Learning Efficient Coding of Natural Images with Maximum Manifold Capacity Representations

Thomas Yerxa [1] **Yilun Kuang** [2 3] **Eero Simoncelli** [1 3 2] **SueYeon Chung** [1 3]

Abstract

Self-supervised Learning (SSL) provides a strategy for constructing useful representations of images without relying on hand-assigned labels. Many such methods aim to map distinct views of the same scene or object to nearby points in the representation space, while employing some constraint to prevent representational collapse. Here we recast the problem in terms of efficient coding by adopting *manifold capacity*, a measure that quantifies the quality of a representation based on the number of linearly separable object manifolds it can support, as the efficiency metric to optimize. Specifically, we adapt the manifold capacity for use as an objective function in a contrastive learning framework, yielding a Maximum Manifold Capacity Representation (MMCR). We apply this method to unlabeled images, each augmented by a set of basic transformations, and find that it learns meaningful features using the standard linear evaluation protocol. Specifically, we find that MMCRs support performance on object recognition comparable to or surpassing that of recently developed SSL frameworks, while providing more robustness to adversarial attacks. Empirical analyses reveal differences between MMCRs and representations learned by other SSL frameworks, and suggest a mechanism by which manifold compression gives rise to class separability.

1. Introduction

Biological visual systems learn complex representations of the world that support a wide range of cognitive behaviors without using a large number of labelled examples. The efficient coding hypothesis (Barlow et al., 1961; Simoncelli & Olshausen, 2001) suggests that this is accomplished by adapting the sensory representation to the statistics of the input signal, so as to reduce redundancy or dimensionality. Visual signals have several clear sources of redundancy. They evolve slowly in time, since temporally adjacent inputs typically correspond to different views of the same scene, which in turn are usually more similar than views of distinct scenes. Moreover, the variations within individual scenes often correspond to variations in a small number of parameters, such as those controlling viewing and lighting conditions, and are thus inherently low dimensional. Many previous results have demonstrated how the computations of neural circuits can be seen as matched to such structures in naturalistic environments (Simoncelli & Olshausen, 2001; Schwartz & Simoncelli, 2001; Laughlin, 1981; Kriegeskorte & Kievit, 2013; Chung & Abbott, 2021; Fairhall et al., 2001). Studies in various modalities have identified geometric structures in neural data that are associated with behavioral tasks (Bernardi et al., 2020; DiCarlo & Cox, 2007; Hénaff et al., 2021; Gallego et al., 2017; Nieh et al., 2021), and explored metrics for quantifying these representation geometries.

Recently, a new theory which connects the geometry (size and dimensionality) of neural representations to the linear decoding capacity of those neural manifolds has been introduced ("Manifold capacity theory" hereafter) (Chung et al., 2018). This theory has been used to evaluate neural representations from biological and artificial neural networks across modalities (Chung & Abbott, 2021). However, these geometrical approaches have remained largely descriptive as a way of evaluating neural data and understanding brain functions, and its constructive usage as a design principle for building model representations has been under explored. Motivated by these observations, we seek to learn a function that represents different views of the same scene with manifolds that are both compact and low-dimensional while simultaneously maximizing the separation between manifolds representing distinct scenes.

Here, we demonstrate for the first time that

- optimizing a network for manifold capacity (MMCR) results in a representation that support high-quality object recognition, when evaluated using the standard linear evaluation paradigm (i.e., applying an optimized

[1]Center for Neural Science, New York University [2]Courant Inst. of Mathematical Sciences, New York University [3]Center for Computational Neuroscience, Flatiron Institute. Correspondence to: Thomas Yerxa <tey214@nyu.edu>.

Preliminary work, under review.

linear classifier to the output of the unsupervised network) (Chen et al., 2020a). Specifically, we show that performance is approximately matched to that of other recently proposed SSL methods.

- examining the learning signal reveals the mechanism underlying the emergence of semantically relevant features from the data
- MMCR renders interpretable geometric properties that result in increased robustness to adversarial attack, relative to that of other recently proposed SSL methods.

1.1. Related Work

Our methodology is closely related to and inspired by recent advances in contrastive self-supervised representation learning (SSL), but has a distinctly different motivation and formulation. Many recent frameworks craft objectives that are designed to maximize the mutual information between representations of different views of the same object (Oord et al., 2018; Chen et al., 2020a; Oord et al., 2018; Tian et al., 2020; Bachman et al., 2019)). However, estimating mutual information in high dimensional feature spaces has proven difficult (Belghazi et al., 2018), and furthermore it is not clear that more closely approximating mutual information in the objective yields improved representations (Wang & Isola, 2020)[1]. By contrast, capacity estimation theories operate in the regime of large ambient dimension as they are derived in the "large N (thermodynamic) limit" (Chung et al., 2018; Bahri et al., 2020). Therefore we examine whether one such measure, which until now had been used only to evaluate representation quality, is useful as an objective function for SSL.

Operationally, a number of existing methods aim to minimize some notion of distance between the representations of different augmented views of the same image, while maximizing the distance between representations of (augmented views of) distinct images either directly or by imposing some constraint on the representation such as feature decorrelation (these are thought of as encouraging alignment and uniformity in the framework of Wang & Isola (2020)). The limitations of using a single pairwise distance comparison have been demonstrated on multiple occasions, notably in the development of the "multi-crop," strategy implemented in SwAV (Caron et al., 2020) and in the contrastive multiview coding approach Tian et al. (2020)). Consistent with this, our formulation is derived from an assumption that different views of an image form a continuous manifold that we aim to compress. Rather than using the mean distance or cosine similarity between pairs of points, we characterize each set of image views with the spectrum of singular values of their representations, using the nuclear norm as a combined measure of the manifold size and dimensionality. The nuclear norm has been previously used to induce or infer low rank structure in the representation of data (Hénaff et al., 2015; Wang et al., 2022; Lezama et al., 2018). In particular, Wang et al. (2022) employ the nuclear norm as a regularizer to supplement an InfoNCE loss. Our approach represents a more radical departure from the traditional InfoNCE loss, as we will detail below. Rather than pair a low-rank prior with a logistic regression-based likelihood, we make the more symmetric choice of employing a *high rank* likelihood. This allows the objective to explicitly discourage dimensional collapse, a well known issue in SSL (Jing et al., 2021).

Another consequence of encouraging maximal rank over the dataset is that the objective encourages the representation to form a simplex equiangular tight frame (sETF). sETFs have been shown to be optimal in terms of cross-entropy loss when features lie on the unit hypersphere (Lu & Steinerberger, 2020), and such representations can be obtained in the supervised setting when optimizing either the traditional cross-entropy loss or a supervised contrastive loss (Papyan et al., 2020; Graf et al., 2021). Recent work has shown that many popular objectives in SSL can be understood as different methods of approximating a loss function whose minima form sETFs (Dubois et al., 2022). Our approach is novel, in that it encourages sETF representations by directly optimizing the distribution of singular values, rather than minimizing a cross-entropy loss.

Recently HaoChen et al. (2021) developed a framework based on spectral decomposition of the "population augmentation graph," which provides theoretical guarantees for the performance of self-supervised learning on downstream tasks under linear probing. This work was extended to provide insights into various other SSL objectives by Balestriero & LeCun (2022), and we show below that leveraging this approach can lead to explicit conditions for the optimality of representation under our proposed objective as well.

2. Maximum manifold capacity representations

2.1. Manifold Capacity Theory

Consider a set of P manifolds embedded in a feature space of dimensionality D, each assigned a class label. Manifold capacity theory is concerned with the question: what is the largest value of $\frac{P}{D}$ such that there exists (with high probability) a hyperplane separating the random dichotomy (Cover, 1965; Gardner, 1988)? Recent theoretical work has

[1] Barlow Twins (Zbontar et al., 2021) notably avoids the curse of dimensionality because the objective effectively estimates information under a Gaussian parameterization rather than doing so non-parametrically as in the InfoNCE loss. Our method also makes use of Gaussian/second order parameterizations, as detailed below.

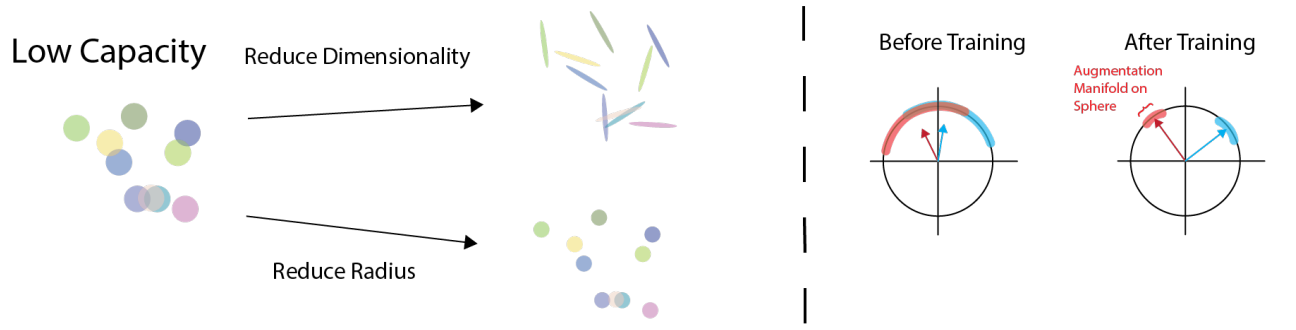

Figure 1. Two dimensional illustrations of high and low capacity representations. Left: the capacity (linear separability) of a random set of elliptical regions can be improved, either by reducing their sizes (while maintaining their dimensionalities), or by reducing their dimensionalities (while maintaining their sizes). Right: the objective proposed in this paper aims to minimize the nuclear norm (product of size and sqrt dimensionality) of normalized data vectors (ie., lying on the unit sphere). Before training the manifolds have a large extent and thus the matrix of their corresponding centroid vectors has low nuclear norm. After training the capacity is increased. The manifolds are compressed and repelled from each other, resulting in centroid matrix with larger nuclear norm and lower similarity.

demonstrated that there exists a critical value, dubbed the manifold capacity α_C, such that when $\frac{P}{D} < \alpha_C$ the probability of finding a separating hyperplane is approximately 1.0, and when $\frac{P}{D} > \alpha_C$ the probability is approximately 0.0 (Chung et al., 2018). Furthermore, α_C can be accurately predicted from three key quantities: (1) the manifold radius R_M, which measures the size of the manifold relative to its distance from the origin, (2) the manifold dimensionality D_M which estimates the number of dimensions along which a manifold has significant extent, and (3) the centroid correlation (if the positions of manifolds are correlated with each other they will be more difficult to separate). In particular, when the centroid correlation is low the manifold capacity can be approximated by $\phi(R_M\sqrt{D_M})$ where $\phi(\cdot)$ is a monotonically decreasing function.

