semi-supervised loss function that encourages embeddings to align with a set of orthonormal prototypes. By explicitly promoting the formation of class-specific orthogonal subspaces, CLOP enhances the expressiveness and separability of learned representations. Empirical results across a range of classification and object detection benchmarks validate the effectiveness of CLOP. The method consistently outperforms strong baselines under both balanced and imbalanced label regimes, and demonstrates robustness across varying learning rates and batch sizes.

Despite these promising results, CLOP assumes a fixed number of well-separated classes and relies on static prototype initialization, which may limit its applicability in fine-grained or evolving-label scenarios. Future work could explore adaptive prototype mechanisms and extend the approach to tasks involving hierarchical or multi-label structures.

References

[1] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

[2] Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L Alexander, David W Jacobs, and Peter N Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2011–2018, 2014.

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, pages 446–461. Springer, 2014.

[4] 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.

[5] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *Advances in neural information processing systems*, 33:8765–8775, 2020.

[6] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018.

[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.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[9] Mark Everingham. The pascal visual object classes challenge,(voc2007) results. *http://pascallin. ecs. soton. ac. uk/challenges/VOC/voc2007/index. html.*, 2007.

[10] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. Technical Report CNS-TR-2004-004, California Institute of Technology, 2004.

[11] Daniel Y Fu, Mayee F Chen, Michael Zhang, Kayvon Fatahalian, and Christopher Ré. The details matter: Preventing class collapse in supervised contrastive learning. In *Computer Sciences & Mathematics Forum*, volume 3, page 4. MDPI, 2022.

[12] Songwei Ge, Shlok Mishra, Simon Kornblith, Chun-Liang Li, and David Jacobs. Hyperbolic contrastive learning for visual representations beyond objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6840–6849, 2023.

[13] Jaidev Gill, Vala Vakilian, and Christos Thrampoulidis. Engineering the neural collapse geometry of supervised-contrastive loss. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7115–7119. IEEE, 2024.

[14] Jamshid Hassanpour, Vinkle Srivastav, Didier Mutter, and Nicolas Padoy. Overcoming dimensional collapse in self-supervised contrastive learning for medical image segmentation. *arXiv preprint arXiv:2402.14611*, 2024.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International conference on machine learning*, pages 4182–4192. PMLR, 2020.

[17] Xinrong Hu, Dewen Zeng, Xiaowei Xu, and Yiyu Shi. Semi-supervised contrastive learning for label-efficient medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, pages 481–490. Springer, 2021.

[18] Tri Huynh, Simon Kornblith, Matthew R Walter, Michael Maire, and Maryam Khademi. Boosting contrastive self-supervised learning with false negative cancellation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2785–2795, 2022.

[19] Nakamasa Inoue and Keita Goto. Semi-supervised contrastive learning with generalized contrastive loss and its application to speaker recognition. In *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1641–1646. IEEE, 2020.

[20] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020.

[21] Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*, 2021.

[22] Se-Ryong Kang, Woncheol Shin, Su Yang, Jo-Eun Kim, Kyung-Hoe Huh, Sam-Sun Lee, Min-Suk Heo, and Won-Jin Yi. Structure-preserving quality improvement of cone beam ct images using contrastive learning. *Computers in Biology and Medicine*, 158:106803, 2023.

[23] 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.

[24] Byoungjip Kim, Jinho Choo, Yeong-Dae Kwon, Seongho Joe, Seungjai Min, and Youngjune Gwon. Selfmatch: Combining contrastive self-supervision and consistency for semi-supervised learning. *arXiv preprint arXiv:2101.06480*, 2021.

[25] Alex Krizhevsky, Geoff Hinton, et al. Convolutional deep belief networks on cifar-10. *Unpublished manuscript*, 40(7):1–9, 2010.

[26] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[27] Huanran Li and Daniel Pimentel-Alarcón. Transfusion: Contrastive learning with transformers. *arXiv preprint arXiv:2403.18681*, 2024.

[28] Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020.

[29] 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.

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

[31] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008.

[32] Luis Patino, Tom Cane, Alain Vallee, and James Ferryman. Pets 2016: Dataset and challenge. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1–8, 2016.

[33] Ralph Peeters and Christian Bizer. Supervised contrastive learning for product matching. In *Companion Proceedings of the Web Conference 2022*, pages 248–251, 2022.

