Journal of Machine Learning Research 26 (2025) 1-42 Submitted 9/22; Revised 6/25; Published 10/25

An Augmentation Overlap Theory of Contrastive Learning

Qi Zhang [1] * ZHANGQ327@STU.PKU.EDU.CN
Yifei Wang [2] * YIFEI_W@MIT.EDU
Yisen Wang [1,3] † YISEN.WANG@PKU.EDU.CN

[1] *State Key Lab of General Artificial Intelligence, School of Intelligence Science and Technology, Peking University*
[2] *CSAIL, MIT*
[3] *Institute for Artificial Intelligence, Peking University*

Editor: Pierre Alquier

Abstract

Recently, self-supervised contrastive learning has achieved great success on various tasks. However, its underlying working mechanism is yet unclear. In this paper, we first provide the tightest bounds based on the widely adopted assumption of conditional independence. Further, we relax the conditional independence assumption to a more practical assumption of augmentation overlap and derive the asymptotically closed bounds for the downstream performance. Our proposed augmentation overlap theory hinges on the insight that the support of different intra-class samples will become more overlapped under aggressive data augmentations, thus simply aligning the positive samples (augmented views of the same sample) could make contrastive learning cluster intra-class samples together. Moreover, from the newly derived augmentation overlap perspective, we develop an unsupervised metric for the representation evaluation of contrastive learning, which aligns well with the downstream performance almost without relying on additional modules. Code is available at https://github.com/PKU-ML/GARC.

Keywords: Contrastive learning, Augmentation overlap, Theoretical understanding, Generalization, Representation evaluation

1. Introduction

Traditionally, supervised learning obtains representations by pulling together samples with the same labels (positive pairs) and pushing apart those with different labels (negative pairs). To improve the effectiveness of representation learning, early works found that the selection of positive and negative pairs plays a crucial role. This led to the development of algorithms such as triplet mining (Schroff et al., 2015) and hard negative sampling (Bucher et al., 2016), which aim to select more informative or challenging pairs from labeled data sets. However, obtaining such labels at scale can be costly. To remove this dependence on labeled data, recent research has focused on identifying surrogate signals for semantic similarity in an unsupervised manner. As a solution, self-supervised contrastive learning (van den Oord

∗. Equal Contribution
†. Corresponding Author

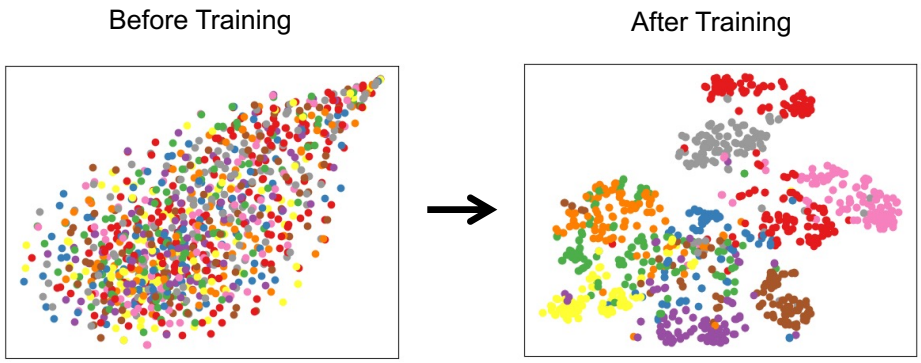
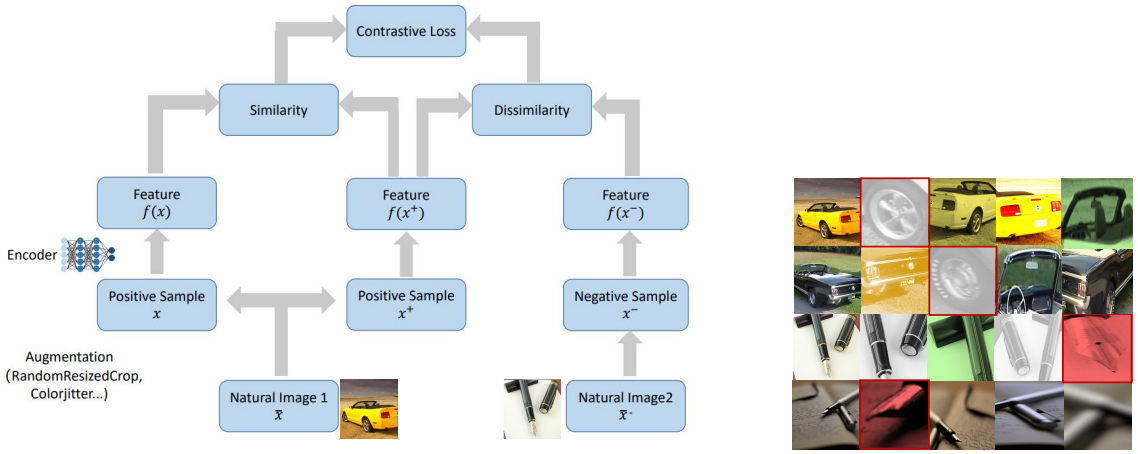


Figure 1: The t-SNE visualization of representations before and after contrastive learning method of SimCLR on CIFAR-10 data set. Each point denotes a sample and its color denotes its class.

et al., 2018) generates positive pairs by applying two independent augmentations to the same input sample, under the assumption that augmented views should preserve semantic identity. Negative samples are typically drawn from other instances in the batch, assuming they are semantically different. When the number of classes is large, the chance of sampling a semantically similar (false negative) instance is low (approximately $1/K$ for K classes), which makes this assumption practical. This technique enables the model to learn class-separated representations without access to labels, as shown in Figure 1, and has become a cornerstone of self-supervised learning. Modern frameworks (Chen et al., 2020; He et al., 2020; Grill et al., 2020; Wang et al., 2021; Oquab et al., 2023; Wang et al., 2023) have further refined this idea and now achieved performance comparable to supervised learning across a range of tasks. However, it is still unclear why contrastive learning can learn a meaningful representation for the downstream tasks and a convincing theoretical analysis is yet wanted.

The core idea of contrastive learning is to learn an encoder by closing the augmented views of the same anchor sample while simultaneously pushing away the augmented views of different anchor samples in the feature space. As shown in Figure 2(a), taking an image sample $\bar{x}$ as an example, contrastive learning first transforms it with data augmentations t, t^+ (e.g., RandomResizedCrop, ColorJitter, GaussianBlur, etc.) to generate a positive pair (x, x^+). Then, the two augmented samples are respectively encoded with the encoder f. In the next step, contrastive learning trains the encoder with the contrastive loss (e.g., InfoNCE loss (van den Oord et al., 2018)) to decrease the distance between positive samples in the feature space and push away other samples. Intuitively, while supervised learning clusters samples based on their labels, contrastive learning performs instance-level discrimination without access to label information. Nevertheless, as shown in Figure 1, when we compare the feature distributions before and after contrastive learning, we observe that the learned representations exhibit clear class separation. That is, samples from the same class are grouped together, whereas those from different classes are pushed apart—even without using any label supervision. Therefore, it is natural to raise several questions:



(a) Frameworks (b) Augmented views

Figure 2: (a) The framework of contrastive learning. (b) The augmented views of four images from ImageNet. The first two rows are cars while the bottom two rows are pens. The ancher samples are shown in the 1st column while the 2-5th columns present its corresponding augmented views.

Why the instance-level contrastive learning can lead to good performance on class-level downstream tasks? How to establish the theoretical relationship between (contrastive learning) pretraining and downstream performance?

Arora et al. (2019) established a theoretical guarantee between (contrastive learning) pretraining and downstream tasks based on the assumption that the augmented views of the same sample are conditionally independent on its label. However, as shown in Figure 2(b), the augmented views of the same sample are input-dependent and their assumption is quite hard to reach in practice. Parallel to that, Wang and Isola (2020) decomposed the objective of contrastive learning to two parts: the alignment of positive samples and the uniformity of negative samples. While we find that only the alignment and the uniformity are not enough to make contrastive learning learn useful representations for the downstream tasks. In this paper, we propose a counter-example (Proposition 5.3) by constructing special positive pairs and prove that the downstream performance can be as bad as a random guess even when the learned representations achieve perfect alignment and uniformity. It means that except for the alignment and uniformity, the selection of positive pairs, i.e., the design of data augmentations, is crucial for contrastive learning, which may be overlooked by the above two kinds of works. Revisiting Figure 2(b), we find that appropriate data augmentations can help intra-class samples generate quite similar augmented views (tires of different cars) while keeping the augmented views of inter-class samples distinguishable (cars and pens), which means that intra-class samples own the support overlap through these similar augmented views generated by data augmentations. As a result, when contrastive learning aligns the positive pairs, the intra-class samples that have overlapped views will be gathered as well.

Motivated by the above observations, we formulate the data augmentations as an augmentation graph to bridge different intra-class samples. Based on which, we further pro-

pose a theory of augmentation overlap to analyze how the intra-class samples are eventually aligned by these overlapped views in contrastive learning. By characterizing the property of the augmentation overlap, we establish sharper upper and lower bounds for the downstream performance of contrastive learning under more practical assumptions. Moreover, we design a theory guided model selection metric via measuring the degree of augmentation overlap, which is totally unsupervised without additional computational costs. The contents of the paper are organized as follows:

- In Section 2, we summarize the related work of understanding self-supervised learning and further introduce the problem setup in Section 3.

- In Section 4, we derive sharper bounds for the downstream performance of contrastive learning by analyzing the influence of negative samples. To focus on negative samples, we adopt the widely used conditional independence assumption on positive samples and show how to improve the error bound on negative samples from the Monte Carlo perspective. Our bounds are the tightest by far.

- In Section 5, we focus on the positive samples and discuss how to relax the conditional independence assumption to a more realistic augmentation overlap assumption, and propose a new theory of augmentation overlap to derive new upper and lower bounds for contrastive learning.

- In Section 6, we present a quantitative analysis on the effect of data augmentation strategy on the augmentation overlap both theoretically and empirically. In brief, strong augmentation strengths are necessary for bridging intra-class samples, which facilitates good downstream performance. However, too strong augmentations can lead to detrimental effects, such as the overlap of views between different classes.

- Lastly, in Section 7, as a proof-of-idea, we demonstrate a practical application of our augmentation overlap theory, that is, we develop a representation evaluation metric for automatic model selection without using supervised downstream data.

Remark. A preliminary work was published at ICLR 2022 (Wang et al., 2022), while we add substantially new results, both theoretically and empirically, in this manuscript, aiming to provide a more comprehensive and in-depth analysis of the generalization property of contrastive learning. Specifically, we add the following key results:

- In Section 4, we establish *sharper generalization bounds* under the conventional conditional independence assumption, which are superior to prior work and near-optimal.

- In Section 5, we generalize our prior bounds under perfect alignment to *much weaker (more practical) assumptions*, and establish its relationship to the spectral properties of the augmentation graph.

- In Section 6, we add a new theoretical discussion on the effect of data augmentation strength based on *random graphs*, as well as empirically verifying these properties on real-world data sets.

- In Section 7.2, we extend our preliminary study on the model selection metric by studying its *various variants*, and evaluate its effectiveness on real-world scenarios by comparison with state-of-the-art methods.

2. Related Work

Self-supervised learning learns representations by designing appropriate surrogate tasks, like rotation prediction (Komodakis and Gidaris, 2018; Wang et al., 2021), masked image reconstruction (He et al., 2022; Zhang et al., 2022; Du et al., 2024), and instance discrimination (He et al., 2020; Zhang et al., 2023). Among the self-supervised methods, contrastive learning is a representative type of framework which learns an encoder by closing the feature distance of positive samples generated by data augmentations (Chen et al., 2020; Grill et al., 2020; He et al., 2020; Zbontar et al., 2021; Caron et al., 2021; Oquab et al., 2023; Cui et al., 2023). Contrastive learning is quickly approaching the performance of supervised learning on different tasks and large-scale data sets (*e.g.,* ImageNet). Despite its huge success on empirical performance, the working mechanism of contrastive learning is still under explorations. In the following, we briefly summarize the related work from two views.

1) View of Pretraining Objectives. InfoNCE (van den Oord et al., 2018) is a commonly used loss function in contrastive learning, which is motivated by NCE (Gutmann and Hyvärinen, 2012) to learn meaningful representations by measuring the mutual information of positive samples. However, some previous works show that the surrogate mutual information estimation of positive samples can not directly ensure the downstream performance of contrastive learning (Kolesnikov et al., 2019; Tschannen et al., 2020). There are other works analyzing contrastive learning from different perspectives. For example, Wang and Isola (2020) think contrastive learning is composed of two properties: alignment of positive samples and uniformity of negative samples. Wang and Liu (2021) regarded InfoNCE as a hardness-aware function and analyzed the important temperature parameter τ. Wen and Li (2021) theoretically proved that contrastive learning can obtain sparse feature representations with appropriate data augmentations. Zimmermann et al. (2021) indicated learned representations by contrastive learning implicitly invert the data generation process. Hu et al. (2023) reveal the connection between contrastive learning and stochastic neighbor embedding while Wang et al. (2023) establish the connection between contrastive learning and graph neural networks. Parulekar et al. (2023) theoretically prove that the representations learned by InfoNCE is well-clustered by the semantics in the data. Besides, some self-supervised learning paradigms without negative samples also achieve impressive performance (Grill et al., 2020). Among them, Tian et al. (2021) theoretically proved that these types of methods will not collapse into trivial representations with the help of their proposed ingredients including Exponential Moving Average (EMA) and stop-gradient, while Zhuo et al. (2023) analyzed it from the perspective of rank differential mechanism.

2) View of Downstream Generalization. Arora et al. (2019) and Lee et al. (2021) theoretically proved that the downstream performance of contrastive learning is close to that of supervised learning and the difference is controlled by an analyzable error term. However, their conclusions are based on an impractical assumption that the positive samples are conditionally independent[1], and the error term will get larger when contrastive learning has

1. Analysis on its impracticality could be found in Appendix B.4.

more negative samples (this is in contrast to empirical results). Following them, Nozawa and Sato (2021) provided a new kind of theoretical analysis without conditionally independent assumption. However, there exists a class collision term in their bounds that can not be ignored. Lei et al. (2023) further establish generalization guarantees that do not rely on the number of negative examples. Besides, there are some other perspectives to theoretically characterize the downstream performance of contrastive learning, including graph theory (HaoChen et al., 2021; Wang et al., 2024b), information theory (Tian et al., 2020; Tsai et al., 2021; Tosh et al., 2020; Ouyang et al., 2025; Cui et al., 2025), kernel method (Li et al., 2021), causal mechanism (Mitrovic et al., 2021), and distributionally robust optimization (Wu et al., 2023).

Other Types of Self-Supervised Learning Methods. Beside contrastive learning, there are other types of self-supervised learning methods. For reconstruction-based methods, Garg and Liang (2020) established theoretical analysis by viewing them as imposing a regularization on the representation via a learnable function, and Cao et al. (2022); Zhang et al. (2022) theoretically analyzed the function of different designs for the reconsturction-based method (He et al., 2022). Wang et al. (2024a) understands reconstruction-based methods from two properties following Wang and Isola (2020), i.e., the alignment and uniformity. Tan et al. (2023) analyze the reconsturtcion-based objectives from the matrix information theory. For auto-regressive based methods, Saunshi et al. (2021) reformulated text classification tasks as sentence completion problems and proposed theoretical guarantees for their downstream performance. Zhang et al. (2024) theoretically compare reconstruction and autoregressive-based methods from a spectral perspective.

3. Problem Setup

Here, we briefly introduce some basic notations and common practice of contrastive learning for the image classification task. Generally, it has two stages, unsupervised pretraining and supervised finetuning. In the first stage, on the unlabeled data $\mathcal{D}_u = \{\bar{x}\}$, we pretrain an encoder $f \in \mathcal{F}$, where $\mathcal{F}$ is a hypothesis class consists of functions mapping from the d-dimensional input space $\mathbb{R}^d$ to the unit hypersphere $\mathbb{S}^{m-1}$ in m-dimensional space. We assume that f is normalized, meaning the representation space $\mathbb{S}^{m-1}$ is bounded. This assumption is consistent with the designs of common contrastive learning frameworks like SimCLR (Chen et al., 2020) and MoCo (He et al., 2020). In the second stage, we evaluate the learned representations z with the labeled data $\mathcal{D}_l = \{(\bar{x}, y)\}$ where label $y \in \{1, 2, \ldots, K\}$. For simplicity, we assume that every sample belongs to a unique class, *i.e.,* $p(y|x)$ being one-hot, and the data set is class balanced, *i.e.,* $p(y = k) = 1/K$.

