Harnessing small projectors and multiple views for efficient vision pretraining

Arna Ghosh [*1] **Kumar Krishna Agrawal** [*2] **Shagun Sodhani** [3]

Adam M. Oberman[† 3] **Blake A. Richards** [† 1 4 5]

Abstract

Recent progress in self-supervised (SSL) visual representation learning has led to the development of several different proposed frameworks that rely on augmentations of images but use different loss functions. However, there are few theoretically grounded principles to guide practice, so practical implementation of each SSL framework requires several heuristics to achieve competitive performance. In this work, we build on recent analytical results to design practical recommendations for competitive and efficient SSL that are grounded in theory. Specifically, recent theory tells us that existing SSL frameworks are actually minimizing the same idealized loss, which is to learn features that best match the data similarity kernel defined by the augmentations used. We show how this idealized loss can be reformulated to a functionally equivalent loss that is more efficient to compute. We study the implicit bias of using gradient descent to minimize our reformulated loss function, and find that using a stronger orthogonalization constraint with a reduced projector dimensionality should yield good representations. Furthermore, the theory tells us that approximating the reformulated loss should be improved by increasing the number of augmentations, and as such using multiple augmentations should lead to improved convergence. We empirically verify our findings on CIFAR, STL and Imagenet datasets, wherein we demonstrate an improved linear readout performance when training a ResNet-backbone using our theoretically grounded recommendations. Remarkably, we also demonstrate that by leveraging these insights, we can reduce the pretraining dataset size by up to $2\times$ while maintaining downstream accuracy simply by using more data augmentations. Taken together, our work provides theoretically grounded recommendations that can be used to improve SSL convergence and efficiency.

1 Introduction

Unsupervised representation learning, i.e., learning features without human-annotated labels, is critical for progress in computer vision. Modern approaches, grouped under the *self-supervised learning (SSL)* umbrella, build on the core insight that similar images should map to nearby points in the learned feature space – often termed as the *invariance criterion*. Current SSL methods can be broadly categorized into contrastive and non-contrastive algorithms, based on whether they formulate their loss functions using negative samples or not, respectively.

[*]Equal Contribution, [†] Co-senior authorship, Correspondence: blake.richards@mcgill.ca
[1]Mila - Quebec AI Institute & Computer Science, McGill University, Montréal, QC, Canada
[2]UC Berkeley, CA, USA, [3] Meta FAIR, Toronto, ON, Canada
[3]Mila - Quebec AI Institute & Mathematics and Statistics, McGill University, Montréal, QC, Canada
[4]Neurology & Neurosurgery and Montreal Neurological Institute, McGill University, Montréal, QC, Canada
[5]CIFAR Learning in Machines & Brains Program, Toronto, ON, Canada

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-1258

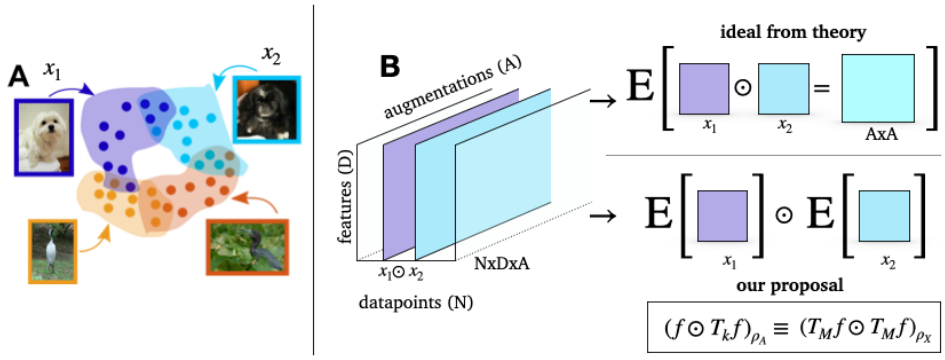

Figure 1: Design of existing SSL algorithms relies on heuristics. **(A)** Augmentation graphs are common in vision pretraining, providing generalizable features for downstream tasks. (B) We propose an equivalent loss function for SSL pretraining that recovers the same eigenfunctions more efficiently than existing approaches.

Despite this difference in their loss formulations, recent theoretical work has established an equivalence between the contrastive and non-contrastive SSL frameworks [19]. This work shows that these different SSL formulations are ultimately minimizing a loss that encourages the learning of features that best match the data similarity kernel defined by the augmentations used. However, this notion of theoretical equivalence holds only in the limit of ideal pretraining settings, i.e. with access to infinite data and compute budget, and the feature learning behavior of different SSL algorithms in practical scenarios is still not well understood. Therefore, researchers often use empirically driven heuristics that are theoretically ungrounded to design successful applications, such as (i) a high-dimensional projector head for non-contrastive SSL or (ii) the use of two augmentations per image [3]. Moreover, existing SSL algorithms are extremely data-hungry, relying on large-scale datasets [33] or data engines [30] to achieve good representations. While this strategy works exceptionally well in data-rich settings (like training on natural-images), it is not viable in data-constrained settings (like medical imaging), where samples are relatively scarce.

With these challenges in mind, the primary focus of this work is to develop theoretically grounded recommendations for improving the effectiveness and efficiency of feature learning, both with respect to the required compute budget as well as data points. Like any unsupervised representation learning algorithm, features learned through SSL depend on three factors: (i) implicit bias of the architecture, (ii) explicit invariance imposed by data augmentations, (iii) implicit bias of the learning rule. While previous works predominantly studied the role of the model architecture capacity and loss function, and their interplay with data augmentations [9, 44], our approach broadens this perspective by also considering the role of the learning rule (gradient descent) in optimizing these loss functions. Specifically, we extend the previous theoretical findings [44] that unified the desiderata of different SSL algorithms. We reformulate the idealized unifying loss to propose a functionally equivalent loss that is more compute-efficient (see Figure 1). Based on our loss formulation, we provide two practical recommendations that can help improve the efficiency of SSL pipelines while maintaining good performance. First, we show that optimizing the reformulated loss using gradient descent can often reduce the orthogonality among the learned embeddings, thereby leading to an inefficient use of the projector network's capacity. Consequently, we recommend using a stronger orthogonalization constraint to eliminate the requirement of high-dimensional projector heads, thereby significantly reducing the parameter overhead of good feature learning. Second, we show that increasing the number of augmentations leads to a better estimate of the data similarity kernel. Consequently, we recommend using more augmentations to improve optimization convergence and learn better features earlier in training.

We empirically verify our theoretically grounded recommendations using the popular ResNet backbone on benchmark datasets: CIFAR, STL and Imagenet. Strikingly, we show that our multi-augmentation approach can learn good features even with *half* of the samples in the pretraining dataset. Our recommendations provide a path towards making SSL pretraining more data and compute-efficient without harming performance and could unlock massive performance gains in data-constrained setups. In summary, our core contributions are as follows:

https://doi.org/10.52202/079017-1258

- **Efficient SSL loss formulation:** We propose an functionally equivalent and compute-efficient formulation of the SSL desiderata that yields the eigenfunctions of the augmentation-defined data similarity kernel.
- **Role of heuristics:** Based on our loss formulation and the implicit bias of gradient descent in optimizing this loss, we provide a mechanistic explanation for the role of projector dimensionality and the number of data augmentations. Consequently, we empirically demonstrate that low-dimensional projector heads are sufficient and that using more augmentations leads to learning better representations.
- **Data efficient SSL:** Leveraging the convergence benefits of the multi-augmentation SSL framework, we empirically demonstrate that we can learn good features with significantly smaller datasets (up to $2\times$) without harming downstream linear probe performance.

2 Preliminaries

Existing SSL approaches in computer vision In recent years machine learning researchers have developed a number of effective approaches for learning from data without labels. The most popular approaches use augmentations of data points as targets for themselves. One of the first was a Simple framework for Contrastive Learning (SimCLR), which relied on an infoNCE loss with augmentations of an image as positive targets and augmentations of other images as negative samples (to construct the contrastive loss) [12]. Other works have relied on non-contrastive approaches, notably BarlowTwins [43] and VICReg [5]. BarlowTwins, which was inspired by the ideas of the neuroscientist Horace Barlow (cite), also uses augmentations of images, but it instead aims to optimize the covariance structure of the representations in order to reduce redundancies in the feature space [43]. Variance Invariance Covariance Regularization (VICReg) was a modification of BarlowTwins that added a variance term in the loss in order to ensure that every feature dimension has a finite variance [5]. In this paper we will focus on non-contrastive methods like BarlowTwins and VICReg, but in line with previous work [44], we also consider how these approaches relate to contrastive methods like SimCLR.

