RETHINKING THE UNIFORMITY METRIC IN SELF-SUPERVISED LEARNING

Xianghong Fang
The Chinese University of Hong Kong, Shenzhen
fangxianghong2@gmail.com

Jian Li
Tencent AI Lab
lijianjack@gmail.com

Qiang Sun *
University of Toronto & MBZUAI
qiang.sun@utoronto.ca

Benyou Wang *
The Chinese University of Hong Kong, Shenzhen & SRIBD
wangbenyou@cuhk.edu.cn

ABSTRACT

Uniformity plays a crucial role in the assessment of learned representations, contributing to a deeper comprehension of self-supervised learning. The seminal work by Wang & Isola (2020) introduced a uniformity metric that quantitatively measures the collapse degree of learned representations. Directly optimizing this metric together with alignment proves to be effective in preventing constant collapse. However, we present both theoretical and empirical evidence revealing that this metric lacks sensitivity to dimensional collapse, highlighting its limitations. To address this limitation and design a more effective uniformity metric, this paper identifies five fundamental properties, some of which the existing uniformity metric fails to meet. We subsequently introduce a novel uniformity metric that satisfies all of these desiderata and exhibits sensitivity to dimensional collapse. When applied as an auxiliary loss in various established self-supervised methods, our proposed uniformity metric consistently enhances their performance in downstream tasks. Our code was released at `WassersteinUniformityMetric`.

1 INTRODUCTION

Self-supervised learning excels in acquiring invariant representations to various augmentations (Chen et al., 2020; He et al., 2020; Caron et al., 2020; Grill et al., 2020; Zbontar et al., 2021). It has been outstandingly successful across a wide range of domains, such as multimodality learning, object detection, and segmentation (Radford et al., 2021; Li et al., 2022; Xie et al., 2021; Wang et al., 2021; Yang et al., 2021; Zhao et al., 2021). To gain a deeper understanding of self-supervised learning, thoroughly evaluating the learned representations is a pragmatic approach (Wang & Isola, 2020; Gao et al., 2021; Tian et al., 2021; Jing et al., 2022).

Alignment, a metric quantifying the similarities between positive pairs, holds significant importance in the evaluation of learned representations (Wang & Isola, 2020). It ensures that samples forming positive pairs are mapped to nearby features, thereby rendering them invariant to irrelevant noise factors (Hadsell et al., 2006; Chen et al., 2020). However, relying solely on alignment proves inadequate for effectively assessing representation quality in self-supervised learning. This limitation becomes evident in the presence of extremely small alignment values in collapsing solutions, as observed in Siamese networks (Hadsell et al., 2006), where all outputs collapse to a single point (Chen & He, 2021), as illustrated in Figure 1. In such cases, the learned representations exhibit optimal alignment but fail to provide meaningful information for any downstream tasks. This underscores the necessity of incorporating additional factors when evaluating learned representations.

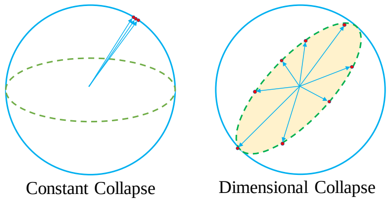


Figure 1: The left figure presents constant collapse, and the right figure visualizes dimensional collapse.

---

[*] Qiang Sun and Benyou Wang are joint corresponding authors.

To better evaluate the learned representations, Wang & Isola (2020) formally introduced a *uniformity* metric by utilizing the logarithm of the average pairwise Gaussian potential (Cohn & Kumar, 2007). Uniformity assesses how feature embeddings are distributed uniformly across the unit hypersphere, and higher uniformity indicates that more information is preserved by the learned representations. Since its invention, uniformity has played a pivotal role in comprehending self-supervised learning and mitigating the issue of constant collapse (Arora et al., 2019; Wang & Isola, 2020; Gao et al., 2021). Nevertheless, the validity of this particular uniformity metric warrants further examination and scrutiny.

To examine the existing uniformity metric (Wang & Isola, 2020), we introduce five principled properties, also known as desiderata, that a desired uniformity metric should fulfill. Guided by these properties, we conduct a theoretical analysis that reveals certain shortcomings of the existing metric, particularly its insensitivity to dimensional collapse (Hua et al., 2021)[1]. We complement our theoretical findings with empirical evidence that underscores the metric's limitations in addressing dimensional collapse. We then introduce a new uniformity metric that satisfies all desiderata. In particular, the proposed metric is sensitive to dimensional collapse and thus is superior to the existing one. Finally, using the proposed uniformity metric as an auxiliary loss within existing self-supervised learning methods consistently improves their performance in downstream tasks.

Our main contributions are summarized as follows. (i) We introduce five desiderata that provide a novel perspective on the design of ideal uniformity metrics. Notably, the existing uniformity metric (Wang & Isola, 2020) does not satisfy all of these desiderata. Specifically, we demonstrate, both theoretically and empirically, its insensitivity to dimensional collapse. (ii) We propose a novel uniformity metric that not only fulfills all of the desiderata but also exhibits sensitivity to dimensional collapse, addressing a crucial limitation of the existing metric. (iii) Our newly proposed uniformity metric can be seamlessly incorporated as an auxiliary loss in a variety of self-supervised methods, consistently improving their performance in downstream tasks.

2 BACKGROUND

2.1 SELF-SUPERVISED REPRESENTATION LEARNING

A common practice of self-supervised learning is to maximize the similarity of representations obtained from different augmentations of one sample. Specifically, given a set of data samples $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$, the Siamese network (Hadsell et al., 2006) takes as input two randomly augmented views $\mathbf{x}_i^a$ and $\mathbf{x}_i^b$ from a input sample $\mathbf{x}_i$. Then the two views are processed by an encoder network f consisting of a backbone (e.g., ResNet (He et al., 2016)) and a projection MLP head (Chen et al., 2020), denoted as g. To enforce invariability to representations of the two views $\mathbf{z}_i^a := g(f(\mathbf{x}_i^a))$ and $\mathbf{z}_i^b := g(f(\mathbf{x}_i^b))$, a natural solution is to maximize the cosine similarity between the representations of two views, and mean square error (MSE) is a widely used loss function to align their ℓ_2-normalized representations on the unit hypersphere:

$$\mathcal{L}_{\text{align}}^{\theta} = \left\| \frac{\mathbf{z}_i^a}{\|\mathbf{z}_i^a\|} - \frac{\mathbf{z}_i^b}{\|\mathbf{z}_i^b\|} \right\|_2^2 = 2 - 2 \cdot \frac{\langle \mathbf{z}_i^a, \mathbf{z}_i^b \rangle}{\|\mathbf{z}_i^a\| \cdot \|\mathbf{z}_i^b\|}, \tag{1}$$

However, a common issue with this approach is that it easily learns an undesired collapsing solution where all representations collapse to one single point, as depicted in Figure 1.

2.2 EXISTING SOLUTIONS TO CONSTANT COLLAPSE

To prevent constant collapse, existing solutions include contrastive learning, asymmetric model architecture, and redundancy reduction.

Contrastive Learning Contrastive learning serves as a valuable technique for mitigating constant collapse. The key idea is to utilize negative pairs. For example, SimCLR (Chen et al., 2020) introduced an in-batch negative sampling strategy that utilizes samples within a batch as negative

---

[1] When dimensional collapse occurs, representations occupy a lower-dimensional subspace instead of the entire embedding space (Jing et al., 2022) and thus some dimensions are not fully utilized; see Figure 1.

samples. However, its effectiveness is contingent on the use of a large batch size. To address this limitation, MoCo (He et al., 2020) used a memory bank, which stores additional representations as negative samples. Recent research endeavors have also explored clustering-based contrastive learning, which combines a clustering objective with contrastive learning techniques (Li et al., 2021; Caron et al., 2020).

Asymmetric Model Architecture The use of asymmetric model architecture represents another approach to combat constant collapse. One plausible explanation for its effectiveness is that such an asymmetric design encourages encoding more information (Grill et al., 2020). To maintain this asymmetry, BYOL (Grill et al., 2020) introduces the concept of using an additional predictor in one branch of the Siamese network while employing momentum updates and stop-gradient operators in the other branch. DINO (Caron et al., 2021), takes this asymmetry a step further by applying it to two encoders, distilling knowledge from the momentum encoder into the other one (Hinton et al., 2015). SimSiam (Chen & He, 2021) removes the momentum update from BYOL, and illustrates that the momentum update may not be essential in preventing collapse. In contrast, Mirror-SimSiam (Zhang et al., 2022a) swaps the stop-gradient operator to the other branch. Its failure challenges the assertion made in SimSiam (Chen & He, 2021) that the stop-gradient operator is the key component for preventing constant collapse. Theoretically, Tian et al. (2021) provides an examination to elucidate why an asymmetric model architecture can effectively avoid constant collapse.

Redundancy Reduction The fundamental principle behind redundancy reduction to mitigate constant collapse is to maximize the information preserved by the representations. The key idea is to decorrelate the learned representations. Barlow Twins (Zbontar et al., 2021) aims to achieve decorrelation by focusing on the cross-correlation matrix, while VICReg (Bardes et al., 2022) focuses on the covariance matrix. Zero-CL (Zhang et al., 2022b) takes a hybrid approach, combining instance-wise and feature-wise whitening techniques to achieve decorrelation.

2.3 COLLAPSE ANALYSIS

While the aforementioned solutions could effectively prevent constant collapse, they are not effective in preventing dimensional collapse, where representations occupy a lower-dimensional subspace instead of the whole space. This phenomenon was observed in contrastive learning by visualizing the singular value spectrum of representations (Jing et al., 2022; Tian et al., 2021). Visualization provides a qualitative assessment but not a quantitive one.

To achieve a quantifiable analysis of collapse degree, To quantitively measure the collapse degree, Wang & Isola (2020) proposed a uniformity metric by utilizing the logarithm of the average pairwise Gaussian potential. Specifically, for a set of representations $\{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n\}$, their uniformity metric is defined as:

$$\mathcal{L}_{\mathcal{U}} := \log \frac{1}{n(n-1)/2} \sum_{i=2}^{n} \sum_{j=1}^{i-1} e^{-t\left\| \frac{\mathbf{z}_i}{\|\mathbf{z}_i\|} - \frac{\mathbf{z}_j}{\|\mathbf{z}_j\|} \right\|_2^2}, \tag{2}$$

Where $t > 0$ is a fixed parameter and is often taken as 2. In this work, we show that this metric is insensitive to dimensional collapse, both theoretically (in Section 3.2) and empirically (in Section 5.2).

3 WHAT MAKES A GOOD UNIFORMITY METRIC?

In this section, we first introduce five desiderata of a good uniformity metric, and then conduct theoretical analysis on the existing uniformity metric $-\mathcal{L}_{\mathcal{U}}$ by Wang & Isola (2020).

3.1 DESIDERATA OF UNIFORMITY

A uniformity metric is a function that maps a set of learned representations in $\mathbb{R}^m$ to a uniformity indicator in $\mathbb{R}$.

$$\mathcal{U} : \{\mathbb{R}^m\}^n \to \mathbb{R}, \tag{3}$$

$\mathcal{D} \in \{\mathbb{R}^m\}^n$ is a set of learned vectors ($\mathcal{D} = \{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n\}$), each vector is the feature representation of an instance, $\mathbf{z}_i \in \mathbb{R}^m$. We first assume that the maximum uniformity metric over the representations should be invariant to the dimensions, which is also satisfied by the metric in (Wang & Isola, 2020).

Assumption 1. *Suppose $\mathcal{D}_1$, $\mathcal{D}_2$ are two sets of learned vectors from Euclidean spaces of different dimensions. Then the maximal uniformity over data is invariant to dimensions:*

$$\sup_{\mathcal{D}_1 \in \{\mathbb{R}^m\}^{n_1}} (\mathcal{U}(\mathcal{D}_1)) = \sup_{\mathcal{D}_2 \in \{\mathbb{R}^k\}^{n_2}} (\mathcal{U}(\mathcal{D}_2)), \quad m \neq k, \forall n_1, n_2, m, k \in \mathbb{R}^+, \tag{4}$$

We then formally propose the five desiderata for an ideal uniformity metric. Intuitively, uniformity should be invariant to the permutation of instances, as the distribution of representations is not affected by permutations.

Property 1 (Instance Permutation Constraint (IPC))**.** *The $\mathcal{U}$ satisfies*

$$\mathcal{U}(\pi(\mathcal{D})) = \mathcal{U}(\mathcal{D}), \tag{5}$$

where π is an instance permutation.

The uniformity should be invariant to the scalings of representations. This copes with modern machine learning practice where scalings are not important in representation learning. For example, recent self-supervised learning methods often learn representations under an ℓ_2 norm constraint (Zbontar et al., 2021; Wang & Isola, 2020; Grill et al., 2020; Chen et al., 2020), and produce embeddings in the form of $\mathcal{D}^s = \{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n\}$, where $\mathbf{s}_i = \mathbf{z}_i / \|\mathbf{z}_i\|_2 \in \mathcal{S}^{m-1}$ is on the unit hypersphere[2].