Contrastive Pretraining. Given a random raw training example $\bar{x} \in \mathcal{D}_u$, we first draw two *positive samples* (x, x^+) by applying two randomly drawn data augmentations $t, t^+ \in \mathcal{T}$ to $\bar{x}$, *i.e.,* $x = t(\bar{x}), x^+ = t^+(\bar{x})$, where $\mathcal{T} = \{t : \mathbb{R}^d \to \mathbb{R}^d\}$ contains all possible predefined data augmentations. Let $p(x, x^+)$ be the joint distribution of positive pairs, and $p(x) = \int p(x, x^+)dx^+$ be the marginal distribution of the augmented data x. Without loss of generality, we take x as the anchor sample, and draw M augmented samples $\{x_i^-\}_{i=1}^M$ independently from the marginal distribution $p(\cdot)$ as its *negative samples*, and denote their corresponding distributions as $p(x_i^-), i = 1, \ldots, M$. Then, the encoder f is learned by the contrastive loss. Researchers have designed various types of contrastive loss like triplet loss

(Schroff et al., 2015), hinge loss (Al-Obaidi et al., 2020), margin loss (Shah et al., 2022), etc. Following the same spirit (pulling in positive pairs and pushing away negative pairs), we mainly consider a most widely used InfoNCE loss (van den Oord et al., 2018) in this paper:

$$\mathcal{L}_{\text{contr}}(f) = \mathbb{E}_{p(x,x^+)\Pi_i p(x_i^-)} \left[-\log \frac{\exp(f(x)^\top f(x^+))}{\sum_{i=1}^M \exp(f(x)^\top f(x_i^-))} \right]. \tag{1}$$

In practice, some variants like SimCLR (Chen et al., 2020) also include the positive pair (x, x^+) in the denominator (inside the logarithm) of the InfoNCE loss. A recent work (Yeh et al., 2021), however, notices that this extra term introduces a coupling between positive and negative samples, which leads to sample deficiency such as requiring a large batch size (*e.g.,* 4096). Instead, both Yeh et al. (2021) and Fürst et al. (2022) show that omitting the term of positive pairs leads to consistent performance improvement. Thus, we adopt this positive-free denominator in the paper.

Evaluation on Downstream Tasks. To measure the quality of representations learned by contrastive pretraining, a popular metric is the prediction accuracy of a linear classifier $g(z) = [w_1^\top, \dots, w_K^\top]z$ learned on top of the learned representations $z = f(x)$. Typically, the linear classifier is trained by the Cross Entropy (CE) loss (Chen et al., 2020) on the labeled data $\mathcal{D}_l$:

$$\mathcal{L}_{\text{CE}}(f) = \mathbb{E}_{p(x,y)} \left[-\log \frac{\exp\left(f(x)^\top w_y\right)}{\sum_{i=1}^K \exp\left(f(x)^\top w_i\right)} \right], \tag{2}$$

where w_y is the optimal linear parameter for class y. For the convenience of theoretical analysis, following the common practice in Arora et al. (2019); Ash et al. (2022); Nozawa and Sato (2021), we directly adopt the mean representation of each class as the classwise weights, *i.e.,* $w_k = \mu_k := \mathbb{E}_{p(x,y)}[f(x)|y = k]$, namely, *mean classifier*. Thus, we obtain the following mean CE loss (mCE):

$$\mathcal{L}_{\text{mCE}}(f) = \mathbb{E}_{p(x,y)} \left[-\log \frac{\exp\left(f(x)^\top \mu_y\right)}{\sum_{k=1}^K \exp\left(f(x)^\top \mu_k\right)} \right]. \tag{3}$$

It is straightforward to conclude that $\mathcal{L}_{\text{CE}}(f) \le \mathcal{L}_{\text{mCE}}(f)$ and their performance is comparable in practice as demonstrated in Arora et al. (2019).

Scale Adjustment. For the theoretical analysis of contrastive learning, our goal is to characterize the gap between the pretraining loss $\mathcal{L}_{\text{contr}}(f)$ (Eq. 1) and the downstream classification loss $\mathcal{L}_{\text{mCE}}(f)$ (Eq. 3). Revisiting Eq. 1 and Eq. 3, we find that their main difference lies on the denominator, *i.e.,* $\mathcal{L}_{\text{contr}}(f)$ is computed over M negative samples while $\mathcal{L}_{\text{mCE}}(f)$ is over K class centers. The scale of M and K could be very different, which further leads to the scale mismatch between $\mathcal{L}_{\text{contr}}(f)$ and $\mathcal{L}_{\text{mCE}}(f)$. Therefore, we adjust the loss by taking the mean score instead of summation in the denominator accordingly:

$$\bar{\mathcal{L}}_{\text{contr}}(f) = \mathbb{E}_{p(x,x^+)\Pi_i p(x_i^-)} \left[-\log \frac{\exp(f(x)^\top f(x^+))}{\frac{1}{M}\sum_{i=1}^M \exp(f(x)^\top f(x_i^-))} \right] = \mathcal{L}_{\text{contr}}(f) + \log(M^{-1}), \tag{4}$$

$$\bar{\mathcal{L}}_{\mathrm{mCE}}(f) = \mathbb{E}_{p(x,y)}\left[-\log \frac{\exp(f(x)^\top \mu_y)}{\frac{1}{K}\sum_{k=1}^{K}\exp(f(x)^\top \mu_k)}\right] = \mathcal{L}_{\mathrm{mCE}}(f) + \log(K^{-1}). \tag{5}$$

In this way, the two normalized objectives are irrelevant to the scale of M and K. Note that the normalization has *no effect* on the learning process as their gradients are equal to the original ones. In the following discussion, we mainly deal with the normalized versions. Additionally, regarding the discrepancy between the two objectives, the scale adjustment above amounts to adding a $\log(K/M)$ term to the contrastive loss:

$$\bar{\mathcal{L}}_{\mathrm{contr}}(f) - \bar{\mathcal{L}}_{\mathrm{mCE}}(f) = \mathcal{L}_{\mathrm{contr}}(f) + \log(K/M) - \mathcal{L}_{\mathrm{mCE}}(f). \tag{6}$$

4. Improved Bounds under Conditional Independence Assumption

In this section, we focus on improving bounds by analyzing the roles of *negative samples*. Without loss of generality, we take the *positive pairs* under the conditional independence assumption (Arora et al., 2019) as an example, which enables us to have a fair comparison to previous bounds and shows our advantages more directly. Specifically, we propose a new technique by showing the negative samples are closely related to the Monte Carlo estimation of downstream class centers, which provides a new insight on the benefits of more negative samples. Further, built upon this technique, we show that *for the first time*, we could obtain an asymptotically closed upper bound on the downstream error. This new finding provides a solid guarantee on the downstream performance of contrastive learning and forms the basis for our further analysis beyond conditional independence in the next section.

Assumption 1 (Conditional Independence (Arora et al., 2019)) *The two positive samples $x, x^+ \sim p(x, x^+)$ are conditionally independent given the label y, i.e., $x \perp\!\!\!\perp x^+ \mid y$.*

In the following, we will show how the InfoNCE loss (Eq. 4) closely upper bounds the mean CE loss (Eq. 5) under the conditional independence assumption. Firstly, we decouple the two objectives into a positive objective and a negative objective, respectively:

$$\bar{\mathcal{L}}_{\mathrm{contr}}(f) = \underbrace{-\mathbb{E}_{p(x,x^+)}[f(x)^\top f(x^+)]}_{\bar{\mathcal{L}}^+_{\mathrm{contr}}(f)} + \underbrace{\mathbb{E}_{\Pi_i p(x_i^-)} \log \frac{1}{M}\sum_{i=1}^{M}\exp(f(x)^\top f(x_i^-))}_{\bar{\mathcal{L}}^-_{\mathrm{contr}}(f)}, \tag{7}$$

$$\bar{\mathcal{L}}_{\mathrm{mCE}}(f) = \underbrace{-\mathbb{E}_{p(x,y)}[f(x)^\top \mu_y]}_{\bar{\mathcal{L}}^+_{\mathrm{mCE}}(f)} + \underbrace{\mathbb{E}_{p(x)} \log\left[\frac{1}{K}\sum_{k=1}^{K}\exp(f(x)^\top \mu_k)\right]}_{\bar{\mathcal{L}}^-_{\mathrm{mCE}}(f)}. \tag{8}$$

It is easy to see that the two positive objectives are equal under conditional independence:

$$\begin{aligned}\bar{\mathcal{L}}^+_{\mathrm{contr}}(f) &= -\mathbb{E}_{p(x,x^+)}[f(x)^\top f(x^+)] = -\mathbb{E}_{p(y)p(x|y)p(x^+|y)}[f(x)^\top f(x^+)] \\ &= -\mathbb{E}_{p(y)p(x|y)}\left[f(x)^\top [\mathbb{E}_{p(x^+|y)} f(x^+)]\right] = -\mathbb{E}_{p(y)p(x|y)}[f(x)^\top \mu_y] = \bar{\mathcal{L}}^+_{\mathrm{mCE}}(f).\end{aligned} \tag{9}$$

Thus, we only need to bound the two negative objectives ($\bar{\mathcal{L}}^-_{\mathrm{contr}}(f)$ and $\bar{\mathcal{L}}^-_{\mathrm{mCE}}(f)$).

Limitations of Previous Analysis. Previous studies of the negative objective (Arora et al., 2019; Nozawa and Sato, 2021; Ash et al., 2022) often adopt a collision-coverage perspective to dissect the negative samples. Specifically, they use the proportion of negative samples belonging to the positive class (class collision) and the probability that there are at least $K-1$ negative samples covering the $K-1$ negative classes (class coverage) to form an upper bound. However, a clear drawback of this analysis is that it actually only requires K negative samples for class coverage, and the other $M-K$ samples are discarded in the analysis, resulting in a very loose bound. In this work, we provide a new technique to analyze the role of M negative samples. In particular, we show that all M negative samples are useful and they all together contribute a tight upper bound that is asymptotically closed with $M \to \infty$. Bao et al. (2022) also adopt this technique to analyze contrastive learning with the InfoNCE loss. Nevertheless, their analysis involves applying a reverse Jensen's inequality *twice*, leading to a large error in the upper bound that is not decreasing w.r.t. M.

Our New View of Negative Samples for Monte Carlo Estimation. For the negative objective of InfoNCE loss, due to the convexity of the logsumexp operator, following Jensen's inequality we have

$$\begin{aligned}
\bar{\mathcal{L}}^{-}_{\mathrm{contr}}(f) &= \mathbb{E}_{\Pi_i p(y_i^-)p(x_i^-|y_i^-)} \log \frac{1}{M}\sum_{i=1}^{M} \exp(f(x)^\top f(x_i^-)) \\
&\geq \mathbb{E}_{\Pi_i p(y_i^-)} \log \frac{1}{M}\sum_{i=1}^{M} \exp(\mathbb{E}_{p(x_i^-|y_i^-)} f(x)^\top f(x_i^-)) \\
&= \mathbb{E}_{\Pi_i p(y_i^-)} \log \frac{1}{M}\sum_{i=1}^{M} \exp(f(x)^\top \mu_{y_i^-}) := \bar{\mathcal{L}}_{\mathrm{MC}}(f).
\end{aligned} \tag{10}$$

As we assume $p(y=k) = 1/K$, it is easy to see that its lower bound $\bar{\mathcal{L}}_{\mathrm{MC}}(f)$ is a (biased) Monte Carlo estimation of the negative objective of the mean CE loss, *i.e.,* $\bar{\mathcal{L}}^{-}_{\mathrm{mCE}}(f) = \log \mathbb{E}_{p(y)} \exp(f(x)^\top \mu_y)$, and the approximation error shrinks to 0 as $M \to \infty$, as characterized in the following lemma.

Lemma 2 *The approximation error of the Monte Carlo estimation $\bar{\mathcal{L}}_{\mathrm{MC}}(f)$ shrinks in the order $\mathcal{O}(M^{-1/2})$, specifically.*

$$|\bar{\mathcal{L}}_{\mathrm{MC}}(f) - \bar{\mathcal{L}}^{-}_{\mathrm{mCE}}(f)| \leq \frac{e}{\sqrt{M}}. \tag{11}$$

Lemma 2 shows that all M negative samples contribute to a better Monte Carlo estimator of the negative objective of the mean CE loss.

Combining Eqs. 9, 10 & 11, we derive an asymptotically closed upper bound on the downstream performance:

$$\begin{aligned}
\bar{\mathcal{L}}_{\mathrm{mCE}}(f) &= \bar{\mathcal{L}}^{+}_{\mathrm{mCE}}(f) + \bar{\mathcal{L}}^{-}_{\mathrm{mCE}}(f) \leq \bar{\mathcal{L}}^{+}_{\mathrm{contr}}(f) + \bar{\mathcal{L}}_{\mathrm{MC}}(f) + \frac{e}{\sqrt{M}} \\
&\leq \bar{\mathcal{L}}^{+}_{\mathrm{contr}}(f) + \bar{\mathcal{L}}^{-}_{\mathrm{contr}}(f) + \frac{e}{\sqrt{M}} = \bar{\mathcal{L}}_{\mathrm{contr}}(f) + \frac{e}{\sqrt{M}}.
\end{aligned} \tag{12}$$

Following a similar routine, we can also derive the lower bound on the downstream performance via a reversed Jensen's inequality (Budimir et al., 2000). Therefore, under the conditional independence assumption, our guarantees for the downstream performance of contrastive learning can be summarized as:

Theorem 3 (Downstream Guarantees under Conditional Independence) *Under Assumption 1, the downstream classification risk $\bar{\mathcal{L}}_{\mathrm{mCE}}(f)$ of any $f \in \mathcal{F}$ can be upper and lower bounded by the contrastive risk $\bar{\mathcal{L}}_{\mathrm{contr}}(f)$ as*

$$\underbrace{\bar{\mathcal{L}}_{\mathrm{contr}}(f) - \frac{1}{2}\mathrm{Var}(f(x) \mid y) - \frac{e}{\sqrt{M}}}_{\textit{lower bound } \mathcal{R}_{\mathrm{L}}} \leq \bar{\mathcal{L}}_{\mathrm{mCE}}(f) \leq \underbrace{\bar{\mathcal{L}}_{\mathrm{contr}}(f) + \frac{e}{\sqrt{M}}}_{\textit{upper bound } \mathcal{R}_{\mathrm{U}}}, \tag{13}$$

where M is the number of negative samples. Furthermore, the conditional variance $\mathrm{Var}(f(x)|y) = \mathbb{E}_{p(x,y)}\|f(x) - \mathbb{E}_{p(\hat{x}|y)}f(\hat{x})\|^2$ is at most 2. As a result, the gap between the upper and the lower bounds can be bounded: $\mathcal{R}_{\mathrm{U}} - \mathcal{R}_{\mathrm{L}} \leq 1 + 2e/\sqrt{M}$.