Formalizing the self-supervised learning problem Now, we will formalize the unsupervised representation learning problem for computer vision. In particular, we assume access to a dataset $\mathcal{D} = \{x_1, x_2, ..., x_n\}$ with $x_i \in \mathbb{R}^p$ consisting of unlabeled images. The objective is to learn a d-dimensional representation ($d < p$) that is useful across multiple downstream applications. We focus on learning the parameters of a deep neural network $f_\theta \in \mathcal{F}_\Theta$, using the multi-augmentation SSL framework, wherein multiple views of an image are used to optimize the pretraining loss function, $\mathcal{L}_{pretrain}(f_\theta, \mathcal{D})$

Non-Contrastive Self-Supervised Learning (NC-SSL) algorithms impose invariance to data augmentations, while imposing regularization on the geometry of the learned feature space. More generally, $\mathcal{L}_{pretrain}$ can be formulated with two terms (i) $\mathcal{L}_{invariance}$: to learn invariance to data augmentations and (ii) $\mathcal{L}_{collapse}$: regularization to prevent collapsing the feature space to a trivial solution.

$$\mathcal{L}_{pretrain} := \mathcal{L}_{invariance} + \beta \mathcal{L}_{collapse} \tag{1}$$

where β denotes a hyperparameter that controls the importance of the collapse-preventing term relative to the invariance term. This formulation separates features that are invariant to the augmentations from those that are sensitive to them. Intuitively, the ideal feature space is more sensitive to semantic attributes (e.g. "that's a dog") and less sensitive to irrelevant attributes (e.g. "direction the dog is facing"), facilitating generalization to new examples.

Data Augmentation graph was introduced by Haochen *et al.* [22] to analyze contrastive losses, like SimCLR [12]. Briefly, we define a graph $\mathcal{G}(\mathcal{A}, \mathcal{W})$ that captures the relationship between images derived from all possible data augmentations. The vertex set $(\mathcal{A}, \rho_A)$ is each augmented sample in a dataset, $\mathcal{X}$, and the adjacency matrix, $\mathcal{W}$, denotes the similarity between pairs of vertices. Let x_0 be an image in $\mathcal{X}$, and let $z = M(x_0) \in \mathcal{A}$ be a random data augmentation of the image, x_0. We define the probability density of reaching z from x_0 via a choice of mapping M:

$$p(z \mid x_0) = \mathbb{P}(z = M(x_0)), \tag{2}$$

Since the mapping is not generally invertible (e.g., cropping), we observe that $p(x_0 \mid z) \neq p(z \mid x_0)$. Using this definition, we now formally define the strength of the edge between nodes $x, z \in \mathcal{A}$ of the

 https://doi.org/10.52202/079017-1258

augmentation graph as the joint probability of generating augmentations x, z from the same image $x_0 \sim \rho_X$. Notably, the edge strength of the (degree-normalized) augmentation graph is equivalent to the data similarity kernel, defined in [44]. Formally,

$$k^{DAF}(x, z) = w_{xz} := \mathbb{E}_{x_0 \sim \rho_X} \left[\frac{p(x \mid x_0)}{p(x)} \frac{p(z \mid x_0)}{p(z)} \right] \tag{3}$$

The magnitude of w_{xz} captures the augmentation-defined similarity between x and z. A higher value of w_{xz} indicates that both patches are more likely to come from the same image and, thereby, are more similar.

The desiderata of different SSL algorithms can be understood as learning features F that best capture $k^{DAF}(x, z)$, i.e. $F(x)^T F(z) \approx k^{DAF}(x, z)$. Recent theoretical work has shows that different SSL losses can be formulated as special cases of the objective function that recovers the top-d eigenfunctions of $k^{DAF}(x, z)$ [44].

$$\mathcal{L}_{ssl}(F) = \mathbf{E}_{x,z \in \mathcal{A}} \left[(k^{DAF}(x, z) - F(x)^T F(z))^2 \right] \tag{4}$$

Note that all rotations of F that don't change its span define an equivalence class of solutions to Equation (4) and make no difference for the downstream generalization of a linear probe. Based on this insight, we define an equivalence among learned feature spaces:

Definition 2.1. Let $F(x) = (f_1(x), \dots f_d(x))$ be a d-dimensional feature vector (a vector of functions). Define the subspace

$$V = V(F) = \{h : X \to \mathbb{R} \mid h(x) = w \cdot F(x), \quad w \in \mathbb{R}^d\} \tag{5}$$

to be the span of the components of F. Given an n-dimensional feature vector, $G(x) = (g_1(x), \dots, g_n(x))$ we say the features G and F are equivalent, if $V(F) = V(G)$.

3 Implicit bias of non-contrastive SSL loss and optimization

We extend the recent theoretical results [44] to propose a compute-efficient reformulation of the loss function of the SSL desiderata that yields equivalent features, i.e. the functions spanning the eigenfunctions of the augmentation-defined data similarity kernel, k^{DAF}. Furthermore, we study the role of gradient descent in optimizing this loss function and uncover a selection and primacy bias in feature learning. Specifically, we find that gradient descent tends to learn the dominant eigenfunctions (eigenfunctions corresponding to larger eigenvalues) earlier during training, and often over-represents these eigenfunctions under weak orthogonalization constraints.

Consequently, we propose employing a stronger orthogonalization constraint during optimization when using a low-dimensional projector to ensure that learned features are equivalent to those learned with a high-dimensional projector. Furthermore, we argue that using more augmentations improves our sample estimate of k^{DAF}, thereby aiding the eigenfunction optimization problem. We dedicate the rest of this section to highlight our key theoretical insights, and practical recommendations that follow them.

3.1 Features in terms of data augmentation kernels

Let us define a kernel operator, T_k, for a positive semi-definite data augmentation kernel, k^{DAF}.

$$T_k f(x) = \mathbb{E}_{z \sim \rho_X}[k(z, x) f(z)] \tag{6}$$

such that Equation (4) can be equivalently written as (Equation 5 of [44])

$$\mathcal{L}_{ssl}(F) = \langle F, (I - T_k) F \rangle_{\rho_{\mathcal{A}}} \tag{7}$$

We can now use Mercer's theorem to factorize k^{DAF} into corresponding spectral features $G : X \to \ell_2$ (where ℓ_2 represents square summable sequences) [15, 16, 31]. However, note that computing k^{DAF} (or T_k) is expensive as it requires computing the overlap among all augmentations of every pair of data points. Instead of computing the eigenfunctions of T_k directly, we propose using an alternative operator T_M:

$$T_M f(x) = \mathbb{E}_{x_0 \sim M(x)} [f(x_0)] = \sum_{x_0} [p(x_0 \mid x) f(x_0)] \tag{8}$$

which averages the values of the function, f, over the augmented images $x_0 = M(x)$ of the data, x. We show that $T_M^T T_M$ is equivalent to T_k, and therefore T_M and T_k have shared eigenfunctions.

Theorem 3.1. *Let $G(x)$ be the infinite Mercer features of the backward data augmentation covariance kernels, k^{DAB}. Let $F(x) = (f_1(x), \ldots, f_{N_k}(x))$ be the features given by minimizing the following data augmentation invariance loss*

$$L(F) = \sum_{i=1}^{N_k} \|T_M f_i - f_i\|^2_{L^2(\rho_X)}, \quad \textit{subject to} \quad (f_i, f_j)_{\rho_X} = \delta_{ij} \tag{9}$$

which includes the orthogonality constraint. Then, $V(F) \subset V(G)$, $\lim_{N_k \to \infty} V(F) = V(G)$.

As shown in the Appendix B, $L(F)$ is equivalent to a constrained optimization formulation of the BarlowTwins loss. Furthermore, $L(F)$ with the additional constraint that $(f_i, f_i) \geq \gamma \; \forall i \in \{1, 2 \ldots N_k\}$ is the constrained optimization formulation of the VICReg loss.

3.2 The implicit bias of gradient descent

Next, we investigate how the use of gradient descent for optimizing $L(F)$ influences the characteristics of the learned feature space, $V(F)$. Given the similarity in its form with that of the BarlowTwins loss, we build on recent findings that demonstrate the sequential nature of learning eigenfunctions when optimizing the BarlowTwins loss under a strong orthogonalization regularization [36]. Since strong orthogonalization is seldom used in practice due to instabilities in training [5, 43], we believe studying the learning dynamics under weak orthogonalization regularization (i.e. low values of β in Equation (1)) is more relevant to provide recommendations for practitioners.