Property 2 (Instance Scaling Constraint (ISC))**.** *The $\mathcal{U}$ satisfies*

$$\mathcal{U}(\{\lambda_1 \mathbf{z}_1, \lambda_2 \mathbf{z}_2, ..., \lambda_n \mathbf{z}_n\}) = \mathcal{U}(\mathcal{D}), \quad \forall \lambda_i \in \mathbb{R}^+. \tag{6}$$

The uniformity metric should be invariant to instance clones, as cloning does not vary the distribution of embeddings[3].

Property 3 (Instance Cloning Constraint (ICC))**.** *The $\mathcal{U}$ satisfies*

$$\mathcal{U}(\mathcal{D} \cup \mathcal{D}) = \mathcal{U}(\mathcal{D}), \tag{7}$$

where $\cup$ indicates the union operator.

The uniformity metric should decrease when cloning features for each instance, as the feature-level cloning will bring redundancy, leading to dimensional collapse (Zbontar et al., 2021; Bardes et al., 2022)[4].

Property 4 (Feature Cloning Constraint (FCC))**.** *The $\mathcal{U}$ satisfies*

$$\mathcal{U}(\mathcal{D} \oplus \mathcal{D}) \leq \mathcal{U}(\mathcal{D}), \tag{8}$$

where $\oplus$ is a feature-level concatenation operator defined as $\mathcal{D} \oplus \mathcal{D} = \{\mathbf{z}_1 \oplus \mathbf{z}_1, \mathbf{z}_2 \oplus \mathbf{z}_2, ..., \mathbf{z}_n \oplus \mathbf{z}_n\}$, and $\mathbf{z}_i \oplus \mathbf{z}_i = [z_{i1}, \cdots, z_{im}, z_{i1}, \cdots, z_{im}]^T \in \mathbb{R}^{2m}$.

The uniformity metric should decrease when adding constant features for each instance, since it introduces uninformative features and results in additional collapsed dimensions[5].

Property 5 (Feature Baby Constraint (FBC))**.** *The $\mathcal{U}$ satisfies*

$$\mathcal{U}(\mathcal{D} \oplus \mathbf{0}^k) \leq \mathcal{U}(\mathcal{D}), \quad k \in \mathbb{N}^+, \tag{9}$$

$\mathcal{D} \oplus \mathbf{0}^k = \{\mathbf{z}_1 \oplus \mathbf{0}^k, \mathbf{z}_2 \oplus \mathbf{0}^k, ..., \mathbf{z}_n \oplus \mathbf{0}^k\}$, and $\mathbf{z}_i \oplus \mathbf{0}^k = [z_{i1}, z_{i2}, ..., z_{im}, 0, 0, ..., 0]^T \in \mathbb{R}^{m+k}$. $\mathcal{U}(\mathcal{D} \oplus \mathbf{0}^k) \leq \mathcal{U}(\mathcal{D})$ if and only if $\mathbf{z}_1 = \mathbf{z}_2 = ... = \mathbf{z}_n = \mathbf{0}^m$.

These five properties are five intuitive yet fundamental properties of an ideal uniformity metric.

---

[2] The vector $\lambda \mathbf{z} (\lambda > 0)$ is located at the same point as the $\mathbf{z}$ on the unit hypersphere.

[3] Although instance cloning enlarges the size of the set $\mathcal{D}$ by repeatedly adding the all the vectors, the probability density function over the entire unit hypersphere is invariant, thus the equality.

[4] Suppose two sets $\mathcal{D} \in R^m$ and $\mathcal{D}_2 \in R^{2m}$ have the maximum uniformity. According to the Assumption 1, we have $\mathcal{U}(\mathcal{D}) = \mathcal{U}(\mathcal{D}_2)$. $\mathcal{D} \oplus \mathcal{D}$ contains redundant information, thus the smaller uniformity than $\mathcal{D}_2$. Then, $\mathcal{U}(\mathcal{D} \oplus \mathcal{D}) \leq \mathcal{U}(\mathcal{D}_2) = \mathcal{U}(\mathcal{D})$.

[5] Suppose two sets $\mathcal{D} \in R^m$ and $\mathcal{D}_2 \in R^{m+k}$ have the maximum uniformity. According to the Assumption 1, we have $\mathcal{U}(\mathcal{D}) = \mathcal{U}(\mathcal{D}_2)$. $\mathcal{D} \oplus \mathbf{0}^k$ contains collapsed dimensions, thus smaller uniformity than $\mathcal{D}_2$. Then, $\mathcal{U}(\mathcal{D}) = \mathcal{U}(\mathcal{D}_2) > \mathcal{U}(\mathcal{D} \oplus \mathbf{0}^k)$. See Appendix O for the more detailed explanation.

3.2 EXAMINING THE UNIFORMITY METRIC $-\mathcal{L}_{\mathcal{U}}$ BY WANG & ISOLA (2020)

We use the desiderata proposed in Section 3.1 to examine the uniformity metric $-\mathcal{L}_{\mathcal{U}}$ in 2. The following claim summarizes the conclusion.

Claim 1. *The uniformity metric $-\mathcal{L}_{\mathcal{U}}$ satisfies Properties 1 and 2, but violates Properties 3, 4, and 5.*

For Properties 1 and 2, we can directly use their definitions to prove. To check the other three properties, see Appendix C.1. The violation of the three properties in particular the Properties 4 and 5 indicates the uniformity metric $-\mathcal{L}_{\mathcal{U}}$ is insensitive to feature redundancy and dimensional collapse. Therefore, we need a new uniformity metric.

4 A NEW UNIFORMITY METRIC

In this section, we introduce a novel uniformity metric designed to capture dimensional collapse.

4.1 THE DISTRIBUTION WITH MAXIMAL UNIFORMITY AND AN APPROXIMATION

Intuitively, the uniform distribution on the unit hypersphere, aka $\text{Unif}(\mathcal{S}^{m-1})$, achieves maximal uniformity. Nevertheless, any distance involving this distribution is hard to calculate. We need the following Fact 1 which establishes an equivalence between the uniform spherical distribution and the normalized isotropic Gaussian distribution.

Fact 1. *If $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_m)$, then $\mathbf{Y} := \mathbf{Z}/\|\mathbf{Z}\|_2$ is uniformly distributed on the unit hypersphere $\mathcal{S}^{m-1}$.*

Because the average length of $\|\mathbf{Z}\|_2$ is roughly $\sigma\sqrt{m}$ (Chandrasekaran et al., 2012), that is,

$$\frac{m}{\sqrt{m+1}} \leq \|\mathbf{Z}\|_2/\sigma \leq \sqrt{m},$$

we expect $\mathbf{Z}/(\sigma\sqrt{m}) \sim \mathcal{N}(0, m^{-1}\mathbf{I}_m)$ to be a reasonably good approximation to $\mathbf{Z}/\|\mathbf{Z}\|_2$, and thus to the uniform spherical distribution. This is rigorously justified in the following theorem.

Theorem 2. *Let Y_i be the i-th coordinate of $\mathbf{Y} = \mathbf{Z}/\|\mathbf{Z}\|_2 \in \mathbb{R}^m$, where $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_m)$. Then the quadratic Wasserstein distance between Y_i and $\widehat{Y}_i \sim \mathcal{N}(0, m^{-1})$ converges to zero as $m \to \infty$:*

$$\lim_{m \to \infty} \mathcal{W}_2(Y_i, \widehat{Y}_i) = 0.$$

The theorem above can be proved by directly utilizing the probability density functions of Y_i and $\widehat{Y}_i$. The detailed proof is deferred to Appendix B. Theorem 2 shows $\mathcal{N}(\mathbf{0}, m^{-1}\mathbf{I}_m)$ approximates the best uniformity metric as $m \to \infty$. We will show empirically that such an approximation is good when m is only moderately large.

4.2 AN EMPIRICAL COMPARISON

In this section, we compare the uniform spherical distribution $\mathbf{Y} \sim \text{Unif}(\mathcal{S}^{m-1})$ and the scaled Gaussian distribution $\widehat{\mathbf{Y}} \sim \mathcal{N}(0, \mathbf{I}_m/m)$ with various m's. Without loss of generality, we compare these two distributions coordinately. The distribution for Y_i is visualized in Figure 7 by binning 200,000 sampled data points, aka samples, into 51 groups. Figure 7 compares the distributions of Y_i and $\widehat{Y}_i$ when $m \in \{2, 4, 8, 16, 32, 64, 128, 256\}$. The quality of the approximation is reasonably good when m is moderately large, e.g., $m \geq 8$ or $m \geq 16$. Figure 8(b) further plots the Wasserstein distance (see definition in Appendix G) between Y_i and $\widehat{Y}_i$ versus increasing m. We observe that the distance converges to zero when m increases, see more details in Appendix E. We also visualize the joint distributions of Y_i and Y_j $(i \neq j)$ in Figure 9(a), $\widehat{Y}_i$ and $\widehat{Y}_j$ $(i \neq j)$ in Figure 9(b), and two distributions could resemble with each other, see more details in Appendix F.

4.3 A NEW METRIC FOR UNIFORMITY

This section proposes to use the distance between the distribution of learned representations and $\mathcal{N}(\mathbf{0}, \mathbf{I}_m/m)$ as a uniformity metric. Specifically, for a set of learned representations

$\{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n\} \subset \mathbb{R}^m$, we first normalize them, and then calculate the mean and covariance matrix:

$$\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbf{z}_i}{\|\mathbf{z}_i\|_2}, \quad \boldsymbol{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbf{z}_i}{\|\mathbf{z}_i\|_2} - \boldsymbol{\mu}\right)^{\mathrm{T}}\left(\frac{\mathbf{z}_i}{\|\mathbf{z}_i\|_2} - \boldsymbol{\mu}\right). \tag{10}$$

To facilitate the computation, we adopt a Gaussian hypothesis to the learned representations and assume they follow $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. With this assumption, we use the quadratic Wasserstein distance[6] to calculate the distance between two distributions; see the definition in Appendix G. We need the following well-known lemma.

Lemma 1 (Wasserstein Distance (Olkin & Pukelsheim, 1982))**.** *Suppose* $\mathbf{Z}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ *and* $\mathbf{Z}_2 \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$*. Then the quadratic Wasserstein distance between* $\mathbf{Z}_1$ *and* $\mathbf{Z}_2$ *is*

$$\sqrt{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2^2 + \mathrm{tr}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2 - 2(\boldsymbol{\Sigma}_2^{\frac{1}{2}}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{\frac{1}{2}})^{\frac{1}{2}}).} \tag{11}$$

The lemma above indicates that the quadratic Wasserstein distance bewteen two Gaussian distributions can be easily computed. Then, we define our proposed uniformity metric as the negative quadratic Wasserstein distance between the learned representations and $\mathcal{N}(0, \mathbf{I}_m/m)$:

$$-\mathcal{W}_2 := -\sqrt{\|\boldsymbol{\mu}\|_2^2 + 1 + \mathrm{tr}(\boldsymbol{\Sigma}) - \frac{2}{\sqrt{m}}\,\mathrm{tr}(\boldsymbol{\Sigma}^{\frac{1}{2}})}, \tag{12}$$

The $-\mathcal{W}_2$ can be used as a metric to evaluate the collapse degree: Smaller $\mathcal{W}_2$ indicates larger uniformity of learned representations. Additionally, our proposed uniformity metric can be used as an auxiliary loss for various existing self-supervised methods since it is differentiable which facilitates the backward pass. In the training phase, the mean and covariance in 10 are calculated using batches.

5 COMPARISON BETWEEN TWO METRICS

5.1 THEORETICAL COMPARISON

We examine the proposed metric $-\mathcal{W}_2$ in terms of the proposed desiderata. The proof is similar to that in Section 3.2 and is collected in Appendix C.2. Table 1 collects the results. Particularly, the proposed metric $-\mathcal{W}_2$ satisfies Properties 3, 4, and 5, while the existing one $-\mathcal{L}_\mathcal{U}$ does not. Consider $\mathcal{D} \oplus \mathbf{0}^k$ versus $\mathcal{D}$. Then, a larger k indicates a more serious dimensional collapse. However, $-\mathcal{L}_\mathcal{U}$ fails to identify this issue since $-\mathcal{L}_\mathcal{U}(\mathcal{D} \oplus \mathbf{0}^k) = -\mathcal{L}_\mathcal{U}(\mathcal{D})$. In sharp contrast, our proposed metric is able to capture this dimensional collapse as $-\mathcal{W}_2(\mathcal{D} \oplus \mathbf{0}^k) < -\mathcal{W}_2(\mathcal{D})$.

Table 1: Theoretical analysis results on the two metrics w.r.t the five desiderata constraints.