Theorem 3 shows that the downstream performance can be upper and lower bounded by the contrastive loss. Thus, minimizing contrastive loss is almost equivalent (with a small surrogate gap) to optimizing the supervised loss, which helps us to understand the empirical effectiveness of contrastive learning. Furthermore, as $M \to \infty$, *i.e.,* adopting more negative samples, the upper and lower bounds can be further narrowed to $\bar{\mathcal{L}}_{\mathrm{contr}}(f) - \frac{1}{2}\mathrm{Var}(f(x) \mid y) \leq \bar{\mathcal{L}}_{\mathrm{mCE}}(f) \leq \bar{\mathcal{L}}_{\mathrm{contr}}(f)$. In the common practice of contrastive learning, a large number of negative samples indeed often brings better downstream performance (Chen et al., 2020). Our theoretical analysis not only echoes with this empirical finding, but also provides new insights on the role of negative samples. According to Lemma 4.2, negative samples from all classes (even the same supervised class as positive samples) contribute to a better Monte Carlo estimation. Thus, we can close the upper bound even in the presence of false negative samples. This implication also aligns well with the practice, as contrastive learning indeed performs comparably or even superiorly to supervised learning on real-world data sets without pruning the false negative samples (Tomasev et al., 2022). While this does not preclude the potential benefits from negative sample mining (as explored in Robinson et al. (2021)) in contrastive learning, as Lemma 4.2 only implies the impact of the size of negative samples for contrastive learning. As for the comparisons of the effects of contrastive learning with and without negative sample mining, it is outside the scope of Lemma 4.2.

Discussion. Prior to ours, several works (Arora et al., 2019; Ash et al., 2022; Nozawa and Sato, 2021; Bao et al., 2022) also propose different versions of guarantees for downstream performance, and we summarize their upper bounds in Table 1. As some works analyze other variants of the pretraining and downstream objectives, *e.g.,* hinge loss (Arora et al., 2019), we generally denote them as $\mathcal{L}_{\mathrm{unsup}}$ and $\mathcal{L}_{\mathrm{sup}}$, respectively. In particular, the seminal work of Arora et al. (2019) draws two important implications from the upper bound: 1) the surrogate gap between the two objectives cannot be closed because of an unavoidable *class collision* errors (τ_M, Col) in the upper bound that account for the proportion of negative samples belonging to the same class of the anchor sample, *i.e.,* the false negative samples; and 2) the overall upper bounds increase with more negative samples (larger M). However, both implications have been shown *contradictory* to the empirical practice, as discussed

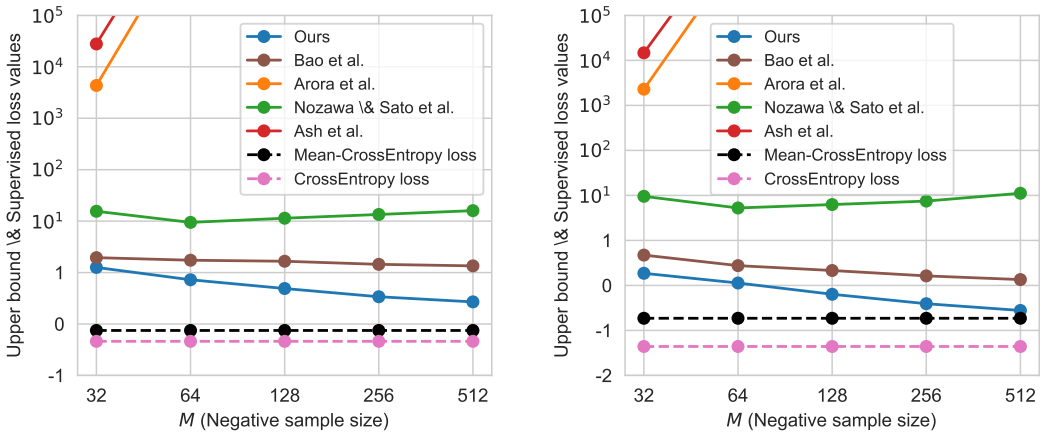
Table 1: A comparison of upper bounds of the supervised downstream loss $\mathcal{L}_{\text{sup}}(f)$ using the unsupervised pretraining loss $\mathcal{L}_{\text{unsup}}(f)$ (under the conditional independence assumption). Here, $\text{Col} = \sum_{m=1}^{M} \mathbb{1}[y_{x_m^-} = y_x]$ denotes the degree of class collision; τ_M denotes the collision probability that at least one of the M negative samples belong to the positive class y_x; v_{M+1} denotes the coverage probability that $M+1$ negative samples contains all classes $k \in [K]$; and $H_{K-1} = \sum_{k=1}^{K-1} 1/k$ is the $(K-1)$-th harmonic number. *Adjusted objective scales for clear comparison.

	Upper Bound	Reference
$\mathcal{L}_{\text{sup}}(f) \leq$	$\frac{1}{(1-\tau_M)v_{M+1}}\Big(\mathcal{L}_{\text{unsup}}(f) - \mathbb{E}\log(\text{Col}+1)\Big)$	Arora et al. (2019)
	$\frac{1}{v_{M+1}}\left(2\mathcal{L}_{\text{unsup}}(f) - \mathbb{E}\log(\text{Col}+1)\right)$	Nozawa and Sato (2021)
	$\frac{2}{1-\tau_M}\left(\frac{2(K-1)H_{K-1}}{M}\right)\Big(\mathcal{L}_{\text{unsup}}(f) - \mathbb{E}\log(\text{Col}+1)\Big)$	Ash et al. (2022)
	$\mathcal{L}_{\text{unsup}}(f) + 2\log(\cosh(1))$	Bao et al. (2022)*
	$\mathcal{L}_{\text{unsup}}(f) + e/\sqrt{M}$	Our work*

above. Subsequential works are devoted to resolving the two issues by eliminating the class collision errors and demonstrating the benefits of a large M. Specifically, Nozawa and Sato (2021) and Ash et al. (2022) manage to show benefit from larger M, and Bao et al. (2022) fully eliminate the class collision error. Nevertheless, their final upper bounds have unavoidable error terms even with $M \to \infty$. Indeed, as shown in Figure 3, as $M \to \infty$, the bounds of previous work are either explosive (Arora et al., 2019; Ash et al., 2022) or roughly constant (Nozawa and Sato, 2021; Bao et al., 2022). In comparison, with the proposed Monte Carlo analysis of negative samples, we take all negative samples into consideration and resolve the two issues completely. In particular, we are the first to show that the upper bound could be asymptotically tight with a large M, as shown theoretically in Table 1 and empirically in Figure 3. Our results suggest that contrastive learning is indeed *almost equivalent* to supervised learning under the conditional independence assumption.

5. From Conditional Independence to Augmentation Overlap

In Section 4, we have improved the bounds on negative samples from the Monte Carlo perspective. While in this section, our focus shifts to elaborating on the effects of positive pairs. We first relax the impractical conditional independence assumption used in Theorem 3 and derive a new bound that is unfortunately with an extra unbounded term of intra-class variance. To further analyze this variance term, we revisit contrastive learning from a graph perspective and build a new augmentation overlap framework. Based on that, we obtain a new guarantee for the gap between the pretraining and downstream performance of contrastive learning.



(a) A randomly initialized encoder f.

(b) An encoder f trained with SimCLR.

Figure 3: Comparison of upper bounds on the downstream loss (measured by mean CE loss) on CIFAR-10. The encoder is a ResNet-18 (He et al., 2016) and we train it using SimCLR (Chen et al., 2020). We calculate the upper bounds using its representations at the (a) initialization and (b) final stages.

5.1 Towards a Relaxation of Conditional Independence

Intuitively, as shown in Figure 2, a positive pair is generated from the same natural image. Consequently, the positive pairs are input-dependent and it is hard to satisfy the conditional independence assumption in practice. So we release the impractical conditional independence assumption and start with a basic assumption on the label consistency between positive samples, that is, any pair of positive samples (x, x^+) should nearly belong to the same class.

Assumption 4 (Label Consistency) *$\forall\ x, x^+ \sim p(x, x^+)$, we denote y and y^+ as their labels. We assume the probability that they have different labels is smaller than α, i.e.,*

$$\mathbb{E}_{x,x^+\sim p(x,x^+)} \mathbb{1}(y \neq y^+) \leq \alpha. \tag{14}$$

This is a basic and natural assumption that is likely to hold in practice. From Figure 2(b), we can see that the widely adopted augmentations (*e.g.,* images cropping, color distortion, and horizontal flipping) will hardly alter the image belonging class.

Theorem 5 (Downstream Guarantees without Conditional Independence) *If Assumption 4 holds, then, for any $f \in \mathcal{F}$, its downstream classification risk $\bar{\mathcal{L}}_{\mathrm{mCE}}(f)$ can be bounded by the contrastive learning risk $\bar{\mathcal{L}}_{\mathrm{contr}}(f)$*

$$\begin{aligned} &\bar{\mathcal{L}}_{\mathrm{contr}}(f) - 2\sqrt{\mathrm{Var}(f(x) \mid y)} - \frac{1}{2}\mathrm{Var}(f(x) \mid y) - 4\sqrt{\alpha} - \frac{e}{\sqrt{M}} \\ &\leq \bar{\mathcal{L}}_{\mathrm{mCE}}(f) \leq \bar{\mathcal{L}}_{\mathrm{contr}}(f) + 2\sqrt{\mathrm{Var}(f(x)|y)} + 4\sqrt{\alpha} + \frac{e}{\sqrt{M}}. \end{aligned} \tag{15}$$

Theorem 5 shows that, even without the conditional independence assumption, we can still derive a bound for the downstream performance. Compared to Theorem 3, there is an additional intra-class variance term $\sqrt{\mathrm{Var}(f(x) \mid y)}$ in the lower and upper bound. Only with the label consistency assumption (Assumption 4), the variance term can not be bounded, which means that the generalization gap in Theorem 5 could be quite large. For example, when the intra-class variance term is large enough, contrastive learning might have inferior performance as shown in Proposition 6.

Proposition 6 *For N training examples of K classes, consider a case that inter-anchor features $\{f(x_i)\}_{i=1}^N$ are randomly distributed in $\mathbb{S}^{m-1}$ while intra-anchor features are perfectly aligned, i.e., $\forall x_i, x_i^+ \sim p(x, x^+), f(x_i) = f(x_i^+)$. In this case, the expectation of the numerator of InfoNCE loss $\mathbb{E}_{p(x_i,x_i^+)}(\exp(f(x_i)^\top f(x_i^+)))$ achieves its maximum while the expectation of the denominator $\mathbb{E}_{\Pi_i p(x_j^-)}(\frac{1}{M}\sum_{j=1}^M \exp(f(x_i)^\top f(x_j^-)))$ achieves its minimum, i.e., both the alignment and uniformity losses (Wang and Isola, 2020) achieve the minimum, thus the InfoNCE loss obtains its minimum. However, the downstream classification accuracy is at most $1/K + \varepsilon$ where ε is close to 0 when N is large enough.*

If given the conditional independence assumption, the alignment between positive samples is equivalent to the alignment between the sample itself and its class center. Thus, the intra-class samples will be finally aligned to the class center, which will not be uniformly distributed such that the above failure case does not exist.

To summarize, the conditional independence assumption is too strong to eliminate the variance term in the theoretical bounds (Theorem 3), while discarding this assumption will make the variance term unbounded and further lose the guarantee on the downstream performance (Theorem 5). In the following part, we will present a new framework to analyze this variance term theoretically.

5.2 A New Theoretical Analysis Framework under Augmentation Overlap

As analyzed in Theorem 5, when contrastive learning has enough negative samples, the gap between pretraining and downstream performance mainly hinges on the variance of features in the same class. However, with current analysis tools, it is quite difficult to analyze the variance of intra-class samples. Therefore, in this section, we build a new augmentation overlap based framework to do this, and further theoretically characterize the generalization of contrastive learning.

Intuitively, contrastive learning is an instance-level task that can not bridge the samples in the same class. However, as shown in Figure 1, the features of the same class will be clustered while the features of different classes will be separated, which indicate the intra-class variance gradually decreases along with the training process. Observing the training process of contrastive learning, we find that the different samples of the same class will generate quite similar views. For example, as shown in Figure 2(b), appropriate data augmentations will make the views of different cars focus on similar tires. Then the training process of contrastive learning will close the feature distance of two cars as they

share a similar view. Thus, the data augmentations play a key role in the understanding of contrastive learning, especially the overlapped views of augmentations.

Accordingly, we formulate the above intuitive understanding of augmentation overlap mathematically via a graph. Assume a common graph $\mathcal{G}(V, E)$ is composed of two components: a set V representing the vertices and a set E representing the edges between the vertices. Combined with a set of augmentations $\mathcal{T} = \{t \mid t : \mathbb{R}^d \to \mathbb{R}^d\}$, an augmentation graph is defined as follows:

Definition 7 (Augmentation Graph) *Given unlabeled natural data $\mathcal{D}_u = \{\bar{x}_i\}_{i=1}^N$ and a collection of augmentations $\mathcal{T} = \{t \mid t : \mathbb{R}^d \to \mathbb{R}^d\}$, an augmentation graph $\mathcal{G}(V, E(\mathcal{T}))$ is defined as*

- *its vertices are the samples, i.e., $V = \{\bar{x}_i\}_{i=1}^N$;*
- *its edge e_{ij} between two vertices $\bar{x}_i$ and $\bar{x}_j$ exists when they have overlapped views, i.e., there exist two augmentations $t_1, t_2 \in \mathcal{T}$ satisfying $t_1(\bar{x}_i) = t_2(\bar{x}_j)$.*

Based on the augmentation graph, we propose to utilize the properties of graph, *i.e.,* connectivity, to replace the impractical conditional independence assumption. Specifically, an augmentation graph is connected when its any two vertices (*e.g.,* $\bar{x}_i, \bar{x}_j$) are connected, *i.e.,* there exists a path $(\bar{x}_i, \cdots, \bar{x}_j)$ between the two vertices. An illustrative example is shown in Figure 4(a).

Assumption 8 *Let $\mathcal{D}_k$ be the set of the samples in the class k of $\mathcal{D}_u$. There exists an appropriate augmentation set $\mathcal{T}$ satisfying that, for $k \in \{1, \cdots, K\}$, the augmentation graph $\mathcal{G}(\mathcal{D}_k, E(\mathcal{T}))$ is connected.*

The assumption says that, for every pair of the intra-class samples, we only assume the path exists between them rather than requiring direct edges, *i.e.,* we only need connected graphs instead of complete graphs, which makes the connected augmentation graph assumption more practical.

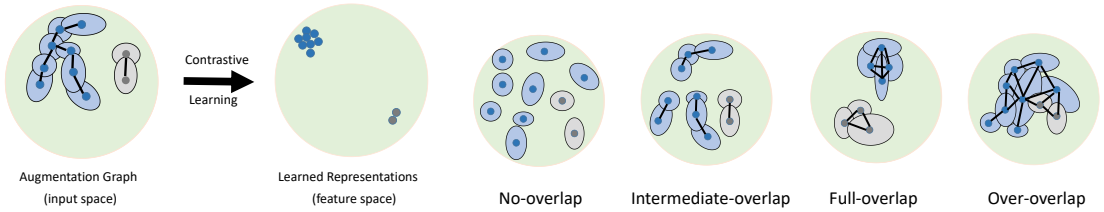
Beside the assumption on intra-class samples, we also need an assumption on the learned encoder of contrastive learning, that is, the distance between the positive samples will gradually converge to a small constant ε during the contrastive learning process as demonstrated in Wang and Isola (2020).

Assumption 9 *$\forall\, x, x^+ \sim p(x, x^+)$, we assume the learned mapping f by contrastive learning is ε-alignment, i.e., $\|f(x) - f(x^+)\| \leq \varepsilon$.*

With the assumptions on the intra-class connectivity of augmentation graph and the characterization on the alignment of positive pairs, we have the following theorem:

Theorem 10 (Guarantees with Connected Augmentation Graph) *If Assumptions 4 and 8 hold, then $\forall f \in \mathcal{F}$ satisfying ε-alignment (Assumption 9), its classification risk can be upper and lower bounded by its contrastive risk as*

$$\begin{aligned} \bar{\mathcal{L}}_{\text{contr}}(f) - (2 + \frac{D\varepsilon}{2})D\varepsilon - 4\sqrt{\alpha} - \frac{e}{\sqrt{M}} \\ \leq \bar{\mathcal{L}}_{\text{mCE}}(f) \leq \bar{\mathcal{L}}_{\text{contr}}(f) + 2D\varepsilon + 4\sqrt{\alpha} + \frac{e}{\sqrt{M}}, \end{aligned} \tag{16}$$



(a) Contrastive learning with an augmentation graph satisfying intra-class connectivity.

(b) Augmentation graph under increasing augmentation strengthes (from left to right).