Theorem 3.2. *(Informal) Let us denote the span of the feature space at initialization as $V(F_0)$ and after training as $V(F_T)$. For small initialization of the network's weights, the alignment of $V(F_T)$ with the eigenfunctions of T_k depend on two factors: (i) alignment of $V(F_0)$ with the eigenfunctions of T_k; (ii) singular values of T_k.*

Under weak orthogonalization constraints, the network tends to learn features that are strongly aligned with eigenfunctions corresponding to large singular values. We refer to this property as the "selection" bias of gradient descent, wherein gradient descent selects certain eigenfunctions based on the corresponding singular values. This selection bias leads to redundancy among the learned feature space, thereby reducing the effective dimensionality of the network's output space compared to its ambient dimensionality. We will leverage this finding to improve the parameter overhead of good feature learning using BarlowTwins and VICReg loss frameworks.

3.3 Takeaway 1: Low-dimensional projectors can yield good representations

Given the proximity of the formulation of Equation (9) to that of BarlowTwins and VICReg losses, we will leverage existing heuristics that have been shown to work in practice. As such, BarlowTwins and VICReg frameworks call for high-dimensional projectors while using a weak orthogonalization regularization to facilitate good feature learning. We know, from Theorem 3.1, that the eventual goal of these frameworks is to learn the eigenfunctions of the underlying data similarity graph. For example, since the intrinsic dimensionality of Imagenet is estimated to be ~ 40 [32], it is not unreasonable to expect that the span of desired features would be of similar dimensionality. It is, thus, intriguing that the current practice would suggest using an ~ 8192-dim projector head to capture the intricacies of the corresponding augmentation-defined data similarity kernel. This discrepancy can be explained by analyzing the learning dynamics, as in Theorem 3.2. Notably, a high-dimensional projector is likelier to have a greater initialization span than its low-dimensional counterpart, thereby increasing the alignment between $V(F_0)$ and relevant eigenfunctions of T_k. We hypothesize that a stronger orthogonalization constraint for low-dimensional projectors can rectify this issue, reducing the redundancy in the network's output space and rendering it sufficient for good feature learning.

3.4 Takeaway 2: Multiple augmentations improve kernel approximation

By comparing the invariance criterion formulation in the standard BarlowTwins and VICReg losses to Equation (7), it can be inferred that current practices use a sample estimate of T_k. Using only two augmentations per sample yields a noisy estimate of T_k, yielding spurious eigenpairs [41] (see

https://doi.org/10.52202/079017-1258

Appendix C). These spurious eigenpairs add stochasticity in the learning dynamics, and coupled with Theorem 3.2, increase the redundancy in the learned feature space [11]. We hypothesize that improving this estimation error by increasing the number of augmentations could alleviate this issue and improve the speed and quality of feature learning.

Of course, increasing the number of augmentations (m) in the standard BarlowTwins and VICReg loss improves the estimate of T_k but comes with added compute costs – a straightforward approach would involve calculating the invariance loss for every pair of augmentations, resulting in $\mathcal{O}(m^2)$ operations. However, Theorem 3.1 proposes an alternative method that uses the sample estimate of T_M, thereby requiring only $\mathcal{O}(m)$ operations, and hence is computationally more efficient while yielding functionally equivalent features (see Appendix B). In summary, Theorem 3.1 establishes a mechanistic role for the number of data augmentations, paving the way for a computationally efficient multi-augmentation framework:

$$\widehat{L}(F) = \mathbb{E}_{x\sim\rho_X}\left[\sum_{i=1}^{N_k}\sum_{j=1}^{m}\|\overline{f_i(x)} - f_i(x_j)\|^2_{L^2(\rho_X)}\right], \quad \text{subject to} \quad (f_i, f_j)_{\rho_X} = \delta_{ij} \tag{10}$$

where $\overline{f_i(x)} = \frac{1}{m}\sum_{j=1}^{m} f_i(x_j)$ is the sample estimate of $T_M f_i(x)$.

4 Experiments

In our experiments, we seek to (i) provide empirical support for our theoretical insights and (ii) present practical primitives for designing efficient SSL routines. Since our proposed loss function is closest to the formulation of BarlowTwins/VICReg, we present empirical evidence comparing our proposal to these baselines. In summary, with extensive experiments across learning algorithms (BarlowTwins & VICReg) and training datasets (CIFAR-10, STL-10 & Imagenet-100), we establish the following:

- **low-dimensional projectors** can yield *good representations*.
- **multi-augmentation** improves downstream accuracy, as well as convergence rate.
- multi-augmentation **improves sample efficiency** in SSL pretraining, i.e., recovering similar performance with significantly fewer unique unlabelled samples.

Experiment Setup: We evaluate the effectiveness of different pretraining approaches using image classification as the downstream task. Across all experiments, we pretrain a Resnet feature encoder backbone for 100 epochs (see Appendix E.1 for longer pretraining results) and use linear probing on the learned representations[1]. All runs are averaged over 3 seeds; error bars indicate standard deviation. Other details related to optimizers, learning rate, etc., are presented in the Appendix D.

4.1 Low-dimensional projectors can yield good representations

pdim	**Barlow Twins**		**VICReg**	
	fixed β	optimal β^*	fixed β	optimal β^*
64	73.6 ± 0.9	82.1 ± 0.2	68.9 ± 0.2	81.9 ± 0.1
256	75.9 ± 0.7	$\mathbf{83.4 \pm 0.4}$	75.3 ± 0.2	81.9 ± 0.3
1024	81.3 ± 1.0	82.9 ± 0.3	79.2 ± 0.9	$\mathbf{82.5 \pm 0.9}$
8192	82.2 ± 0.4	82.2 ± 0.4	80.4 ± 1.5	80.4 ± 1.5

Table 1: Optimizing for orthogonality appropriately allows low-dimensional projectors to match the performance (on CIFAR-10) of much higher-dimensional projectors.

Existing works recommend using high-dimensional MLPs as projectors (e.g., d=8192 for Imagenet in [5, 43]), and show significant degradation in performance when using lower-dimensional projectors for a fixed redundancy coefficient (β). To reproduce this result, we run a grid search to find the optimal coefficient (β^*_{8192}) for $d = 8192$ and show that performance progressively degrades for lower d if the same coefficient β^*_{8192} is reused for $d \in \{64, 128, 256, 512, 1024, 2048, 4096, 8192\}$.

[1]Code: https://github.com/kumarkrishna/fastssl

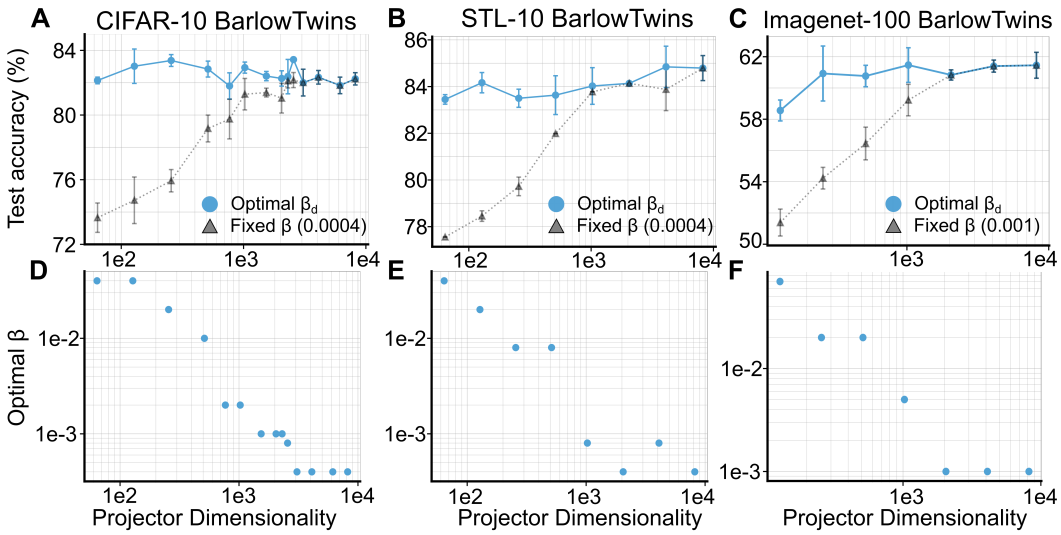

Figure 2: Low-dimensional projectors can yield good representations. We demonstrate that using a higher orthogonality constraint, β, for lower projector dimensionality can achieve similar performance over a wide range of projector dimensions (d).