For manifolds of arbitrary geometry calculating the manifold radii and dimensionalities involves an iterative process that alternates between determining the set of "anchor points" on each manifold that are relevant for the classification problem, and computing the statistics of random projections of these anchor points (Cohen et al., 2020). This process is both computationally costly and non-differentiable, and therefore not suitable for use as an objective function. For more detail on the general theory see Appendx C. However, if the submanifolds are assumed to be elliptical in shape there is an analytical expression for each of these,

$$R_M = \sqrt{\textstyle\sum_i \lambda_i^2}, \qquad D_M = \frac{(\sum_i \lambda_i)^2}{\sum_i \lambda_i^2}, \tag{1}$$

where the λ_i^2 are the eigenvalues of the covariance matrix of points on the manifold. For reference, for a batch of 100 128-D manifolds with 100 points sampled from each, computing these elliptical-assuming measures is approximately 500 times faster in terms of wall-clock time.

Using these definitions for manifold radius and dimensionality we can write the capacity as $\alpha_C = \phi(\sum_i \sigma_i)$ where σ_i are the *singular values* of a matrix containing points on the manifold (equivalently, the square roots of the eigenvalues of the covariance matrix). In this form, the sum is the L_1 norm of the singular values, known as the *Nuclear Norm* of the matrix. When used as an objective function, this measure will prefer sparse solutions (i.e., a small number of non-zero singular values) corresponding to low dimensionality. It is worth comparing this objective to another natural candidate for quantifying size: the determinant of the covariance matrix. The determinant is equal to the product of the eigenvalues (which captures the squared volume of the corresponding ellipse), but lacks the preference for lower dimensionality that comes with the Nuclear Norm. Specifically, since the determinant is zero whenever one (or more) eigenvalue is zero, it cannot distinguish zero-volume manifolds of different dimensionality. In Yu et al. (2020), lossy coding rate (entropy) is used as a measure of compactness, which simplifies to the log determinant under a Gaussian model Ma et al. (2007). In that work, the identity matrix is added to a multiple of the feature covariance matrix before evaluating the determinant, which solves the dimensionality issue described above.

2.2. Optimizing Manifold Capacity

Now we construct an SSL objective function based on manifold capacity. For each input image (notated as a vector $\mathbf{x}_b \in \mathbb{R}^{\mathbb{D}}$) we generate k samples from the corresponding manifold by applying a set of random augmentations (each drawn from the same distribution), yielding a manifold sample matrix $\tilde{\boldsymbol{X}}_b \in \mathbb{R}^{D\times k}$. Each augmented image is transformed by a Deep Neural Network, which computes nonlinear function $f(\mathbf{x}_b;\theta)$ parameterized by θ, and the d-

dimensional responses are projected onto the unit sphere yielding manifold response matrix $\boldsymbol{Z}_b \in \mathbb{R}^{d\times k}$. The centroid $\boldsymbol{c}_b$ is approximated by averaging across the columns (response vectors). For a set of images $\{\mathbf{x}_1, ..., \mathbf{x}_B\}$ we compute normalized response matrices $\{\boldsymbol{Z}_1, ..., \boldsymbol{Z}_B\}$ and assemble their corresponding centroids into matrix $\boldsymbol{C} \in \mathbb{R}^{d\times B}$.

Given the responses and their centroids, the loss function is expressed as:

$$\mathcal{L} = -||\boldsymbol{C}||_* + \lambda \left(\frac{1}{B} \sum_{b=1}^{B} ||\boldsymbol{Z}_b||_* \right) \tag{2}$$

where $||\cdot||_*$ indicates the nuclear norm and λ is a tradeoff parameter. The first term maximizes the extent of the "centroid manifold" to encourage separability while the second term encourages object manifold compression .

Compression by Maximizing Centroid Nuclear Norm Alone Interestingly, the first term also has a compressive effect. This is because each centroid vector, as a mean of unit vectors, has norm that is linearly related to the average cosine similarity of vectors of said unit vectors. Specifically,

$$||\boldsymbol{c}_b||^2 = \frac{1}{K} + \frac{2}{K^2} \sum_{k=1}^{K} \sum_{l=1}^{k-1} \boldsymbol{z}_{b,k}^T \boldsymbol{z}_{b,l} \tag{3}$$

Here $\boldsymbol{z}_{b,i}$ denotes the representation of the i^{th} augmentation of $\boldsymbol{x}_b$. Then because the nuclear norm is bounded below by the Frobenius norm (Recht et al., 2010), $||\boldsymbol{C}||_* \geq ||\boldsymbol{C}||_F = \sqrt{\sum_{b=1}^{B} ||\boldsymbol{c}_b||^2}$, maximizing the centroid nuclear norm optimizes an upper bound on the norms of centroid vectors, thus encouraging intra-object manifold similarity. We can gain further insight by considering how the distribution of singular vectors of a matrix depends on the norms and pairwise similarities of the constituent column vectors. While no closed form solution exists for the singular values of an arbitrary matrix, the case where the matrix is composed of two column vectors can provide useful intuition. If $\boldsymbol{C} = [\boldsymbol{c}_1, \boldsymbol{c}_2]$, $\boldsymbol{Z}_1 = [\boldsymbol{z}_{1,1}, \boldsymbol{z}_{1,2}]$, $\boldsymbol{Z}_2 = [\boldsymbol{z}_{2,1}, \boldsymbol{z}_{2,2}]$, the singular values of $\boldsymbol{C}$ and $\boldsymbol{Z}_i$ are:

$$\boldsymbol{\sigma}(\boldsymbol{C}) = \frac{1}{\sqrt{2}}(||\boldsymbol{c}_1||^2 + ||\boldsymbol{c}_2||^2 \pm ((||\boldsymbol{c}_1||^2 - ||\boldsymbol{c}_2||^2)^2 + 4(\boldsymbol{c}_1^T \boldsymbol{c}_2)^2)^{1/2}$$

$$\boldsymbol{\sigma}(\boldsymbol{Z}_i) = \sqrt{1 \pm \boldsymbol{z}_{i,1}^T \boldsymbol{z}_{i,2}}$$

So, $||\boldsymbol{\sigma}(\boldsymbol{C})||_1 = ||\boldsymbol{C}||_*$ is maximized when the centroid vectors have maximal norms (bounded above by 1, since they are the centroids of unit vectors), and are orthogonal to each other. As we saw above the centroid norms is a linear function of within-manifold similarity. Similarly, $||\boldsymbol{\sigma}(\boldsymbol{Z}_i)||_1 = ||\boldsymbol{Z}_i||_*$ is minimized when the within-manifold similarity is maximal. So, both terms in the objective encourage object manifold compression (in the simple case described above the effect is nearly mathematically equivalent). Surprisingly, this implies the first term alone encapsulates both of the key ingredients of a contrastive learning framework, and we do observe that simply maximizing $||\boldsymbol{C}||_*$ is sufficient to learn a useful representation. This is because the compressive role of "positives" in contrastive learning is carried out by forming the centroid vectors, so the objective is not positive-free even with $\lambda = 0$. For example, if only a single view is used the objective lacks a compressive component and fails to produce a useful representation. In Appendix F we demonstrate empirically that this implicit form effectively reduces $||\boldsymbol{Z_b}||_*$ So, all three factors which determine the manifold capacity (radius, dimensionality, and centroid correlations) can be elegantly expressed in an objective function with a single term, $-||\boldsymbol{C}||_*$; therefore we simply drop the second term from Eq. 2 entirely.

2.3. Conditions for Optimal Embeddings

Here we introduce a simplified version of the framework developed by HaoChen et al. (2021), which is a slight modification of the formulation in Balestriero & LeCun (2022) in order to derive conditions for the optimal embeddings under the proposed loss. Given a dataset $\boldsymbol{X'} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_N]^T \in \mathbb{R}^{N\times D'}$ we construct a new dataset by creating k randomly augmented views of the original data, $\boldsymbol{X} = [\text{view}_1(\boldsymbol{X'}), ..., \text{view}_k(\boldsymbol{X'})] \in \mathbb{R}^{Nk\times D}$. The advantage of doing so is that we can now leverage the knowledge that different views of the same underlying datapoint are *semantically related*. We can express this notion of similarity in the symmetric matrix $\boldsymbol{G} \in \{0,1\}^{Nk\times Nk}$ with $\boldsymbol{G}_{ij} = 1$ if augmented datapoints i and j are semantically related (and $\boldsymbol{G}_{ii} = 1$ as any datapoint is related to itself). We can normalize $\boldsymbol{G}$ such that its rows and columns sum to 1 (so rows of $\boldsymbol{G}$ are k-sparse with nonzero entries equal to $1/k$).