Figure 4: Illustrative examples of augmentation graphs, where each dot denotes a sample $x \in \mathcal{D}_u$ and its color denotes its class. The lighter disks denote the support of the positive samples $p(x^+|x)$. We draw a solid edge for each pair that has an edge.

where D denotes the maximal radius of the intra-class augmentation graphs $\{\mathcal{G}_k, k = 1, \dots, K\}$.

Comparing Eq. 15 and Eq. 16, we can see that the intra-class variance term $\sqrt{\mathrm{Var}(f(x) \mid y)}$ is replaced by $D\varepsilon$, where D is the radius of augmentation graph and ε measures the feature distance between positive samples. Intuitively, for samples in each class k, if its corresponding augmentation graph $\mathcal{G}_k$ is connected, there exists a path between any two samples of the same class and aligning the adjacent sample pairs in the paths can eventually align all the samples of the same class. As a result, the intra-class variance can be controlled. Therefore, to ensure the generalization of contrastive learning, we need to meet two conditions: 1) choosing appropriate data augmentations to construct a connected augmentation graph with a smaller radius (smaller D); and 2) selecting a proper objective to align the positive samples better (smaller ϵ). In Figure 4(a), we visualize the training process of contrastive learning under augmentation graph.

Corollary 11 *If Assumptions 4 and 8 hold, then, for f satisfying 0-alignment, its classification risk can be upper and lower bounded by its contrastive risk as*

$$\bar{\mathcal{L}}_{\mathrm{contr}}(f) - 4\sqrt{\alpha} - \frac{e}{\sqrt{M}} \leq \quad \bar{\mathcal{L}}_{\mathrm{mCE}}(f) \leq \bar{\mathcal{L}}_{\mathrm{contr}}(f) + 4\sqrt{\alpha} + \frac{e}{\sqrt{M}}. \tag{17}$$

The corollary shows that the intra-class variance term will varnish when the positive samples are perfectly aligned ($\epsilon = 0$). Furthermore, if enough negative samples are given ($M \to \infty$) and we design appropriate augmentations that do not change the labels of samples ($\alpha = 0$), the gap between pretraining and downstream performance will be totally closed, *i.e.,* the unsupervised contrastive learning is equivalent to a supervised task.

5.3 Further Discussion on the Connectivity of Augmentation Graph

Theorem 10 has provided a view of graph radius for the characterization of the connectivity of augmentation graph. In this part, we additionally provide another new perspective, *i.e.,* spectral property of augmentation graph.

Corollary 12 *If Assumptions 4 and 8 hold, then $\forall f \in \mathcal{F}$ satisfying ε-alignment, its classification risk can be upper and lower bounded by its contrastive risk as*

$$\begin{aligned}\bar{\mathcal{L}}_{\mathrm{contr}}(f) - \left(2 + \frac{\log((1-\omega^2)/\omega^2)}{2\log(|\lambda_1|/|\lambda_2|)}\varepsilon\right)\frac{\log((1-\omega^2)/\omega^2)}{\log(|\lambda_1|/|\lambda_2|)}\varepsilon - 4\sqrt{\alpha} - \frac{e}{\sqrt{M}} \\ \leq \bar{\mathcal{L}}_{\mathrm{mCE}}(f) \leq \bar{\mathcal{L}}_{\mathrm{contr}}(f) + \frac{2\log((1-\omega^2)/\omega^2)}{\log(|\lambda_1|/|\lambda_2|)}\varepsilon + 4\sqrt{\alpha} + \frac{e}{\sqrt{M}},\end{aligned} \tag{18}$$

where $\omega = \min_{i,k}|(\mu_{1k})_i|$, $\lambda_1 = \min_k|\lambda_{1k}|$ and $\lambda_2 = \max_k|\lambda_{2k}|$. Among them, $\mu_{1k}, \mu_{2k}, \cdots$ are the orthonormal eigenvectors with eigenvalues $\lambda_{1k}, \lambda_{2k}, \cdots$ ($|\lambda_{1k}| \geq |\lambda_{2k}| \geq \cdots$) of the adjacent matrix of intra-class subgraph $\mathcal{G}_k$.

Comparing Eq. 16 and Eq. 18, we find that the graph radius D is replaced by $\frac{\log((1-\omega^2)/\omega^2)}{\log(|\lambda_1|/|\lambda_2|)}$ that is controlled by the largest and second largest eigenvalues of the augmentation graph. The eigenvalues are widely discussed in the spectral graph theory and a smaller $|\lambda_2|$ implies the stronger connectivity of the graph (Chung, 1989). From the corollary, we note that the gap between the pretraining and downstream performance is narrowed when λ_2 decreases, which further verifies that contrastive learning needs the augmentation graph to be closely connected.

Previously, there is another work analyzing the downstream performance of contrastive learning from a spectral graph perspective (HaoChen et al., 2021), but the difference can be obviously observed: 1) our analysis is applicable for the widely adopted InfoNCE and CE losses, while theirs is developed for their own spectral loss; 2) ours starts from the alignment and uniformity perspective while theirs starts from the matrix decomposition perspective; and 3) there are several kinds of cases that their method fails to analyze while ours can. As the former two points are easily verified, we only provide the detailed comparison on the third point below.

Lemma 13 (The main results in HaoChen et al. (2021)) *Let $f^\star \in \mathcal{F} : \mathbb{R}^d \to \mathbb{S}^{m-1}$ be a minimizer of $\mathcal{L}_{\mathrm{contr}}(f)$ and $m \geq 2K$. The downstream error is denoted as $Err(f^\star) = \mathbb{E}_{p(x,y)}(\hat{y}(x) \neq y)$ where $\hat{y}(x) = \arg\max g(f^\star(x))$ is the predicted label with the downstream classifier g. Then, we have*

$$Err(f^\star) \leq \mathcal{O}\left(\frac{\alpha}{\rho^2_{\lfloor m/2 \rfloor}}\right), \tag{19}$$

where m is the representation dimension, K is the number of classes, α is the minimum error under a set of labeling functions where the augmented views and the anchor sample have different labels, and ρ_q is the sparsest q-partition of augmentation graph.

Their bounds (Lemma 13) and ours (Corollary 12) are both based on the connectivity of augmentation graph but with different measures. We use the first and the second largest eigenvalues of the adjacent matrix of the augmentation graph while they use the sparsest q-partition that is estimated by the q-smallest eigenvalues of the normalized Laplacian matrix of the augmentation graph. Thus, these two bounds are not identical: 1) ours provide both the upper and lower bounds while they only provide the upper bound; and 2) there are some cases their method fails to analyze while ours could as shown below.

Case I: If Assumption 8 holds and $\lfloor m/2 \rfloor \leq K$, then $\rho_{\lfloor m/2 \rfloor} = 0$. When Assumption 4 holds, then $\alpha = 0$. The bound in Lemma 13 becomes a $\frac{0}{0}$ term which fails to be analyzed.

Case II: If we adopt the same setting in Proposition 6, *i.e.,* there exists no augmentation overlap for inter-anchor samples, the bound in Lemma 13 becomes a $\frac{0}{0}$ term again.

While for our bounds, the above two cases can still be characterized (proofs are shown in Appendix). This is because that our bounds are not dependent on the output dimension m and can still analyze the difference between representations when $\alpha = 0$. That is, we can analyze the downstream performance of contrastive learning in a more fine-grained way with less restrictions.

6. Analysis of Augmentation Strategy on Augmentation Overlap

Based on the above analysis, we can find that the gap between pretraining and downstream tasks hinges on the connectivity of the augmentation graph $\mathcal{G}(V, E(T))$. In this section, we will further analyze how the data augmentation strategy in contrastive learning influences the graph connectivity (augmentation overlap).

6.1 Degree of Augmentation Overlap

To simplify our analysis, here we consider the following setting. For each class k, there is a cluster center c_k on a hypersphere $\mathbb{S}^d$ where N anchor samples are uniformly distributed around c_k. The positive samples are obtained by adding a uniform noise sampling from a uniform distribution $U(0, r)^d$ to the anchor samples. Intuitively, when the augmentation strength r is too weak, there exist no overlapped views across different samples. When the augmentation is too strong, samples of different classes may be aligned together. Formally, based on the random graph theory, we have the following theoretical results.

Theorem 14 *For N random samples taken from a class, when gradually increasing the augmentation strength r, we have the following degree of augmentation overlap:*

(a) ***No-overlap.*** *When $0 \leq r < r_1 = \frac{[(d/2)!]^{\frac{1}{d}}}{\sqrt{\pi}}(\frac{1}{d})!(\frac{S}{N-1})^{\frac{1}{d}}[1 - \frac{1/d+1/d^2}{2(N-1)} + O(\frac{1}{(N-1)^2})]$ where S is the surface area of sample distribution and r_1 is the minimal pairwise distance among N samples, all samples (vertices) in the augmentation graph are isolated. As a result, the learned features could be totally random as in Proposition 6.*

(b) ***Over-overlap.*** *When $r \geq r_3 = \frac{1}{2}\min_{i,j} \|c_i - c_j\|$ where r_3 is the (asymptotic) minimal distance between samples from different classes, the label consistency assumption is no longer guaranteed.*

(c) ***Full-overlap.*** *When $r_2 \leq r < r_3$ where $r_2 = \frac{[(d/2)!]^{\frac{1}{d}}}{\sqrt{\pi}} \frac{(N-2+1/d)!}{(N-2)!}(\frac{S}{N-1})^{\frac{1}{d}}[1 - \frac{1/d+1/d^2}{2(N-1)} + O(\frac{1}{(N-1)^2})]$ is the maximal pairwise distance among N samples, all samples from the same class are directly connected while samples from the different classes are not connected. As a result, the augmentation graph for each class is a complete graph.*

(d) ***Intermediate-overlap.*** *When $r_1 \leq r < r_2$, there exists at least two samples that are connected, while the whole augmentation graph is not a complete graph.*

Under cases (c) and (d), the classwise connectivity in Assumption 8 is guaranteed.

The above Theorem 14 indicates a trade-off on the augmentation strength. On the one hand, we need the augmentation to be strong enough to align the samples from the same class. On the other hand, we need to avoid too strong augmentations that generate overlapped views for inter-class samples. In Figure 4(b), we visualize the relationship between different degrees of overlap and the connectivity of augmentation graph. According to Theorem 10, the guarantees about the generalization of contrastive learning only need the augmentation graph to be connected (may not need to be a complete graph). Thus, there is a proper (perfect) augmentation strength from 'intermediate-overlap' to 'full-overlap', named 'perfect-overlap', which can be described through the following theorem.

Theorem 15 *Denote the minimal augmentation strength needed for connectivity as:* $r_{mc} = \inf\{r_1 \leq r \leq r_2 : \mathcal{G}(V, E(\mathcal{T})) \text{ is connected}\}$ *and the volume of unit hyperball as* V_u*, we obtain:*

$$\text{For } d \geq 2, r_{mc} = O\left((2\frac{(1-1/d)S\log N}{V_u N^2})^{\frac{1}{d}}\right) \text{ as } N \to \infty. \tag{20}$$

Theorem 15 implies that the perfect augmentation strength (making augmentation graph be connected) is strongly related to the sample number (N) and sample dimension (d). As the sample dimension increases, the needed augmentation strength to meet 'perfect-overlap' will increase in an exponential order. Since the dimension of the samples in real-world data sets is quite large, stronger data augmentations are usually needed. However, simple data augmentations like Gaussian noise with strong variance (strong data augmentations) will change the label of images, which contradicts the label consistency assumption. Thus, we need to design data augmentations that are strong enough while do not change the label of samples simultaneously. This is in accordance with the practical fact that common contrastive learning methods usually use complex data augmentations like RandomResizedCrop and ColorJitter instead of simple Gaussian noise.

6.2 Empirical Understanding of Augmentation Strength

Based on the setting in Section 6.1, we conduct a series of experiments on a synthetic data set to verify the influence of the augmentation strength. Consider a binary classification task ($k = 2$) with the InfoNCE loss. Data are generated from two uniform distributions on a unit ball $\mathbb{S}^2$ in the 3-dimensional space (one center is $(0, 0, 1)$ and another is $(0, 0, -1)$). More details can be found in Appendix B.1.

As shown in Figure 5, we find that with the enhancement of the augmentation strength r, the accuracy of the classification tasks increases first and then falls to 50%. From the learned representations with different augmentation strengths, we find that when the augmentation strength is weak, the features of the intra-class samples are more isolated. When the augmentation strength is too strong, the features of the different classes are mixed. In both cases, it is almost impossible to obtain a linear classifier with good downstream performance.

To further understand the relationship between the strength of augmentations and the connectivity of augmentation graph, we visualize the augmentation graphs with different

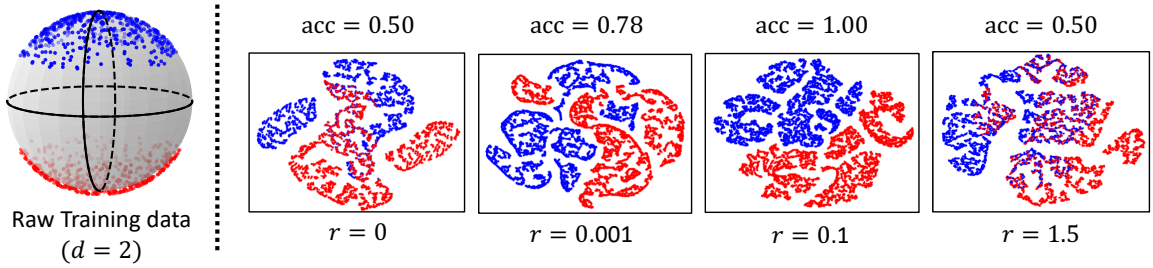
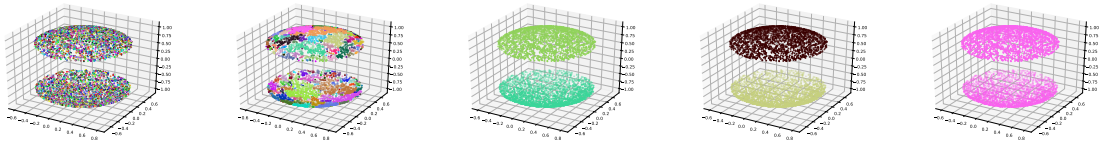


Figure 5: t-SNE visualization of features learned with different augmentation strength r on the random augmentation graph experiment. Each dot denotes a sample and its color denotes its class.

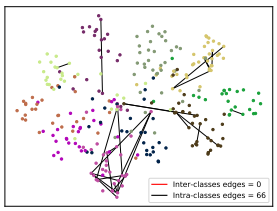


(a) $r = 0.0$, acc = 0.50, $D = \inf$. (b) $r = e^{-3}$, acc = 0.78, $D = \inf$. (c) $r = 0.08$, acc = 0.96, $D = 5$. (d) $r = 0.5$, acc = 1.00, $D = 2$. (e) $r = 1.5$, acc = 0.50, $D = 1$

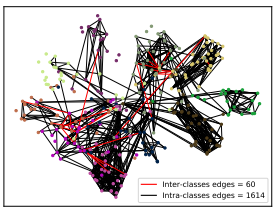
Figure 6: Visualization of the augmentation graph with different augmentation strength r on the synthetic data. Each color denotes a connected component. The number of the connected components and the maximal intra-class radius D decrease with the increase of the augmentation strength.

augmentation strengths on the above synthetic data set and present the maximal intra-class radius D of them in Figure 6. We observe that when augmentations are not applied on the data (Figure 6(a)), all the samples are isolated ($D = \inf$) and contrastive learning becomes a simple instance-level discriminative task. The degree of intra-class overlap aligns with the increase of the augmentation strength. At the point of the highest linear accuracy where $r = 0.5$ (Figure 6(d)), all the intra-class samples are closely connected ($D = 2$) while all the inter-class samples are separated, which satisfies Assumption 8. When the augmentation is too strong (Figure 6(e)), we find that the two subgraphs of the augmentation graph are connected, which means the inter-class samples may be wrongly aligned, and as a result, the linear accuracy dramatically decreases to 50%.