Our insights in Section 3.3 suggest low-dimensional projectors should recover similar performance with appropriate orthogonalization. To test this, we find the best β by performing a grid search independently for each $d \in \{64, 128, 256, 512, 1024, 2048, 4096, 8192\}$. As illustrated in Figure 2, using low-dimensional projectors yield features with similar downstream task performance, compared to the features obtained using high-dimensional projectors. Strikingly, we also observe that the optimal $\beta_d \propto 1/d$, which aligns with our theoretical insights.

Recommendation: Start with low-dimensional projector, using $\beta = \mathcal{O}(\frac{1}{d})$, and sweep over $(pdim = d, \beta = \mathcal{O}\left(\frac{1}{d}\right))$ if needed.

4.2 Multiple Augmentations Improve Performance and Convergence

Although some SSL pretraining approaches, like SWaV [10], incorporate more than two views, the most widely used heuristic in non-contrastive SSL algorithms involves using two views jointly encoded by a shared backbone. In line with this observation, our baselines for examining the role of multiple augmentations use two views for computing the cross-correlation matrix.

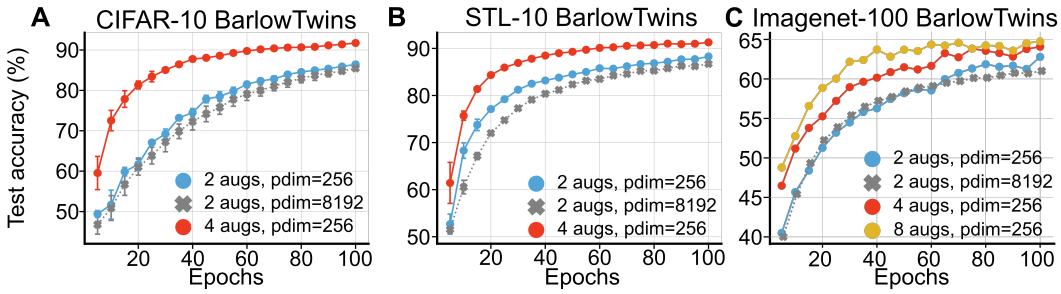

Figure 3: Using multiple augmentations improves representation learning performance and convergence. (A-C) Across BarlowTwins for CIFAR-10, STL-10 and Imagenet-100 pretraining, using 4 augmentations instead of 2 helps improve performance. Please see Appendix E.3 for more results.

To demonstrate the role of multiple augmentations in pretraining, we adapt the invariance criterion of BarlowTwins/VICReg to be in line with Equation (10). In particular, for $\#augs \in \{2, 4, 8\}$, we

https://doi.org/10.52202/079017-1258

augs	pdim	**BarlowTwins** Time (min)	**VICReg** Time (min)
2	8192	99.36 ± 0.01	94.36 ± 0.01
2	256	62.34 ± 0.06	51.73 ± 0.04
4	256	$\mathbf{43.09 \pm 0.20}$	$\mathbf{39.02 \pm 0.05}$

Table 2: Using multiple augmentations yields faster convergence, with reduced time to reach baseline performance on CIFAR-10, i.e. performance of feature encoder pretrained with an 8192-dim projector and 2 augmentations.

pretrain a Resnet-50 encoder with our proposed loss. Building on the insight from the previous section, we use a 256-dimensional projector head for all multi-augmentation experiments.

In Figure 3, we track the downstream performance of the pretrained models across training epochs. For performance evaluation, we use the linear evaluation protocol as outlined by [13]. Figure 3(A-C) shows that pretraining with multiple augmentations outperforms the 2-augmentation baseline. Furthermore, we observe that the four-augmentation pretrained models converge faster (both in terms of the number of epochs and wall-clock time) than their two-augmentation counterparts (see Figure 3(D-F)). Additionally, we show in Appendix E.2 that our framework can also be applied to multi-augmentation settings like SWaV, where not all augmentations are of the same resolution.

Recommendatation: Using multiple augmentations (> 2) is likely to improve convergence as well as downstream accuracy.

4.3 Sample Efficient Multi-augmentation Learning

Data Augmentation can be viewed as a form of data inflation, where the number of training samples is increased by k (for k augmentations). In this section, we examine the role of multi-augmentation in improving sample efficiency. In particular, we are interested in understanding if the same performance can be achieved with a fraction of the pretraining dataset, simply by using more augmentations.

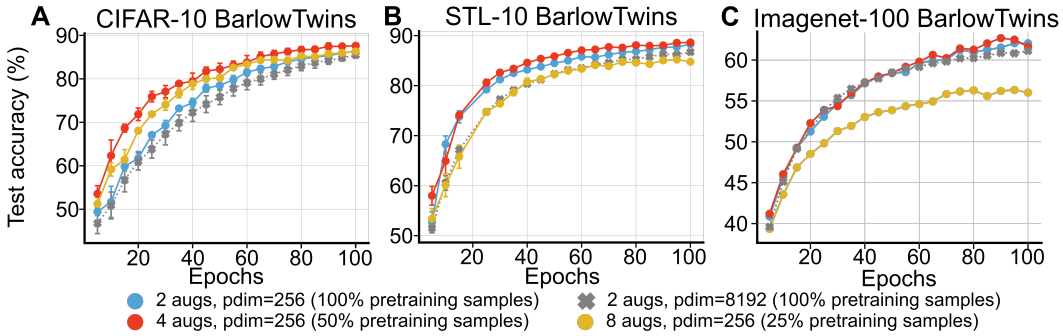

Figure 4: Multi-augmentation improves sample efficiency, recovering similar performance with significantly fewer unique samples in the pretraining dataset. Across BarlowTwins pretraining on CIFAR-10, STL-10 and Imagenet-100 for the same effective dataset size ($\#augs \times \#unique\_samples$), using more patches improves performance at the same epoch (A-C). However, a tradeoff exists wherein more data augmentations fail to improve performance in the scarce data regime.

augs	pdim	Percentage of Dataset	**BarlowTwins** Time (min)	**VICReg** Time (min)
2	8192	100 %	63.43 ± 0.02	66.05 ± 0.01
2	256	100 %	39.52 ± 0.04	40.64 ± 0.04
4	256	50 %	28.25 ± 0.01	$\mathbf{32.39 \pm 0.01}$
8	256	25 %	$\mathbf{27.74 \pm 0.01}$	34.76 ± 0.01

Table 3: Time required to pass 80% accuracy on CIFAR-10 when pretraining on fraction of the dataset, while using multiple augmentations. See Figure 5 for further discussion.

To examine the relation between the number of augmentations and sample efficiency, we fixed the effective size of the inflated dataset. This is achieved by varying the fraction of the unique samples in the pretraining dataset depending on the number of augmentations $k \in \{2, 4, 8\}$, e.g., we use 50% of the dataset for 4 views. We then evaluate the performance of the pretrained models on the downstream task, where the linear classifier is trained on the same set of labeled samples. Strikingly, Figure 4 shows that using multiple augmentations can achieve similar (sometimes even better) performance with lesser pretraining samples, thereby indicating that more data augmentations can be used for feature learning to compensate for smaller pretraining datasets.

Recommendation: In a low-data regime, using diverse & multiple augmentations can be as effective as acquiring more unique samples.

5 Related Work

Self-Supervised Pretraining requires significant compute resources and most practitioners rely on empirical heuristics (see SSL cookbook [3] for a summary). While recent advances in SSL theory explore learning dynamics in linear (or shallow) models [39, 40], with a focus on understanding dimensionality collapse [20, 24], the theoretical underpinnings of most of the heuristics considered essential for good feature learning, are missing.

Contrastive SSL has received more theoretical attention, owing to its connection with metric learning and noise contrastive estimation [4, 25, 29]. In particular, HaoChen *et al.* [22] provide a theoretical framework for the SimCLR loss from an augmentation graph perspective, which leads to practical recommendations. Subsequently, Garrido *et al.* [19] establish a duality between contrastive and non-contrastive learning objectives, further bridging the gap between theory and practice.

Non-contrastive SSL algorithms' theoretical foundations have received more attention recently [9, 44]. Prior works [2, 18, 19] have demonstrated that with modified learning objectives, low-dimensional projectors yield representations with good downstream performance. Similarly, previous works have demonstrated notable performance boosts when using a multi-patch framework in contrastive [17] and non-contrastive SSL [10, 42]. However, the theoretical basis for the benefits and trade-offs of either low-dimensional projectors or multiple augmentations is largely unclear. It is worth noting that Schaeffer *et al.* [34] present an information-theoretic perspective of the recently proposed non-contrastive SSL loss that leverages multiple augmentations, namely MMCR [42], but the computational advantages of using multiple augmentations on the learning dynamics is an active area of research.