Now let $\boldsymbol{Z} \in \mathbb{R}^{Nk\times d}$ be an embedding of the augmented dataset. Then we have $\boldsymbol{GZ} = [\boldsymbol{C}, ..., \boldsymbol{C}]^T$ where $\boldsymbol{C}$ is the matrix of centroid vectors introduced above, and the number of repetitions of $\boldsymbol{C}$ is k. Then because $\boldsymbol{\sigma}([\boldsymbol{C}, ..., \boldsymbol{C}]) = \sqrt{k}\boldsymbol{\sigma}(\boldsymbol{C})$ we can write MMCR loss function as,

$$\mathcal{L} = -||\boldsymbol{GZ}||_* \tag{4}$$

This connection allows us to make the following statements about the optimal embeddings $\boldsymbol{Z}$ under our loss, which we prove in Appendix A:

Theorem: Under the proposed loss, the left singular vectors of an optimal embedding, $\boldsymbol{Z}^*$, are the eigenvectors of $\boldsymbol{G}$, and the singular values of $\boldsymbol{Z}$ are proportional to the top d

eigenvalues of $\boldsymbol{G}$.

Computational Complexity: Evaluating the loss for our method involves computing a singular value decomposition of $\boldsymbol{C} \in \mathbb{R}^{d \times B}$ which has complexity $\mathcal{O}(Bd \times \min(B, d))$, where B is the batch size and d is the dimensionality of the output. By comparison contrastive methods that compute all pairwise distances in a batch have complexity $\mathcal{O}(B^2 d)$ and non-contrastive methods that involve regularizing the covariance structure have complexity $\mathcal{O}(Bd^2)$. Additionally, the complexity of our method is constant with respect to the number of views used (though the feature extraction phase is linear in the number of views), while pairwise similarity metrics will have quadratic complexity with the number of views.

3. Results

We tested our method on datasets of different size, including CIFAR-10, STL-10, and CIFAR-100, ImageNet-1k and ImageNet-100 (Krizhevsky et al., 2009; Coates et al., 2011). We used a standard linear evaluation technique (Chen et al., 2020a) to verify that our method extracts semantically relevant features from the data. We report the top-1 accuracy of linear classifiers for ImageNet-1k in 3.2; results for other datasets can be found in Appendix J. Additionally we conduct a series of analyses to understand how learning to compress object manifolds gives rise to class manifold separability. Finally we investigate how the geometrical differences between representations trained according to different contrastive SSL methods impact adversarial robustness. To reduce the computational requirements, this set of analyses is carried out on models trained on the CIFAR-10 dataset. We primarily compare our method to SimCLR and Barlow Twins, popular examples from the contrastive and "non-contrastive," categories of self-supervised learning (Chen et al., 2020a; Zbontar et al., 2021). For details on each specific analysis see Appendix E.

3.1. Implementation Details

Architecture. For all experiments we use ResNet-50 (He et al., 2016) as a backbone architecture (for variants trained on CIFAR we removed max pooling layers). Following Chen et al. (2020a), we append a small perceptron with one hidden layer to the output of the average pooling layer of the ResNet so that $z_i = g(h(x_i))$, where h is the ResNet and g is the MLP. For ImageNet-1k/100 we used an MLP with dimensions $[8192, 8192, 512]$ and for smaller datasets we used $[512, 128]$.

Optimization We employ the set of augmentations proposed in (Zbontar et al., 2021). For ImageNet we used the LARS optimizer with a learning rate of 4.8, linear warmup during the first 10 epochs and cosine decay thereafter, with a batchsize of 2048. For smaller CIFAR-10 we used a smaller batch size and the Adam optimizer with fixed learning rate. See Appendix D for exact details.

Table 1. Top-1 classification accuracies of linear classifiers for representations trained with various objective functions for 100 epochs on ImageNet-1k. Results for other methods come from Ozsoy et al. (2022), except for SwAV which is copied from Dubois et al. (2022)

Method	Accuracy (%)
Barlow Twins Zbontar et al. (2021)	68.7
SimCLR (Chen et al., 2020a)	66.5
SimSiam (Chen & He, 2021)	68.1
BYOL (Grill et al., 2020)	69.3
MoCo-V2 (Chen et al., 2020b)	67.4
VICReg (Bardes et al., 2021)	68.7
SwAV (Caron et al., 2020)	64.6
CorInfoMax (Ozsoy et al., 2022)	69.1
SwAV (w/ multi-crop)(Caron et al., 2020)	69.5
W-MSE (4 views) (Ermolov et al., 2021)	69.4
MMCR (2 views)	68.4
MMCR (4 views)	70.2
MMCR (8 views)	**71.5**

3.2. Performance

Figure 10 details the evolution of the representation during the course of training. The centroid nuclear norm (Fig. 10b) increases steadily as the centroids become increasingly orthogonal to each other (Fig. 10c) and grow in norm (Fig. 10d). The compression of individual augmentation manifolds is reflected in Fig. 2a. The downstream classification accuracies are reported in Table 4. Note that though we report results using a default batch size of 2048, a batch size as low as 256 can be used to obtain reasonable results (1.2% drop compared to batch 2048), see Appendix K for a sweep of batch size parameter.

3.3. Representation geometric analysis

In figure 2 we show that our representation, which is optimized using an objective that assumes elliptical manifold geometry, nevertheless yields representations with high mean field manifold capacity (relative to baseline methods). For completeness we also analyzed the geometries of class manifolds, whose points are the representations of different examples from the same class. This analysis provided further evidence that learning to maximize augmentation manifold capacity compresses and separates class manifolds, leading to a useful representation. Interestingly MMCRs seem to use a different strategy than the baseline methods to increase the capacity, namely MMCRs produce class/augmentation

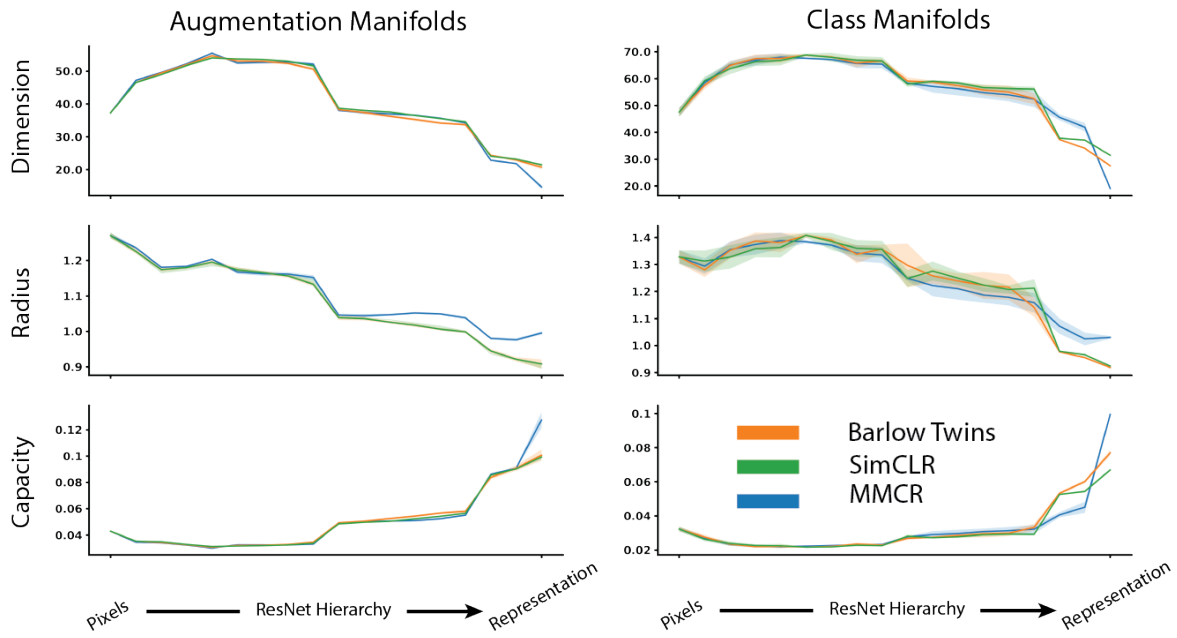

Figure 2. Mean Field Manifold Capacity Analysis. The shared x-axis of all plots is the representational hierarchy, the leftmost entries represent the inputs (pixels) and the rightmost the output of the encoder/learned representation. The top row shows the manifold radius, the middle the dimensionality, and the bottom the resultant capacity. Shaded regions indicate a 95% confidence interval around the mean (analysis was conducted with 5 different random samples from the dataset, see E).

manifolds with larger radii, but lower dimensionality (Fig. 2)

3.4. Emergence of neural manifolds via gradient coherence

We hypothesize the class separability in MMCRs arises because augmentation manifolds corresponding to examples from the same class are optimally compressed by more similar transformations than those stemming from distinct classes. To investigate this empirically, we evaluate the gradient of the objective function for inputs belonging to the same class. We can then check whether gradients obtained from (distinct) batches of the same class are more similar to each other than those obtained from different classes, which would suggest that the strategy for compressing augmentation manifolds from the same class are relatively similar to each other. Figure 3 demonstrates that this is the case: within class gradient coherence, as measured by cosine similarity, is consistently higher than across class coherence across both training epochs and model hierarchy.