Properties	IPC	ISC	ICC	FCC	FBC
$-\mathcal{L}_\mathcal{U}$	✔	✔	✗[7]	✗	✗
$-\mathcal{W}_2$	✔	✔	✔	✔	✔

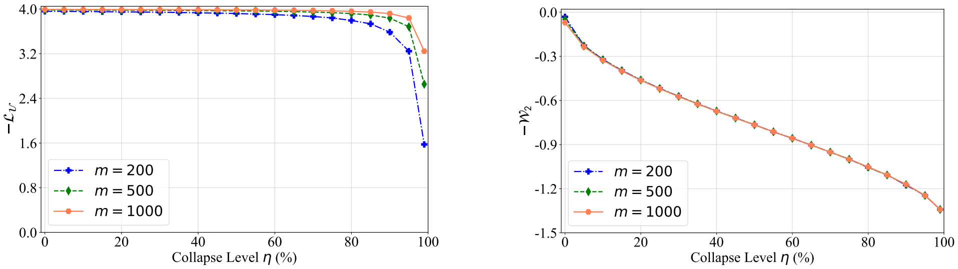
5.2 EMPIRICAL COMPARISON VIA SYNTHETIC DATA

We further conduct synthetic experiments to investigate two uniformity metrics. An empirical study on the correlation between these metrics reveals that data points following a standard Gaussian distribution achieve maximum uniformity compared to those from various other distributions, as detailed in Appendix I. Furthermore, we generate data vectors from this distribution to facilitate a comprehensive comparison between the two metrics.

---

[6] We discuss using other distribution distances as uniformity metrics, such as Kullback-Leibler Divergence and Bhattacharyya Distance over Gaussian distribution. See more details in Appendix H

[7] $-\mathcal{L}_\mathcal{U}$ could satisfy ICC if and only if $\mathbf{z}_1 = \mathbf{z}_2 = ... = \mathbf{z}_n$. However, this is a trivial case that indicates serious constant collapse.

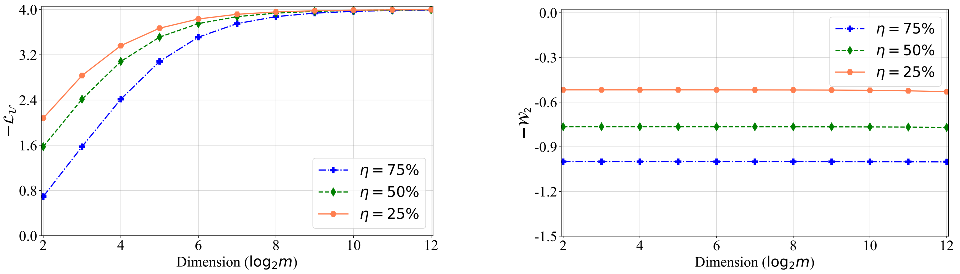


(a) Collapse analysis via $-\mathcal{L}_\mathcal{U}$

(b) Collapse analysis via $-\mathcal{W}_2$

Figure 2: Analysis on dimensional collapse degrees. $-\mathcal{W}_2$ is more sensitive to collapse degrees than $-\mathcal{L}_\mathcal{U}$.

On Dimensional Collapse Degrees To synthesize data exhibiting varying degrees of dimensional collapse, we sample data vectors from the standard Gaussian distribution and mask certain dimensions with zero vectors. This approach demonstrates that an increase in the number of masked dimensions leads to more significant dimensional collapse. The percentage of zero-value coordinates in the masked vectors is η while that of non-zero coordinates is $1 - \eta$. As shown in Figure 2(a) and Figure 2(b), $-\mathcal{W}_2$ is capable of capturing different collapse degrees, while $-\mathcal{L}_\mathcal{U}$ stays the same even with 80% collapse ($\eta = 80\%$), indicating that $-\mathcal{L}_\mathcal{U}$ is insensitive to dimensional collapse degrees.



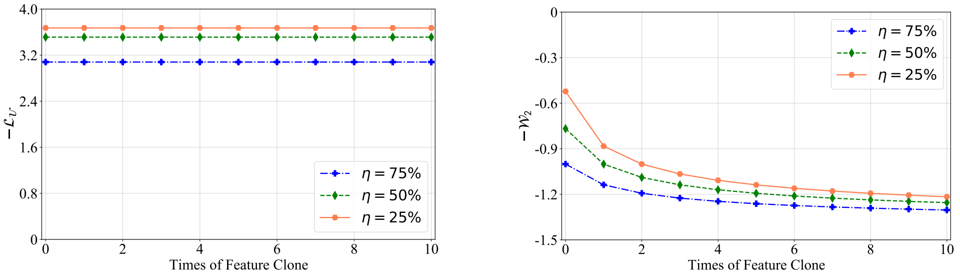
(a) Collapse analysis via $-\mathcal{L}_\mathcal{U}$

(b) Collapse analysis via $-\mathcal{W}_2$

Figure 3: Dimensional collapse w.r.t various dimensions. $-\mathcal{L}_\mathcal{U}$ fails to identify collapse degrees with a large dimension, while $-\mathcal{W}_2$ is able to identify collapse degrees no matter how great/small m is.

On Sensitiveness of Dimensions Moreover, Figure 3 shows that $-\mathcal{L}_\mathcal{U}$ can not distinguish different degrees of dimensional collapse ($\eta = 25\%, 50\%$, and 75%) when the dimension m becomes large (e.g., $m \geq 2^8$). In sharp contrast, $-\mathcal{W}_2$ only depends on the degree of dimensional collapse and is independent of the dimension when the dimensional collapse degree is fixed.

To complement the theoretical comparisons between the two metrics discussed in Section 5.1, we also conduct empirical comparisons in terms of FCC and FBC. ICC comparisons see in Appendix J.

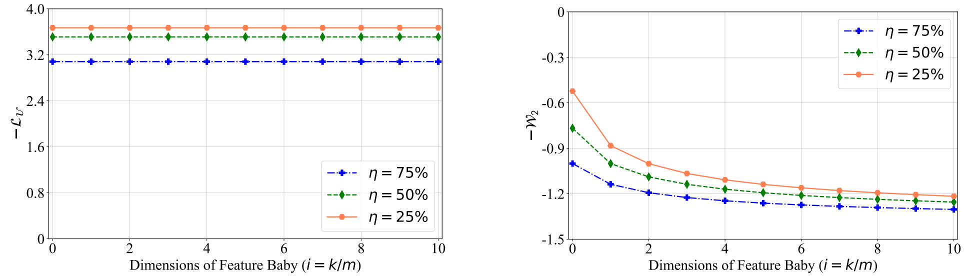


(a) FCC analysis via $-\mathcal{L}_\mathcal{U}$

(b) FCC analysis via $-\mathcal{W}_2$

Figure 4: FCC analysis w.r.t the times of feature clone.

On Feature Cloning Constraint (FCC) We further investigate the impact of feature cloning by creating multiple feature clones of the dataset, e.g., $\mathcal{D} \oplus \mathcal{D}$ and $\mathcal{D} \oplus \mathcal{D} \oplus \mathcal{D}$, corresponding to one and two times cloning, respectively. Figure 4(a) demonstrates that the value of $-\mathcal{L}_\mathcal{U}$ remains constant as the number of clones increases, which violates the inequality constraint in Equation 8. In contrast, in Figure 4(b), our proposed metric $-\mathcal{W}_2$ decreases, satisfying the constraint.



(a) FBC analysis via $-\mathcal{L}_\mathcal{U}$

(b) FBC analysis via $-\mathcal{W}_2$

Figure 5: FBC analysis w.r.t the times of feature baby.

On Feature Baby Constraint (FBC) We finally analyze the effect of feature baby, where we insert k dimension zero-value vectors into the dataset $\mathcal{D}$. This created dataset is denoted as $\mathcal{D} \oplus \mathbf{0}^k$, and we examine the impact of k in both metrics. Figure 5(a) shows that the value of $-\mathcal{L}_\mathcal{U}$ remains constant as k increases, violating the inequality constraint in Equation 9. In contrast, Figure 5(b) illustrates that our proposed metric $-\mathcal{W}_2$ decreases, satisfying the constraint.

In summary, our empirical results align with our theoretical analysis, confirming that our proposed metric $-\mathcal{W}_2$ performs better than the existing metric $-\mathcal{L}_\mathcal{U}$ in capturing feature redundancy and dimensional collapse.

6 EXPERIMENTS

In this section, we impose the proposed uniformity metric as an auxiliary loss term for various existing self-supervised methods, and conduct experiments on CIFAR-10 and CIFAR-100 datasets to demonstrate its effectiveness.

Models We conduct experiments on a series of self-supervised representation learning models: (i) AlignUniform (Wang & Isola, 2020), whose loss objective consists of an alignment objective and a uniform objective. (ii) three contrastive methods, i.e., SimCLR (Chen et al., 2020), MoCo (He et al., 2020), and NNCLR (Dwibedi et al., 2021). (iii) two asymmetric models, i.e., BYOL (Grill et al., 2020) and SimSiam (Chen & He, 2021). (iv) two methods via redundancy reduction, i.e., BarlowTwins (Zbontar et al., 2021) and Zero-CL (Zhang et al., 2022b). To study the behavior of proposed Wasserstein distance in the self-supervised representation learning, we impose it as an auxiliary loss term to the following models: MoCo v2, BYOL, BarlowTwins, and Zero-CL. To facilitate better use of Wasserstein distance, we also design a linear decay for weighting Wasserstein distance during the training phase, i.e., $\alpha_t = \alpha_{\max} - t\,(\alpha_{\max} - \alpha_{\min})/T$, where t, T, $\alpha_{\max}$, $\alpha_{\min}$, α_t are current epoch, maximum epochs, maximum weight, minimum weight, and current weight, respectively. More detailed experiments setting see in Appendix K.

Metrics We evaluate the above methods from two perspectives: one is linear evaluation accuracy measured by Top-1 accuracy (Acc@1) and Top-5 accuracy (Acc@5); another is representation capacity. According to (Arora et al., 2019; Wang & Isola, 2020), alignment and uniformity are the two most important properties to evaluate self-supervised representation learning. We use two metrics $\mathcal{L}_\mathcal{U}$ and $\mathcal{W}_2$ to measure the uniformity, and a metric $\mathcal{A}$ to measure the alignment between the positive pairs (Wang & Isola, 2020). More details about the alignment metric see in Appendix L.

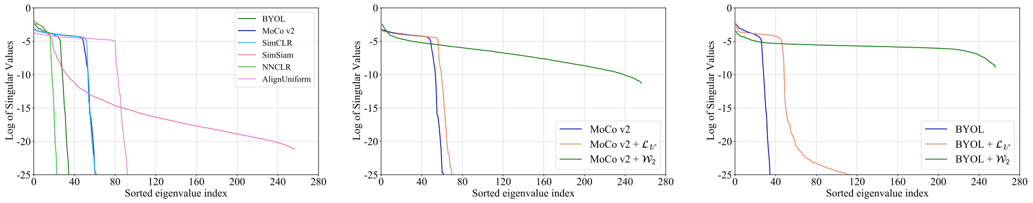
Main Results As shown in Table 2, we could observe that by imposing $\mathcal{W}_2$ as an additional loss it consistently outperforms the performance than that without the loss or that imposes $\mathcal{U}$ as the additional loss. Interestingly, although it slightly harms alignment, it usually results in improvement in uniformity and finally leads to better accuracy. This demonstrates the effectiveness of $\mathcal{W}_2$ as a uniformity metric. Note imposing an additional loss during training does not affect the training or inference efficiency; therefore, adding $\mathcal{W}_2$ as loss is beneficial without any tangible costs.

Convergence Analysis We test the Top-1 accuracy of these models on CIFAR-10 and CIFAR-100 via linear evaluation protocol (as described in Appendix K) when training them in different epochs. As

Table 2: Main comparison on CIFAR-10 and CIFAR-100 datasets. Proj. and Pred. are the hidden dimension in projector and predictor. $\uparrow$ and $\downarrow$ mean gains and losses, respectively.