[34] Chenfan Qu, Chongyu Liu, Yuliang Liu, Xinhong Chen, Dezhi Peng, Fengjun Guo, and Lianwen Jin. Towards robust tampered text detection in document image: New dataset and new solution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5937–5946, 2023.

[35] Evgenia Rusak, Patrik Reizinger, Roland S Zimmermann, Oliver Bringmann, and Wieland Brendel. Content suppresses style: dimensionality collapse in contrastive learning. In *NeurIPS 2022 Workshop: Self-Supervised Learning-Theory and Practice*, volume 3, 2022.

[36] Xiang Shen, Yinge Sun, Yao Zhang, and Mani Najmabadi. Semi-supervised intent discovery with contrastive learning. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 120–129, 2021.

[37] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

[38] Li Tao, Xueting Wang, and Toshihiko Yamasaki. An improved inter-intra contrastive learning framework on self-supervised video representation. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(8):5266–5280, 2022.

[39] Yang Tao, Kai Guo, Yizhen Zheng, Shirui Pan, Xiaofeng Cao, and Yi Chang. Breaking the curse of dimensional collapse in graph contrastive learning: A whitening perspective. *Information Sciences*, 657:119952, 2024.

[40] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020.

[41] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020.

[42] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[43] Haoqing Wang, Xun Guo, Zhi-Hong Deng, and Yan Lu. Rethinking minimal sufficient representation in contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16041–16050, 2022.

[44] Yuyang Wang, Rishikesh Magar, Chen Liang, and Amir Barati Farimani. Improving molecular contrastive learning via faulty negative mitigation and decomposed fragment contrast. *Journal of Chemical Information and Modeling*, 62(11):2713–2725, 2022.

[45] Haiping Wu and Xiaolong Wang. Contrastive learning of image representations with cross-video cycle-consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10149–10159, 2021.

[46] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.

[47] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010.

[48] Teng Xiao, Huaisheng Zhu, Zhengyu Chen, and Suhang Wang. Simple and asymmetric graph contrastive learning without augmentations. *Advances in Neural Information Processing Systems*, 36, 2024.

[49] Tete Xiao, Xiaolong Wang, Alexei A Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. *arXiv preprint arXiv:2008.05659*, 2020.

[50] Lingling Xu and Weiming Wang. Improving aspect-based sentiment analysis with contrastive learning. *Natural Language Processing Journal*, 3:100009, 2023.

[51] 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 *International Conference on Machine Learning*, pages 38938–38970. PMLR, 2023.

[52] Chun Yang, Jianxiao Zou, JianHua Wu, Hongbing Xu, and Shicai Fan. Supervised contrastive learning for recommendation. *Knowledge-Based Systems*, 258:109973, 2022.

[53] Fan Yang, Kai Wu, Shuyi Zhang, Guannan Jiang, Yong Liu, Feng Zheng, Wei Zhang, Chengjie Wang, and Long Zeng. Class-aware contrastive semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14421–14430, 2022.

[54] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3973–3981, 2015.

[55] Chun-Hsiao Yeh, Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu, Yubei Chen, and Yann LeCun. Decoupled contrastive learning. In *European Conference on Computer Vision*, pages 668–684. Springer, 2022.

[56] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.

[57] Yuhang Zhang, Xiaopeng Zhang, Jie Li, Robert C Qiu, Haohang Xu, and Qi Tian. Semi-supervised contrastive learning with similarity co-calibration. *IEEE Transactions on Multimedia*, 25:1749–1759, 2022.

[58] Mingkai Zheng, Shan You, Lang Huang, Chen Luo, Fei Wang, Chen Qian, and Chang Xu. Simmatchv2: Semi-supervised learning with graph consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16432–16442, 2023.

[59] Mingkai Zheng, Shan You, Lang Huang, Fei Wang, Chen Qian, and Chang Xu. Simmatch: Semi-supervised learning with similarity matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14471–14481, 2022.

[60] 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.