3.5. Manifold subspace alignment

Within-class gradient coherence constitutes a plausible mechanistic explanation for the emergence of class separability, but it does not explain why members of the same class share similar compression strategies. To begin answering this question we examine the geometric properties of augmentation manifolds in the pixel domain. Here we observe small but measurable differences between the distributions of within-class similarity and across-class similarity, as demonstrated in the top row of figure 4. The subtle difference in the geometric properties of augmentation manifolds in the pixel domain in turn leads to the increased gradient coherence observed above, which over training leads to a representation that rearranges and reshapes augmentation manifolds from the same class in a similar fashion (bottom row of figure 4), thus allowing for the linear separation of classes. Not only are centroids of same-class-manifolds in more similar regions of the representation space than those coming from distinct classes (Fig 4 third column bottom row) but additionally same-class-manifolds have more similar shapes to each other (Fig 4 bottom row columns 1 and 2 show same-class-manifolds occupy subspaces with lower relative angles and share more variance).

We next ask how the representation learned according to the MMCR objective differs from those optimized for other self supervised loss functions. While MMCR encourages centroids to be as close to orthogonal to each other, the InfoNCE loss employed in Chen et al. (2020a) benefits when negative pairs are as dissimilar as possible, which is achieved when the two points lie in opposite regions of the same subspace rather than in distinct (orthogonal) subspaces. The Barlow Twins (Zbontar et al., 2021) loss is not an explicit function

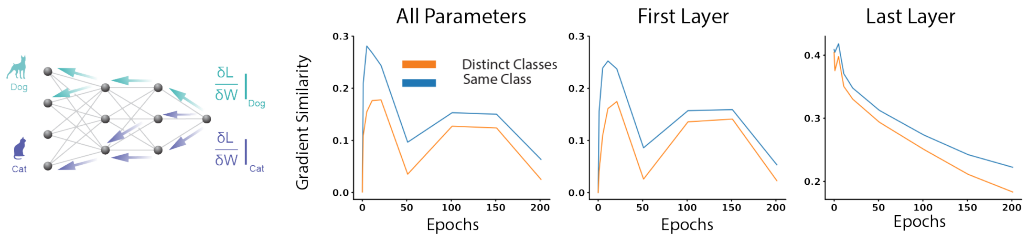

Figure 3. Gradient cosine similarity for pairs of single-class batches. We plot the mean pairwise similarity for pairs of gradients for for different subsets of the model parameters (all parameters, and the first and last linear operators) obtained from single-class-batches coming from the same or distinct classes over the course of training. To the left is a visualization of the fact that single-class gradients flow backward through the model in more similar directions.

of feature vector similarities, but instead encourages individual features to be correlated across the batch dimension and distinct features to be uncorrelated. In 5 we demonstrate that these intuitions are borne out empirically: the MMCR representation produces augmentation manifold centroids that are significantly less similar to each other than the two baseline methods.

3.6. Adversarial Robustness

Previous work using similar geometrically-motivated loss functions such as the orthogonal low-rank embedding (Lezama et al., 2018) and maximal coding rate reduction (Yu et al., 2020) have reported have increased inter-class margins and robustness to label noise. We therefore tested whether the increased tendency to orthogonalize in MMCR models leads to any benefit in terms of adversarial robustness. In Fig. 6 we show that the MMCR model (and attached classifier) is indeed more robust than either Barlow Twins or SimCLR trained models against PGD attacks with a range of strengths (Madry et al., 2018). We found similar results using the stronger AutoAttack protocol (Croce & Hein, 2020), see H. Note that Barlow Twins models seem to be more robust than SimCLR models. We speculate that this is a result of the decorrelation encouraged by the Barlow Twins objective (as opposed to the anti-correlation encouraged by SimCLR) .

4. Discussion

We present a novel self-supervised learning algorithm inspired by manifold capacity theory. Most existing SSL methods can be categorized as either "contrastive," or "non-contrastive," depending on whether they avoid collapse by imposing constraints on the embedding gram or covariance matrix, respectively. Our framework strikes a compromise, optimizing the singular values of the embedding matrix itself, leading to a "best of both worlds," effect. Learning MMCRs requires neither large batch size (as is typical of instance constrastive methods), nor large embedding dimension (as is typical of feature constrastive methods). Finally our method extends naturally to the multi-view case, offering improved performance (more samples leads to a better estimate of each centroid vector) with minimal increases in computational cost (evaluation time for our objective does not grow with the number of views).

When trained on several datasets of unlabelled images, our method yields representations whose downstream task performance is comparable to or better than existing SSL methods. Furthermore we conducted a gradient-based analysis to examine why the self-supervised learning signal is capable of producing useful representations. Finally, motivated by an empirical exploration of the geometrical differences between the representations produced by the three considered methods, we demonstrate that MMCRs can offer improved robustness to adversarial attacks.

Our formulation relies on a restricted form of manifold capacity, in which we approximate the manifold geometries as elliptical. This significantly reduces the computation required to calculate the geometric properties that dictate capacity, allowing its efficient use as a SSL learning objective. Although representational manifold geometries are generally not elliptical, we have demonstrated that this approximation can nonetheless produce a useful learning signal, and leads to networks with high manifold capacity (Fig. 2). Nevertheless, it would be interesting to consider other reductions of the mean field manifold capacity that can capture non-elliptical structure of individual manifolds, perhaps by computing higher order statistics of constituent points.

We were able to leverage manifold capacity analysis in its full generality to gain insight into the geometry of the MMCR networks. Intriguingly, our method produces augmentation and class manifolds with lower dimensionality but larger radius than either Barlow Twins or SimCLR (Fig. 2). Future work will seek to understand why this is the case, but more generally this suggests that capacity analysis can be a fruitful way to understand the different encoding strategies encouraged by various SSL paradigms. Another

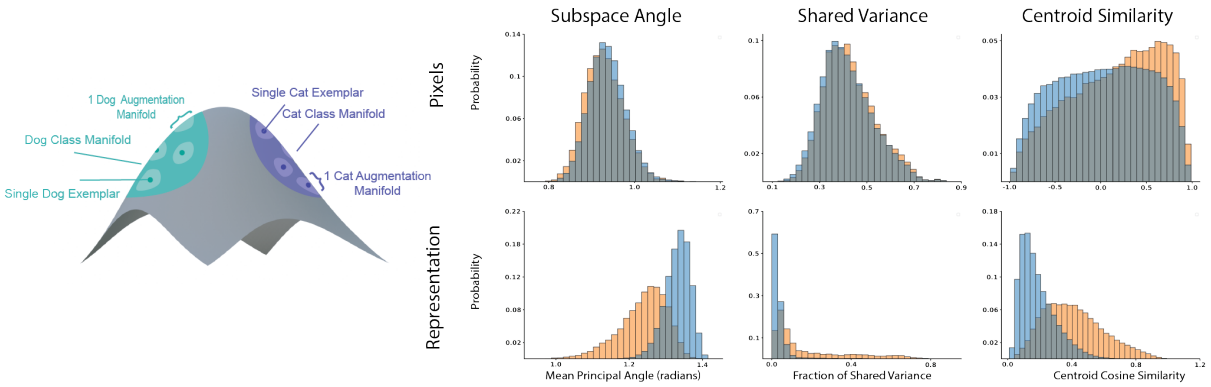

Figure 4. The distributions of various similarity metrics for augmentation manifolds from either the same and distinct classes. In the top row we consider augmentation manifolds in the pixel domain, and in the bottom row we observe how these distributions are transformed by the learned representation. To the left a schematic shows details the exemplar-augmentation manifold-class manifold structure of the learned representation.

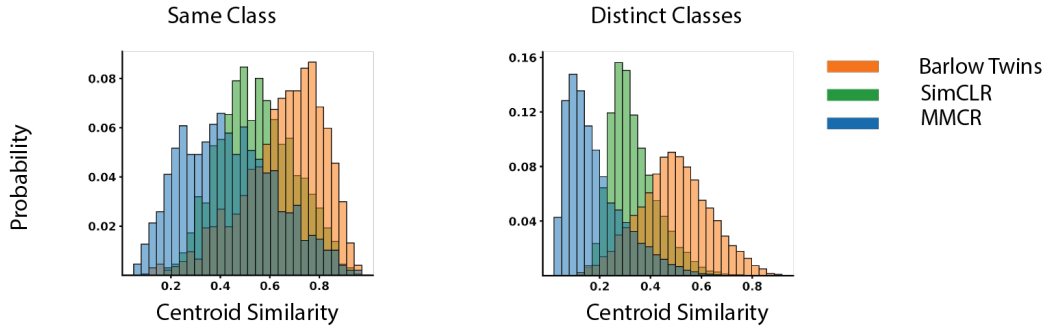

Figure 5. Centroid similarities for models trained according to different SSL objectives. The left panel shows the distribution of centroid cosine similarities for augmentation manifolds for examples of the same class, while the right shows the same distribution for examples from distinct classes.

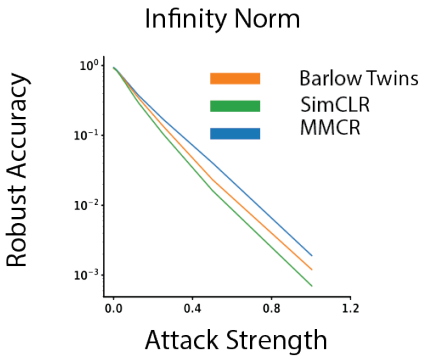

Figure 6. Adversarial Robustness of SSL Models under PGD Attack. For each of the three SSL models with trained classifiers, we apply Projected Gradient Descent (PGD) with ℓ_∞-norm perturbation under 50 attack iterations. Inputs were scaled such that their standard deviation was 1.0, so we report the raw attack strengths on the x-axis. Additional details can be found in H

factor that distinguishes MMCRs from other models is a tendency to orthogonalize augmentation manifold centroids and thus form a representation that is globally high dimensional. Given that the recent observations that the representations in visual cortex and high performing models of visual cortex are surprisingly high dimensional (Stringer et al., 2019; Elmoznino & Bonner, 2022), it may be interesting to test how well MMCRs can predict neural activity.