Methods	Proj.	Pred.	CIFAR-10					CIFAR-100				
			Acc@1$\uparrow$	Acc@5$\uparrow$	$\mathcal{W}_2 \downarrow$	$\mathcal{L}_{\mathcal{U}} \downarrow$	$\mathcal{A} \downarrow$	Acc@1$\uparrow$	Acc@5$\uparrow$	$\mathcal{W}_2 \downarrow$	$\mathcal{L}_{\mathcal{U}} \downarrow$	$\mathcal{A} \downarrow$
SimCLR	256	✗	89.85	99.78	1.04	-3.75	0.47	63.43	88.97	1.05	-3.75	0.50
NNCLR	256	256	87.46	99.63	1.23	-3.12	0.38	54.90	83.81	1.23	-3.18	0.43
SimSiam	256	256	86.71	99.67	1.19	-3.33	0.39	56.10	84.34	1.21	-3.29	0.42
AlignUniform	256	✗	90.37	99.76	0.94	-3.82	0.51	65.08	90.15	0.95	-3.82	0.53
MoCo v2	256	✗	90.65	99.81	1.06	-3.75	0.51	60.27	86.29	1.07	-3.60	0.46
MoCo v2 + $\mathcal{L}_{\mathcal{U}}$	256	✗	90.98 $\uparrow_{0.33}$	99.67	0.98 $\uparrow_{0.08}$	-3.82	0.53 $\downarrow_{0.02}$	61.21 $\uparrow_{0.94}$	87.32	0.98 $\uparrow_{0.09}$	-3.81	0.52 $\downarrow_{0.06}$
MoCo v2 + $\mathcal{W}_2$	256	✗	91.41 $\uparrow_{0.76}$	99.68	0.33 $\uparrow_{0.73}$	-3.84	0.63 $\downarrow_{0.12}$	63.68 $\uparrow_{3.41}$	88.48	0.28 $\uparrow_{0.79}$	-3.86	0.66 $\downarrow_{0.20}$
BYOL	256	256	89.53	99.71	1.21	-2.99	**0.31**	63.66	88.81	1.20	-2.87	**0.33**
BYOL + $\mathcal{L}_{\mathcal{U}}$	256	✗	90.09 $\uparrow_{0.56}$	99.75	1.09 $\uparrow_{0.12}$	-3.66	0.40 $\downarrow_{0.09}$	62.68 $\downarrow_{0.98}$	88.44	1.08 $\uparrow_{0.12}$	-3.70	0.51 $\downarrow_{0.18}$
BYOL + $\mathcal{W}_2$	256	256	90.31 $\uparrow_{0.78}$	99.77	0.38 $\uparrow_{0.83}$	-3.90	0.65 $\downarrow_{0.34}$	65.16 $\uparrow_{1.50}$	89.25	0.36 $\uparrow_{0.84}$	-3.91	0.69 $\downarrow_{0.36}$
BarlowTwins	256	✗	91.16	99.80	0.22	-3.91	0.75	68.19	90.64	0.23	-3.91	0.75
BarlowTwins + $\mathcal{L}_{\mathcal{U}}$	256	✗	91.38 $\uparrow_{0.22}$	99.77	0.21 $\uparrow_{0.01}$	-3.92	0.76 $\downarrow_{0.01}$	68.41 $\uparrow_{0.22}$	90.99	0.22 $\uparrow_{0.01}$	-3.91	0.76 $\downarrow_{0.01}$
BarlowTwins + $\mathcal{W}_2$	256	✗	**91.43** $\uparrow_{0.27}$	99.78	0.19 $\uparrow_{0.03}$	-3.92	0.76 $\downarrow_{0.01}$	68.47 $\uparrow_{0.28}$	90.64	0.19 $\uparrow_{0.04}$	-3.91	0.79 $\downarrow_{0.04}$
Zero-CL	256	✗	91.35	99.74	0.15	**-3.94**	0.70	68.50	90.97	0.15	-3.93	0.75
Zero-CL + $\mathcal{L}_{\mathcal{U}}$	256	✗	91.28 $\downarrow_{0.07}$	99.74	0.15	**-3.94**	0.72 $\downarrow_{0.02}$	68.44 $\downarrow_{0.06}$	90.91	0.15	-3.93	0.74 $\uparrow_{0.01}$
Zero-CL + $\mathcal{W}_2$	256	✗	91.42 $\uparrow_{0.07}$	**99.82**	**0.14** $\uparrow_{0.01}$	**-3.94**	0.71 $\downarrow_{0.01}$	**68.55** $\uparrow_{0.05}$	**91.02**	**0.14** $\uparrow_{0.01}$	**-3.94**	0.76 $\downarrow_{0.01}$

shown in Figure 12. By imposing $\mathcal{W}_2$ as an additional loss for these models, it converges faster than the raw models, especially for MoCo v2 and BYOL with serious collapse problem. Our experiments show that imposing the proposed uniformity metric as an auxiliary penalty loss could largely improve uniformity but damage alignment. We further conduct uniformity and alignment analyses through all the training epochs in Figure 13 and Figure 14 respectively, see Appendix N.



(a) Singular Value Spectrum (b) MoCo v2 (c) BYOL

Figure 6: Dimensional collapse analysis on CIFAR-100 dataset.

Dimensional Collapse Analysis We visualize singular value spectrum of the representations (Jing et al., 2022) of various models, where the spectrum contains the singular values of the covariance matrix of representations from CIFAR-100 dataset in sorted order and logarithmic scale. As shown in Figure 6(a), most singular values collapse to zero in most models (exclude BarlowTwins and Zero-CL), indicating a large number of collapsed dimensions occur in most models. To further understand how the additional loss $\mathcal{W}_2$ benefits the alleviation of the dimensional collapse, we impose $\mathcal{W}_2$ as an additional loss for Moco v2 and BYOL models, as shown in Figure 6(b) and Figure 6(c), the number of collapsed dimensions almost decrease to zero, indicating $\mathcal{W}_2$ can effectively address the dimensional collapse issue. In contrast, the additional loss $\mathcal{L}_{\mathcal{U}}$ has a minimal effect in preventing dimensional collapse.

7 CONCLUSION

In this paper, we have identified five fundamental properties that an ideal uniformity metric should adhere to. However, the existing uniformity metric introduced by Wang & Isola (2020) falls short of satisfying three of these properties, underscoring its inability to account for dimensional collapse. Empirical studies further support this finding. To address this limitation, we have introduced a novel uniformity metric that successfully satisfies all of these desiderata, with a notable capability to capture dimensional collapse. When integrated as an auxiliary loss in various well-established self-supervised methods, our proposed uniformity metric consistently enhances their performance in downstream tasks. A possible limitation of our work is that the five identified desiderata may not provide a complete characterization of an ideal uniformity metric. A pursuit of complete characterization of an ideal uniformity metric is in order.

ACKNOWLEDGEMENT

This work is supported by the Shenzhen Science and Technology Program (JCYJ20220818103001002), Shenzhen Doctoral Startup Funding (RCBS20221008093330065), and Tianyuan Fund for Mathematics of National Natural Science Foundation of China (NSFC) (12326608).

REFERENCES

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

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *ICLR*, 2022.

A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, 1943.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Herv'e J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. The convex geometry of linear inverse problems. *Foundations of Computational mathematics*, 12:805–849, 2012.

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

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021.

Henry Cohn and Abhinav Kumar. Universally optimal distribution of points on spheres. *Journal of the American Mathematical Society*, 2007.

Victor Guilherme Turrisi da Costa, Enrico Fini, Moin Nabi, N. Sebe, and Elisa Ricci. Solo-learn: A library of self-supervised methods for visual representation learning. *JMLR*, 2022.

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.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *ArXiv*, 2021.

Jean-Bastien Grill, Florian Strub, Florent Altch'e, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, 2020.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006.

Kaiming He, X. 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 B. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531, 2015.

Tianyu Hua, Wenxiao Wang, Zihui Xue, Sucheng Ren, Yue Wang, and Hang Zhao. On feature decorrelation in self-supervised learning. In *ICCV*, 2021.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *ICLR*, 2022.

Junnan Li, Pan Zhou, Caiming Xiong, Richard Socher, and Steven C. H. Hoi. Prototypical contrastive learning of unsupervised representations. In *ICLR*, 2021.

Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *ICLR*, 2022.

David Lindley and Solomon Kullback. Information theory and statistics. *Journal of the American Statistical Association*, 54:825, 1959.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*, 2017.

Ingram Olkin and Friedrich Pukelsheim. The distance between two random vectors with given dispersion matrices. *Linear Algebra and its Applications*, 48:257–263, 1982.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

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

Ramon Van Handel. Probability in high dimension. *Lecture Notes (Princeton University)*, 2016.

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

Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *CVPR*, 2021.

Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. In *ICCV*, 2021.

Ceyuan Yang, Zhirong Wu, Bolei Zhou, and Stephen Lin. Instance localization for self-supervised detection pretraining. In *CVPR*, 2021.

Yang You, Igor Gitman, and Boris Ginsburg. Scaling sgd batch size to 32k for imagenet training. *ArXiv*, 2017.

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

Chaoning Zhang, Kang Zhang, Chenshuang Zhang, Trung X. Pham, Chang D. Yoo, and In So Kweon. How does simsiam avoid collapse without negative samples? a unified understanding with self-supervised contrastive learning. In *ICLR*, 2022a.

Shaofeng Zhang, Feng Zhu, Junchi Yan, Rui Zhao, and Xiaokang Yang. Zero-CL: Instance and feature decorrelation for negative-free symmetric contrastive learning. In *ICLR*, 2022b.

Xiangyu Zhao, Raviteja Vemulapalli, P. A. Mansfield, Boqing Gong, Bradley Green, Lior Shapira, and Ying Wu. Contrastive learning for label efficient semantic segmentation. In *ICCV*, 2021.

A PROBABILITY DENSITY FUNCTION OF $\mathbf{Y}_i$

Lemma 2. *For a random variable* $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_m)$*, and* $\mathbf{Z} \in \mathbb{R}^m$*, for the* ℓ_2*-normalized form* $\mathbf{Y} = \mathbf{Z}/\|\mathbf{Z}\|_2$*, the probability density function (pdf) of a variable* Y_i *in the* i*-th coordinate is:*

$$f_{Y_i}(y_i) = \frac{\Gamma(m/2)}{\sqrt{\pi}\Gamma((m-1)/2)}(1-y_i^2)^{(m-3)/2}, \ \ \forall\, y_i \in [-1,1].$$

Proof. $\mathbf{Z} = [Z_1, Z_2, \cdots, Z_m] \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_m)$, then $Z_i \sim \mathcal{N}(0, \sigma^2), \forall i \in [1, m]$. We denote the variable $U = Z_i/\sigma \sim \mathcal{N}(0,1)$, $V = \sum_{j\neq i}^m (Z_j/\sigma)^2 \sim \mathcal{X}^2(m-1)$, then U and V are independent with each other. For the variable $T = \frac{U}{\sqrt{V/(m-1)}}$, it obeys the Student's t-distribution with $m-1$ degrees of freedom, and its probability density function (pdf) is:

$$f_T(t) = \frac{\Gamma(m/2)}{\sqrt{(m-1)\pi}\Gamma((m-1)/2)}(1+\frac{t^2}{m-1})^{-m/2}.$$

For the variable $Y_i = \frac{Z_i}{\sqrt{\sum_{i=1}^m Z_i^2}} = \frac{Z_i}{\sqrt{Z_i^2+\sum_{j\neq i}^m Z_j^2}} = \frac{Z_i/\sigma}{\sqrt{(Z_i/\sigma)^2+\sum_{j\neq i}^m (Z_j/\sigma)^2}} = \frac{U}{\sqrt{U^2+V}}$, then $T = \frac{U}{\sqrt{V/(m-1)}} = \frac{\sqrt{m-1}Y_i}{\sqrt{1-Y_i^2}}$ and $Y_i = \frac{T}{\sqrt{T^2+m-1}}$, the relation between the cumulative distribution function (cdf) of T and that of Y_i can be formulated as follows:

$$\begin{aligned}
F_{Y_i}(y_i) = P(\{Y_i \le y_i\}) &= \begin{cases} P(\{Y_i \le y_i\}) & y_i \le 0 \\ P(\{Y_i \le 0\}) + P(\{0 < Y_i \le y_i\}) & y_i > 0 \end{cases} \\
&= \begin{cases} P(\{\frac{T}{\sqrt{T^2+m-1}} \le y_i\}) & y_i \le 0 \\ P(\{\frac{T}{\sqrt{T^2+m-1}} \le 0\}) + P(\{0 < \frac{T}{\sqrt{T^2+m-1}} \le y_i\}) & y_i > 0 \end{cases} \\
&= \begin{cases} P(\{\frac{T^2}{T^2+m-1} \ge y_i^2, T \le 0\}) & y_i \le 0 \\ P(\{T \le 0\} + P(\{\frac{T^2}{T^2+m-1} \le y_i^2, T > 0\}) & y_i > 0 \end{cases} \\
&= \begin{cases} P(\{T \le \frac{\sqrt{m-1}y_i}{\sqrt{1-y_i^2}}\}) & y_i \le 0 \\ P(\{T \le 0\} + P(\{0 < T \le \frac{\sqrt{m-1}y_i}{\sqrt{1-y_i^2}}\}) & y_i > 0 \end{cases} \\
&= P(\{T \le \frac{\sqrt{m-1}y_i}{\sqrt{1-y_i^2}}\}) = F_T(\frac{\sqrt{m-1}y_i}{\sqrt{1-y_i^2}}).
\end{aligned}$$

Therefore, the pdf of Y_i can be derived as follows:

$$\begin{aligned}
f_{Y_i}(y_i) &= \frac{d}{dy_i}F_{Y_i}(y_i) = \frac{d}{dy_i}F_T(\frac{\sqrt{m-1}y_i}{\sqrt{1-y_i^2}}) \\
&= f_T(\frac{\sqrt{m-1}y_i}{\sqrt{1-y_i^2}})\frac{d}{dy_i}(\frac{\sqrt{m-1}y_i}{\sqrt{1-y_i^2}}) \\
&= [\frac{\Gamma(m/2)}{\sqrt{(m-1)\pi}\Gamma((m-1)/2)}(1-y_i^2)^{m/2}][\sqrt{m-1}(1-y_i^2)^{-3/2}] \\
&= \frac{\Gamma(m/2)}{\sqrt{\pi}\Gamma((m-1)/2)}(1-y_i^2)^{(m-3)/2}.
\end{aligned}$$