A Proof of Lemma 1

Proof of Lemma 1. Consider $\mathcal{L}_i$ as the i-th loss term of $\mathcal{L}_{\text{InfoNCE}}$, defined by the following expression:

$$\mathcal{L}_i := -\log \mathbb{P}_i$$

where $\mathbb{P}_i$ denotes the probability that i-th embedding choose its positive pair as closest neighbor:

$$\mathbb{P}_i := \frac{\exp(\mathbf{z}_i^\top \mathbf{z}_{j(i)}/\tau)}{\exp(\mathbf{z}_i^\top \mathbf{z}_{j(i)}/\tau) + \sum_{a \notin \{i,j(i)\}} \exp(\mathbf{z}_i^\top \mathbf{z}_a/\tau)}$$

As detailed in [55], the gradient of $\mathcal{L}_i$ with respect to $\mathbf{z}_i$, $\mathbf{z}_{j(i)}$, and $\mathbf{z}_a$ can be derived as follows:

$$-\frac{\partial \mathcal{L}_i}{\partial \mathbf{z}_i} := (1-\mathbb{P}_i)/\tau \left(\mathbf{z}_{j(i)} - \sum_{a \notin \{i,j(i)\}} \frac{\exp(\mathbf{z}_i^\top \mathbf{z}_a/\tau)}{\sum_{b \notin \{i,j(i)\}} \exp(\mathbf{z}_i^\top \mathbf{z}_b/\tau)} \mathbf{z}_a \right)$$

$$-\frac{\partial \mathcal{L}_i}{\partial \mathbf{z}_{j(i)}} := \frac{(1-\mathbb{P}_i)}{\tau} \mathbf{z}_i$$

$$-\frac{\partial \mathcal{L}_i}{\partial \mathbf{z}_a} := -\frac{(1-\mathbb{P}_i)}{\tau} \frac{\exp(\mathbf{z}_i^\top \mathbf{z}_a/\tau)}{\sum_{b \notin \{i,j(i)\}} \exp(\mathbf{z}_i^\top \mathbf{z}_b/\tau)} \mathbf{z}_i$$

In the standard setup of self-supervised learning, for any sample, there is one positive pair among I and the remainder are all negative pairs. By aggregating all the gradient respect to a single sample, we have the gradient of InfoNCE respect to $\mathbf{z}_i$:

$$-\frac{\partial \mathcal{L}_{\text{InfoNCE}}}{\partial \mathbf{z}_i} := \frac{(1-\mathbb{P}_i) + (1-\mathbb{P}_{j(i)})}{\tau} \mathbf{z}_{j(i)} - \sum_{a \notin \{i,j(i)\}} \frac{(1-\mathbb{P}_i)}{\tau} \frac{\exp(\mathbf{z}_i^\top \mathbf{z}_a/\tau)}{\sum_{b \notin \{i,j(i)\}} \exp(\mathbf{z}_i^\top \mathbf{z}_b/\tau)} \mathbf{z}_a - \sum_{a \notin \{i,j(i)\}} \frac{(1-\mathbb{P}_a)}{\tau} \frac{\exp(\mathbf{z}_i^\top \mathbf{z}_a/\tau)}{\sum_{b \notin \{a,j(a)\}} \exp(\mathbf{z}_a^\top \mathbf{z}_b/\tau)} \mathbf{z}_a$$

Now, considering the first scenario, where all embeddings equal, that means that $\mathbf{z}_i = \mathbf{z}_{j(i)} = \mathbf{z}_a = \mathbf{z}^*$ for all $a \in I$, the loss terms $\mathbb{P}_i$, $\mathbb{P}_{j(i)}$, and $\mathbb{P}_a$ converge to a constant $\mathbb{P}^*$, given by:

$$\mathbb{P}_i = \mathbb{P}_{j(i)} = \mathbb{P}_a = -\log \frac{1}{|I|-1} := \mathbb{P}^*$$

Consequently, the gradient of $\mathcal{L}_{\text{InfoNCE}}$ with respect to $\mathbf{z}_i$ under this assumption reduces to zero, aligning with our expectations:

$$-\frac{\partial \mathcal{L}_{\text{InfoNCE}}}{\partial \mathbf{z}_i} = \frac{2(1-\mathbb{P}^*)}{\tau} \mathbf{z}^* - 2(|I|-2) \frac{(1-\mathbb{P}^*)}{\tau} \frac{1}{|I|-2} \mathbf{z}^* = 0$$

We establish the existence of local minima in scenarios where all embeddings are identical. Now, we consider the second scenario where all embeddings generated reside within the same rank-1 subspace. Denoting $\mathbf{z}^*$ as their unit basis, we can represent each embedding $\mathbf{z}_i$ as:

$$\mathbf{z}_i = \alpha \mathbf{z}^*, \quad \alpha \in \{-1, 1\}, \quad \forall i$$

The gradient of the loss function $\mathcal{L}_{\text{InfoNCE}}$ with respect to $\mathbf{z}_i$ simplifies to:

$$-\frac{\partial \mathcal{L}}{\partial \mathbf{z}_i} = \beta \mathbf{z}_i$$

Here, β is a scalar that aggregates contributions from all relevant weights.