In this study, we introduced one specific model of learning using metrics based on a specific theory of representations: self-supervised learning via maximizing neural manifold capacity. With recent trends in neuroscience focused on representation geometric observations in neural data, we believe that this work lays foundations for future studies in learning based on representation geometries informed by new discoveries in neuroscience.

References

Bachman, P., Hjelm, R. D., and Buchwalter, W. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32, 2019.

Bahri, Y., Kadmon, J., Pennington, J., Schoenholz, S. S., Sohl-Dickstein, J., and Ganguli, S. Statistical mechanics of deep learning. *Annual Review of Condensed Matter Physics*, 11(1), 2020.

Balestriero, R. and LeCun, Y. Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods. *arXiv preprint arXiv:2205.11508*, 2022.

Bardes, A., Ponce, J., and LeCun, Y. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Barlow, H. B. et al. Possible principles underlying the transformation of sensory messages. *Sensory communication*, 1(01), 1961.

Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, D. Mutual information neural estimation. In *International conference on machine learning*, pp. 531–540. PMLR, 2018.

Bernardi, S., Benna, M. K., Rigotti, M., Munuera, J., Fusi, S., and Salzman, C. D. The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell*, 183(4): 954–967, 2020.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020a.

Chen, X. and He, K. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021.

Chen, X., Fan, H., Girshick, R., and He, K. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

Chung, S. and Abbott, L. Neural population geometry: An approach for understanding biological and artificial neural networks. *Current opinion in neurobiology*, 70: 137–144, 2021.

Chung, S., Lee, D. D., and Sompolinsky, H. Classification and geometry of general perceptual manifolds. *Physical Review X*, 8(3):031003, 2018.

Coates, A., Ng, A., and Lee, H. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

Cohen, U., Chung, S., Lee, D. D., and Sompolinsky, H. Separability and geometry of object manifolds in deep neural networks. *Nature communications*, 11(1):1–13, 2020.

Cover, T. M. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE transactions on electronic computers*, (3):326–334, 1965.

Croce, F. and Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

da Costa, V. G. T., Fini, E., Nabi, M., Sebe, N., and Ricci, E. solo-learn: A library of self-supervised methods for visual representation learning. *J. Mach. Learn. Res.*, 23: 56–1, 2022.

DiCarlo, J. J. and Cox, D. D. Untangling invariant object recognition. *Trends in cognitive sciences*, 11(8):333–341, 2007.

Dubois, Y., Hashimoto, T., Ermon, S., and Liang, P. Improving self-supervised learning by characterizing idealized representations. *arXiv preprint arXiv:2209.06235*, 2022.

Elmoznino, E. and Bonner, M. F. High-performing neural network models of visual cortex benefit from high latent dimensionality. *bioRxiv*, 2022.

Ermolov, A., Siarohin, A., Sangineto, E., and Sebe, N. Whitening for self-supervised representation learning. In *International Conference on Machine Learning*, pp. 3015–3024. PMLR, 2021.

Fairhall, A. L., Lewen, G. D., Bialek, W., and de Ruyter van Steveninck, R. R. Efficiency and ambiguity in an adaptive neural code. *Nature*, 412(6849):787–792, 2001.

Gallego, J. A., Perich, M. G., Miller, L. E., and Solla, S. A. Neural manifolds for the control of movement. *Neuron*, 94(5):978–984, 2017.

Gardner, E. The space of interactions in neural network models. *Journal of physics A: Mathematical and general*, 21(1):257, 1988.

Gerl, F. and Krey, U. Storage capacity and optimal learning of potts-model perceptrons by a cavity method. *Journal of Physics A: Mathematical and General*, 27(22):7353, 1994.

Graf, F., Hofer, C., Niethammer, M., and Kwitt, R. Dissecting supervised constrastive learning. In *International Conference on Machine Learning*, pp. 3821–3830. PMLR, 2021.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

HaoChen, J. Z., Wei, C., Gaidon, A., and Ma, T. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34:5000–5011, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Hénaff, O. J., Rabinowitz, N., Ballé, J., and Simoncelli, E. P. The local low-dimensionality of natural images. In *Int'l Conf on Learning Representations (ICLR)*, San Diego, CA, May 2015. URL `http://arxiv.org/abs/1412.6626`.

Hénaff, O. J., Bai, Y., Charlton, J. A., Nauhaus, I., Simoncelli, E. P., and Goris, R. L. Primary visual cortex straightens natural video trajectories. *Nature communications*, 12(1):1–12, 2021.

Jing, L., Vincent, P., LeCun, Y., and Tian, Y. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*, 2021.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Knyazev, A. V. and Argentati, M. E. Principal angles between subspaces in an a-based scalar product: algorithms and perturbation estimates. *SIAM Journal on Scientific Computing*, 23(6):2008–2040, 2002.

Kriegeskorte, N. and Kievit, R. A. Representational geometry: integrating cognition, computation, and the brain. *Trends in cognitive sciences*, 17(8):401–412, 2013.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Laughlin, S. A simple coding procedure enhances a neuron's information capacity. *Zeitschrift für Naturforschung c*, 36(9-10):910–912, 1981.

Lezama, J., Qiu, Q., Musé, P., and Sapiro, G. Ole: Orthogonal low-rank embedding-a plug and play geometric loss for deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8109–8118, 2018.

Lu, J. and Steinerberger, S. Neural collapse with cross-entropy loss. *arXiv preprint arXiv:2012.08465*, 2020.

Ma, Y., Derksen, H., Hong, W., and Wright, J. Segmentation of multivariate mixed data via lossy data coding and compression. *IEEE transactions on pattern analysis and machine intelligence*, 29(9):1546–1562, 2007.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=rJzIBfZAb`.

Marshall, A. W., Olkin, I., and Arnold, B. C. *Inequalities: theory of majorization and its applications*, volume 143. Springer, 1979.

Nieh, E. H., Schottdorf, M., Freeman, N. W., Low, R. J., Lewallen, S., Koay, S. A., Pinto, L., Gauthier, J. L., Brody, C. D., and Tank, D. W. Geometry of abstract learned knowledge in the hippocampus. *Nature*, 595(7865):80–84, 2021.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Ozsoy, S., Hamdan, S., Arik, S. Ö., Yuret, D., and Erdogan, A. T. Self-supervised learning with an information maximization criterion. *arXiv preprint arXiv:2209.07999*, 2022.

Papyan, V., Han, X., and Donoho, D. L. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

Recht, B., Fazel, M., and Parrilo, P. A. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

Schwartz, O. and Simoncelli, E. P. Natural signal statistics and sensory gain control. *Nature neuroscience*, 4(8): 819–825, 2001.

Simoncelli, E. P. and Olshausen, B. A. Natural image statistics and neural representation. *Annual review of neuroscience*, 24(1):1193–1216, 2001.

Stringer, C., Pachitariu, M., Steinmetz, N., Carandini, M., and Harris, K. D. High-dimensional geometry of population responses in visual cortex. *Nature*, 571(7765): 361–365, 2019.

Tian, Y., Krishnan, D., and Isola, P. Contrastive multiview coding. In *European conference on computer vision*, pp. 776–794. Springer, 2020.

Wang, T. and Isola, P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pp. 9929–9939. PMLR, 2020.

Wang, Y., Lin, J., Cai, Q., Pan, Y., Yao, T., Chao, H., and Mei, T. A low rank promoting prior for unsupervised contrastive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Wu, Z., Xiong, Y., Yu, S. X., and Lin, D. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3733–3742, 2018.

Yu, Y., Chan, K. H. R., You, C., Song, C., and Ma, Y. Learning diverse and discriminative representations via the principle of maximal coding rate reduction. *Advances in Neural Information Processing Systems*, 33:9422–9434, 2020.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pp. 12310–12320. PMLR, 2021.

A. Optimal Embeddings

Recall the setting of self supervised learning as described in Balestriero & LeCun (2022): given a dataset $\boldsymbol{X'} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_N]^T \in \mathbb{R}^{N \times D'}$ we construct a new dataset by creating k randomly augmented views of the original data, $\boldsymbol{X} = [\text{view}_1(\boldsymbol{X'}), ..., \text{view}_k(\boldsymbol{X'})] \in \mathbb{R}^{Nk \times D}$. The advantage of doing so is that we can now leverage the knowledge that different views of the same underlying datapoint are *semantically related*. We can express this notion of similarity in the symmetric matrix $\boldsymbol{G} \in \{0, 1\}^{Nk \times Nk}$ with $\boldsymbol{G}_{ij} = 1$ if augmented datapoints i and j are semantically related (and $\boldsymbol{G}_{ii} = 1$ as any datapoint is related to itself). We can normalize $\boldsymbol{G}$ such that its rows and columns sum to 1 (so rows of $\boldsymbol{G}$ are k-sparse with nonzero entries equal to $1/k$).