□

B PROOF OF THE THEOREM 2

Proof. According to the Lemma 2, the pdf of Y_i and $\widehat{Y}_i \sim \mathcal{N}(0, \frac{1}{m})$ are:

$$f_{Y_i}(y) = \frac{\Gamma(m/2)}{\sqrt{\pi}\Gamma((m-1)/2)}(1-y^2)^{(m-3)/2}, \quad f_{\widehat{Y}_i}(y) = \sqrt{\frac{m}{2\pi}}\exp\{-\frac{my^2}{2}\}.$$

Then the Kullback-Leibler divergence between Y_i and $\widehat{Y}_i$ can be formulated as:

$$
\begin{aligned}
D_{\mathrm{KL}}(Y_i \| \widehat{Y}_i) &= \int_{-1}^{1} f_{Y_i}(y)[\log f_{Y_i}(y) - \log f_{\widehat{Y}_i}(y)]dy \\
&= \int_{-1}^{1} f_{Y_i}(y)[\log \frac{\Gamma(m/2)}{\sqrt{\pi}\Gamma((m-1)/2)} + \frac{m-3}{2}\log(1-y^2) - \log\sqrt{\frac{m}{2\pi}} + \frac{my^2}{2}]dy \\
&= \log\sqrt{\frac{2}{m}}\frac{\Gamma(m/2)}{\Gamma((m-1)/2)} + \int_{-1}^{1} f_{Y_i}(y)[\frac{m-3}{2}\log(1-y^2) + \frac{my^2}{2}]dy.
\end{aligned}
$$

We set $\mu = y^2$, we have $y = \sqrt{\mu}$, and $dy = \frac{1}{2}\mu^{-\frac{1}{2}}du$. Then:

$$
\begin{aligned}
\mathcal{A} &:= \int_{-1}^{1} f_{Y_i}(y)[\frac{m-3}{2}\log(1-y^2) + \frac{my^2}{2}]dy \\
&= 2\int_{0}^{1} \frac{\Gamma(m/2)}{\sqrt{\pi}\Gamma((m-1)/2)}(1-y^2)^{\frac{m-3}{2}}[\frac{m-3}{2}\log(1-y^2) + \frac{my^2}{2}]dy \\
&= \frac{\Gamma(m/2)}{\sqrt{\pi}\Gamma((m-1)/2)}\int_{0}^{1}(1-\mu)^{\frac{m-3}{2}}[\frac{m-3}{2}\log(1-\mu) + \frac{m}{2}\mu]\mu^{-\frac{1}{2}}d\mu \\
&= \frac{\Gamma(m/2)}{\sqrt{\pi}\Gamma((m-1)/2)}\frac{m-3}{2}\int_{0}^{1}(1-\mu)^{\frac{m-3}{2}}\mu^{-\frac{1}{2}}\log(1-\mu)d\mu \\
&+ \frac{\Gamma(m/2)}{\sqrt{\pi}\Gamma((m-1)/2)}\frac{m}{2}\int_{0}^{1}(1-\mu)^{\frac{m-3}{2}}\mu^{\frac{1}{2}}d\mu.
\end{aligned}
$$

By using the property of Beta distribution, and the inequality $\frac{-\mu}{1-\mu} \leq \log(1-\mu) \leq -\mu$, we have:

$$
\begin{aligned}
\mathcal{A}_1 &:= \frac{\Gamma(m/2)}{\sqrt{\pi}\Gamma((m-1)/2)}\frac{m-3}{2}\int_{0}^{1}(1-\mu)^{\frac{m-3}{2}}\mu^{-\frac{1}{2}}\log(1-\mu)d\mu \\
&\leq -\frac{\Gamma(m/2)}{\sqrt{\pi}\Gamma((m-1)/2)}\frac{m-3}{2}\int_{0}^{1}(1-\mu)^{\frac{m-3}{2}}\mu^{\frac{1}{2}}d\mu \\
&= -\frac{\Gamma(m/2)}{\sqrt{\pi}\Gamma((m-1)/2)}\frac{m-3}{2}B(\frac{3}{2}, \frac{m-1}{2}) \\
\mathcal{A}_2 &:= \frac{\Gamma(m/2)}{\sqrt{\pi}\Gamma((m-1)/2)}\frac{m}{2}\int_{0}^{1}(1-\mu)^{\frac{m-3}{2}}\mu^{\frac{1}{2}}d\mu \\
&= \frac{\Gamma(m/2)}{\sqrt{\pi}\Gamma((m-1)/2)}\frac{m}{2}B(\frac{3}{2}, \frac{m-1}{2}).
\end{aligned}
$$

Then, we have:

$$
\begin{aligned}
\mathcal{A} = \mathcal{A}_1 + \mathcal{A}_2 &\leq -\frac{\Gamma(m/2)}{\sqrt{\pi}\Gamma((m-1)/2)}\frac{m-3}{2}B(\frac{3}{2}, \frac{m-1}{2}) + \frac{\Gamma(m/2)}{\sqrt{\pi}\Gamma((m-1)/2)}\frac{m}{2}B(\frac{3}{2}, \frac{m-1}{2}) \\
&= \frac{3}{2}\frac{\Gamma(m/2)}{\sqrt{\pi}\Gamma((m-1)/2)}B(\frac{3}{2}, \frac{m-1}{2}) = \frac{3}{2}\frac{\Gamma(m/2)}{\sqrt{\pi}\Gamma((m-1)/2)}\frac{\Gamma(3/2)\Gamma((m-1)/2)}{\Gamma((m+2)/2)} \\
&= \frac{3}{2}\frac{\Gamma(3/2)\Gamma(m/2)}{\sqrt{\pi}\Gamma((m+2)/2)} = \frac{3}{2}\frac{(\sqrt{\pi}/2)\Gamma(m/2)}{\sqrt{\pi}\Gamma((m+2)/2)} = \frac{3}{4}\frac{\Gamma(m/2)}{\Gamma((m+2)/2)}.
\end{aligned}
$$

According to the Stirling formula, we have $\Gamma(x+\alpha) \to \Gamma(x)x^{\alpha}$ as $x \to \infty$, therefore:

$$
\begin{aligned}
\lim_{m\to\infty} D_{\mathrm{KL}}(Y_i \| \widehat{Y}_i) &= \lim_{m\to\infty} \log\sqrt{\frac{2}{m}}\frac{\Gamma(m/2)}{\Gamma((m-1)/2)} + \lim_{m\to\infty}\mathcal{A} \\
&\leq \lim_{m\to\infty} \log\sqrt{\frac{2}{m}}\frac{\Gamma((m-1)/2)(\frac{m-1}{2})^{1/2}}{\Gamma((m-1)/2)} + \lim_{m\to\infty}\frac{3}{4}\frac{\Gamma(m/2)}{\Gamma((m+2)/2)} \\
&= \lim_{m\to\infty} \log\sqrt{\frac{2}{m}}\sqrt{\frac{m-1}{2}} + \frac{3}{4}\frac{\Gamma(m/2)}{\Gamma(m/2)m} = \lim_{m\to\infty}\log\sqrt{\frac{m-1}{m}} + \frac{3}{4m} = 0.
\end{aligned}
$$

We further use T_2 inequality (Van Handel, 2016, Theorem 4.31) to derive the quadratic Wasserstein metric (Van Handel, 2016, Definition 4.29) as:

$$\lim_{m\to\infty} \mathcal{W}_2(Y_i, \widehat{Y}_i) \leq \lim_{m\to\infty} \sqrt{\frac{2}{m} D_{\mathrm{KL}}(Y_i \| \widehat{Y}_i)} = 0.$$

□

C EXAMINING THE DESIDERATA FOR TWO UNIFORMITY METRICS

C.1 PROOF FOR $-\mathcal{L}_{\mathcal{U}}$ ON DESIDERATA

The first two properties (Property 1 and 2) could be easily proved using the definition. We here examine the rest three properties one by one for the existing uniformity metric $-\mathcal{L}_{\mathcal{U}}$.

Proof. Firstly, we prove that the baseline metric $-\mathcal{L}_{\mathcal{U}}$ cannot satisfy the Property 3. According to the definition of $\mathcal{L}_{\mathcal{U}}$ in Equation 2, we have:

$$\begin{aligned}\mathcal{L}_{\mathcal{U}}(\mathcal{D} \cup \mathcal{D}) &:= \log \frac{1}{2n(2n-1)/2}\Big(4\sum_{i=2}^{n}\sum_{j=1}^{i-1} e^{-t\|\frac{\mathbf{z}_i}{\|\mathbf{z}_i\|} - \frac{\mathbf{z}_j}{\|\mathbf{z}_j\|}\|_2^2} + \sum_{i=1}^{n} e^{-t\|\frac{\mathbf{z}_i}{\|\mathbf{z}_i\|} - \frac{\mathbf{z}_i}{\|\mathbf{z}_i\|}\|_2^2}\Big) \\ &= \log \frac{1}{2n(2n-1)/2}\Big(4\sum_{i=2}^{n}\sum_{j=1}^{i-1} e^{-t\|\frac{\mathbf{z}_i}{\|\mathbf{z}_i\|} - \frac{\mathbf{z}_j}{\|\mathbf{z}_j\|}\|_2^2} + n\Big).\end{aligned}$$

We set $G = \sum_{i=2}^{n}\sum_{j=1}^{i-1} e^{-t\|\frac{\mathbf{z}_i}{\|\mathbf{z}_i\|} - \frac{\mathbf{z}_j}{\|\mathbf{z}_j\|}\|_2^2}$, and then, we have:

$$G = \sum_{i=2}^{n}\sum_{j=1}^{i-1} e^{-t\|\frac{\mathbf{z}_i}{\|\mathbf{z}_i\|} - \frac{\mathbf{z}_j}{\|\mathbf{z}_j\|}\|_2^2} \leq \sum_{i=2}^{n}\sum_{j=1}^{i-1} e^{-t\|\frac{\mathbf{z}_i}{\|\mathbf{z}_i\|} - \frac{\mathbf{z}_i}{\|\mathbf{z}_i\|}\|_2^2} = n(n-1)/2.$$

$G = n(n-1)/2$ if and only if $\mathbf{z}_1 = \mathbf{z}_2 = ... = \mathbf{z}_n$.

$$\begin{aligned}\mathcal{L}_{\mathcal{U}}(\mathcal{D} \cup \mathcal{D}) - \mathcal{L}_{\mathcal{U}}(\mathcal{D}) &= \log \frac{4G+n}{2n(2n-1)/2} - \log \frac{G}{n(n-1)/2} \\ &= \log \frac{(4G+n)n(n-1)/2}{2nG(2n-1)/2} = \log \frac{(4G+n)(n-1)}{4nG-2G} \\ &= \log \frac{4nG - 4G + n^2 - n}{4nG - 2G} \geq \log 1 = 0.\end{aligned}$$

$\mathcal{L}_{\mathcal{U}}(\mathcal{D} \cup \mathcal{D}) = \mathcal{L}_{\mathcal{U}}(\mathcal{D})$ if and only if $G = n(n-1)/2$, which requires $\mathbf{z}_1 = \mathbf{z}_2 = ... = \mathbf{z}_n$ (a trivial case that all representations collapse to a constant point. We exclude this trivial case for consideration in the paper, and we have $-\mathcal{L}_{\mathcal{U}}(\mathcal{D} \cup \mathcal{D}) < -\mathcal{L}_{\mathcal{U}}(\mathcal{D})$. Therefore, the baseline metric $-\mathcal{L}_{\mathcal{U}}$ cannot satisfy the Property 3.