Deep Learning theory has made significant strides in understanding the optimization landscape and dynamics of supervised learning [1]. In concurrent works [9, 44], the interplay between the inductive bias of data augmentations, architectures, and generalization has been explored from a purely theoretical perspective, establishing an equivalence among different SSL losses [44]. Furthermore, Simon *et al.* [36] used a more straightforward formulation of the BarlowTwins loss and investigated the learning dynamics in linearized models for the case when the invariance and orthogonalization losses have equal penalties. Although such a setting rarely used in practice, their approach serves as an inspiration for our work in studying the learning dynamics of non-contrastive SSL losses.

6 Discussion

Summary: Our work builds on existing theoretical results that establish an equivalence among different SSL frameworks, and proposes a compute-efficient reformulation of the common SSL loss. Using this loss reformulation and a study of the optimization dynamics, we proposed practical recommendations to improve the sample and compute efficiency of SSL algorithms. Specifically, we recommended low-dimensional projectors with increased orthogonality constraints and multi-augmentation frameworks, and we verified the effectiveness of these recommendations empirically. It is worth noting that our multi-augmentation formulation improves the efficiency of learning without altering the desiderata of SSL, i.e. the network learns the same feature space using our proposed multi-augmentation framework as with the original SSL formulation in the limit of infinite pretraining budget. To demonstrate this equivalence between the original SSL loss and our proposed version, we show in Appendix E.1 that longer pretraining on the 2-augmentation loss leads to similar downstream performance as the multi-augmentation versions (4 and 8 augmentations).

 https://doi.org/10.52202/079017-1258

We also showed that the multi-augmentation framework can be used to learn good features from fewer unique samples in the pretraining dataset simply by improving the estimation of the data augmentation kernel. This result has direct implications on improving the Pareto frontier of samples-vs-performance for SSL pretraining, wherein we can achieve better downstream performance when limited number of samples are available in the pretraining dataset.

Pareto Optimal SSL In the context of sample efficiency, training a model using two augmentations with different fractions of the dataset leads to a natural Pareto frontier, i.e., training on the full dataset achieves the best error but takes the most time (**Baseline (2-Aug)**). Our extensive experiments demonstrate that using more than two augmentations improves the overall Pareto frontier, i.e., achieves better convergence while maintaining accuracy (**Multi-Aug**). Strikingly, as shown in Figure 5, we observe that we can either use a larger pretraining dataset or more augmentations for a target error level. Therefore, the number of augmentations can be used as a knob to control the sample efficiency of the pretraining routine.

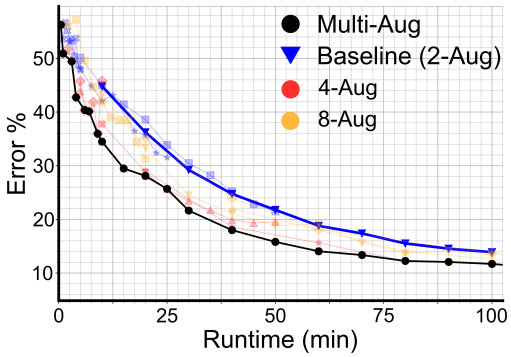

Figure 5: Using > 2 augmentations with a fraction of the dataset improves overall Pareto frontier, speeding runtime up to $\sim 2\times$.

Connections to Downstream performance: While our core theoretical results are aimed at accelerating convergence of the SSL loss itself, our empirical results highlight an improved downstream task performance earlier during pretraining. While this discrepancy might seem counter-intuitive at first, it is worth noting that the SSL loss inherent influences downstream performance as it encourages clustering of semantically similar images in the representation space. Such clustering properties in the representation space facilitates easier classification through methods k-nearest neighbors or linear decoding for a large number of tasks that rely on the semantic content of images. Previous works [2, 18, 20, 37] have discussed in detail how certain geometric properties of the learned representation space are connected to the linear classification performance for arbitrary decision boundaries, in expectation. However, an in-depth analysis of downstream tasks that are more amenable to linear decoding from the learned SSL representation space requires framing metrics of alignment between the pretraining objective (SSL desiderate) and the downstream task labels, and is an active area of research.

Open Questions: Looking ahead, it would be exciting to extend this analysis to other categories of SSL algorithms, such as Masked AutoEncoders (MAE). Furthermore, our insights provide opportunities to explore sample-efficient methods that rely on less data, which is particularly important in critical domains such as medical imaging, where data is often relatively scarce and expensive. On a different note, it is intriguing that animals often spend extended periods of time exploring novel objects, likely to gain multiple views of the object [6, 28]. Given the theoretical underpinnings of the computational benefits of multi-augmentation SSL outlined in our work, it would be exciting to develop models of biological learning that leverage these insights and enable sample-efficient continual learning in similar environments.

Limitations: Our algorithm relies on multiple augmentations of the same image to improve the estimation of the data-augmentation kernel. Though this approach speeds up the learning process, it also adds some extra computational overhead, which means that the impact of faster learning on wall-clock time is less than might be hoped for. One way to mitigate the effects of this limitation would be to scale up to a multi-GPU setting, since the computations for each augmentation can be run on a separate GPU in parallel. This could help ensure that the improved speed of learning directly translates to a significantly reduced wall-clock time for training.

Impact Statement: The goal of our work is to advance the general field of visual representation learning. Although there are potential downstream societal consequences of our work, we feel there are no direct consequences that must be specifically highlighted here.

https://doi.org/10.52202/079017-1258

Acknowledgements

The authors would like to thank Arnab Kumar Mondal for insightful discussions that helped shape the project's scope, Colleen Gillon for aesthetic contribution to the figures and Florian Bordes for helping setup the FFCV-SSL library. The authors are also grateful to Aidan Sirbu, Chen Sun, Jonathan Cornford, Roman Pogodin, Zahraa Chorghay and the anonymous reviewers whose comments and suggestions have significantly enhanced the quality and presentation of the results. This research was generously supported by Vanier Canada Graduate Scholarship (A.G.); NSERC (Discovery Grant: RGPIN-2020-05105; RGPIN-2018-04821; Discovery Accelerator Supplement: RGPAS-2020-00031; Arthur B. McDonald Fellowship: 566355-2022), Healthy Brains, Healthy Lives (New Investigator Award: 2b-NISU-8), and CIFAR Learning in Machines & Brains Program (B.A.R.); Canada CIFAR AI Chair program (A.O. & B.A.R.). The authors also acknowledge the material support of NVIDIA in the form of computational resources, as well as the compute resources, software and technical help provided by Mila (mila.quebec).

References

[1] Madhu S Advani, Andrew M Saxe, and Haim Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020.

[2] Kumar K Agrawal, Arnab Kumar Mondal, Arna Ghosh, and Blake Richards. α-req: Assessing representation quality in self-supervised learning by measuring eigenspectrum decay. *Advances in Neural Information Processing Systems*, 35:17626–17638, 2022.

[3] Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, et al. A cookbook of self-supervised learning. *arXiv preprint arXiv:2304.12210*, 2023.

[4] Randall Balestriero and Yann LeCun. Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods. In *NeurIPS*, 2022.

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

[6] Daniel E Berlyne. Novelty and curiosity as determinants of exploratory behaviour. *British journal of psychology*, 41(1):68, 1950.

[7] Florian Bordes, Randall Balestriero, and Pascal Vincent. Towards democratizing joint-embedding self-supervised learning, 2023.

[8] Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[9] Vivien Cabannes, Bobak Kiani, Randall Balestriero, Yann LeCun, and Alberto Bietti. The ssl interplay: Augmentations, inductive bias, and generalization. In *International Conference on Machine Learning*, pages 3252–3298. PMLR, 2023.

[10] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 132–149, 2018.

[11] Feng Chen, Daniel Kunin, Atsushi Yamamura, and Surya Ganguli. Stochastic collapse: How gradient noise attracts sgd dynamics towards simpler subnetworks. *arXiv preprint arXiv:2306.04251*, 2023.

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

[13] Yubei Chen, Adrien Bardes, Zengyi Li, and Yann LeCun. Intra-instance vicreg: Bag of self-supervised image patch embedding. *arXiv preprint arXiv:2206.08954*, 2022.

 https://doi.org/10.52202/079017-1258

[14] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

[15] Zhijie Deng, Jiaxin Shi, Hao Zhang, Peng Cui, Cewu Lu, and Jun Zhu. Neural eigenfunctions are structured representation learners. *arXiv preprint arXiv:2210.12637*, 2022.