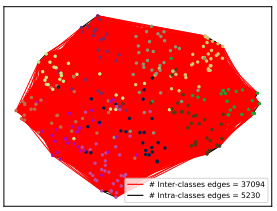
Besides the synthetic data set, we also verify the practicality of our theoretical results on the real-world data set CIFAR-10. While the augmentation graph is defined in terms of connections in the input space, directly measuring semantic similarity in this high-dimensional space presents significant challenges. To overcome this, we adopt a well-established surrogate metric by using pretrained representations to approximate the similarity of input images, following Dwibedi et al. (2021). Specifically, for characterizing the connectivity of two samples, we augment each sample 20 times and then compute their cosine similar-



(a) Almost no-overlap augmentation graph (r=(0.99,1), acc=0.25).



(b) Proper-overlap augmentation graph (r=(0.08,1), acc=0.75).



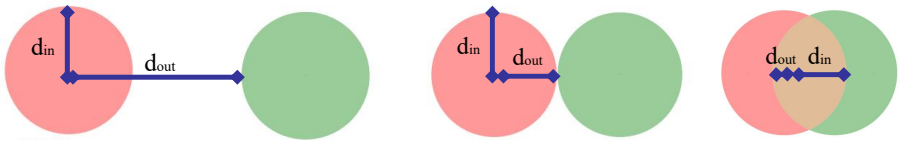
(c) Over-overlap augmentation graph (r=(0.01,0.03), acc=0.29).

Figure 7: The augmentation graph of CIFAR-10 with different strength r of RandomResizedCrop. We choose a random subset of test images and randomly augment each one 20 times. Then, we calculate the sample distance in the representation space as in prior work like FID (Heusel et al., 2017) and draw edges for image pairs whose smallest view distance is below a small threshold. Afterwards, we visualize the samples with t-SNE and color intra-class edges in **black** and inter-class edges in **blue** and report their frequencies.

ity. An edge is considered to exist between two samples if the cosine similarity between their augmented counterparts exceeds a threshold. We use the RandomResizedCrop as the data augmentation and regard the scale of the crop (a, b) as the augmentation strength r, *i.e.,* $r = (1 - a) + (1 - b)$. The connectivity of the augmentation graph on CIFAR-10 is visualized in Figure 7, we find that when the augmentation strength is too weak where the scale of the crop is (0.99,1) (Figure 7(a)), the linear accuracy is 25% and most samples are isolated. When we adopt appropriate augmentation strength (Figure 7(b)), the linear accuracy increases to 75% and at the same time, the edges between intra-class samples take 96.4% of all the edges. When the augmentation strength is too strong (Figure 7(c)), the linear accuracy falls down to 29% and the edges of the inter-class samples take 87.6% of the edges. The empirical findings in the real-world data sets are quite close to the results on the synthetic data sets. The above results further verify that the connectivity of augmentation graph decides the learned representation of contrastive learning, *i.e.,* too weak or too strong data augmentations will hurt its downstream performance.

7. Applications of Augmentation Overlap Theory

Although contrastive learning has received impressive empirical success in many downstream tasks, how to quickly evaluate the quality of the learned representation is still under-explored. In practice, the most common method is to train a linear classifier following the pretrained encoder, *i.e.,* linear evaluation (Chen et al., 2020; He et al., 2020; Grill et al., 2020). However, linear evaluation needs supervised label information and extra training time which are quite expensive for large-scale real-world data sets. Recently, some works try to evaluate the performance of contrastive learning without supervised labels, for



(a) $d_{out} > d_{in}$ (b) $d_{out} = d_{in}$ (c) $d_{out} < d_{in}$

Figure 8: Measuring support overlap with the help of the distance between different augmented samples.

example, Reed et al. (2021) find rotation prediction accuracy of the learned representation is strongly related to the classification accuracy in downstream tasks. However, their method still needs additional training time for the rotation classifier. In this section, we propose an unsupervised metric based on our augmentation overlap theory which can evaluate the performance of contrastive learning with almost no additional computational cost.

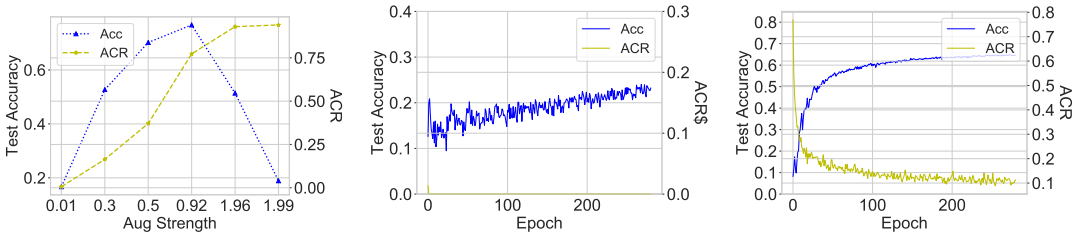
7.1 An Unsupervised Metric for Representation Evaluation

Based on our augmentation overlap theory, generating overlapped views of intra-class samples is a critical step of contrastive learning and the connectivity of the augmentation graph is strongly related to the downstream performance. Intuitively, when justifying whether two samples are adjacent in the augmentation graph, we need to enumerate all their augmented views to find whether they share an augmented view. However, going through the augmentation space needs huge computational costs which can hardly be implemented in practice, especially for the real-world data sets. To be specific, common contrastive learning paradigms adopt different kinds of data augmentations, including RandomResizedCrop, ColorJitter, GrayScale, and GaussianBlur. Each data augmentation has different continuous parameters, for example, Colorjitter controls the brightness, contrast, saturation, and hue of the images, which all can be random positive values. As the result, it is impractical to search the space of data augmentations with current computing resources.

Therefore, we propose an approximation metric to measure the connectivity of augmentation graph with the help of the nearest neighbours around the sample. For each sample $x_i \in \mathcal{D}_u$, we random augment it for C times and get a support set $\hat{\mathcal{D}}_u = \{x_{ij}, 0 \le i \le N, 0 \le j \le C\}$. We define $d_{in}(x_{ij}, f) = \max\{\|f(x_{ij}) - f(x_{ip})\|, 0 \le p \le C\}$ as the maximal distance of the augmented samples of the same anchor and $d_{out}(x_{ij}, f) = \min\{\|f(x_{ij}) - f(x_{lp})\|, 0 \le i, j \le N, 0 \le p \le C, i \ne l\}$ as the minimal distance of the augmented samples of different anchors in the feature space. As shown in Figure 8, we find that when the different samples have support overlap, d_{in} is larger than d_{out}. Therefore we can measure the degree of augmentation overlap by comparing d_{in} and d_{out}. Following that, we define a Average Confusion Ratio (ACR) as the ratio of the intra-anchor distance is larger then the inter-anchor distance:

$$\mathrm{ACR}(f) = \mathbb{E}_{x_{ij} \in \hat{\mathcal{D}}_u} \mathbb{I}(d_{out}(x_{ij}, f) \le d_{in}(x_{ij}, f)). \tag{21}$$

Note that when $C \to \infty$ and the support set includes all anchors, ACR is exactly the ratio of the anchors that have overlapped views with other anchors. Intuitively, ACR



(a) ACR *v.s.* aug-strength r. (b) ACR while training (r=0.01). (c) ACR while training (r=0.92).

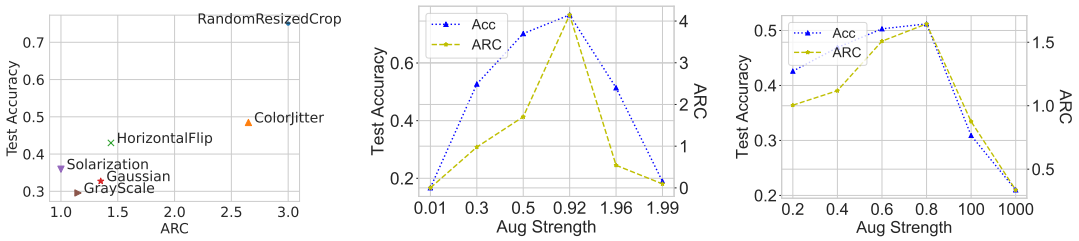
Figure 9: (a) Average Confusion Rate (ACR) and downstream accuracy *v.s.* different augmentation strength (before training). (b,c): ACR and downstream accuracy while training.

approximately measures the ratio that augmented views of different anchors can be closer than the augmented views of the same sample, which means that different samples generate the overlapped views. To show the performance of ACR on real-world data sets, we first take the RandomResizedCrop operator with different strengths on CIFAR-10 as an example. The augmentation strength r of RandomResizedCrop is defined as the scale range $[a, b]$ of crop, *i.e.,* $r = (1 - a) + (1 - b)$. As shown in Figure 9(a), we find that ACR increases with stronger augmentations. The empirical results verify that the metric ACR is a closed approximator of the degree of the connectivity of the augmentation graph. However, Figure 9(a) also shows that the downstream accuracy increases first and then falls down while ACR keeps increasing with stronger data augmentations, which is consistent with our analysis that when augmentation strength is too strong, there occurs over-overlap. As a result, only measuring ACR can not be a satisfying evaluation method. Figure 9(b) and Figure 9(c) show the change of ACR during the training process of contrastive learning. We find that when the augmentation strength is too weak, contrastive learning becomes a quite simple task, the initial ACR is low and it decreases to 0 in a short period, and when we adopt appropriate augmentation strength, the initial ACR is high and it will decrease mildly. Inspired by the empirical findings, we think the change process of ACR may be more strongly related to the downstream performance of contrastive learning, so we propose a new metric to estimate the downstream accuracy of contrastive learning without label messages, named ARC (Average Relative Confusion):

$$\mathrm{ARC} = \frac{1 - \mathrm{ACR}(f_{\mathrm{final}})}{1 - \mathrm{ACR}(f_{\mathrm{init}})}. \tag{22}$$

To evaluate the effectiveness of ARC, we conduct comprehensive experiments on CIFAR-10 to demonstrate the relationship between ARC and downstream performance on the representations trained with different data augmentation and different strengths. The details can be found in Appendix B.2.

Different kinds of data augmentations. We test 6 kinds of augmentations used in SimCLR (Chen et al., 2020), *i.e.,* RandomResizedCrop, ColorJitter, GrayScale, HorizontalFlip, GaussianBlur and Solarization. Each augmentation is singly applied with the default parameter in SimCLR. The results are presented in Figure 10(a), which shows that



(a) Different augmentations (b) RandomResizedCrop (c) Colorjitter

Figure 10: Average Relative Confusion (ARC) and downstream accuracy *v.s.* different data augmentations adopted in SimCLR with different strengths.

ARC aligns well with the linear accuracy of models trained by different kinds of augmentations. We also find that RandomResizedCrop and ColorJitter are the most two powerful data augmentations used in SimCLR, so we then focus on these two augmentations for further analysis.

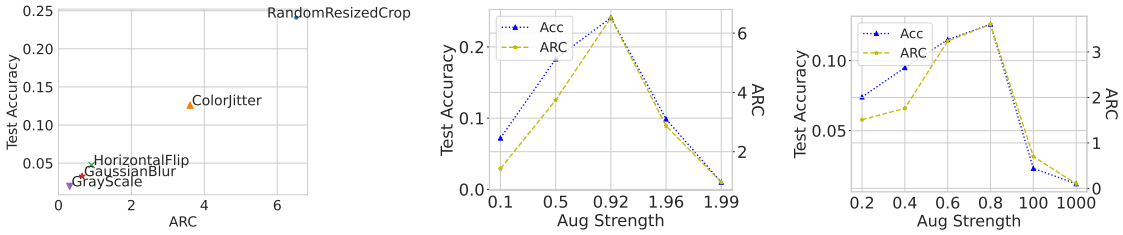
RandomResizedCrop with different strengths. The strength r of RandomResizedCrop is still defined as the scale $[a, b]$ of crop, *i.e.,* $r = (1-a)+(1-b)$. From Figure 10(b), we find that the ARC curve and the linear accuracy curve have almost the same pace, *i.e.,* increasing first and then falling down, which indicates that our ARC metric has a close relationship with downstream performance.

ColorJitter with different strengths. The strength of ColorJitter is directly controlled by four parameters, *i.e.,* brightness, contrast, saturation, and hue. Brightness, contrast, and saturation can be any positive values while hue is a positive value that is not larger than 0.5. So we set the parameter of ColorJitter as $(brightness, contrast, saturation, hue) = (r, r, r, \min(0.5, 0.25 * r))$. From Figure 10(c), the curves of ARC and linear accuracy perform at almost the same pace again, which implies that our ARC metric is a good measure for the downstream performance evaluation.

ARC on the large-scale data set. Besides CIFAR-10, we also evaluate the effectiveness of ARC on the large-scale data set ImageNet. The experiments are conducted on ImageNet with various data augmentations and various strengths including 1) different types of augmentations with default parameters in SimCLR, 2) RandomResizedCrop with different strengths, 3) Colorjitter with different strengths. As shown in Figure 11, the proposed ARC aligns well with the linear accuracy, which indicates that the ARC metric also performs well on large-scale data sets.

7.2 Further Analysis of the Proposed Metric ARC

The proposed metric ARC is based on ACR, while in practice ACR may have outliers to mislead its calculation. For example in Figure 12, x_{i1} and x_{i2} are two augmented views of a car while x_{j1} and x_{j2} are two augmented views of a horse. Due to aggressive augmentations, x_{i2} is more similar to x_{j2} than x_{i1}. So the inter-anchor distance $d_{out}(x_{i1})$ is smaller than the intra-anchor distance $d_{in}(x_{i1})$, which results in the wrong increase on ACR, because



(a) Different augmentations (b) RandomResizedCrop (c) ColorJitter

Figure 11: Average Relative Confusion (ARC) and downstream accuracy *v.s.* different data augmentations adopted in SimCLR with different strengths on ImageNet.

Figure 12: Outliers of data augmentations will mislead the calculation of ARC defined in Section 7.1 ($\|(x_{i2} - x_{i1}\| > \|x_{i2} - x_{j2}\|$).

higher ACR in fact indicates the higher overlap degree while here the car and horse have no semantic overlap.