It is important to note that $\mathbf{z}^i$ represents the normalized output of the function f, with $\mathbf{x}_i$ denoting the original, unnormalized embedding. This implies the following relation:

$$-\frac{\partial \mathcal{L}}{\partial \mathbf{x}_i} = -\frac{\partial \mathcal{L}}{\partial \mathbf{z}_i} \frac{\partial \mathbf{z}_i}{\partial \mathbf{x}_i} = \frac{1}{\|\mathbf{x}_i\|_2} \left(\mathbb{I} - \frac{\mathbf{x}_i \mathbf{x}_i^\top}{\|\mathbf{x}_i\|_2^2} \right) \beta \mathbf{z}_i = \frac{\beta}{\|\mathbf{x}_i\|_2} \left(\mathbf{z}_i - \frac{\mathbf{x}_i}{\|\mathbf{x}_i\|_2} \right) = 0,$$

where $\mathbb{I}$ represents the identity matrix.

□

B Gradient Analysis of Repulsive Force from InfoNCE Loss

Recall that

$$\mathcal{L}_{\text{infoNCE}} = -\sum_{i \in I} \log \frac{\exp(\mathbf{z}_i^\top \mathbf{z}_{j(i)}/\tau)}{\sum_{a \neq i} \exp(\mathbf{z}_i^\top \mathbf{z}_a/\tau)}$$

Assume that the model can successfully merge the positive pair into the same embedding, the loss of repulsive becomes:

$$\mathcal{L}_{\text{repulsive}} = -\sum_{i \in I} \log \frac{\exp(1/\tau)}{\sum_{a \neq i, j(i)} \exp(\mathbf{z}_i^\top \mathbf{z}_a/\tau) + \exp(1/\tau)}$$

We start by rewriting the loss for a given sample i:

$$\ell_i = -\log \frac{\exp(1/\tau)}{\exp(1/\tau) + \sum_{a \in \mathcal{N}(i)} \exp\Big(\mathbf{z}_i^\top \mathbf{z}_a/\tau\Big)},$$

where we denote

$$\mathcal{N}(i) := \{\, a : a \neq i \text{ and } a \neq j(i) \,\}.$$

Define the denominator

$$D_i := \exp(1/\tau) + \sum_{a \in \mathcal{N}(i)} \exp\Big(\mathbf{z}_i^\top \mathbf{z}_a/\tau\Big).$$

Then

$$\ell_i = \log\Big[D_i\Big] - \frac{1}{\tau}.$$

Define $\mathbb{P}_{ia}$ as the probability that sample i choose sample a as closest neighbor,

$$\mathbb{P}_{ia} := \frac{\exp\Big(\mathbf{z}_i^\top \mathbf{z}_a/\tau\Big)}{\exp(1/\tau) + \sum_{a \in \mathcal{N}(i)} \exp\Big(\mathbf{z}_i^\top \mathbf{z}_a/\tau\Big)} = \frac{\exp\Big(\mathbf{z}_i^\top \mathbf{z}_a/\tau\Big)}{D_i}$$

Taking the gradient of ℓ_i with respect to $\mathbf{z}_i$,

$$\frac{\partial}{\partial \mathbf{z}_i} \exp\Big(\mathbf{z}_i^\top \mathbf{z}_a/\tau\Big) = \frac{1}{\tau} \exp\Big(\mathbf{z}_i^\top \mathbf{z}_a/\tau\Big)\mathbf{z}_a.$$

Thus, differentiating the log term gives:

$$\frac{\partial \ell_i}{\partial \mathbf{z}_i} = \frac{1}{D_i} \sum_{a \in \mathcal{N}(i)} \frac{1}{\tau} \exp\Big(\mathbf{z}_i^\top \mathbf{z}_a/\tau\Big)\mathbf{z}_a.$$