[16] Zhijie Deng, Jiaxin Shi, and Jun Zhu. Neuralef: Deconstructing kernels by deep neural networks. In *International Conference on Machine Learning*, pages 4976–4992. PMLR, 2022.

[17] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9588–9597, 2021.

[18] Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann Lecun. Rankme: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *International Conference on Machine Learning*, pages 10929–10974. PMLR, 2023.

[19] Quentin Garrido, Yubei Chen, Adrien Bardes, Laurent Najman, and Yann LeCun. On the duality between contrastive and non-contrastive self-supervised learning. In *The Eleventh International Conference on Learning Representations*, 2023.

[20] Arna Ghosh, Arnab Kumar Mondal, Kumar Krishna Agrawal, and Blake Richards. Investigating power laws in deep representation learning. *arXiv preprint arXiv:2202.05808*, 2022.

[21] I Gohberg. Classes of linear operator theory. *Advances and Applications*, 49, 1990.

[22] Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34:5000–5011, 2021.

[23] Lajos Horváth and Piotr Kokoszka. *Inference for functional data with applications*, volume 200. Springer Science & Business Media, 2012.

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

[25] Daniel D. Johnson, Ayoub El Hanchi, and Chris J. Maddison. Contrastive learning can find an optimal basis for approximately view-invariant functions. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[26] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *Technical Report*, 2009.

[27] Guillaume Leclerc, Andrew Ilyas, Logan Engstrom, Sung Min Park, Hadi Salman, and Aleksander Madry. ffcv. https://github.com/libffcv/ffcv/, 2022. commit 3a12966.

[28] Marianne Leger, Anne Quiedeville, Valentine Bouet, Benoît Haelewyn, Michel Boulouard, Pascale Schumann-Bard, and Thomas Freret. Object recognition test in mice. *Nature protocols*, 8(12):2531–2537, 2013.

[29] Yazhe Li, Roman Pogodin, Danica J Sutherland, and Arthur Gretton. Self-supervised learning with kernel dependence maximization. *Advances in Neural Information Processing Systems*, 34:15543–15556, 2021.

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

[31] David Pfau, Stig Petersen, Ashish Agarwal, David GT Barrett, and Kimberly L Stachenfeld. Spectral inference networks: Unifying deep and spectral learning. *arXiv preprint arXiv:1806.02215*, 2018.

[32] Phil Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2021.

[33] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[34] Rylan Schaeffer, Victor Lecomte, Dhruv Bhandarkar Pai, Andres Carranza, Berivan Isik, Alyssa Unell, Mikail Khona, Thomas Yerxa, Yann LeCun, SueYeon Chung, et al. Towards an improved understanding and utilization of maximum manifold capacity representations. *arXiv preprint arXiv:2406.09366*, 2024.

[35] Kendrick Shen, Robbie M Jones, Ananya Kumar, Sang Michael Xie, Jeff Z HaoChen, Tengyu Ma, and Percy Liang. Connect, not collapse: Explaining contrastive learning for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 19847–19878. PMLR, 2022.

[36] James B Simon, Maksis Knutins, Liu Ziyin, Daniel Geisz, Abraham J Fetterman, and Joshua Albrecht. On the stepwise nature of self-supervised learning. *arXiv preprint arXiv:2303.15438*, 2023.

[37] Vimal Thilak, Chen Huang, Omid Saremi, Laurent Dinh, Hanlin Goh, Preetum Nakkiran, Joshua M Susskind, and Etai Littwin. Lidar: Sensing linear probing performance in joint embedding ssl architectures. In *The Twelfth International Conference on Learning Representations*, 2024.

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

[39] Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *International Conference on Machine Learning*, pages 10268–10278. PMLR, 2021.

[40] Yuandong Tian, Lantao Yu, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning with dual deep networks. *arXiv preprint arXiv:2010.00578*, 2020.

[41] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[42] Thomas Yerxa, Yilun Kuang, Eero Simoncelli, and SueYeon Chung. Learning efficient coding of natural images with maximum manifold capacity representations. *Advances in Neural Information Processing Systems*, 36:24103–24128, 2023.

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

[44] Runtian Zhai, Bingbin Liu, Andrej Risteski, Zico Kolter, and Pradeep Ravikumar. Understanding augmentation-based self-supervised representation learning via rkhs approximation. *arXiv preprint arXiv:2306.00788*, 2023.

 https://doi.org/10.52202/079017-1258

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: Our main claims are backed by theoretical and empirical results. Theorem 3.1 presents our functionally-equivalent compute-efficient formulation of the SSL objective, and Theorem 3.2 demonstrates the implicit bias of gradient descent during optimizing the SSL loss. Our empirical results demonstrate the utility of our theoretical insights in improving the parameter overhead of good feature learning, optimization convergence and the sample efficiency.

2. **Limitations**

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

   Answer: [Yes]

   Justification: We have a section on limitations in the discussion.

3. **Theory Assumptions and Proofs**

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

   Answer: [Yes]

   Justification: The formal statements alongside proofs are presented in the supplementary material (Appendices A to C).

4. **Experimental Result Reproducibility**

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

   Answer: [Yes]

   Justification: We provide implementation details in the supplementary material (Appendix D). We have also released our code base in the public github repo, FastSSL, to facilitate the implementation of our proposed framework.

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

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

   Answer: [Yes] .

   Justification: We use open-access datasets, like CIFAR, STL and Imagenet. Our code base can be found in the public github repo, FastSSL

   Guidelines:

6. **Experimental Setting/Details**

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

   Answer: [Yes]

   Justification: We present details of the experiment setup and results in Section 4 of the main paper, and additional implementation details in Appendix D.

7. **Experiment Statistical Significance**

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

   Answer: [Yes]

Justification: We report standard error bars, computed over 3 seeds, for all result plots and tables.

8. **Experiments Compute Resources**

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

   Answer: [Yes]

   Justification: All our CIFAR and STL experiments were done on a single 48-GB RTX8000 GPU and all Imagenet experiments were performed on 2 40-GB A100 GPUs. All experiments were performed on the Mila cluster, aided by compute resources, software and technical help provided by Mila (mila.quebec).

9. **Code Of Ethics**

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

   Answer: [Yes]

   Justification: We adhere to the NeurIPS Code of Ethics in our work, and research in general.

10. **Broader Impacts**

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

    Answer: [Yes]

    Justification: We add a statement on societal impact in the Discussion section. Since the goal of our work is to advance the general field of visual representation learning, we feel there are no direct consequences that must be specifically highlighted here. Although we recognize that there might be potential downstream consequences that warrant attention while building intelligent systems that leverage this work.

11. **Safeguards**

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

    Answer: [NA]

    Justification: We do not release any new data or state-of-the-art models.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: We acknowledge and cite the datasets and model architectures used in this work. Our codebase is publicly available on our github repo, FastSSL. Moreover, our codebase relies on the Python packages of PyTorch, FFCV [27] and FFCV-SSL [7], which are referred to in the github repo README.

13. **New Assets**

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

    Answer: [NA]

    Justification: No new assets are released.

14. **Crowdsourcing and Research with Human Subjects**

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

    Answer: [NA]

https://doi.org/10.52202/079017-1258

Justification: Our work does not involve crowdsourcing nor research with human subjects.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

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

    Answer: [NA]

    Justification: Our work does not involve research with human subjects.

A Hilbert Space of functions

A.1 Functions and inner product space

Definition A.1. Given X, ρ_X, and $f, g : X \to \mathbb{R}$, define the $L^2(\rho_X)$ inner product and norm, respectively,

$$(f,g)_{\rho_X} = \int f(x)g(x)d\rho_X(x), \quad \|f\|^2_{\rho_X} = (f,f)_{\rho_X} \tag{11}$$

Define

$$L^2(\rho, X) = \left\{ f : X \to \mathbb{R} \mid \|f\|^2_{\rho_X} < \infty \right\}$$

to be the (equivalence class) of functions with finite ρ_X norm.

A.2 Spectral theory

In this section we quote the relevant (abstract) Hilbert Space theory.