Then, we prove that the baseline metric $-\mathcal{L}_{\mathcal{U}}$ cannot satisfy the Property 4. Given $\mathbf{z}_i = [z_{i1}, z_{i2}, ..., z_{im}]^T$, and $\mathbf{z}_j = [z_{j1}, z_{j2}, ..., z_{jm}]^T$, and we set $\widehat{\mathbf{z}}_i = \mathbf{z}_i \oplus \mathbf{z}_i$ and $\widehat{\mathbf{z}}_j = \mathbf{z}_j \oplus \mathbf{z}_j$, we have:

$$\mathcal{L}_{\mathcal{U}}(\mathcal{D} \oplus \mathcal{D}) := \log \frac{1}{n(n-1)/2} \sum_{i=2}^{n}\sum_{j=1}^{i-1} e^{-t\|\frac{\widehat{\mathbf{z}}_i}{\|\widehat{\mathbf{z}}_i\|} - \frac{\widehat{\mathbf{z}}_j}{\|\widehat{\mathbf{z}}_j\|}\|_2^2}.$$

As $\widehat{\mathbf{z}}_i = [z_{i1}, z_{i2}, ..., z_{im}, z_{i1}, z_{i2}, ..., z_{im}]^T$ and $\widehat{\mathbf{z}}_j = [z_{j1}, z_{j2}, ..., z_{jm}, z_{j1}, z_{j2}, ..., z_{jm}]^T$, then $\|\widehat{\mathbf{z}}_i\| = \sqrt{2}\|\mathbf{z}_i\|$, $\|\widehat{\mathbf{z}}_j\| = \sqrt{2}\|\mathbf{z}_j\|$, and $\langle \widehat{\mathbf{z}}_i, \widehat{\mathbf{z}}_j \rangle = 2\langle \mathbf{z}_i, \mathbf{z}_j \rangle$, we have:

$$\|\frac{\widehat{\mathbf{z}}_i}{\|\widehat{\mathbf{z}}_i\|} - \frac{\widehat{\mathbf{z}}_j}{\|\widehat{\mathbf{z}}_j\|}\|_2^2 = 2 - 2\frac{\langle \widehat{\mathbf{z}}_i, \widehat{\mathbf{z}}_j \rangle}{\|\widehat{\mathbf{z}}_i\|\|\widehat{\mathbf{z}}_j\|} = 2 - 2\frac{2\langle \mathbf{z}_i, \mathbf{z}_j \rangle}{\sqrt{2}\|\mathbf{z}_i\|\sqrt{2}\|\mathbf{z}_j\|} = \|\frac{\mathbf{z}_i}{\|\mathbf{z}_i\|} - \frac{\mathbf{z}_j}{\|\mathbf{z}_j\|}\|_2^2.$$

Therefore, $-\mathcal{L}_{\mathcal{U}}(\mathcal{D} \oplus \mathcal{D}) = -\mathcal{L}_{\mathcal{U}}(\mathcal{D})$, indicating that the baseline metric $-\mathcal{L}_{\mathcal{U}}$ cannot satisfy the Property 4.

Finally, we prove that the baseline metric $-\mathcal{L}_{\mathcal{U}}$ cannot satisfy the Property 5. Given $\mathbf{z}_i = [z_{i1}, z_{i2}, ..., z_{im}]^T$, and $\mathbf{z}_j = [z_{j1}, z_{j2}, ..., z_{jm}]^T$, and we set $\widehat{\mathbf{z}}_i = \mathbf{z}_i \oplus \mathbf{0}^k$ and $\widehat{\mathbf{z}}_j = \mathbf{z}_j \oplus \mathbf{0}^k$, we have:

$$\mathcal{L}_{\mathcal{U}}(\mathcal{D} \oplus \mathbf{0}^k) := \log \frac{1}{n(n-1)/2} \sum_{i=2}^{n} \sum_{j=1}^{i-1} e^{-t\| \frac{\widehat{\mathbf{z}}_i}{\|\widehat{\mathbf{z}}_i\|} - \frac{\widehat{\mathbf{z}}_j}{\|\widehat{\mathbf{z}}_j\|} \|_2^2}.$$

As $\widehat{\mathbf{z}}_i = [z_{i1}, z_{i2}, ..., z_{im}, 0, 0, ..., 0]^T$, and $\widehat{\mathbf{z}}_j = [z_{j1}, z_{j2}, ..., z_{jm}, 0, 0, ..., 0]^T$, then $\|\widehat{\mathbf{z}}_i\| = \|\mathbf{z}_i\|$, $\|\widehat{\mathbf{z}}_j\| = \|\mathbf{z}_j\|$, and $\langle \widehat{\mathbf{z}}_i, \widehat{\mathbf{z}}_j \rangle = \langle \mathbf{z}_i, \mathbf{z}_j \rangle$, therefore:

$$\| \frac{\widehat{\mathbf{z}}_i}{\|\widehat{\mathbf{z}}_i\|} - \frac{\widehat{\mathbf{z}}_j}{\|\widehat{\mathbf{z}}_j\|} \|_2^2 = 2 - 2\frac{\langle \widehat{\mathbf{z}}_i, \widehat{\mathbf{z}}_j \rangle}{\|\widehat{\mathbf{z}}_i\| \|\widehat{\mathbf{z}}_j\|} = 2 - 2\frac{\langle \mathbf{z}_i, \mathbf{z}_j \rangle}{\|\mathbf{z}_i\| \|\mathbf{z}_j\|} = \| \frac{\mathbf{z}_i}{\|\mathbf{z}_i\|} - \frac{\mathbf{z}_j}{\|\mathbf{z}_j\|} \|_2^2.$$

Therefore, $-\mathcal{L}_{\mathcal{U}}(\mathcal{D} \oplus \mathbf{0}^k) = -\mathcal{L}_{\mathcal{U}}(\mathcal{D})$, indicating that the baseline metric $-\mathcal{L}_{\mathcal{U}}$ cannot satisfy the Property 5. $\square$

C.2 PROOF FOR $-\mathcal{W}_2$ ON DESIDERATA

The first two properties (Property 1 and 2) could be easily proved using the definition. We here to examine the rest three properties one by one for the proposed uniformity metric $-\mathcal{W}_2$.

Proof. Firstly, we prove that our proposed metric $-\mathcal{W}_2$ could satisfy the Property 3. As $\mathcal{D} \cup \mathcal{D} = \{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n, \mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n\}$, then its mean vector and covariance matrix can be formulated as follows:

$$\widehat{\boldsymbol{\mu}} = \frac{1}{2n} \sum_{i=1}^{n} 2\mathbf{z}_i/\|\mathbf{z}_i\| = \boldsymbol{\mu}, \quad \widehat{\mathbf{\Sigma}} = \frac{1}{2n} \sum_{i=1}^{n} 2(\mathbf{z}_i/\|\mathbf{z}_i\| - \widehat{\boldsymbol{\mu}})^T (\mathbf{z}_i/\|\mathbf{z}_i\| - \widehat{\boldsymbol{\mu}}) = \mathbf{\Sigma}.$$

Then we have:

$$\mathcal{W}_2(\mathcal{D} \cup \mathcal{D}) := \sqrt{\|\widehat{\boldsymbol{\mu}}\|_2^2 + 1 + \operatorname{tr}(\widehat{\mathbf{\Sigma}}) - \frac{2}{\sqrt{m}} \operatorname{tr}(\widehat{\mathbf{\Sigma}}^{1/2})} = \mathcal{W}_2(\mathcal{D}).$$

Therefore, $-\mathcal{W}_2(\mathcal{D} \cup \mathcal{D}) = -\mathcal{W}_2(\mathcal{D})$, indicating that our proposed metric $-\mathcal{W}_2$ could satisfy the Property 3.
Then, we prove that our proposed metric $-\mathcal{W}_2$ could satisfy the Property 4. Given $\mathbf{z}_i = [z_{i1}, z_{i2}, ..., z_{im}]^T$, and $\widehat{\mathbf{z}}_i = \mathbf{z}_i \oplus \mathbf{z}_i = [z_{i1}, z_{i2}, ..., z_{im}, z_{i1}, z_{i2}, ..., z_{im}]^T \in \mathbb{R}^{2m}$, for the set: $\mathcal{D} \oplus \mathcal{D}$, its mean vector and covariance matrix can be formulated as follows:

$$\widehat{\boldsymbol{\mu}} = \begin{pmatrix} \boldsymbol{\mu}/\sqrt{2} \\ \boldsymbol{\mu}/\sqrt{2} \end{pmatrix}, \quad \widehat{\mathbf{\Sigma}} = \begin{pmatrix} \mathbf{\Sigma}/2 & \mathbf{\Sigma}/2 \\ \mathbf{\Sigma}/2 & \mathbf{\Sigma}/2 \end{pmatrix}.$$

As $\widehat{\mathbf{\Sigma}}^{1/2} = \begin{pmatrix} \mathbf{\Sigma}^{1/2}/2 & \mathbf{\Sigma}^{1/2}/2 \\ \mathbf{\Sigma}^{1/2}/2 & \mathbf{\Sigma}^{1/2}/2 \end{pmatrix}$, $\operatorname{tr}(\widehat{\mathbf{\Sigma}}) = \operatorname{tr}(\mathbf{\Sigma})$ and $\operatorname{tr}(\widehat{\mathbf{\Sigma}}^{1/2}) = \operatorname{tr}(\mathbf{\Sigma}^{1/2})$, Then we have:

$$\begin{aligned} \mathcal{W}_2(\mathcal{D} \oplus \mathcal{D}) &:= \sqrt{\|\widehat{\boldsymbol{\mu}}\|_2^2 + 1 + \operatorname{tr}(\widehat{\mathbf{\Sigma}}) - \frac{2}{\sqrt{2m}} \operatorname{tr}(\widehat{\mathbf{\Sigma}}^{1/2})} \\ &= \sqrt{\|\boldsymbol{\mu}\|_2^2 + 1 + \operatorname{tr}(\mathbf{\Sigma}) - \frac{2}{\sqrt{2m}} \operatorname{tr}(\mathbf{\Sigma}^{1/2})} \\ &> \sqrt{\|\boldsymbol{\mu}\|_2^2 + 1 + \operatorname{tr}(\mathbf{\Sigma}) - \frac{2}{\sqrt{m}} \operatorname{tr}(\mathbf{\Sigma}^{1/2})} = \mathcal{W}_2(\mathcal{D}). \end{aligned}$$

Therefore, $-\mathcal{W}_2(\mathcal{D} \oplus \mathcal{D}) < -\mathcal{W}_2(\mathcal{D})$, indicating that our proposed metric $-\mathcal{W}_2$ could satisfy the Property 4.
Finally, we prove that our proposed metric $-\mathcal{W}_2$ could satisfy the Property 5. Given $\mathbf{z}_i =$

$[z_{i1}, z_{i2}, ..., z_{im}]^T$, and $\widehat{\mathbf{z}}_i = \mathbf{z}_i \oplus \mathbf{0}^k = [z_{i1}, z_{i2}, ..., z_{im}, 0, 0, ..., 0]^T \in \mathbb{R}^{m+k}$, for the set: $\mathcal{D} \oplus \mathbf{0}^k$, its mean vector and covariance matrix can be formulated as follows:

$$\widehat{\boldsymbol{\mu}} = \begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{0}^k \end{pmatrix}, \quad \widehat{\boldsymbol{\Sigma}} = \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{0}^{m\times k} \\ \mathbf{0}^{k\times m} & \mathbf{0}^{k\times k} \end{pmatrix}.$$

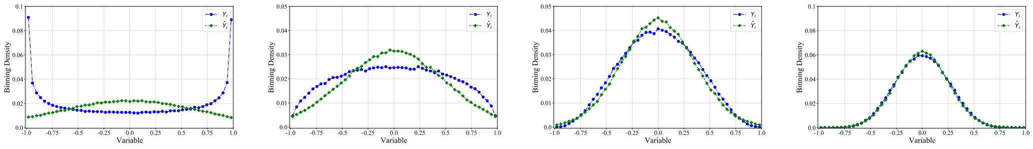
Therefore, $\operatorname{tr}(\widehat{\boldsymbol{\Sigma}}) = \operatorname{tr}(\boldsymbol{\Sigma})$, and $\operatorname{tr}(\widehat{\boldsymbol{\Sigma}}^{1/2}) = \operatorname{tr}(\boldsymbol{\Sigma}^{1/2})$:

$$\begin{aligned} \mathcal{W}_2(\mathcal{D} \oplus \mathbf{0}^k) &:= \sqrt{\|\widehat{\boldsymbol{\mu}}\|_2^2 + 1 + \operatorname{tr}(\widehat{\boldsymbol{\Sigma}}) - \frac{2}{\sqrt{m+k}} \operatorname{tr}(\widehat{\boldsymbol{\Sigma}}^{1/2})} \\ &= \sqrt{\|\boldsymbol{\mu}\|_2^2 + 1 + \operatorname{tr}(\boldsymbol{\Sigma}) - \frac{2}{\sqrt{m+k}} \operatorname{tr}(\boldsymbol{\Sigma}^{1/2})} \\ &> \sqrt{\|\boldsymbol{\mu}\|_2^2 + 1 + \operatorname{tr}(\boldsymbol{\Sigma}) - \frac{2}{\sqrt{m}} \operatorname{tr}(\boldsymbol{\Sigma}^{1/2})} = \mathcal{W}_2(\mathcal{D}). \end{aligned}$$

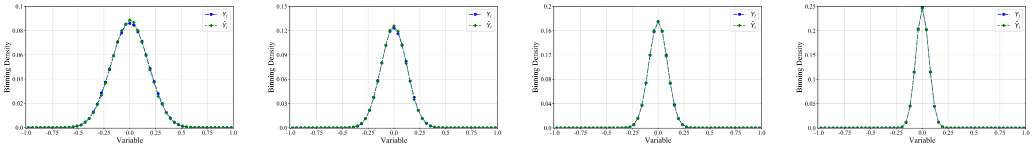
Therefore, $-\mathcal{W}_2(\mathcal{D} \oplus \mathbf{0}^k) < -\mathcal{W}_2(\mathcal{D})$, indicating that our proposed metric $-\mathcal{W}_2$ could satisfy the Property 5. □

D BINNING DENSITIES OF Y_i AND $\widehat{Y}_i$

We compare the distributions of Y_i and $\widehat{Y}_i$ coordinately. To estimate the distributions of Y_i and $\widehat{Y}_i$, we bin 200,000 sampled data points, aka samples, into 51 groups. Figure 7 compares the binning densities of Y_i and $\widehat{Y}_i$ when $m \in \{2, 4, 8, 16, 32, 64, 128, 256\}$. We can observe that two distributions are highly overlapped when m is moderately large, e.g., $m \geq 8$ or $m \geq 16$.