Rewriting in terms of the softmax probabilities,

$$\frac{\partial \ell_i}{\partial \mathbf{z}_i} = \frac{1}{\tau} \sum_{a \in \mathcal{N}(i)} \mathbb{P}_{ia}\, \mathbf{z}_a,$$

Taking the gradient of ℓ_i with respect to $\mathbf{z}_a$,

$$\frac{\partial}{\partial \mathbf{z}_a} \exp\Big(\mathbf{z}_i^\top \mathbf{z}_a/\tau\Big) = \frac{1}{\tau} \exp\Big(\mathbf{z}_i^\top \mathbf{z}_a/\tau\Big)\mathbf{z}_i.$$

Thus, differentiating the log term gives:

$$\frac{\partial \ell_i}{\partial \mathbf{z}_a} = \frac{1}{D_i} \frac{1}{\tau} \exp\Big(\mathbf{z}_i^\top \mathbf{z}_a/\tau\Big)\mathbf{z}_i.$$

Rewriting in terms of the softmax probabilities,

$$\frac{\partial \ell_i}{\partial \mathbf{z}_a} = \frac{1}{\tau} \mathbb{P}_{ia}\, \mathbf{z}_i.$$

Therefore, the gradient of $\mathcal{L}_{\text{repulsive}}$ is:

$$\frac{\partial \mathcal{L}_{\text{repulsive}}}{\partial \mathbf{z}_i} = \frac{\partial \ell_i}{\partial \mathbf{z}_i} + \sum_{a \in \mathcal{N}(i)} \frac{\partial \ell_a}{\partial \mathbf{z}_i} = \frac{1}{\tau} \sum_{a \in \mathcal{N}(i)} \mathbb{P}_{ia}\, \mathbf{z}_a + \frac{1}{\tau} \sum_{a \in \mathcal{N}(i)} \mathbb{P}_{ai}\, \mathbf{z}_a = \frac{1}{\tau} \sum_{a \in \mathcal{N}(i)} \left(\mathbb{P}_{ia} + \mathbb{P}_{ai}\right) \mathbf{z}_a$$

Since the probability matrix $\mathbb{P}$ is obviously symmetric,

$$\frac{\partial \mathcal{L}_{\text{repulsive}}}{\partial \mathbf{z}_i} = \frac{2}{\tau} \sum_{a \in \mathcal{N}(i)} \mathbb{P}_{ia} \mathbf{z}_a$$

Thus, for the gradient of the before the normalization,

$$\frac{\partial \mathcal{L}}{\partial \mathbf{x}_i} = \frac{\partial \mathcal{L}}{\partial \mathbf{z}_i} \frac{\partial \mathbf{z}_i}{\partial \mathbf{x}_i} = \frac{1}{\|\mathbf{x}_i\|_2} \left(\mathbb{I} - \frac{\mathbf{x}_i(\mathbf{x}_i)^\top}{\|\mathbf{x}_i\|_2^2}\right) \frac{2}{\tau} \sum_{a \in \mathcal{N}(i)} \mathbb{P}_{ia} \mathbf{z}_a = \frac{1}{\|\mathbf{x}_i\|_2} \left(\mathbb{I} - \mathbf{z}_i(\mathbf{z}_i)^\top\right) \frac{2}{\tau} \sum_{a \in \mathcal{N}(i)} \mathbb{P}_{ia} \mathbf{z}_a$$

Given the first iteration of points as $\{\mathbf{x}_i^0\}$, then, after the first iteration of descent, the next embeddings are:

$$\mathbf{x}_i^1 = \mathbf{x}_i^0 - \eta \frac{1}{\|\mathbf{x}_i^0\|_2} \left(\mathbb{I} - \mathbf{z}_i^0(\mathbf{z}_i^0)^\top\right) \frac{2}{\tau} \sum_{a \in \mathcal{N}(i)} \mathbb{P}_{ia}\, \mathbf{z}_a^0$$

Thus, the mean of $\mathbf{x}_i^1$ can be calculated as:

$$\boldsymbol{\mu}^1 = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i^1 = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i^0 - \frac{2\eta}{N\tau} \sum_{i=1}^N \sum_{a \in \mathcal{N}(i)} \frac{1}{\|\mathbf{x}_i^0\|_2} \mathbb{P}_{ia} \left(\mathbb{I} - \mathbf{z}_i^0(\mathbf{z}_i^0)^\top\right) \mathbf{z}_a^0$$