Now let $\boldsymbol{Z} \in \mathbb{R}^{Nk \times d}$ be an embedding of the augmented dataset. Then we have $\boldsymbol{GZ} = [\boldsymbol{C}, ..., \boldsymbol{C}]^T$ where $\boldsymbol{C}$ is the matrix of centroid vectors introduced above, and the number of repetitions of $\boldsymbol{C}$ is k. Then because $\boldsymbol{\sigma}([\boldsymbol{C}, ..., \boldsymbol{C}]) = \sqrt{k}\boldsymbol{\sigma}(\boldsymbol{C})$ we can write MMCR loss function as,

$$\begin{aligned} \mathcal{L} &= -||\boldsymbol{GZ}||_* \\ &= -||\boldsymbol{Q\Lambda Q}^T\boldsymbol{USV}^T||_* \\ &= -||\boldsymbol{\Lambda Q}^T\boldsymbol{US}||_* \end{aligned} \tag{5}$$

Where we have taken the eigendecomposition of $\boldsymbol{G}$ which is real and symmetric and the SVD of $\boldsymbol{Z}$, and then used the fact that the singular value spectrum is invariant under left or right orthogonal transformations. We now show that a global optima of this objective is achieved when the left singular vectors of $\boldsymbol{Z}$ are the eigenvectors of $\boldsymbol{G}$ and the singular values of $\boldsymbol{Z}$ are proportional to the eigenvalues of $\boldsymbol{G}$. Throughout we will assume that the size of the dataset is greater than the dimensionality of the embedddings, $N > d$, as is the case in practical applications. First we prove a simple lemma about the spectrum of matrices who are extended by zeros (i.e. embedded in a higher dimensional space).

Lemma A.1: For $\boldsymbol{A} \in \mathbb{R}^{N \times N}$, $\boldsymbol{B} \in \mathbb{R}^{N \times d}$ with $d < N$, $||\boldsymbol{AB}||_* = ||\boldsymbol{A}\tilde{\boldsymbol{B}}||_*$ where $\tilde{\boldsymbol{B}} = [\boldsymbol{B}, \boldsymbol{0}] \in \mathbb{R}^{N \times N}$.

Proof: First note that $\boldsymbol{A}\tilde{\boldsymbol{B}} = [\boldsymbol{AB}, \boldsymbol{0}]$ so it suffices to show that for arbitrary $\boldsymbol{X}$ that $\boldsymbol{\sigma}(X) = \boldsymbol{\sigma}([\boldsymbol{X}, \boldsymbol{0}])$. Taking the SVD of $\boldsymbol{X}$,

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{U} & \widetilde{\boldsymbol{U}} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma} \\ \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{V}^T \end{bmatrix} = \boldsymbol{U\Sigma V}^T$$

Then a valid singular value decomposition for $\tilde{X}$ is

$$\tilde{\boldsymbol{X}} = \begin{bmatrix} \boldsymbol{U} & \widetilde{\boldsymbol{U}} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{V}^T & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix}$$

Clearly then, $||\boldsymbol{X}||_* = ||\tilde{\boldsymbol{X}}||_*$

Theorem: The proposed loss achieves a global minimum when the left singular vectors of $\boldsymbol{Z}$ are the eigenvectors of $\boldsymbol{G}$, and the singular values of $\boldsymbol{Z}$ are proportional to the top d eigenvalues of $\boldsymbol{G}$.

Proof: Let $\tilde{\boldsymbol{Z}} = [\boldsymbol{Z}, \boldsymbol{0}] \in \mathbb{R}^{N \times N}$. By Lemma A.1 we have $||\boldsymbol{GZ}||_* = ||\boldsymbol{G}\tilde{\boldsymbol{Z}}||_*$. Von Neumann's trace inequality can be used to show $||\boldsymbol{G}\tilde{\boldsymbol{Z}}||_* \leq \sum_{i=1}^{Nk} \boldsymbol{\sigma}_i(\boldsymbol{G})\boldsymbol{\sigma}_i(\tilde{\boldsymbol{Z}})$ (see Marshall et al. (1979) for proof). Examining (4) it is clear that this bound is achieved when $\boldsymbol{U} = \boldsymbol{Q}$. The problem can therefore be reduced to the constrained optimization problem,

$$\begin{aligned} &\min_{\boldsymbol{\sigma}_i(\tilde{\boldsymbol{Z}})} \sum_{i=1}^{Nk} \boldsymbol{\sigma}_i(\boldsymbol{G})\boldsymbol{\sigma}_i(\tilde{\boldsymbol{Z}}) \\ &\text{subject to} \sum_{i=1}^{Nk} \boldsymbol{\sigma}_i(\tilde{\boldsymbol{Z}})^2 = Nk \end{aligned}$$

where the constraint comes from the fact that columns of $\boldsymbol{Z}$ are unit vectors. Intuitively, we are maximizing the inner product between a fixed vector $\boldsymbol{\sigma}(\boldsymbol{G})$ and a vector with fixed L2 norm. The solution of course is to align the two vectors as

closely as possible, i.e. when $\boldsymbol{\sigma}_i(\tilde{\boldsymbol{Z}}) \propto \boldsymbol{\sigma}_i(\boldsymbol{G})$ for $i = 1, ..., d$. It is worth noting that by construction $\boldsymbol{\sigma}_i(\tilde{\boldsymbol{Z}}) = 0$ for $i > d$ and the columns of $\boldsymbol{U}$ associated with these zero valued singular values are unconstrained.

B. Pytorch Style Pseudocode for MMCR

```
# h: encoder
# g: projection head
# B: batch size
# K: number of augmentations
# D: projector output dimensionality
#
# lmbda: trade-off parameter
for x in loader:
    # K randomly augmented views
    x = multi_augment(x) # B x K x H x W

    # push through encoder and projector
    z = g(h(x)) # B x K x D

    # project onto unitsphere
    z = normalize(z, dim=-1)

    # calculate centroids (mean over augmentation axis)
    c = z.mean(dim=1) # B x D

    # calculate singular values
    U_z, S_z, V_z = svd(z) # batch svd
    U_c, S_c, V_c = svd(c)

    # calculate loss
    loss = -1.0 * sum(S_c) + lmbda * sum(S_z) / B

    # backward pass and optimization step
    loss.backward()
    optim.step()
```

C. Mean Field Theory Manifold Capacity Background Information

Mean Field Theory Recall the problem setting for manifold capacity analysis: given a set of P manifolds embedded in a feature space of dimensionality D, each assigned a random binary class label (Chung et al., 2018). Manifold capacity theory is concerned with the question: what is the largest value of $\frac{P}{D}$ such that there exists (with high probability) a hyperplane separating the two classes? In the thermodynamic limit, where $P, D \to \infty$ but $\frac{P}{D}$ remains finite, the inverse capacity can be written exactly,

$$\alpha_M^{-1} = \mathbb{E}_{\vec{T}}[F(\vec{T})] \tag{6}$$

where, $F(\vec{T}) = \min_{\vec{V}} \left\{ \|\vec{V} - \vec{T}\|^2 \mid g_{\mathcal{S}}(\vec{V}) \geq 0 \right\}$, $\mathcal{S}$ is the set defining the manifold geometry (i.e. the set of vectors $\vec{S}$ that are points on an individual manifold), $\vec{T}$ are random vectors drawn from a white multivariate Gaussian distribution, and $g_{\mathcal{S}}(\vec{V}) = \min_{\vec{S}} \{\vec{V} \cdot \vec{S} \mid \vec{S} \in \mathcal{S}\}$, is the concave support function.

The KKT equations for this convex optimization problem are:

$$\begin{aligned} \vec{V} - \vec{T} - \lambda \tilde{S}(\vec{T}) &= 0 \\ \lambda &\geq 0 \\ g_{\mathcal{S}}(\vec{V}) - \kappa &\geq 0 \\ \lambda \left[g_{\mathcal{S}}(\vec{V}) - \kappa \right] &= 0. \end{aligned} \tag{7}$$

, where $\tilde{S}(\vec{T})$ is a subgradient of the support function. When the support function is differentiable, the subgradient is unique and equal to the gradient,

$$\tilde{S}(\vec{T}) = \nabla g_{\mathcal{S}}(\vec{V}) = \arg\min_{\vec{S}\in\mathcal{S}} \vec{V}\cdot\vec{S} \tag{8}$$

$\tilde{S}(\vec{T})$ is the unique point in the convex hull of $\mathcal{S}$ that satisfies the first KKT equation, and is called the "anchor point" for $\mathcal{S}$ induced by the random vector $\vec{T}$.

Equivalent Interpretation of Anchor Points For a given dichotomy (random binary class labelling) the weight vector of the maximum margin separating hyperplane can be decomposed into a sum of at most P vectors, with each manifold contributing a single vector, which lies within the convex hull of the manifold. The position of said point point is a function of the manifolds position relative to all of the other manifolds in the space and depends on the particular set of random labels. Thus there exists a distribution of separating-hyperplane-determining-points for each individual manifold. Using the "cavity" method it can be shown that these points are none other than the anchor points that are involved in solving the optimization problem described above (Gerl & Krey, 1994).

Numerical Solution To solve the mean field equations numerically, one samples several random Gaussian vectors $\vec{T}$, and then for each $\vec{T}$, $\vec{V}$ and $\vec{S}$ are determined by solving the quadratic programming program given above. The capacity is then estimated as the mean value of F or the samples $\vec{T}$.

Manifold Geometries The way the capacity varies in terms of the statistics of the anchor points can be simplified by introducing two key quantities, the manifold radius R_M and manifold dimensionality R_M:

$$\begin{aligned} R_M^2 &= \mathbb{E}_{\vec{T}}[||\tilde{S}(\vec{T})||^2] \\ D_M &= \mathbb{E}_{\vec{T}}[\vec{T}\cdot\hat{S}(\vec{T})] \end{aligned} \tag{9}$$

where $\hat{S}(\vec{T})$ is a unit-vector in the direction of the anchor point $\tilde{S}$. In particular as discussed in the main text, the manifold capacity can be approximated by $\phi(R_M\sqrt{D_M})$ where ϕ is a monotonically decreasing function.