Generalized ARC. To avoid these outliers, we generalize the definition of ACR. For an augmented sample x_{ij}, we first compute its distance in the features space to the augmented views of any anchor sample x_l: $\mathcal{Q}(x_{ij}, x_l, f) = \{\|f(x_{ij}) - f(x_{lp})\|^2, 0 \leq p \leq C\}$. Then we respectively use statistic $\mathcal{A}_1$ to compute the distance of the augmented samples from the same anchor, *i.e.,* $d_{in}(x_{ij}, \mathcal{A}_1, f) = \mathcal{A}_1(\mathcal{Q}(x_{ij}, x_i, f))$ and use statistic $\mathcal{A}_2$ to compute the distance of the augmented samples from different anchors, *i.e.,* $d_{out}(x_{ij}, \mathcal{A}_2, f) = \{\mathcal{A}_2(\mathcal{Q}(x_{ij}, x_l, f)), l \neq i\}$. Finally we compare $d_{in}(x_{ij}, \mathcal{A}_1, f)$ and the k-smallest distance $d_{out-k}(x_{ij}, \mathcal{A}_2, f, k)$ of $d_{out}(x_{ij}, \mathcal{A}_2)$. The generalized ACR is defined as

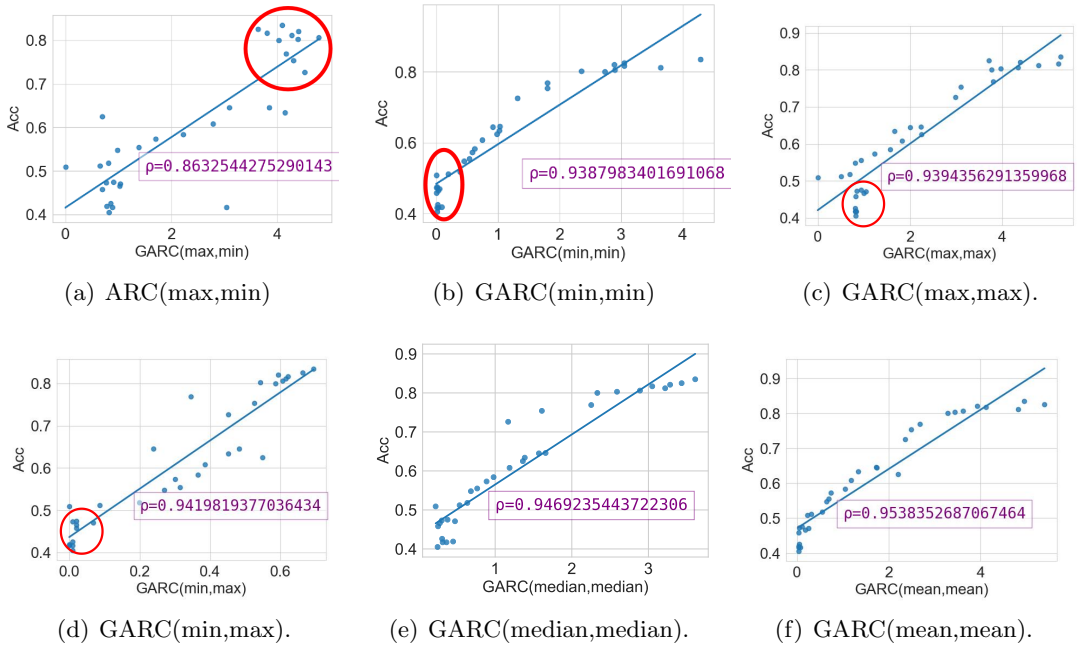
$$\text{GACR}(f, \mathcal{A}_1, \mathcal{A}_2, k) = \mathbb{E}_{x_{ij} \in \hat{\mathcal{D}}_u} \mathbb{I}\left(d_{out-k}(x_{ij}, \mathcal{A}_2, f, k) \leq d_{in}(x_{ij}, \mathcal{A}_1, f)\right) \tag{23}$$

Similarly, based on the generalized ACR, the generalized ARC is defined as

$$\text{GARC}(\mathcal{A}_1, \mathcal{A}_2, k) = \frac{1 - \text{GACR}(f_{\text{final}}, \mathcal{A}_1, \mathcal{A}_2, k)}{1 - \text{GACR}(f_{\text{init}}, \mathcal{A}_1, \mathcal{A}_2, k)}. \tag{24}$$

Note that when $\mathcal{A}_1$ is set to the maximal value, $\mathcal{A}_2$ is set to the minimum value, and k is set to 1, GARC degrades to ARC discussed in Section 7.1.

Variants of GARC. Following the definition of GARC, here we consider its six different variants, *i.e.,* GARC(min, min), GARC(max,max), GARC(max, min), GARC(min,



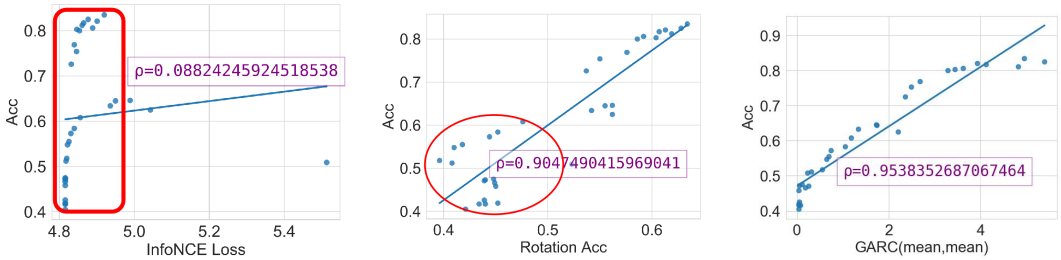
(a) ARC(max,min) (b) GARC(min,min) (c) GARC(max,max).

(d) GARC(min,max). (e) GARC(median,median). (f) GARC(mean,mean).

Figure 13: The relationship between linear evaluation accuracy and different variants of GARC.

max), GARC(median, median), and GARC(mean, mean). The experiments are conducted on CIFAR-10 with different data augmentations like 1) RandomResizedCrop with different scales of the crop, 2) ColorJitter with different parameters (contrast, brightness, saturation, hue), and 3) composition of all augmentations used in SimCLR with different parameters. The relationship between the downstream accuracy of these models trained with different augmentations and our proposed GARC metric is shown in Figure 13. We find that all the six variants of GARC have a close relationship with the linear accuracy of learned representations. Note that GARC(max,min) is the ARC defined in Section 7.1. As shown in Figure13(a), the performance of GARC(max,min) decreases when the linear accuracy is high, which is consistent with our findings that aggressive data augmentations will generate outliers and mislead its calculation (marked by red). While GARC(min,min), GARC(max,max), and GARC(min,max) can solve this issue and measures the representations trained with aggressive augmentations more accurately. However, they do not perform well on the representations with low linear accuracy (marked by red). This is because they will ignore some overlapped views, especially with weak data augmentations. In contrast, GARC(median,median) and GARC(mean, mean) can get rid of the influence of these outliers and keeps a strong relationship with the downstream performance of various representations.

Comparison to other metrics. To further evaluate the performance of our proposed metric, we compare GARC with other unsupervised evaluation metrics such as rotation prediction accuracy (Reed et al., 2021). They train a linear classifier following the pretrained



(a) InfoNCE (b) Rotation (c) GARC(mean,mean).

Figure 14: The relationship between linear evaluation accuracy and different unsupervised metrics.

encoder to predict the rotation angle of the samples and find that the prediction accuracy is strongly related to the downstream classification accuracy. We compare the three different unsupervised metrics, *i.e.,* GARC, rotation accuracy, and InfoNCE loss, in Figure 14 on CIFAR-10 with the same settings above. The details can be found in Appendix B.3. Figure 14 shows that the lower InfoNCE loss does not imply better downstream performance while both our proposed GARC and rotation accuracy are strongly related to the downstream classification accuracy. As the Pearson correlation coefficient ρ (Pearson, 1895) between GARC and linear accuracy is 0.954 while rotation is only 0.905, our proposed metric outperforms rotation accuracy and has a closer relationship to the performance of downstream classification. To be specific, in Figure 14 (a), InfoNCE loss does not change a lot with different augmentations (4.8 to 5.0) in most cases while the linear accuracy will increase from 40% to 80%, thus InfoNCE is not a good indicator for the downstream performance. Figure 14 (b) shows that rotation accuracy performs better in well-clustered representations (right upper corner of the figure) while it performs poorly when the linear accuracy is lower (left lower corner of the figure). Conversely, our proposed GARC keeps a strong relationship with downstream performance of various representations as shown in Figure 14 (c). For the computational cost, compared to linear evaluation, rotation accuracy can be obtained without access to supervised labels. However, evaluating rotation accuracy still needs an extra linear classifier and expensive computation costs. In contrast, our proposed GARC directly uses the statistical information of the pretrained models with little additional overheads. In summary, GARC is a well-performed unsupervised model selection metric that is strongly related to the classification performance of the representations and can be obtained with negligible costs.

8. Conclusion

In this paper, we proposed a new understanding of contrastive learning through a revisiting of the role of data augmentations. In particular, we notice the aggressive data augmentations applied in contrastive learning can significantly increase the augmentation overlap between intra-class samples, and as a result, by aligning positive samples, we can also cluster inter-class samples together. Based on this insight, we develop a new augmentation

overlap theory that could guarantee the downstream performance of contrastive learning without relying on the conditional independence assumption. With this perspective, we also characterize how different augmentation strength affects downstream performance on both random graphs and real-world data sets. Last but not least, we develop a new surrogate metric for evaluating the learned representations of contrastive learning without labels and show that it aligns well with downstream performance. Overall, we believe that we pave a new way for understanding contrastive learning with insights on the designing of contrastive learning methods and evaluation metrics.

Acknowledgments

Yisen Wang was supported by National Key R&D Program of China (2022ZD0160300), Beijing Natural Science Foundation (L257007), Beijing Major Science and Technology Project under Contract no. Z251100008425006, National Natural Science Foundation of China (92370129, 62376010), Beijing Nova Program (20230484344, 20240484642), and State Key Laboratory of General Artificial Intelligence.

Appendix A. Proofs

A.1 Proof of Lemma 2

Proof First, we have

$$\begin{aligned}&\mathbb{E}_{p(x,z_i)}\left[\log\frac{1}{M}\sum_{i=1}^{M}\exp(f(x)^\top g(z_i)) - \log\mathbb{E}_{p(z_i)}\exp(f(x)^\top g(z_i))\right]\\ \leq& e\mathbb{E}_{p(x,z_i)}\left[\frac{1}{M}\sum_{i=1}^{M}\exp(f(x)^\top g(z_i)) - \mathbb{E}_{p(z_i)}\exp(f(x)^\top g(z_i))\right] = \frac{e}{\sqrt{M}},\end{aligned}$$

where the first inequality follows the Intermediate Value Theorem and e (the natural number) is the upper bound of the absolute derivative of log between two points when $|f(x)^\top g(z_i)| \leq 1$. And the second inequality uses the following inequality, given the bounded support of $\exp(f(x)^\top g(z_i))$ as following: for i.i.d random variables Y_i with mean $\mathbb{E}Y$ and bounded variance $\sigma_Y^2 < \sigma^2$, we have:

$$\begin{aligned}&\mathbb{E}\left[\left|\frac{1}{M}\sum_{i=1}^{M}Y_i - \mathbb{E}Y_i\right|\right]\\ &\leq\sqrt{\mathbb{E}\left[\left|\frac{1}{M}\sum_{i=1}^{M}Y_i - \mathbb{E}Y_i\right|^2\right]}\\ &=\frac{\sigma}{\sqrt{M}},\end{aligned}$$

where the first inequality follows Jensen's inequality and the second equation follows the property that Y_i is i.i.d random variables. Here, we set $Y_i = \exp(f(x)^\top g(z_i))$. As $g(z)$ is the mean classifier and $f(x)$ is normalized, $|f(x)^\top g(z_i)| \leq 1$, $|Y_i| \leq e$. With Popoviciu's inequality, Y_i has bounded variance $(e)^2$.

■

A.2 Proof of Theorem 3

We will prove the upper and lower bounds separately as follows.

A.2.1 The Upper Bound

We first prove that the mean CE loss can be upper bounded by the InfoNCE loss with Lemma 2.

Proof Denote $p(x, x^+, y)$ as the joint distribution of the positive pairs x, x^+ and the label y. Denote the M independently negative smaples as $\{x_i^-\}_{i=1}^M$. Denote μ_y as the center of features of class y, $y = 1, \ldots, K$. Then we have the following lower bounds of the InfoNCE

loss:

$$
\begin{aligned}
\bar{\mathcal{L}}_{\text{contr}}(f) &= -\mathbb{E}_{p(x,x^+)} f(x)^\top f(x^+) + \mathbb{E}_{p(x)} \mathbb{E}_{p(x_i^-)} \log \frac{1}{M} \sum_{i=1}^{M} \exp(f(x)^\top f(x_i^-)) \\
&\overset{(1)}{\geq} -\mathbb{E}_{p(x,x^+)} f(x)^\top f(x^+) + \mathbb{E}_{p(x)} \log \frac{1}{M} \mathbb{E}_{p(x_i^-)} \sum_{i=1}^{M} \exp(f(x)^\top f(x_i^-)) - \frac{e}{\sqrt{M}} \\
&= -\mathbb{E}_{p(x,x^+)} f(x)^\top f(x^+) + \mathbb{E}_{p(x)} \log \mathbb{E}_{p(x^-)} \exp(f(x)^\top f(x^-)) - \frac{e}{\sqrt{M}} \\
&= -\mathbb{E}_{p(x,x^+,y)} f(x)^\top f(x^+) + \mathbb{E}_{p(x)} \log \mathbb{E}_{p(y^-)} \mathbb{E}_{p(x^-|y^-)} \exp(f(x)^\top f(x^-)) - \frac{e}{\sqrt{M}} \\
&\overset{(2)}{\geq} -\mathbb{E}_{p(x,x^+,y)} f(x)^\top f(x^+) + \mathbb{E}_{p(x)} \log \mathbb{E}_{p(y^-)} \exp(\mathbb{E}_{p(x^-|y^-)} \left[f(x)^\top f(x^-) \right]) - \frac{e}{\sqrt{M}} \\
&\overset{(3)}{=} -\mathbb{E}_{p(x,y)} f(x)^\top \mu_y + \mathbb{E}_{p(x)} \log \mathbb{E}_{p(y^-)} \exp(\mathbb{E}_{p(x^-|y^-)} \left[f(x)^\top f(x^-) \right]) - \frac{e}{\sqrt{M}} \\
&= -\mathbb{E}_{p(x,y)} f(x)^\top \mu_y + \mathbb{E}_{p(x)} \log \mathbb{E}_{p(y^-)} \exp(f(x)^\top \mu_{y^-}) - \frac{e}{\sqrt{M}} \\
&= -\mathbb{E}_{p(x,y)} f(x)^\top \mu_y + \mathbb{E}_{p(x)} \log \frac{1}{K} \sum_{k=1}^{K} \exp(f(x)^\top \mu_k) - \frac{e}{\sqrt{M}} \\
&= \mathbb{E}_{p(x,y)} \big[-f(x)^\top \mu_y + \log \frac{1}{K} \sum_{k=1}^{K} \exp(f(x)^\top \mu_k) \big] - \frac{e}{\sqrt{M}} \\
&= \bar{\mathcal{L}}_{\text{mCE}}(f) - \frac{e}{\sqrt{M}},
\end{aligned}
$$

which is equivalent to our desired results. In the proof above, (1) follows Lemma 2; (2) follows the Jensen's inequality for the convex function $\exp(\cdot)$; (3) follows the conditional independence assumption. ■

A.2.2 The Lower Bound

In this part, we further show a lower bound on the downstream performance.

Lemma 16 (Budimir et al. (2000) Corollary 3.5 (restated)) *Let $g : \mathbb{R}^m \to \mathbb{R}$ be a differentiable convex mapping and $z \in \mathbb{R}^m$. Suppose that g is L- smooth with the constant*

$L > 0$, i.e., $\forall x, y \in \mathbb{R}^m, \|\nabla g(x) - \nabla g(y)\| \leq L\|x - y\|$. Then we have

$$\begin{aligned}
0 &\leq \mathbb{E}_{p(z)} g(z) - g\left(\mathbb{E}_{p(z)} z\right) \\
&\leq L\left[\mathbb{E}_{p(z)}\|z\|^2 - \|\mathbb{E}_{p(z)} z\|^2\right] \\
&= L[\sum_{j=1}^{m} \mathbb{E}_{p(z)}\|z^{(j)}\|^2 - \sum_{j=1}^{m} \|\mathbb{E}_{p(z)} z^{(j)}\|^2] \\
&= L[\sum_{j=1}^{m} \mathbb{E}_{p(z^{(j)})}\|z^{(j)}\|^2 - \sum_{j=1}^{m} \|\mathbb{E}_{p(z^{(j)})} z^{(j)}\|^2] \\
&= L\sum_{j=1}^{m} \mathrm{Var}(z^{(j)})
\end{aligned} \tag{25}$$

where $x^{(j)}$ denotes the j-th dimension of x.

With the lemma above, we can derive the lower bound of the downstream performance.