Since the negative pair relationship is symmetric, ie. $a \in \mathcal{N}(i) \iff i \in \mathcal{N}(a)$, then

$$\sum_{i=1}^N \sum_{a \in \mathcal{N}(i)} \mathbb{P}_{ia} \frac{1}{\|\mathbf{x}_i^0\|_2} \left(\mathbb{I} - \mathbf{z}_i^0(\mathbf{z}_i^0)^\top\right) \mathbf{z}_a^0 = \sum_{a=1}^N \sum_{i \in \mathcal{N}(a)} \mathbb{P}_{ia} \frac{1}{\|\mathbf{x}_i^0\|_2} \left(\mathbb{I} - \mathbf{z}_i^0(\mathbf{z}_i^0)^\top\right) \mathbf{z}_a^0$$

Thus,

$$\begin{aligned}
\boldsymbol{\mu}^1 &= \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i^0 - \frac{2\eta}{N\tau} \sum_{a=1}^N \sum_{i \in \mathcal{N}(a)} \mathbb{P}_{ia} \frac{1}{\|\mathbf{x}_i^0\|_2} \left(\mathbb{I} - \mathbf{z}_i^0(\mathbf{z}_i^0)^\top\right) \mathbf{z}_a^0 \\
&= \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i^0 - \frac{2\eta}{N\tau} \sum_{i=1}^N \sum_{a \in \mathcal{N}(i)} \frac{\mathbb{P}_{ia}}{\|\mathbf{x}_a^0\|_2} \left(\mathbb{I} - \mathbf{z}_a^0(\mathbf{z}_a^0)^\top\right) \mathbf{z}_i^0 \\
&= \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i^0 - \frac{2\eta}{N\tau} \sum_{i=1}^N \sum_{a \in \mathcal{N}(i)} \frac{\mathbb{P}_{ia}}{\|\mathbf{x}_a^0\|_2} \mathbf{z}_i^0 + \frac{2\eta}{N\tau} \sum_{i=1}^N \sum_{a \in \mathcal{N}(i)} \frac{\mathbb{P}_{ia}}{\|\mathbf{x}_a^0\|_2} \mathbf{z}_a^0(\mathbf{z}_a^0)^\top \mathbf{z}_i^0 \\
&= \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i^0 - \frac{2\eta}{N\tau} \sum_{i=1}^N \sum_{a \in \mathcal{N}(i)} \frac{\mathbb{P}_{ia}}{\|\mathbf{x}_a^0\|_2} \mathbf{z}_i^0 + \frac{2\eta}{N\tau} \sum_{i=1}^N \sum_{a \in \mathcal{N}(i)} \frac{\mathbb{P}_{ia}}{\|\mathbf{x}_i^0\|_2} \mathbf{z}_i^0(\mathbf{z}_i^0)^\top \mathbf{z}_a^0 \\
&= \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i^0 - \frac{2\eta}{N\tau} \sum_{i=1}^N \sum_{a \in \mathcal{N}(i)} \mathbb{P}_{ia} \frac{1}{\|\mathbf{x}_a^0\|_2} \mathbf{z}_i^0 + \frac{2\eta}{N\tau} \sum_{i=1}^N \sum_{a \in \mathcal{N}(i)} \mathbb{P}_{ia} \frac{(\mathbf{z}_i^0)^\top \mathbf{z}_a^0}{\|\mathbf{x}_i^0\|_2} \mathbf{z}_i^0 \\
&= \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i^0 - \frac{2\eta}{N\tau} \sum_{i=1}^N \sum_{a \in \mathcal{N}(i)} \mathbb{P}_{ia} \left(\frac{1}{\|\mathbf{x}_a^0\|_2} - \frac{(\mathbf{z}_i^0)^\top \mathbf{z}_a^0}{\|\mathbf{x}_i^0\|_2}\right) \frac{1}{\|\mathbf{x}_i^0\|_2} \mathbf{x}_i^0 \\
&= \frac{1}{N} \sum_{i=1}^N \left[\left(1 - \frac{2\eta}{\tau}\right) \frac{1}{\|\mathbf{x}_i^0\|_2} \sum_{a \in \mathcal{N}(i)} \mathbb{P}_{ia} \left(\frac{1}{\|\mathbf{x}_a^0\|_2} - \frac{(\mathbf{z}_i^0)^\top \mathbf{z}_a^0}{\|\mathbf{x}_i^0\|_2}\right)\right] \mathbf{x}_i^0 \\
&= \left(1 - \frac{2\eta}{\tau}\right) \frac{1}{N} \sum_{i=1}^N \left[\frac{1}{\|\mathbf{x}_i^0\|_2} \sum_{a \in \mathcal{N}(i)} \mathbb{P}_{ia} \left(\frac{1}{\|\mathbf{x}_a^0\|_2} - \frac{(\mathbf{z}_i^0)^\top \mathbf{z}_a^0}{\|\mathbf{x}_i^0\|_2}\right)\right] \mathbf{x}_i^0
\end{aligned}$$