(a) $m = 2$ (b) $m = 4$ (c) $m = 8$ (d) $m = 16$



(e) $m = 32$ (f) $m = 64$ (g) $m = 128$ (h) $m = 256$

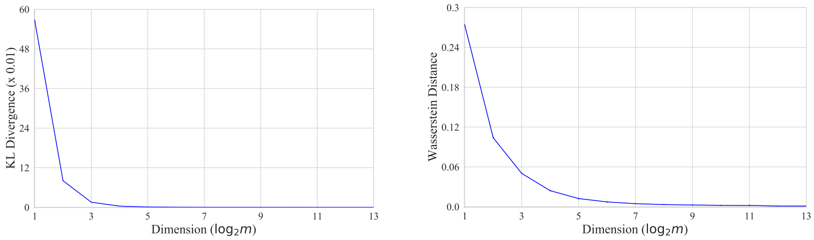
Figure 7: Comparing the binning densities of Y_i and $\widehat{Y}_i$ with various dimensions. For 2d visualization, see Figure 9 Appendix F.

E DISTANCES BETWEEN Y_i AND $\widehat{Y}_i$

We employ the KL Divergence and Wasserstein distance (see definition in Appendix G) to quantitatively measure the distribution distance between Y_i and $\widehat{Y}_i$. Specifically, we estimate the distributions of Y_i and $\widehat{Y}_i$ by bin 200,000 sampled data points, aka samples, into 51 groups. Then, we instantiate $\mathbb{P}_r$ and $\mathbb{P}_g$ with the binning densities of Y_i and $\widehat{Y}_i$, and finally calculate $D_{\mathrm{KL}}(\mathbb{P}_r \| \mathbb{P}_g)$ and $W_1(\mathbb{P}_r, \mathbb{P}_g)$ (see definition in Appendix G) ten times and average them, as visualized in Figure 8.

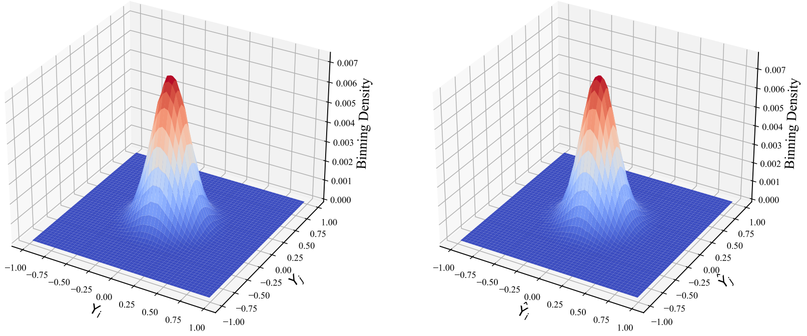
F A TWO-DIMENSIONAL VISUALIZATION FOR $\mathbf{Y}$ AND $\widehat{\mathbf{Y}}$

By binning 2000000 data samples into 51×51 groups in two-axis, we also analyze the joint binning density and present 2D joint binning density of two arbitrary individual dimensions, Y_i and Y_j $(i \neq j)$ in Figure 9(a), and $\widehat{Y}_i$ and $\widehat{Y}_j$ $(i \neq j)$ in Figure 9(b). Even if m is relatively small (i.e., 32), it looks like the density of the two distributions is close.



(a) KL Divergence (b) Wasserstein Distance

Figure 8: Two distances between Y_i and $\widehat{Y}_i$ w.r.t various dimensions.



(a) Density for Y_i and Y_j (b) Density for $\widehat{Y}_i$ and $\widehat{Y}_j$

Figure 9: Visualization of two arbitrary dimensions for $\mathbf{Y}$ and $\widehat{\mathbf{Y}}$ when $m = 32$.

G THE DEFINITION OF WASSERSTEIN DISTANCE

Definition 1. Wasserstein Distance or Earth-Mover Distance with p norm is defined as below:

$$W_p(\mathbb{P}_r, \mathbb{P}_g) = \Big(\inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y)\sim\gamma} \big[\|x - y\|^p \big] \Big)^{1/p} \ . \tag{13}$$

where $\Pi(\mathbb{P}_r, \mathbb{P}_g)$ denotes the set of all joint distributions $\gamma(x, y)$ whose marginals are respectively $\mathbb{P}_r$ and $\mathbb{P}_g$. Intuitively, when viewing each distribution as a unit amount of earth/soil, Wasserstein Distance or Earth-Mover Distance takes the minimum cost of transporting "mass" from x to y to transform the distribution $\mathbb{P}_r$ into the distribution $\mathbb{P}_g$.

H OTHER DISTRIBUTION DISTANCES OVER GAUSSIAN DISTRIBUTION

In this section, besides Wasserstein distance over Gaussian distribution, as shown in Lemma 1, we also discuss using other distribution distances as uniformity metrics and make comparisons with Wasserstein distance. As provided Kullback-Leibler Divergence and Bhattacharyya Distance over Gaussian distribution in Lemma 3 and in Lemma 4, both calculations require the covariance matrix to be a full rank matrix, making them hard to be used to conduct dimensional collapse analysis. On the contrary, our proposed uniformity metric via Wasserstein distance is free from such requirements on the covariance matrix, making it easier to be widely used in practical scenarios.

Lemma 3 (Kullback-Leibler divergence (Lindley & Kullback, 1959))**.** *Suppose two random variables* $\mathbf{Z}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ *and* $\mathbf{Z}_2 \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ *obey multivariate normal distributions, then Kullback-Leibler divergence between* $\mathbf{Z}1$ *and* $\mathbf{Z}_2$ *is:*

$$D_{\mathrm{KL}}(\mathbf{Z}_1, \mathbf{Z}_2) = \frac{1}{2}((\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \mathrm{tr}(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1 - \mathbf{I}) + \ln \frac{\det \boldsymbol{\Sigma}_2}{\det \boldsymbol{\Sigma}_1}).$$

Lemma 4 (Bhattacharyya Distance (Bhattacharyya, 1943))**.** *Suppose two random variables* $\mathbf{Z}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ *and* $\mathbf{Z}_2 \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ *obey multivariate normal distributions,* $\boldsymbol{\Sigma} = \frac{1}{2}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)$*, then bhattacharyya distance between* $\mathbf{Z}1$ *and* $\mathbf{Z}_2$ *is:*

$$\mathcal{D}_B(\mathbf{Z}_1, \mathbf{Z}_2) = \frac{1}{8}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \frac{1}{2} \ln \frac{\det \boldsymbol{\Sigma}}{\sqrt{\det \boldsymbol{\Sigma}_1 \det \boldsymbol{\Sigma}_2}}.$$

I CORRELATION BETWEEN $-\mathcal{L}_{\mathcal{U}}$ AND $-\mathcal{W}_2$.

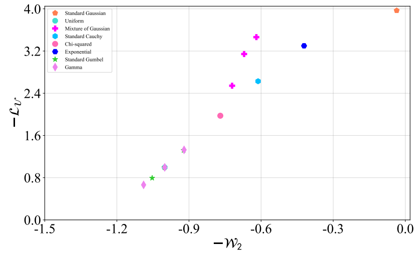


Figure 10: Uniformity analysis on distributions via two metrics.

We employ synthetic experiments to study uniformity metrics across different distributions. In detail, we manually sample 50000 data vectors from different distributions, such as standard Gaussian distribution $\mathcal{N}(\mathbf{0}, I)$, uniform Distribution $U(\mathbf{0}, \mathbf{1})$, the mixture of Gaussian, etc. Based on these data vectors, we estimate the uniformity of different distributions by two metrics. As shown in Fig. 10, standard Gaussian distribution achieves the maximum values by both $-\mathcal{W}_2$ and $-\mathcal{L}_{\mathcal{U}}$, which indicates that standard Gaussian distribution could achieve larger uniformity than other distributions. This empirical result is consistent with the Fact 1 that standard Gaussian distribution achieves the maximum uniformity.

J EMPIRICAL COMPARISON ON INSTANCE CLONING CONSTRAINT (ICC)

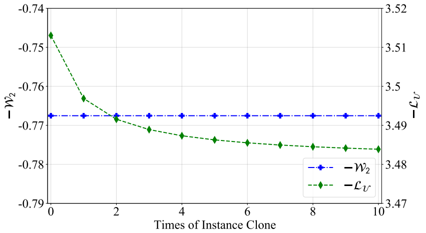


Figure 11: ICC analysis.

We randomly sample 10,000 data vectors from a standard Gaussian distribution and mask 50% of their dimensions with zero-vectors, forming the dataset $\mathcal{D}$. To investigate the impact of instance cloning, we create multiple clones of the dataset, denoted as $\mathcal{D} \cup \mathcal{D}$ and $\mathcal{D} \cup \mathcal{D} \cup \mathcal{D}$, which correspond to one and two times cloning, respectively. We evaluate two metrics on the cloned vectors. Figure 11 shows that the value of $-\mathcal{L}_{\mathcal{U}}$ slightly decreases as the number of clone instances increases, indicating that $-\mathcal{L}_{\mathcal{U}}$ violates the equality constraint in Equation 7. In contrast, our proposed metric $-\mathcal{W}_2$ remains constant, satisfying the constraint.

K EXPERIMENTS SETTING IN THE EXPERIMENTS

Setting To make a fair comparison, we conduct all experiments in Sec. 6 on a single 1080 GPU. Also, we adopt the same network architecture for all models, i.e., ResNet-18 (He et al., 2016) as the encoder, a three-layer MLP as the projector, and a three-layer MLP as the projector, respectively. Besides, We use LARS optimizer (You et al., 2017) with a base learning rate 0.2, along with a cosine decay learning rate schedule (Loshchilov & Hutter, 2017) for all models. We evaluate all models under a linear evaluation protocol. Specifically, models are pre-trained for 500 epochs and evaluated by adding a linear classifier and training the classifier for 100 epochs while keeping the learned representations unchanged. We also deploy the same augmentation strategy for all models, which is the composition of a series of data augmentation operations, such as color distortion, rotation, and cutout. Following (da Costa et al., 2022), we set temperature $t = 0.2$ for all contrastive methods. As for MoCo (He et al., 2020) and NNCLR (Dwibedi et al., 2021) that require an extra queue to save negative samples, we set the queue size to 2^{12}. For the linear decay for weighting Wasserstein distance, detailed parameter settings are shown in Table 3.

Table 3: Parameter setting for various models in experiments.

Models	MoCo v2	BYOL	BarlowTwins	Zero-CL
α_{max}	1.0	0.2	30.0	30.0
α_{min}	1.0	0.2	0	30.0

L ALIGNMENT METRIC FOR SELF-SUPERVISED REPRESENTATION LEARNING

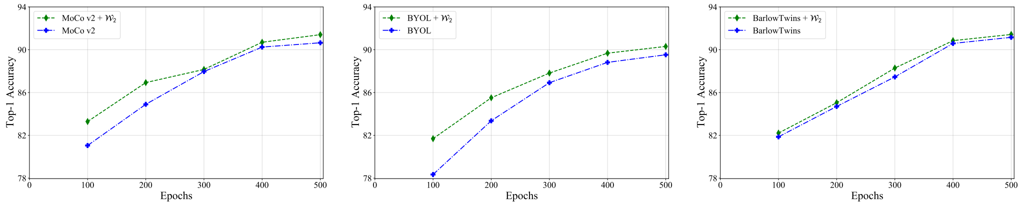
As one of the important indicators to evaluate representation capacity, the alignment metric measures the distance among semantically similar samples in the representation space, and smaller alignment generally brings better representation capacity. Wang et al (Wang & Isola, 2020) propose a simpler approach by calculating the average distance between the positive pairs as alignment, and it can be formulated as:

$$\mathcal{A} := \mathbb{E}_{(\mathbf{z}_i^a, \mathbf{z}_i^b) \sim p_{\mathbf{z}}^{pos}} \left[\left\| \frac{\mathbf{z}_i^a}{\|\mathbf{z}_i^a\|} - \frac{\mathbf{z}_i^b}{\|\mathbf{z}_i^b\|} \right\|_2^\beta \right]. \tag{14}$$

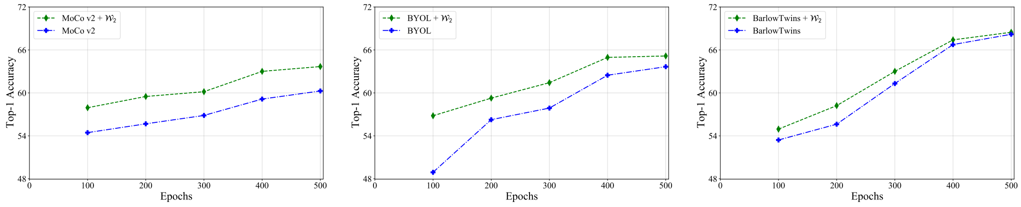
where $(\mathbf{z}_i^a, \mathbf{z}_i^b)$ is a positive pair as discussed in Sec 2.1. We set $\beta = 2$ in the experiments.