Definition A.2 (Spectral Operator)**.** Given orthogonal functions, $\Phi = (\phi_i)_{i\in I}$ in $L^2(\rho_X)$, and non-negative $\Lambda = (\lambda_i)_{i\in I}$, with $\|\Lambda\|_2^2 = \sum_{i\in I} \lambda_i^2 < \infty$. Call (Φ, Λ) a spectral pair and define the corresponding spectral operator by

$$T_{\Phi,\Lambda}(h) = \sum_{j=1}^{\infty} \lambda_j \left(h, \phi_j\right) \phi_j, \tag{12}$$

Theorem A.3 (Spectral Decomposition)**.** *Suppose H is a Hilbert space. A symmetric positive-definite Hilbert-Schmidt operator $T : \mathbb{H} \to \mathbb{H}$ admits the spectral decomposition equation 12 with orthonormal ϕ_j which are the eigenfunctions of T, i.e. $T(\phi_j) = \lambda_j \phi_j$. The ϕ_j can be extended to a basis by adding a complete orthonormal system in the orthogonal complement of the subspace spanned by the original ϕ_j.*

Remark A.4. The ϕ_j in equation 12 can thus be assumed to form a basis, but some λ_j may be zero.

We defer the reader to [21, 23] for an in-depth discussion and proof of Theorem A.3.

Denote by $\mathcal{L}$ the space of bounded (continuous) linear operators on $\mathbb{H}$ with the norm

$$\|T\|_{\mathcal{L}} = \sup\{\|T(x)\| \mid \|x\| \leq 1\}.$$

Definition A.5 (Compact Operators)**.** An operator $T \in \mathcal{L}$ is said to be compact if there exist two orthonormal bases $\{g_j\}$ and $\{f_j\}$, and a real sequence $\{\lambda_j\}$ converging to zero, such that

$$T(h) = \sum_{j=1}^{\infty} \lambda_j (h, g_j) f_j, \quad h \in \mathbb{H}, \tag{Compact}$$

The λ_j may be assumed positive. The existence of representation equation Compact is equivalent to the condition: T maps every bounded set into a compact set. Compact operators are also called completely continuous operators. Representation equation Compact is called the singular value decomposition.

Definition A.6 (Hilbert-Schmidt Operators)**.** A compact operator admitting representation equation Compact is said to be a Hilbert-Schmidt operator if $\sum_{j=1}^{\infty} \lambda_j^2 < \infty$. The space $\mathcal{S}$ of Hilbert-Schmidt operators is a separable Hilbert space with the scalar product

$$\langle T_1, T_2 \rangle_{\mathcal{S}} = \sum_{i=1}^{\infty} \left(T_1\left(f_i\right), T_2\left(f_i\right)\right), \tag{13}$$

where $\{f_i\}$ is an arbitrary orthonormal basis. Note the value of equation 13 is independent of the basis. The corresponding norm is

$$\|T\|^2_{\mathcal{S}} = \sum_{j\geq 1} \lambda_j^2 \tag{HS}$$

One can show that

$$\|T\|_{\mathcal{L}} \leq \|T\|_{\mathcal{S}}$$

https://doi.org/10.52202/079017-1258

Definition A.7. An operator $T \in \mathcal{L}$ is said to be symmetric if

$$\langle T(f), g\rangle = \langle f, T(g)\rangle, \quad f, g \in \mathbb{H},$$

and positive-definite if

$$\langle T(f), f\rangle \geq 0, \quad f \in \mathbb{H}.$$

(An operator with the last property is sometimes called positive semidefinite, and the term positive-definite is used when the inequality is strict.)

B Data augmentation kernel perspective of non-contrastive SSL

Theorem B.1. *Let $G(x)$ be the infinite Mercer features of the backward data augmentation covariance kernels, k^{DAB}. Let $F(x) = (f_1(x), f_2(x), \ldots, f_k(x))$ be the features given by minimizing the following data augmentation invariance loss*

$$L(F) = \sum_{i=1}^{N_k} \|T_M f_i - f_i\|^2_{L^2(\rho_X)}, \quad \textit{subject to} \quad (f_i, f_j)_{\rho_X} = \delta_{ij} \tag{14}$$

which includes the orthogonality constraint. Then, $V(F) \subset V(G)$, $V(F) \to V(G)$ as $N_k \to \infty$.

The idea of the proof uses the fact that, as linear operators, $T_{k^{DAB}} = T_M^\top T_M$ and that $T_{k^{DAF}} = T_M T_M^\top$. Then we use spectral theory of compact operators, which is analogue of the Singular Value Decomposition in Hilbert Space, to show that eigenfunctions of $T_M^\top T_M$ operator are the same as those obtained from optimizing $L(F)$. A similar result can be obtained using k^{DAF} and $T_M^\top$.

Note that $L(F)$ is the constrained optimization formulation of the BarlowTwins loss. Furthermore, $L(F)$ with the additional constraint that $(f_i, f_i) \geq \gamma \; \forall i \in \{1, 2 \ldots N_k\}$ is the constrained optimization formulation of the VICReg loss.

B.1 Proof of theorem 3.1

We show we can factor the linear operator, leading to a practical algorithm. Here, we show that we can capture the backward data augmentation kernel with the forward data augmentation averaging operator

Lemma B.2. *Using the definitions above, and with k in equation 6 given by k^{DAB},*

$$T_k = T_M^\top T_M$$

Proof. First, define the non-negative definite bilinear form

$$B^{VAR}(f, g) = (T_M f, T_M g)_{\rho_X} \tag{15}$$

Given the backwards data augmentation covariance kernel, k^{DAB}, define

$$B^{DAB}(f, g) = (T_k f, g)_{\rho_X}$$

We claim, that

$$B^{VAR} = B^{DA,B} \tag{16}$$

This follows from the following calculation,

$$\begin{aligned}
B^{DA,B}(f, g) &= (T_k f, g)_{\rho_X} && (17)\\
&= \mathbb{E}_x[T_k f(x), g(x)] = \mathbb{E}_x \mathbb{E}_z[k_{DA,B}(z, x) f(z) g(x)] && (18)\\
&= \mathbb{E}_x \mathbb{E}_z \mathbb{E}_{x_0} \left[\frac{p(x_0 \mid x)}{\rho(x_0)} \frac{p(x_0 \mid z)}{\rho(x_0)} f(z) g(x) \right] && (19)\\
&= \mathbb{E}_{x_0} \left[\sum_x \left(\frac{\rho(x) p(x_0 \mid x)}{\rho(x_0)} g(x) \right) \sum_z \left(\frac{\rho(z) p(x_0 \mid z)}{\rho(x_0)} f(z) \right) \right] && (20)\\
&= \mathbb{E}_{x_0} \left[\sum_x (p(x \mid x_0) g(x)) \sum_z (p(z \mid x_0) f(z)) \right] \quad \text{[Using Bayes' rule]} && (21)\\
&= \mathbb{E}_{x_0} \left[T_M f(x_0) T_M g(x_0) \right] = (T_M f, T_M g)_{\rho_X} = B^{VAR}(f, g) && (22)
\end{aligned}$$

□

For implementations, it is more natural to consider *invariance* to data augmentations.

Theorem B.3 (equivalent eigenfunctions)**.** *Assume that T_M is a compact operator. Define the invariance bilinear form*

$$B^{INV}(f,g) = (T_M f - f, T_M g - g) \tag{23}$$

Then B^{INV}, B^{VAR} share the same set of eigenfunctions. Moreover, these are the same as the eigenfunctions of $B^{DA,B}$. In particular, for any eigenfunction f_j of B^{VAR}, with eigenvalue λ_j, then f_j is also and eigenfunction of B^{INV}, with the corresponding eigenvalue given by $(\sqrt{\lambda_j} - 1)^2$.

Proof. Define T_{MM} by,

$$T_{MM} f = T_M^\top T_M f \tag{24}$$

Define

$$T_{MS} = (T_M - I)^\top (T_M - I) \tag{25}$$

Note, by the assumption of compactness, T_M has the Singular Value Decomposition, (see the Hilbert Space section for equation SVD),

$$T_M(h) = \sum_{j=1}^{\infty} \lambda_j (h, g_j) f_j \tag{SVD}$$

Let f_j be any right eigenvector of T_M, with eigenvalue μ_j. Then f_j is also a right eigenvector $T_M - I$, with eigenvalue $\mu_j - 1$. So we see that T_{MM} has f_j as an eigenvector, with eigenvalue $\lambda_j = \mu_j^2$ and T_{MS} has f_j as an eigenvector, with eigenvalue $(\sqrt{\lambda_j} - 1)^2$. Finally, the fact that there are no other eigenfunctions also follows from equation SVD.