Since $-1 \leq \mathbf{z}_i^\top \mathbf{z}_a \leq 1$, Denote

$$\begin{aligned}\beta_{ia} &:= \left[\frac{1}{\|\mathbf{x}_i^0\|_2} \sum_{a\in\mathcal{N}(i)} \mathbb{P}_{ia} \left(\frac{1}{\|\mathbf{x}_a^0\|_2} - \frac{(\mathbf{z}_i^0)^\top \mathbf{z}_a^0}{\|\mathbf{x}_i^0\|_2}\right)\right] \\ &\leq \frac{1}{\|\mathbf{x}_i^0\|_2} \sum_{a\in\mathcal{N}(i)} \mathbb{P}_{ia} \left(\frac{1}{\|\mathbf{x}_a^0\|_2} + \frac{1}{\|\mathbf{x}_i^0\|_2}\right) \\ &\leq \frac{2}{\min_j \|\mathbf{x}_j^0\|_2^2} \sum_{a\in\mathcal{N}(i)} \mathbb{P}_{ia}\end{aligned}$$

And we can lowerbound sum of probability by

$$\sum_{a\in\mathcal{N}(i)} \mathbb{P}_{ia} = \frac{\sum_{a\in\mathcal{N}(i)} \exp\Big((\mathbf{z}_i^0)^\top (\mathbf{z}_a^0)/\tau\Big)}{\exp(1/\tau) + \sum_{a\in\mathcal{N}(i)} \exp\Big((\mathbf{z}_i^0)^\top (\mathbf{z}_a^0)/\tau\Big)} \leq \frac{|\mathcal{N}| \exp(1/\tau)}{\exp(1/\tau) + |\mathcal{N}| \exp(1/\tau)} = \frac{|\mathcal{N}|}{1+|\mathcal{N}|}$$

Thus, denote $\sigma := \min_j \|\mathbf{x}_j^0\|_2$, we show that

$$\beta_{ia} \leq \frac{2}{\sigma^2} \frac{|\mathcal{N}|}{1+|\mathcal{N}|}$$

Therefore,

$$\|\boldsymbol{\mu}_1\|_2 \leq \left|\left(1 - \frac{2\eta}{\tau}\right) \frac{2}{\sigma^2} \frac{|\mathcal{N}|}{1+|\mathcal{N}|}\right| \|\boldsymbol{\mu}_0\|_2$$

If we want to limit the increase of mean, we want the coefficient to be less than 1, which means that

$$\begin{aligned}-\frac{\sigma^2(1+|\mathcal{N}|)}{2|\mathcal{N}|} &< 1 - \frac{2\eta}{\tau} < \frac{\sigma^2(1+|\mathcal{N}|)}{2|\mathcal{N}|} \\ -1 - \frac{\sigma^2(1+|\mathcal{N}|)}{2|\mathcal{N}|} &< -\frac{2\eta}{\tau} < -1 + \frac{\sigma^2(1+|\mathcal{N}|)}{2|\mathcal{N}|} \\ 1 + \frac{\sigma^2(1+|\mathcal{N}|)}{2|\mathcal{N}|} &> \frac{2\eta}{\tau} > 1 - \frac{\sigma^2(1+|\mathcal{N}|)}{2|\mathcal{N}|} \\ \frac{\tau}{2}\left(1 + \frac{\sigma^2(1+|\mathcal{N}|)}{2|\mathcal{N}|}\right) &> \eta > \frac{\tau}{2}\left(1 - \frac{\sigma^2(1+|\mathcal{N}|)}{2|\mathcal{N}|}\right)\end{aligned}$$