Proof Similar to the proof of the upper bound, we have

$$\begin{aligned}
&\bar{\mathcal{L}}_{\mathrm{mCE}}(f) = -\mathbb{E}_{p(x,y)} f(x)^\top \mu_y + \mathbb{E}_{p(x)} \log \frac{1}{K} \sum_{i=1}^{K} \exp(f(x)^\top \mu_i) \\
=& -\mathbb{E}_{p(x,y)} f(x)^\top \mu_y + \mathbb{E}_{p(x)} \log \frac{1}{K} \sum_{i=1}^{K} \exp(f(x)^\top \mu_i) \\
=& -\mathbb{E}_{p(x,y)} f(x)^\top \mu_y + \mathbb{E}_{p(x)} \log \mathbb{E}_{p(y_i^-)} \exp(f(x)^\top \mu_{y_i}) \\
\overset{(1)}{\geq}& -\mathbb{E}_{p(x,y)} f(x)^\top \mu_y + \mathbb{E}_{p(x)} \mathbb{E}_{p(y_i^-)} \log \frac{1}{M} \sum_{i=1}^{M} \exp(f(x)^\top \mu_{y_i}) - \frac{e}{\sqrt{M}} \\
\overset{(2)}{=}& -\mathbb{E}_{p(x,x^+)} f(x)^\top f(x^+) + \mathbb{E}_{p(x)} \mathbb{E}_{p(y_i^-)} \log \frac{1}{M} \sum_{i=1}^{M} \exp(\mathbb{E}_{p(x_i^-|y_i^-)} f(x)^\top f(x_i^-)) - \frac{e}{\sqrt{M}} \\
\overset{(3)}{\geq}& -\mathbb{E}_{p(x,x^+)} f(x)^\top f(x^+) \\
&+ \mathbb{E}_{p(x)} \mathbb{E}_{p(y_i^-)} \mathbb{E}_{p(x_i^-|y)} \left[\log \frac{1}{M} \sum_{i=1}^{M} \exp(f(x)^\top f(x^-))\right] - \frac{1}{2}\sum_{j=1}^{m} \mathrm{Var}(f_j(x^-) \mid y) - \frac{e}{\sqrt{M}} \\
=& -\mathbb{E}_{p(x,x^+)} \left[f(x)^\top f(x^+) + \mathbb{E}_{p(x_i^-)} \log \sum_{i=1}^{M} \exp(f(x)^\top f(x^-))\right] \\
&\qquad - \frac{1}{2}\sum_{j=1}^{m} \mathrm{Var}(f_j(x^-) \mid y) - \frac{e}{\sqrt{M}} \\
=& \bar{\mathcal{L}}_{\mathrm{contr}}(f) - \frac{1}{2}\mathrm{Var}(f(x) \mid y) - \frac{e}{\sqrt{M}},
\end{aligned}$$

which is our desired result. In the proof, (1) we adopt the Monte Carlo estimate with M samples from $p(y)$ and bound the approximation error with Lemma 2; (2) follows the conditional

independence assumption; (3) we first show that the convex function logsumexp is L-smooth as a function of $f(x_j^-)$ in our scenario. Because $\|f(X)\| \leq 1$, we have $\forall f(x_{j_1}), f(x_{j_2}) \in \mathbb{R}^m$, the following bound on the difference of their gradients holds

$$
\begin{aligned}
&\left\| \frac{\partial \log[\exp(f(x)^\top f(x_{j_1}^-) + \sum_{i\neq j} \exp(f(x)^\top f(x_i^-)))]}{\partial f(x_{j_1}^-)} - \frac{\partial \log[\exp(f(x)^\top f(x_{j_2}^-) + \sum_{i\neq j} \exp(f(x)^\top f(x_i^-)))]}{\partial f(x_{j_2}^-)} \right\| \\
=&\left\| \left(\frac{\exp(f(x)^\top f(x_{j_1}^-))}{\exp(f(x)^\top f(x_{j_1}^-) + \sum_{i\neq j} \exp(f(x_i^-)))} - \frac{\exp(f(x)^\top f(x_{j_2}^-))}{\exp(f(x)^\top f(x_{j_2}^-) + \sum_{i\neq j} \exp(f(x)^\top f(x_i^-)))} \right) f(x) \right\| \\
\leq&\left| \frac{(\sum_{i\neq j} \exp(f(x)^\top f(x_i^-)) \exp(f(x_{j_1}^-)) - \sum_{i\neq j} \exp(f(x)^\top f(x_i^-)) \exp(f(x)^\top f(x_{j_2}^-))}{(\exp(f(x)^\top f(x_{j_1}^-)) + \sum_{i\neq j} \exp(f(x)^\top f(x_i^-)))(\exp(f(x)^\top f(x_{j_2}^-)) + \sum_{i\neq j} \exp(f(x)^\top f(x_i^-)))} \right| \cdot \|f(x)\| \\
\leq&\|f(x)\| \leq \frac{1}{2} \left\| f(x_{j_1}^-) - f(x_{j_2}^-) \right\|.
\end{aligned}
$$

So here the logsumexp is L-smooth for $L = \frac{1}{2}$. Then, we can apply the reversed Jensen's inequality in Lemma 16.

■

A.3 Proof of Theorem 5 with Labeling Error

Similar to Theorem 3, we will prove the upper and lower bounds separately as follows.

A.3.1 The Upper Bound

With Lemma 2, we show that upper bounds the mean CE loss with the InfoNCE loss.

Proof We denote $p(x, x^+, y)$ as the joint distribution of the positive pairs x, x^+ and the label y of x. Denote the M independently negative smaples as $\{x_i^-\}_{i=1}^M$. Denote μ_y as the

center of features of class y, $y = 1, \ldots, K$. Similar to the proof of Theorem 3, we have

$$
\begin{aligned}
&\bar{\mathcal{L}}_{\text{contr}}(f) = -\mathbb{E}_{p(x,x^+)} f(x)^\top f(x^+) + \mathbb{E}_{p(x)} \mathbb{E}_{p(x_i^-)} \log \frac{1}{M} \sum_{i=1}^{M} \exp(f(x)^\top f(x_i^-)) \\
\overset{(1)}{\geq} & - \mathbb{E}_{p(x,x^+,y)} f(x)^\top (\mu_y + f(x^+) - \mu_y) + \mathbb{E}_{p(x)} \log \mathbb{E}_{p(y^-)} \exp(\mathbb{E}_{p(x^-|y^-)} \left[f(x)^\top f(x^-) \right]) - \frac{e}{\sqrt{M}} \\
= & - \mathbb{E}_{p(x,x^+,y)} [f(x)^\top \mu_y + f(x)^\top (f(x^+) - \mu_y)] + \mathbb{E}_{p(x)} \log \mathbb{E}_{p(y^-)} \exp(f(x)^\top \mu_{y^-}) - \frac{e}{\sqrt{M}} \\
\overset{(2)}{\geq} & - \mathbb{E}_{p(x,x^+,y)} \left[f(x)^\top \mu_y + \|(f(x^+) - \mu_y)\| \right] + \mathbb{E}_{p(x)} \log \mathbb{E}_{p(y^-)} \exp(f(x)^\top \mu_{y^-}) - \frac{e}{\sqrt{M}} \\
\overset{(3)}{\geq} & - \mathbb{E}_{p(x,y)} f(x)^\top \mu_y - \sqrt{\mathbb{E}_{p(x,y)} \|f(x^+) - \mu_y\|^2} + \mathbb{E}_{p(x)} \log \mathbb{E}_{p(y^-)} \exp(f(x)^\top \mu_{y^-}) - \frac{e}{\sqrt{M}} \\
= & - \mathbb{E}_{p(x,y)} f(x)^\top \mu_y - \sqrt{\mathbb{E}_{p(x,y)} \|f(x^+) - \mu_{y^+} + \mu_{y^+} - \mu_y\|^2} + \mathbb{E}_{p(x)} \log \frac{1}{K} \sum_{k=1}^{K} \exp(f(x)^\top \mu_k) - \frac{e}{\sqrt{M}} \\
\geq & \mathbb{E}_{p(x,y)} \big[- f(x)^\top \mu_y + \log \frac{1}{K} \sum_{k=1}^{K} \exp(f(x)^\top \mu_k) \big] - 2\sqrt{\text{Var}(f(x^+) \mid y(x^+))} - 2\sqrt{\|\mu_{y^+} - \mu_y\|^2} - \frac{e}{\sqrt{M}} \\
\overset{(4)}{\geq} & \mathbb{E}_{p(x,y)} \big[- f(x)^\top \mu_y + \log \frac{1}{K} \sum_{k=1}^{K} \exp(f(x)^\top \mu_k) \big] - 2\sqrt{\text{Var}(f(x^+) \mid y(x^+))} - 4\sqrt{\alpha} - \frac{e}{\sqrt{M}} \\
= & \bar{\mathcal{L}}_{\text{mCE}}(f) - 2\sqrt{\text{Var}(f(x^+) \mid y^+)} - 4\sqrt{\alpha} - \frac{e}{\sqrt{M}} \\
= & \bar{\mathcal{L}}_{\text{mCE}}(f) - 2\sqrt{\text{Var}(f(x) \mid y)} - 4\sqrt{\alpha} - \frac{e}{\sqrt{M}},
\end{aligned}
$$

which is equivalent to our desired results. In the proof above, (1) follows Lemma 2, (2) follows from the fact that because $f(x) \in \mathbb{S}^{m-1}$, we have

$$
f(x)^\top (f(x^+) - \mu_y) \leq \left(\frac{f(x^+) - \mu_y}{\|f(x^+) - \mu_y\|} \right)^\top (f(x^+) - \mu_y) = \|f(x^+) - \mu_y\|; \tag{26}
$$

(3) follows the Jensen inequality and the fact that because $p(x, x^+) = p(x^+, x)$ holds, x, x^+ have the same marginal distribution and (4) follows assumption 4 and $f(x)$ is normalized. We note that when we throw the conditional independence assumption, we do not have $\mathbb{E}_{p(x,x^+)} f(x)^\top = \mathbb{E}_{p(x,y)} f(x)^\top \mu_y$ and there exists an additional variance term $\sqrt{\text{Var}(f(x^+) \mid y^+)}$. ■

A.3.2 The Lower Bound

With the Lemma 16, we can derive the lower bound of the downstream performance.

Proof Similar to the proof of the upper bound, we have

$$
\begin{aligned}
&\bar{\mathcal{L}}_{\mathrm{mCE}}(f) = -\mathbb{E}_{p(x,y)} f(x)^\top \mu_y + \mathbb{E}_{p(x)} \log \frac{1}{K} \sum_{i=1}^{K} \exp(f(x)^\top \mu_i) \\
=& -\mathbb{E}_{p(x,y)} f(x)^\top \mu_y + \mathbb{E}_{p(x)} \log \mathbb{E}_{p(y_i^-)} \exp(f(x)^\top \mu_{y_i}) \\
\overset{(1)}{\geq}& -\mathbb{E}_{p(x,y)} [f(x)^\top f(x^+) + f(x)^\top (\mu_y - f(x^+))] + \mathbb{E}_{p(x)} \mathbb{E}_{p(y_i^-)} \log \frac{1}{M} \sum_{i=1}^{M} \exp(f(x)^\top \mu_{y_i}) - \frac{e}{\sqrt{M}} \\
\overset{(2)}{\geq}& -\mathbb{E}_{p(x,x^+)} f(x)^\top f(x^+) - \mathbb{E}_{p(x^+,y)} \| f(x^+)^\top - \mu_y \| \\
&\qquad + \mathbb{E}_{p(x)} \mathbb{E}_{p(y_i^-)} \log \frac{1}{M} \sum_{i=1}^{M} \exp(\mathbb{E}_{p(x_i^-|y_i^-)} f(x)^\top f(x_i^-)) - \frac{e}{\sqrt{M}} \\
\overset{(3)}{\geq}& -\mathbb{E}_{p(x,x^+)} f(x)^\top f(x^+) - 2\sqrt{\mathrm{Var}(f(x^+) \mid y^+)} - 4\sqrt{\alpha} \\
&+ \mathbb{E}_{p(x)} \mathbb{E}_{p(y_i^-)} \mathbb{E}_{p(x_i^-|y)} \left[\log \frac{1}{M} \sum_{i=1}^{M} \exp(f(x)^\top f(x^-)) \right] - \frac{1}{2} \sum_{j=1}^{m} \mathrm{Var}(f_j(x^-) \mid y) - \frac{e}{\sqrt{M}} \\
=& \bar{\mathcal{L}}_{\mathrm{contr}}(f) - 2\sqrt{\mathrm{Var}(f(x) \mid y)} - \frac{1}{2} \mathrm{Var}(f(x) \mid y) - \frac{e}{\sqrt{M}} - -4\sqrt{\alpha},
\end{aligned}
$$

which is our desired result. In the proof, (1) we adopt the Monte Carlo estimate with M samples from $p(y)$ and bound the approximation error with Lemma 2; (2) follows the same deduction in the upper bound; (3) the second term is derived following the Jensen inequality for the alignment term. As for the third term, we prove that the convex function logsumexp is L-smooth as a function of $f(x_j^-)$ in the proof of Theorem 3. Then, we can apply the reversed Jensen's inequality in Lemma 16. Similar to the upper bound, when we throw conditional independence assumption, the additional variance term $\sqrt{\mathrm{Var}(f(x) \mid y)}$ arises.

■

A.4 Proof of Proposition 6

Proof We only need to give a counterexample that satisfies the desired classification accuracy. We consider the case where any pair of samples from $\{x_i\}_{i=1}^N$ will not be aligned, which is easily achieved if we adopt a small enough data augmentation. In this scenario, the perfect alignment of positive samples (x_i, x_i^+) could have no effect on the other samples. Therefore, when the features $\{f(x_i)\}_{i=1}^N$ are uniformly distributed in $\mathbb{S}^{m-1}$, according to the law of large number, for any measurable set $\mathcal{U} \in \mathbb{S}^{m-1}$, when N is large enough, there will be almost equal size of features from each class in $\mathcal{U}$. Consequently, any classifier g that classifies $\mathcal{U}$ to class k will only have $1/K$ accuracy asymptotically. ■

A.5 Proof of Theorem 10

Proof Consider any pair of samples (x, x') from the same class y, and the positive sample of x as x^+. As intra-class connectivity holds, x and x' are connected, and the maximal length of the path from x to x' is D. Therefore, we can bound the representation distance between x and x' by the triangular inequality under the ε-alignment

$$\|f(x) - f(x')\| \leq D \sup_{(x,x^+)\sim p(x,x^+)} \|f(x) - f(x_+)\| \leq D\varepsilon \tag{27}$$

With the inequality above, we can bound the variance terms in Theorem 5. In particular, the conditional variance can be bounded as

$$\begin{aligned} &\operatorname{Var}(f(x) \mid y) \\ =&\mathbb{E}_{p(y)}\mathbb{E}_{p(x|y)}\|f(x) - \mathbb{E}_{x'} f(x')\|^2 \\ \leq&\mathbb{E}_{p(y)}\mathbb{E}_{p(x|y)}\mathbb{E}_{p(x'|y)}\|f(x) - f(x')\|^2 \\ \leq&\mathbb{E}_{p(y)} \max_{x,x'\sim p(x|y)} \|f(x) - f(x')\|^2 \\ \overset{(1)}{\leq}&\mathbb{E}_{p(y)} D^2\varepsilon^2 = D^2\varepsilon^2, \end{aligned} \tag{28}$$

where (1) follows Eq. 27. Then, we can bound the variance items in Theorem 5 with Eq. 28. ■

A.6 Proof of Corollary 11

Proof Combing with Theorem 10 and $\epsilon = 0$, the intra-class variance term varnishes. So we can directly obtain this result. ■

A.7 Proof of Corollary 12

Proof $\mu_{1k}, \mu_{2k}, \cdots$ are orthonormal eigenvectors of adjacent matrix of the intra-class graph $\mathcal{G}_k$ with eigenvalues $\lambda_{1k}, \lambda_{2k}, \cdots$ ($|\lambda_{1k}| \geq |\lambda_{2k}| \geq \cdots$). Let M denote the adjacency matrix of graph $\mathcal{G}_k$ and $(M)_{r,s}$ denote the row r and column s element of M, we have $M = \sum_i \lambda_{ik}\mu_{ik}\mu_{ik}^\top$

$$
\begin{aligned}
(M^q)_{r,s} &= \sum_i \lambda_{ik}^q (\mu_{ik}\mu_{ik}^\top)_{r,s} \\
&= \lambda_{1k}^q (\mu_{1k}\mu_{1k}^\top)_{r,s} + \sum_{i>1} \lambda_{ik}^q (\mu_{ik}\mu_{ik}^\top)_{r,s} \\
&\geq |\lambda_{1k}|^q \omega_k^2 - |\lambda_{2k}|^q \sum_{i>1} |(\mu_{ik}\mu_{ik}^\top)_{r,s}| \\
&= |\lambda_{1k}|^q \omega_k^2 - |\lambda_{2k}|^q \sum_{i>1} |(\mu_{ik})_r||(\mu_{ik})_s| \\
&\overset{(1)}{\geq} |\lambda_{1k}|^q \omega_k^2 - |\lambda_{2k}|^q \left(\sum_{i>1} |(\mu_{ik})_r|^2\right)^{\frac{1}{2}} \left(\sum_{i>1} |(\mu_{ik})_s|^2\right)^{\frac{1}{2}} \\
&= |\lambda_{1k}|^q \omega_k^2 - |\lambda_{2k}|^m \left(1-(u_{1k})_r^2\right)^{\frac{1}{2}} \left(1-(u_{1k})_s^2\right)^{\frac{1}{2}} \\
&\geq |\lambda_{1k}|^q \omega_k^2 - |\lambda_{2k}|^q \left(1-\omega_k^2\right),
\end{aligned}
\tag{29}
$$

where (1) employs the Cauchy–Schwarz inequality. So, if $q \geq \frac{\log((1-\omega_k^2)/\omega_k^2)}{\log(|\lambda_{1k}|/|\lambda_{2k}|)}$, every element of $(M)^q > 0$, *i.e.,* $D_k \leq \frac{\log((1-\omega_k^2)/\omega_k^2)}{\log(|\lambda_{1k}|/|\lambda_{2k}|)}$. For the maximal radius D, we have $D \leq \frac{\log((1-\omega^2)/\omega^2)}{\log(|\lambda_1|/|\lambda_2|)}$, where $\omega = \min_{i,k}|(\mu_{1k})_i|, \lambda_1 = \min_k|\lambda_{1k}|, \lambda_2 = \max_k|\lambda_{2k}|$. ■

With this Lemma, we can directly obtain corollary 12.