Elliptical Geometries In the case where the manifolds exhibit elliptical symmetries, the manifold radius and dimensionality can be written in terms of the eigenvalues of the covariance matrix of the anchor points:

$$\begin{aligned} R_M^2 &= \sum_i \lambda_i^2 \\ D_M &= \frac{\left(\sum_i \lambda_i\right)^2}{\sum_i \lambda_i^2} \end{aligned} \tag{10}$$

So, in this case R_M is the total variability of the anchor points, and D_M is a generalized participation ratio of the anchor point covariance, a well known soft measure of dimensionality.

D. Additional Pre-training information

Settings for CIFAR/STL-10 We take the parameters of each augmentation directly from Zbontar et al. (2021), but for these lower resolution images we omitted Gaussian blurring and solarization augmentations. All models were trained for 500 epochs using the Adam optimizer (Kingma & Ba, 2014) with a learning rate of $1e-3$ and weight decay of $1e-6$. For all three methods we used a one hidden layer MLP with hidden dimension of 512 and output dimension of 128 for the projector head g. We swept batch size for each method and chose the one that resulted in the highest downstream task performance. For both SimCLR and Barlow Twins we found that a batch size of 128 was optimal (among 32, 64, 128, 256, and 512) for all 3 datasets. For MMCR there is a trade-off between batch size and the number of augmentations used, and the optimal value of that trade-off is highly dataset dependent. For CIFAR-10 and CIFAR-100 we used batch size of 32 and 40 views, and for STL-10 we used a batch of 64 with 20 views For Barlow Twins we used $\lambda = \frac{1}{128}$ which normalizes for the number of elements in the on-diagonal and off-diagonal terms in the loss. For SimCLR we used the recommended setting of $\tau = 0.5$. The overall performance of both baseline methods (and likely MMCR as well) could be increased with a more thorough

hyperparameter search and by employing methodology that more closely matches the original works. For example, both methods would likely benefit from the combination of larger batch size, the use of the LARS optimizer (which is designed for large batch optimization), a learning rate scheduler consisting of linear warm-up followed by cosine annealing, longer training, and the use of more diverse augmentations (i.e. including solarization and gaussian blur). Additionally Barlow Twins reports that the representation can benefit from using a much larger projector network than we use. Because our goal was primarily to demonstrate that MMCR can produce representations that are comparable to these baselines rather than to produce state-of-the-art results on small scale datasets we opted for simplifications wherever possible (using off the shelf Adam for optimization with a fixed learning rate, and fixing architectural hyperparameters like the projector dimensionality).

Settings for ImageNet-100 For ImageNet we more closely match the pre-training procedures of previous works. We use a batch size of 2048 and a smaller number of views for MMCR (4), and also use the full suite of augmentations from Zbontar et al. (2021). For the sake of efficiency we train for a reduced number of epochs (200). For MMCR and SimCLR we modified the projector hidden dimensionality to be 4096 for the projector head, following the original work (Chen et al., 2020a). For Barlow Twins we used the recommended 2-layer MLP with hidden and output dimensions of 8192, and set $\lambda = 5e - 3$, however these hyperparameters were optimal for the full ImageNet dataset, and not neccesarrily for ImageNet-100. We were unable to achieve better downstream performance using a ResNet-50 backbone than what has previously been reported in the literature for this dataset with a ResNet-18 backbone, therefore we report the ResNet-18 performance reported in (da Costa et al., 2022). For SimCLR we use $\tau = 0.1$ which is the recommended setting for larger batch sizes.

Settings for ImageNet-1k: For ImageNet-1k we use mostly identical settings to ImageNet-100, but we increased the capacity of the projector network (using a 2 hidden layer MLP with hidden dimenisons of 8192 and output dimension of 512). We scaled the learning rate linearly with batch size: $\text{lr} = 0.6 \times \frac{\text{batch size}}{256}$. Additionally we reduce the number of pretraining epochs to 100.

E. Details of Representational Analyses

E.1. Manifold Capacity Analysis

For each pre-trained model, we extract layer activations across the ResNet hierarchy after a forward pass of a set of images. For class manifold analysis, the set of images contain 10 classes, where each class has 100 examples. Augmentation manifolds instead have 100 exemplars with 100 examples each. Following (Cohen et al., 2020), we take activations from all convolutional layers in ResNet-50 after a ReLU non-linearity. The specific extracted layers highlighted in bold fonts are given by Table 2. The final analysis results are averaged over five data samplings with different random seeds and random projections of intermediate features to lower-dimension spaces (default 5000 dimensions).

E.2. Gradient Coherence Analysis

In Fig 3, for each of the classes of CIFAR-10, we generate 100 batches of 32 augmentation manifolds of samples from a specific class (with 40 augmentations each). We then measure the gradient of the loss function for each batch during different stages of training, and compute the cosine similarity between every pair of gradients. Across all stages of training the mean cosine similarity between gradients generated from batches of the same class is larger than those from distinct classes (left column). This observation remains true when isolating the gradients of parameters from different stages of in the resnet-50 hierarchy (center and right columns, respectively).

E.3. Manifold Subspace Alignment

For Fig. 4 we generated 100 samples from the augmentation manfiolds of 500 images in the CIFAR-10 dataset. We then measure the mean subspsace angle (left column), fraction of shared variance (middle column) and centroid cosine similarity between each pair of manifolds. The same procedure was used for generating the data for 5.

Subspace Angle. Besides measuring the size and dimensionality of individual object manifolds we also wish to characterize the degree of overlap between pairs of manifolds. For this, we measure the angle between their subspaces (Knyazev & Argentati, 2002), which is a generalization of the notion of angles that applies to subspaces of arbitrary dimension.

Shared Variance. Object manifolds will generally have a lower intrinsic dimensionality then the space they are embedded in. Therefore, the data will have low variance along several of the principal vectors used to calculate the set of subspace

Table 2. A Total of 18 Extracted ResNet-50 Layers (in **Bold**) for MFTMA Analysis

Layer	Type	Conv2d Size (H × W × C)
pixel	**Input**	None
conv1	$\begin{bmatrix}\text{Conv2d}\\ \text{BatchNorm}\\ \textbf{ReLU}\end{bmatrix} \times 1$	$\begin{bmatrix}7 \times 7 \times 64\end{bmatrix} \times 1$
conv2_x	$\begin{bmatrix}\begin{bmatrix}\text{Conv2d}\\ \text{BatchNorm}\end{bmatrix} \times 3\\ \textbf{ReLU}\end{bmatrix} \times 3$	$\begin{bmatrix}1 \times 1 \times 64\\ 3 \times 3 \times 64\\ 1 \times 1 \times 256\end{bmatrix} \times 3$
conv3_x	$\begin{bmatrix}\begin{bmatrix}\text{Conv2d}\\ \text{BatchNorm}\end{bmatrix} \times 3\\ \textbf{ReLU}\end{bmatrix} \times 4$	$\begin{bmatrix}1 \times 1 \times 128\\ 3 \times 3 \times 128\\ 1 \times 1 \times 512\end{bmatrix} \times 4$
conv4_x	$\begin{bmatrix}\begin{bmatrix}\text{Conv2d}\\ \text{BatchNorm}\end{bmatrix} \times 3\\ \textbf{ReLU}\end{bmatrix} \times 6$	$\begin{bmatrix}1 \times 1 \times 256\\ 3 \times 3 \times 256\\ 1 \times 1 \times 1024\end{bmatrix} \times 6$
conv5_x	$\begin{bmatrix}\begin{bmatrix}\text{Conv2d}\\ \text{BatchNorm}\end{bmatrix} \times 3\\ \textbf{ReLU}\end{bmatrix} \times 3$	$\begin{bmatrix}1 \times 1 \times 512\\ 3 \times 3 \times 512\\ 1 \times 1 \times 2048\end{bmatrix} \times 3$

angles, and so many of the principal angles will have little meaning. To address this limitation we also compute the shared variance between the linear subspaces that contain object manifolds.

F. Implicit MMCR Effectively Reduces Augmentation Manifold Nuclear Norm

To test whether or not implicit manifold compression actually reduces the mean augmentation manifold nuclear norm, we can vary the value of λ. Below we see the evolution of bother terms of the loss for several different values of lambda during training on CIFAR-10. For these experiments the batch size was 64 and the number of augmentations per image was 4.0. As shown in 7, the level of compression of individual manifolds is nearly the same across all values of the tradeoff parameter tested.

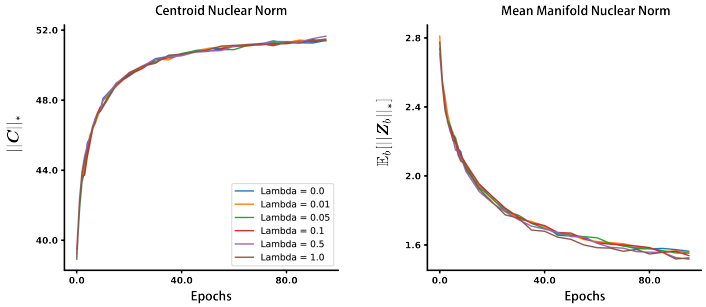

Figure 7. Validation loss values for different values of λ

G. Classification Evaluation Procedure

CIFAR and STL-10: During pre-training all models were monitored with a k-nearest neighbor classifier (k=200) and checkpointed every 5 epochs. After pre-training, we trained linear classifiers on all checkpoints whose monitor accuracy was within 1% of the highest observed accuracy, and select the model that achieves the highest linear classification accuracy. Linear classifiers were trained using the Adam optimizer with batch size of 1024 and an initial learning rate of 0.1, which decayed according to a cosine scheduler over the course of 50 epochs. For the linear classifier training, at train time we use the same set of augmentations as during unsupervised pretraining, at test time we only use center cropping and random horizontal flipping.