M CONVERGENCE ANALYSIS ON TOP-1 ACCURACY

Here we show the change of Top-1 accuracy through all the training epochs in Fig 12. During training, we take the model checkpoint after finishing each epoch to train linear classifier, and then evaluate the Top-1 accuracy on the unseen images of the test set (in either CIFAR-10 or CIFAR-100). In both CIFAR-10 and CIFAR-100, we could obverse that imposing the proposed uniformity metric as an auxiliary penalty loss could largely improve the Top-1 accuracy, especially in the early stage.



(a) MoCo v2 in CIFAR-10 (b) BYOL in CIFAR-10 (c) BarlowTwins in CIFAR-10



(d) MoCo v2 in CIFAR-100 (e) BYOL in CIFAR-100 (f) BarlowTwins in CIFAR-100

Figure 12: Convergence analysis on Top-1 accuracy during training.

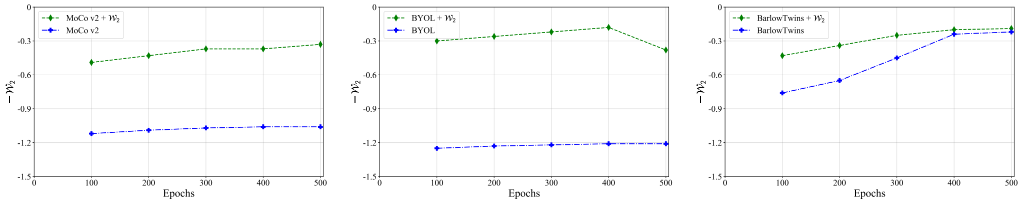
N ANALYSIS ON UNIFORMITY AND ALIGNMENT

Here we show the change of uniformity and alignment through all the training epochs in Figure 13 and Figure 14 respectively. During training, we take the model checkpoint after finishing each epoch to evaluate the uniformity (i.e., using the proposed metric $\mathcal{W}_2$) and alignment (Wang & Isola, 2020) on the unseen images of the test set (in either CIFAR-10 or CIFAR-100). In both CIFAR-10 and CIFAR-100, we could obverse that imposing the proposed uniformity metric as an auxiliary penalty loss could largely improve its uniformity. Consequently, it also lightly damages the alignment (*the smaller, the better-aligned*) since a better uniformity usually leads to worse alignment by definition.

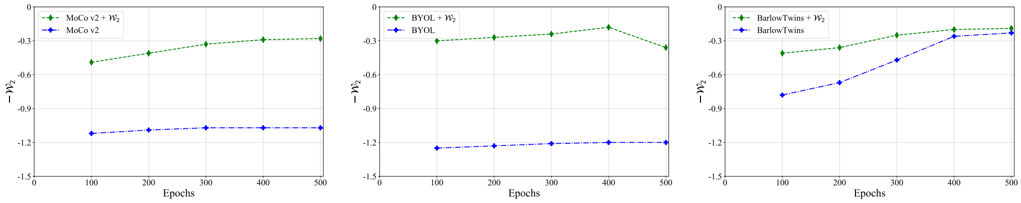
O THE EXPLANATION FOR PROPERTY 5

Here we explain why the Property 5 is an inequality instead of an equality by case study. Suppose a set of data vectors ($\mathcal{D}$) defined in Sec. 3.1 is with the maximum uniformity. When more dimensions with zero-value are inserted to $\mathcal{D}$, the set of new data vectors ($\mathcal{D} \oplus \mathbf{0}^k$) cannot achieve maximum uniformity anymore, as they only occupy a small space on the surface of the unit hypersphere. Therefore, the uniformity would decrease significantly with large k.

To further illustrate the inequality, we visualize sampled data vectors. In Figure 15(a), we visualize 400 data vectors ($\mathcal{D}_1$) sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I}_2)$, and they almost uniformly distribute on the $\mathcal{S}^1$. We insert one dimension with zero-value to $\mathcal{D}_1$, and denote it as $\mathcal{D}_1 \oplus \mathbf{0}^1$, as shown in Figure 15(b). In comparison with $\mathcal{D}_2$ where 400 data vectors are sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I}_3)$, as visualized in Figure 15(c),

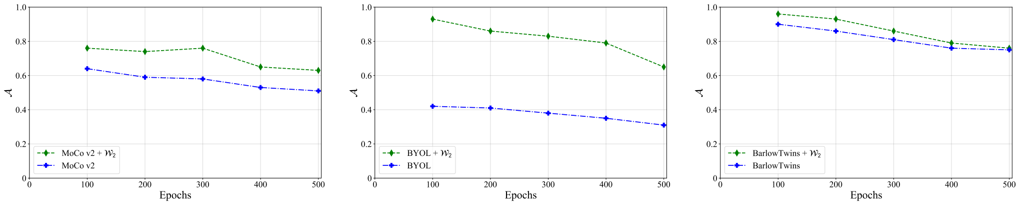


(a) MoCo v2 in CIFAR-10 (b) BYOL in CIFAR-10 (c) BarlowTwins in CIFAR-10

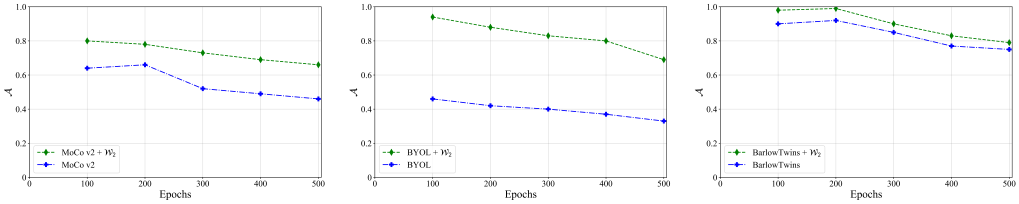


(d) MoCo v2 in CIFAR-100 (e) BYOL in CIFAR-100 (f) BarlowTwins in CIFAR-100

Figure 13: Visualization on uniformity during training

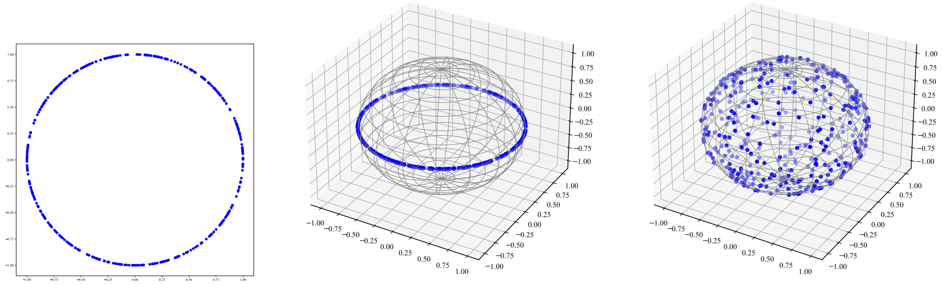


(a) MoCo v2 in CIFAR-10 (b) BYOL in CIFAR-10 (c) BarlowTwins in CIFAR-10



(d) MoCo v2 in CIFAR-100 (e) BYOL in CIFAR-100 (f) BarlowTwins in CIFAR-100

Figure 14: Visualization of alignment during training.



(a) Two-dimensional visualization with no collapsed dimension (b) Three-dimensional visualization with one collapsed dimension (c) Three-dimensional visualization with no collapsed dimension

Figure 15: Case study for Property 5 and blue point are data vectors.

$\mathcal{D}_1 \oplus \mathbf{0}^1$ only occupy a ring on the $\mathcal{S}^2$, while $\mathcal{D}_2$ almost uniformly distribute on the $\mathcal{S}^2$. Therefore, $\mathcal{U}(\mathcal{D}_2) > \mathcal{U}(\mathcal{D}_1 \oplus \mathbf{0}^1)$. According to the Assumption 1, no matter how great/small m, the maximum uniformity over various dimensions m should be equal, then we have $\mathcal{U}(\mathcal{D}_1) = \mathcal{U}(\mathcal{D}_2) > \mathcal{U}(\mathcal{D}_1 \oplus \mathbf{0}^1)$. The Property 5 should be an inequality and can be used to identify the capacity that captures sensitivity to the dimensional collapse.

P WHY PROPERTY 5 RELATES THE DIMENSIONAL COLLPASE

Intuitively, Increasing the value of k in Property 5 would exacerbate the degree of dimensional collapse. To illustrate this, suppose a set of data vectors ($\mathcal{D}$) defined in Sec. 3.1 are sampled from an isotropic Gaussian distribution, $\mathcal{N}(\mathbf{0}, \mathbf{I}_m)$. These data vectors are uniformly distributed on the unit hypersphere, resulting in a collapse degree of 0%. However, when inserting m dimension zero-value vectors to $\mathcal{D}$, denoted as $\mathcal{D} \oplus 0^m$, half of the dimensions become collapsed. As a result, the collapse degree increases to 50%. Figure 16 visualizes such collapse of $\mathcal{D} \oplus \mathbf{0}^k$ by the singular value spectrum of the representations. We can observe that a larger k would lead to a more serious dimensional collapse.

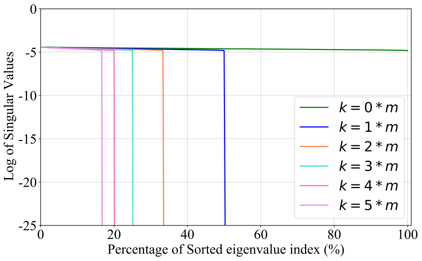
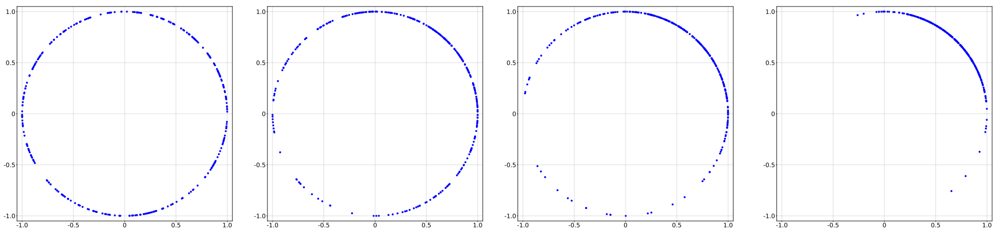


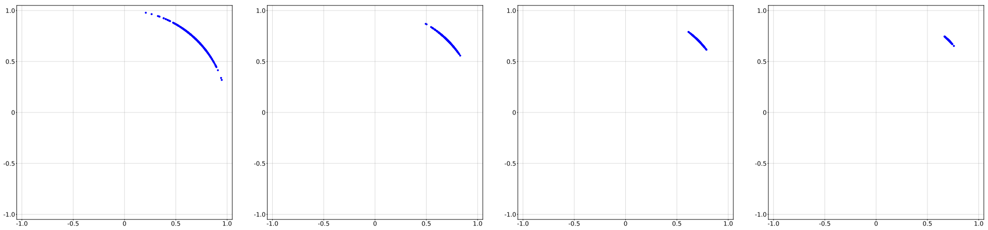
Figure 16: Singular value spectrum of $\mathcal{D} \oplus \mathbf{0}^k$.

In summary, Property 5 is closely related to the dimensional collapse.

Q EXCESSIVELY LARGE MEAN CAUSE COLLAPSED REPRESENTATION



(a) $\mathcal{N}(\mathbf{0}, \mathbf{I}_2)$ (b) $\mathcal{N}(\mathbf{0.5}, \mathbf{I}_2)$ (c) $\mathcal{N}(\mathbf{1.0}, \mathbf{I}_2)$ (d) $\mathcal{N}(\mathbf{2.0}, \mathbf{I}_2)$



(e) $\mathcal{N}(\mathbf{4.0}, \mathbf{I}_2)$ (f) $\mathcal{N}(\mathbf{8.0}, \mathbf{I}_2)$ (g) $\mathcal{N}(\mathbf{16.0}, \mathbf{I}_2)$ (h) $\mathcal{N}(\mathbf{32.0}, \mathbf{I}_2)$

Figure 17: Visualization ℓ_2 normalized Gaussian distribution

We assume $\mathbf{X}$ follows a Gaussian distribution, $\mathbf{X} \sim \mathcal{N}(0, I_2)$. By adding an additional vector to change its mean, we obtain $\mathbf{Y}$, where $\mathbf{Y} = \mathbf{X} + k\mathbf{I}$ and $\mathbf{Y} \sim \mathcal{N}(k, I_2)$. $\mathbf{I}$ is a vector of all ones, and k is a constant. We vary k from 0 to 32 and visualize ℓ_2-normalized $\mathbf{Y}$ in Figure 17. It is evident that an excessively large mean will cause representations to collapse to a single point even if the covariance matrix is an identity matrix.