A.8 Proof of Cases That HaoChen et al. (2021) Fail to Analyze

When Assumption 8 holds, the error of labeling function $\alpha = 0$.

Case I: If Assumption 8 holds and $\lfloor m/2 \rfloor \leq K$, then $\rho_{\lfloor m/2 \rfloor} = 0$. When Assumption 4 holds, then $\alpha = 0$. The bound in Lemma 13 becomes a $\frac{0}{0}$ term which fails to be analyzed.
Proof We give a solution of the sparsest partition of augmentation graph. $\forall t \leq \lfloor m/2 \rfloor - 1$, we set $S_t = \mathcal{G}_t$, where $\mathcal{G}_t$ is the subgraph of latent class t, and $S_{\lfloor m/2 \rfloor} = \mathcal{G} - (S_1 \bigcap S_2 \bigcap \cdots \bigcap S_{\lfloor m/2 \rfloor - 1})$. As Assumption 8 holds, there exists no edge between different latent classes, *i.e.,* $\phi_{\mathcal{G}}(S_1) = \phi_{\mathcal{G}}(S_1) = \cdots = \phi_{\mathcal{G}}(S_{\lfloor m/2 \rfloor - 1}) = 0$, where ϕ is Dirichlet conductance of augmentation graph. So we have $\rho_{\lfloor m/2 \rfloor} = 0$. ■

Case II: If we adopt the same settings in Proposition 6, *i.e.,* there exists no support overlapping, the bound in Lemma 13 becomes a $\frac{0}{0}$ term again.
Proof When there exist no overlap between intra-class samples as analyzed in Proposition 6, there exist no edge between different intra-class samples. So $\phi_{\mathcal{G}}(S_1) = \phi_{\mathcal{G}}(S_1) = \cdots = \phi_{\mathcal{G}}(S_{\lfloor m/2 \rfloor - 1}) = 0$. So we have $\rho_{\lfloor m/2 \rfloor} = 0$. ■

A.9 Proof of Theorem 14

Proof From definition and notation in section 5. We can construct an augmentation Graph $\mathcal{G}(V, E(\mathcal{T}))$ given N random samples. We define D_k as the distance from a random point to its k-th nearest neighbour. Percus and Martin (1998) discuss D_k in random grpah and

give the estimation of that:

$$D_k \approx \frac{[(d/2)!]^{\frac{1}{d}}}{\sqrt{\pi}} \frac{(k-1+1/d)!}{(k-1)!} (\frac{S}{N-1})^{\frac{1}{d}} [1 - \frac{1/d + 1/d^2}{2(N-1)} + O(\frac{1}{(N-1)^2})], \tag{30}$$

where d is the dimension of hypersphere, S is the surface area of sample distribution and N is the number of random points. When $r < D_1$ there is no edge in the graph. So the class is separated. When $r > D_{N-1}$, any pair of vertexes have an edge between them, so the graph is full connected. With this conclusion, we can directly have Theorem 14. ■

A.10 Proof of Theorem 15

Proof Denote

$$r_{mc} = \inf\{r_i > 0 : G_N(V, E, r_i) \text{is connected}\}. \tag{31}$$

With Theorem 1.1 from Penrose (1999) and features are uniformly distributed in the surface of unit hypersphere, we have

$$\lim_{N\to\infty} (r_{mc}^d \frac{N^2}{\log N}) = 2\frac{(1-\frac{1}{d})S}{V_u}, d \geq 2, \tag{32}$$

where V_u denotes to the volume of unit hypershpere. ■

Appendix B. Additional Experimental Details

B.1 Simulation on Random Augmentation Graph

Following our setting in Section 6.1, we consider a binary classification task with InfoNCE loss. We generate data from two uniform distributions on a unit ball $\mathbb{S}^2$ in the 3-dimensional space. One center is $(0, 0, 1)$ and another is $(0, 0, -1)$. The area of both parts are 1. We take 5000 samples as train set and 1000 samples as test set. For the encoder class $\mathcal{F}$, a single-hidden-layer neural network with softmax activation and 256 output dimensions is adopted, which is trained by InfoNCE loss.

B.2 The Calculation Process of ARC

Our experiments are mainly conducted on the real-world data set CIFAR-10. We use SimCLR as the training framework and ResNet18 as the backbone network. When calculating ARC, we set the number of intra-anchor augmented views $C = 6$ and $k = 1$. We train the network for 200 epochs and use the encoder trained with 200 epochs as the final encoder while the encoder trained with 1 epoch as the initial encoder. When we present the relationship between ARC and linear downstream accuracy trained by different augmentations, we adapt log operator on ARC.

B.3 The Comparison between Different Unsupervised Metrics

The experiments are conducted on CIFAR-10 data set with different data augmentations including 1) RandomResizedCrop with 12 different scales of the crop, 2) ColorJitter with 10 different groups of parameters (contrast, brightness, saturation, hue), 3) composition of all augmentations used in SimCLR with 11 groups of different parameters. For computing ARC, we follow the default settings described in Appendix B.2. For rotation accuracy, we train a 4-dimension linear classifier following the fixed encoder to predicted the angles of the rotation $(0°, 90°, 180°, 270°)$.

B.4 Analysis on the Impracticality of the Conditional Independence Assumption

Intuitively, as illustrated in Figure 2, a positive pair is constructed by applying different augmentations to the same natural image. As a result, the two resulting inputs are inherently dependent, making the conditional independence assumption impractical in real-world contrastive learning scenarios.

To further demonstrate this, we conduct an empirical analysis to quantify the deviation from the conditional independence assumption. Since the exact conditional probabilities $p(x, x^+ \mid y)$ and $p(x \mid y)p(x^+ \mid y)$ are inaccessible, we approximate them using a pretrained encoder f and the cosine similarity between features. Specifically, we use $\frac{f(x)^\top f(x^+)}{\sum_{y(x')=y(x)} f(x')^\top f(x'^+)}$ to estimate $p(x, x^+|y)$ and use $\sum_{x^+} \frac{f(x)^\top f(x^+)}{\sum_{y(x')=y(x)} f(x')^\top f(x'^+)}$ to estimate $p(x|y)$. We then compute the ratio between the estimated joint and marginal probabilities.

In our experiment, we use a ResNet-18 encoder pretrained on CIFAR-10 using SimCLR (Chen et al., 2020). As shown in Table 2, the joint probability is significantly larger than the product of the marginals (ratio ≈ 3.83), indicating a notable dependence between the inputs.

These results support the claim that the conditional independence assumption is impractical in realistic scenarios.

Table 2: Comparison between the estimated joint and marginal probabilities. Estimation is performed using an encoder f (ResNet-18 pretrained on CIFAR-10 with SimCLR) and cosine similarity between features.

$p(x, x^+\|y)$	$p(x\|y)p(x^+\|y)$	ratio$(\frac{p(x,x^+\|y)}{p(x\|y)p(x^+\|y)})$
0.0023	0.0006	3.83

References

Sumia Abdulhussien Razooqi Al-Obaidi, Davood Zabihzadeh, and Hamideh Hajiabadi. Robust metric learning based on the rescaled hinge loss. *International Journal of Machine Learning and Cybernetics*, 11(11):2515–2528, 2020.

Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. In *ICML*, 2019.

Jordan T Ash, Surbhi Goel, Akshay Krishnamurthy, and Dipendra Misra. Investigating the role of negatives in contrastive representation learning. In *AISTATS*, 2022.

Han Bao, Yoshihiro Nagano, and Kento Nozawa. On the surrogate gap between contrastive and supervised losses. In *ICML*, 2022.

Maxime Bucher, Stéphane Herbin, and Frédéric Jurie. Hard negative mining for metric learning based zero-shot classification. In *ECCV 2016 Workshops*, 2016.

Ivan Budimir, Sever S Dragomir, and Josep Pecaric. Further reverse results for jensen's discrete inequality and applications in information theory. *Research Group in Mathematical Inequalities and Applications*, 2000.

Shuhao Cao, Peng Xu, and David A Clifton. How to understand masked autoencoders. *arXiv preprint arXiv:2202.03670*, 2022.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

F. R. K. Chung. Diameters and eigenvalues. *American Mathematical Society*, 1989.

Jingyi Cui, Weiran Huang, Yifei Wang, and Yisen Wang. Rethinking weak supervision in helping contrastive learning. In *ICML*, 2023.

Jingyi Cui, Hongwei Wen, and Yisen Wang. An augmentation-aware theory for self-supervised contrastive learning. In *ICML*, 2025.

Tianqi Du, Yifei Wang, and Yisen Wang. On the role of discrete tokenization in visual representation learning. In *ICLR*, 2024.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *ICCV*, 2021.

Andreas Fürst, Elisabeth Rumetshofer, Viet Tran, Hubert Ramsauer, Fei Tang, Johannes Lehner, David Kreil, Michael Kopp, Günter Klambauer, Angela Bitto-Nemling, et al. Cloob: Modern hopfield networks with infoloob outperform clip. In *ICLR*, 2022.

Siddhant Garg and Yingyu Liang. Functional regularization for representation learning: A unified theoretical perspective. In *NeurIPS*, 2020.

Jean-Bastien Grill, Florian Strub, Florent Altché, C. Tallec, Pierre H. Richemond, Elena Buchatskaya, C. Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, B. Piot, K. Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, 2020.

Michael U Gutmann and Aapo Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 2012.

Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. In *NeurIPS*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017.

Tianyang Hu, Zhili Liu, Fengwei Zhou, Wenjia Wang, and Weiran Huang. Your contrastive learning is secretly doing stochastic neighbor embedding. In *ICLR*, 2023.

Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *CVPR*, 2019.

Nikos Komodakis and Spyros Gidaris. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018.

Jason D Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. Predicting what you already know helps: Provable self-supervised learning. In *NeuIPS*, 2021.

Yunwen Lei, Tianbao Yang, Yiming Ying, and Ding-Xuan Zhou. Generalization analysis for contrastive representation learning. In *ICML*, 2023.

Yazhe Li, Roman Pogodin, Danica J Sutherland, and Arthur Gretton. Self-supervised learning with kernel dependence maximization. In *NeurIPS*, 2021.

Jovana Mitrovic, Brian McWilliams, Jacob C Walker, Lars Holger Buesing, and Charles Blundell. Representation learning via invariant causal mechanisms. In *ICLR*, 2021.

Kento Nozawa and Issei Sato. Understanding negative samples in instance discriminative self-supervised representation learning. In *NeurIPS*, 2021.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Zhuo Ouyang, Kaiwen Hu, Qi Zhang, Yifei Wang, and Yisen Wang. Projection head is secretly an information bottleneck. *arXiv preprint arXiv:2503.00507*, 2025.

Advait Parulekar, Liam Collins, Karthikeyan Shanmugam, Aryan Mokhtari, and Sanjay Shakkottai. Infonce loss provably learns cluster-preserving representations. In *COLT*, 2023.

Karl Pearson. Vii. note on regression and inheritance in the case of two parents. In *RSPL*, 1895.

Mathew D. Penrose. A Strong Law for the Longest Edge of the Minimal Spanning Tree. *The Annals of Probability*, 1999.

Allon G Percus and Olivier C Martin. Scaling universalities of kth-nearest neighbor distances on closed manifolds. *Advances in Applied Mathematics*, 1998.

Colorado J Reed, Sean Metzger, Aravind Srinivas, Trevor Darrell, and Kurt Keutzer. Self-augment: Automatic augmentation policies for self-supervised learning. In *CVPR*, 2021.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *ICLR*, 2021.

Nikunj Saunshi, Sadhika Malladi, and Sanjeev Arora. A mathematical exploration of why language models help solve downstream tasks. In *ICLR*, 2021.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

Anshul Shah, Suvrit Sra, Rama Chellappa, and Anoop Cherian. Max-margin contrastive learning. In *AAAI*, 2022.

Zhiquan Tan, Jingqin Yang, Weiran Huang, Yang Yuan, and Yifan Zhang. Information flow in self-supervised learning. *arXiv preprint arXiv:2309.17281*, 2023.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. In *NeurIPS*, 2020.

Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *ICML*, 2021.

Nenad Tomasev, Ioana Bica, Brian McWilliams, Lars Buesing, Razvan Pascanu, Charles Blundell, and Jovana Mitrovic. Pushing the limits of self-supervised resnets: Can we outperform supervised learning without labels on imagenet? *arXiv preprint arXiv:2201.05119*, 2022.

Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive estimation reveals topic posterior information to linear models. *arXiv preprint arXiv:2003.02234*, 2020.

Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Self-supervised learning from a multi-view perspective. In *ICLR*, 2021.

Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. In *ICLR*, 2020.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *CVPR*, 2021.

Liang Wang, Xiang Tao, Qiang Liu, and Shu Wu. Rethinking graph masked autoencoders through alignment and uniformity. In *AAAI*, 2024a.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, 2020.

Yifei Wang, Zhengyang Geng, Feng Jiang, Chuming Li, Yisen Wang, Jiansheng Yang, and Zhouchen Lin. Residual relaxation for multi-view representation learning. In *NeurIPS*, 2021.

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 *ICLR*, 2022.

Yifei Wang, Qi Zhang, Tianqi Du, Jiansheng Yang, Zhouchen Lin, and Yisen Wang. A message passing perspective on learning dynamics of contrastive learning. In *ICLR*, 2023.

Yifei Wang, Jizhe Zhang, and Yisen Wang. Do generated data always help contrastive learning? In *ICLR*, 2024b.

Zixin Wen and Yuanzhi Li. Toward understanding the feature learning process of self-supervised contrastive learning. In *ICML*, 2021.

Junkang Wu, Jiawei Chen, Jiancan Wu, Wentao Shi, Xiang Wang, and Xiangnan He. Understanding contrastive learning via distributionally robust optimization. In *NeurIPS*, 2023.

Chun-Hsiao Yeh, Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu, Yubei Chen, and Yann LeCun. Decoupled contrastive learning. *arXiv preprint arXiv:2110.06848*, 2021.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *ICML*, 2021.

Qi Zhang, Yifei Wang, and Yisen Wang. How mask matters: Towards theoretical understandings of masked autoencoders. In *NeurIPS*, 2022.

Qi Zhang, Yifei Wang, and Yisen Wang. On the generalization of multi-modal contrastive learning. In *ICML*, 2023.

Qi Zhang, Tianqi Du, Haotian Huang, Yifei Wang, and Yisen Wang. Look ahead or look around? a theoretical comparison between autoregressive and masked pretraining. In *ICML*, 2024.

Zhijian Zhuo, Yifei Wang, Jinwen Ma, and Yisen Wang. Towards a unified theoretical understanding of non-contrastive learning via rank differential mechanism. In *ICLR*, 2023.

Roland S Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *ICML*, 2021.