C Extra Experiments on CLOP

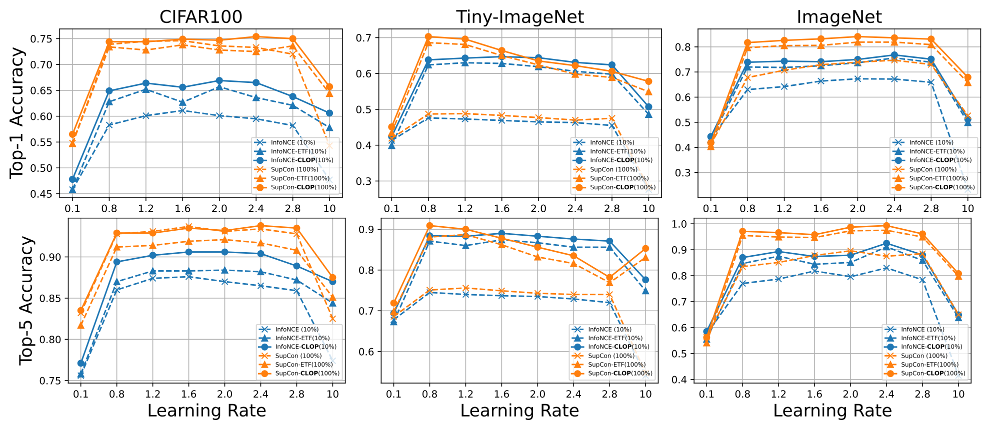


Figure 3: Top-1 classification accuracy across different learning rates. The percentage of labels used for supervised training is indicated in the legend.

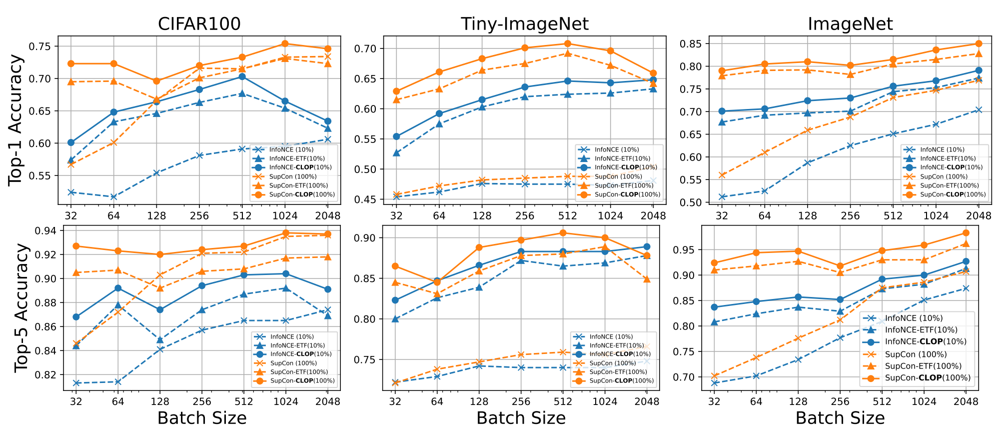


Figure 4: Top-1 classification accuracy across different batch sizes. The percentage of labels used for supervised training is indicated in the legend.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Refer to Abstract.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Refer to Section 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Refer to Section 3, where the range of the learning rate is derived under the assumption that positive sample pairs are successfully merged into the same embedding. Proof can be found in Appendix A and Appendix B.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The formulation is explained in Section 4, and the model's backbone structure is explained in Section 5.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Refer the the supplemental codes and readme file.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Refer to Section 5.3 for ablation studies on the choice of model configuration.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Refer to Section 5, Appendix C.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please refer to the last paragraph of Section 4. This paper adopts the SimCLR framework with minor modifications that do not introduce significant computational overhead or affect memory usage and execution time; see [4] for further information.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This research is conducted following the code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This paper aims to improve current semi-supervised learning accuracy in vision-related tasks. The authors do not anticipate any direct societal impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The release of the model does not have a high risk of misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Refer to Section 5.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The released code is well documented.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about answerNA (if any)?

    Answer: [NA]

    Justification: No crowdsourcing experiment is involved.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No human subjects are involved.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [No]

    Justification: No important, original, or non-standard component of the core methods in this research is involved with LLM usage.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.