ImageNet-1k/100: For ImageNet datasets we closely followed the most widely adopted evaluation procedure. Following pre-training we freeze the encoder weights and train a linear layer in a supervised fashion using SGD with a batch size of 256, learning rate of 0.3, and weight decay of 1e-6 for 100 epochs. During linear classifier training the only data augmentations are random cropping and random horizontal flips, and during evaluation inputs are center cropped.

H. Additional Details for Adversarial Robustness Analyses

In Figure 6, we choose 50 iterations for the PGD ℓ_∞-norm since it guarantees a robust accuracy value not far away from asymptotically larger PGD attack iterations (Madry et al., 2018; Croce & Hein, 2020). In our experiment, we have shown that the PGD attack indeed converges in a similar fashion (See Figure 8). However, the robust accuracy for MMCR tends to converge at larger PGD attack iterations.

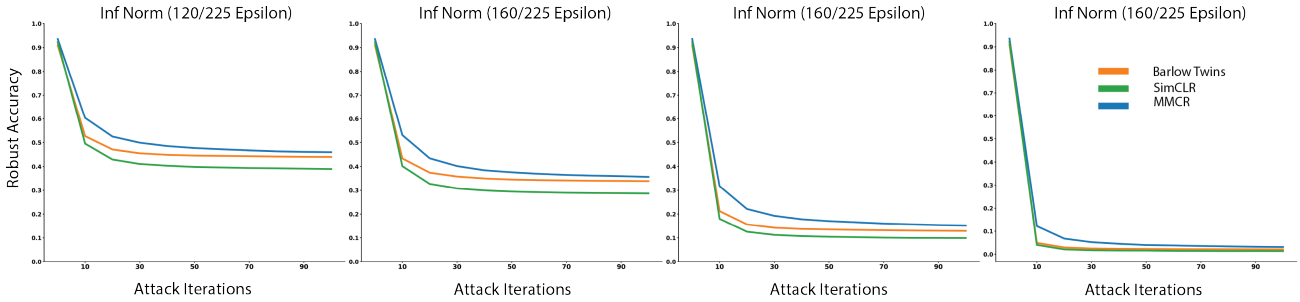

Figure 8. Convergence for different settings of adversarial attack strengths

We therefore also analyzed the robust accuracies for the three SSL methods with varying iterations across all epsilon attack strength. Figure 9 shows MMCR exhibits a significantly higher robust accuracy compared to Barlow-Twins and SimCLR in the low iterations regime.

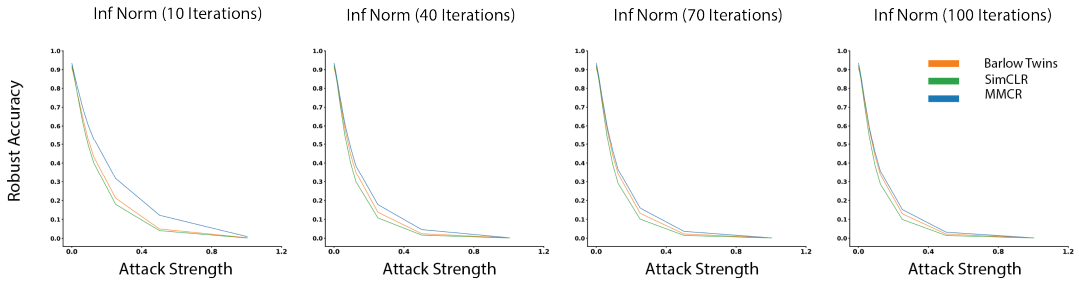

Figure 9. PGD ℓ_∞-norm attack with varying iterations.

Aside from the standard PGD adversarial attack, we also tested three SSL methods under the AutoAttack protocol. The ℓ_∞-norm AutoAttack accuracy is given by Table 3.

Table 3. AutoAttack ℓ_∞-norm Robust Accuracy

Method	Clean Accuracy	Eps = 40/255	Eps = 160/255
Barlow Twins (our repro.)	90.91	74.55	31.53
SimCLR (our repro.)	92.22	72.48	26.37
Implicit MMCR (ours)	93.53	**75.88**	**32.47**

Table 4. Top-1 classification accuracies of linear classifiers for representations trained with various datasets and objective functions. Note: for Barlow Twins on ImageNet-100 we report the result from da Costa et al. (2022) which uses a ResNet-18 backbone, as we were unable to obtain better performance. For MMCR on ImageNet-100 we tested both 2 views (matched to baselines) and 4 views, results are formatted (2-view)/(4-view)

[t]

Method	CIFAR-10	CIFAR-100	STL-10	ImageNet-100
Barlow Twins (our repro.)	90.91	67.91	89.96	80.38*
SimCLR (our repro.)	92.22	70.04	91.11	79.64
MMCR ($\lambda = 0.0$)	93.53	69.87	90.62	81.52/82.88
MMCR ($\lambda = 0.01$)	93.39	70.94	90.77	81.28/82.56

I. Training Metrics

In the Fig. 10 below we monitor the evolution of both the objective (second panel), the mean augmentation manifold nuclear norm, the centroid norm, and the mean centroid similarity evaluated on the test set over the course of training.

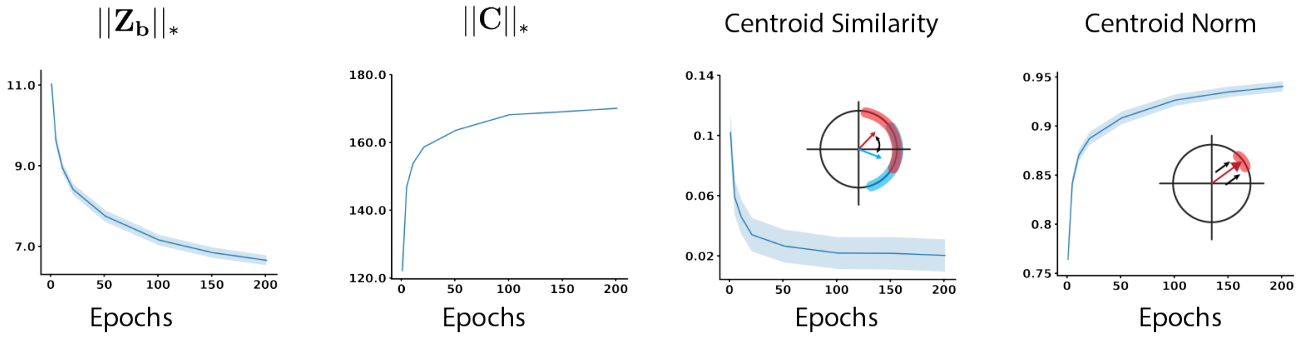

Figure 10. Evolution of various metrics during training. Geometric measures are evaluated on a set of 200 manifolds, each defined by an image drawn from the CIFAR-10 dataset, along with 16 augmentations. Shaded regions indicate a 95% confidence interval around the mean.

J. Classification Performance on Smaller Datasets

In Table 4 below we report the performance of both our method as well as Barlow Twins and SimCLR when trained using a ResNet-50 backbone on smaller datasets.

K. Batch Size Dependence

One of the most cited drawbacks of contrastive SSL methods has been that training with large batch sizes is necessary to achieve strong performance on downstream tasks, while non-contrastive methods such as VICReg and Barlow Twins (Zbontar et al., 2021; Bardes et al., 2021) that place constraints on the cross-correlation/covariance matrices of the embeddings are much more amenable to smaller batch training. It is also worth noting that the need for large batch sizes in contrastive methods can be alleviated in various ways such as by maintaining a memory bank as in Wu et al. (2018) or by employing a slowly updating momentum encoder as in He et al. (2020). Given that our method is neither wholly contrastive nor non-

contrastive as it acts on the spectrum of the embedding matrix directly we wondered whether its performance as a function of training batch size would exhibit more similarity to one category of SSL or another in terms of batch size dependency. We pretrained on ImageNet-1k using batch size in $256, 512, 1024, 2048, 4096$ and evaluate the linear classification accuracy for each. Encouragingly we observe only a modest decrease in performance for the smallest batch size tested, but strangely there is a dip at batch size of 512. Given this we tested two additional batch sizes of 320 and 768. The results of this sweep in comparison to Barlow Twins and SimCLR are shown in in Fig. 11 Future work should endeavor to better understand the impact of various hyperparameters on the quality of learned representations. Note that for these runs we used two views and the linear learning rate scaling as described in Appendix D.

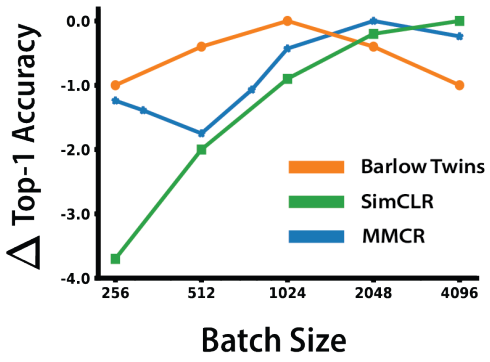

Figure 11. Plotted on the y-axis is the relative drop in performance from the top setting for three methods (viewing downstream accuaracy as a function of batch size, plotted is: Accuracy(Batch Size) - Max(Accuracy(Batch Size))). Data for both Barlow Twins and SimCLR are inherited from Zbontar et al. (2021).