The final part follows from the previous lemma. □

Equivalence of Barlow Twins loss to Equation (9). The BarlowTwins loss from [43] is as follows:

$$\mathcal{L}_{BT} = \sum_i (C_{ii} - 1)^2 + \beta \sum_i \sum_{j \neq i} C_{ij}^2 \tag{26}$$

where C is the cross-correlation matrix computed between the outputs of the network to two different augmentations. First, the BarlowTwins loss can be seen as the unconstrained optimization form of the following constrained optimization objective:

$$\mathcal{L}_{BT} = \sum_i (C_{ii} - 1)^2 \quad \text{, subject to} \quad C_{ij} = 0 \quad \forall j \neq i \tag{27}$$

where β is the Lagrangian multiplier [8]. In [43], the cross-correlation matrix C is computed by a dot product between normalized functions f_i's such that $(f_i, f_i)_{\rho_X} = 1 \; \forall i$. The network output for one augmentation of x, a, can be thought of as a Monte-Carlo estimate (with one sample) of $T_M f_i(x)$, where f_i is the i^{th} dimension of the network's output. Therefore, the BarlowTwins loss can be written in its following equivalent form:

$$\hat{L}^{BT}(F) = \sum_{i=1}^{N_k} \left((T_M f_i, T_M f_i)_{\rho_X} - 1\right)^2 \quad \text{, subject to} \quad (f_i, f_j)_{\rho_X} = \delta_{ij} \tag{28}$$

As shown by [44], the eigenvalues of $T_M^T T_M$ are always less than 1. Therefore, we do not need the square in Equation (28). Rewriting it, we get the following:

$$\hat{L}^{BT}(F) = \sum_{i=1}^{N_k} (T_M f_i, T_M f_i)_{\rho_X} \quad \text{, subject to} \quad (f_i, f_j)_{\rho_X} = \delta_{ij} \tag{29}$$

Using Theorem B.3, we show that the loss recovers the equivalent eigenfunctions for the following reason. We can rewrite the loss as

$$\begin{aligned} \hat{L}^{BT}(F) &= \sum_{i=1}^{N_k} ((T_M - I) f_i, (T_M - I) f_i)_{\rho_X} \quad \text{, subject to} \quad (f_i, f_j)_{\rho_X} = \delta_{ij} \\ \implies \hat{L}^{BT}(F) &= \sum_{i=1}^{N_k} \|T_M f_i - f_i\|^2_{L^2(\rho_X)} \quad \text{, subject to} \quad (f_i, f_j)_{\rho_X} = \delta_{ij} \end{aligned} \tag{30}$$

https://doi.org/10.52202/079017-1258

which recovers the loss Equation (9). Note that the VICReg loss [5], in addition to the constraints imposed by the BarlowTwins loss, ensures that the norm of f_i's are more than some threshold. This can be easily incorporated into the constraint with a constant along with δ_{ij}. In conclusion, both BarlowTwins and VICReg losses can be seen as equivalent forms of the loss Equation (9).

Theorem B.4. *(Informal) Let us denote the span of the feature space at initialization as $V(F_0)$ and after training as $V(F_T)$. For small initialization of the network's weights, the alignment of $V(F_T)$ with the eigenfunctions of $\mathcal{T}$ depend on two factors: (i) alignment of $V(F_0)$ with the eigenfunctions of $\mathcal{T}$; (ii) singular values of $\mathcal{T}$.*

Theorem B.4. (Formal) *Let $\Gamma = V\Lambda V^T$ represent the eigendecomposition of Γ, and define z as the projection of the weight vectors in W onto singular vectors of Γ, V. Formally, $z = WV$. Assuming small initialization (as in Simon et al. (2023), i.e. $|z_{pi}(0)| << 1$ for all p, i, we can derive the following conclusions:*

1. $sign(\frac{\Delta z_{pi}(t)}{z_{pi}(t)}) = sign(\lambda_i)$
2. *For all $\lambda_i, \lambda_j > 0$, $\frac{z_{pi}(t)}{z_{pi}(0)} = (\frac{z_{pj}(t)}{z_{pj}(0)})^{\frac{\lambda_i}{\lambda_j}}$ where λ_i denotes the i^{th} singular value, i.e. i^{th} element of diagonal matrix Λ.*

Proof. We will first show that the above holds for a linear network, i.e. the output of the network with weights $W \in \mathbf{R}^{m\times n}$ is WX for some input $X \in \mathbf{R}^{n\times b}$, where m is the output dimensionality, n is the input dimensionality and b is the batch size.
Let us first analytically compute the cross-correlation matrix C following [36].

$$C = WXX'^TW^T = W\mathcal{T}W^T$$
$$C_{pq} = \sum_{i,j} W_{pi}\mathcal{T}_{ij}W_{qj} \quad , \quad C_{pp} = \sum_{i,j} W_{pi}\mathcal{T}_{ij}W_{pj}$$

where X and X' are matrices $\in \mathbf{R}^{n\times b}$ containing two augmentations of a each image in a batch of images. Also, we have defined $\mathcal{T} = XX'^T$, i.e. the augmentation-defined data correlation matrix. Rewriting the BarlowTwins loss function from [43]:

$$\mathcal{L}_{BT} = \sum_i (C_{ii} - 1)^2 + \beta \sum_i \sum_{j\neq i} C_{ij}^2$$

To study the learning dynamics, we need to compute the gradient of $\mathcal{L}_{BT}$ w.r.t. the parameters W.

$$\frac{dW_{pq}}{dt} = -\eta \frac{\partial \mathcal{L}_{BT}}{\partial W_{pq}} = -2\eta \sum_i (C_{ii} - 1)\frac{\partial C_{ii}}{\partial W_{pq}} - 2\eta\beta \sum_i \sum_{j\neq i} C_{ij}\frac{\partial C_{ij}}{\partial W_{pq}} \tag{31}$$

Let us now analytically compute the derivatives of C_{ii} and C_{ij} w.r.t W_{pq} to simplify each of the terms in Equation (31).

$$\begin{aligned}
\frac{\partial C_{ii}}{\partial W_{pq}} &= \frac{\partial}{\partial W_{pq}} \sum_{j,k} W_{ij}\mathcal{T}_{jk}W_{ik} = \frac{\partial}{\partial W_{pq}} \sum_{j,k} W_{ij}\mathcal{T}_{jk}W_{ik}\delta_{pi} \\
&= \left(\sum_{j,k} \mathcal{T}_{jk}W_{pk}\delta_{jq} + \sum_{j,k} W_{pj}\mathcal{T}_{jk}\delta_{kq} \right) \delta_{pi} \\
&= \left(\sum_{k} \mathcal{T}_{qk}W_{pk} + \sum_{j} W_{pj}\mathcal{T}_{jq} \right) \delta_{pi} \\
&= 2\left[W\mathcal{T}\right]_{pq} \delta_{pi} \\
\implies \sum_i (C_{ii} - 1)\frac{\partial C_{ii}}{\partial W_{pq}} &= 2(C_{pp} - 1)\left[W\mathcal{T}\right]_{pq}
\end{aligned} \tag{32}$$

Using similar algebra steps, we can simplify the second term:

$$\begin{aligned}
\frac{\partial C_{ii}}{\partial W_{pq}} &= [W\mathcal{T}]_{jq}\,\delta_{pi} + [W\mathcal{T}]_{iq}\,\delta_{pj} \\
\implies \sum_i \sum_{j\neq i} C_{ij}\frac{\partial C_{ii}}{\partial W_{pq}} &= \sum_i \sum_{j\neq i} C_{ij}\left([W\mathcal{T}]_{jq}\,\delta_{pi} + [W\mathcal{T}]_{iq}\,\delta_{pj}\right) \\
&= \sum_{j\neq q} C_{pj}\,[W\mathcal{T}]_{jq} + \sum_{i\neq q} C_{ip}\,[W\mathcal{T}]_{iq} \\
&= 2\,[(C-I)W\mathcal{T}]_{pq} - 2(C_{pp}-1)\,[W\mathcal{T}]_{pq} \qquad (33)
\end{aligned}$$

Substituting Equations (32) and (33) into Equation (31), we get:

$$\begin{aligned}
\frac{dW_{pq}}{dt} &= -\eta\frac{\partial \mathcal{L}_{BT}}{\partial W_{pq}} = -4\eta(C_{pp}-1)\,[W\mathcal{T}]_{pq} - 4\eta\beta\,[(C-I)W\mathcal{T}]_{pq} + 4\eta\beta(C_{pp}-1)\,[W\mathcal{T}]_{pq} \\
&= -4\eta(1-\beta)(C_{pp}-1)\,[W\mathcal{T}]_{pq} - 4\eta\beta\,[(C-I)W\mathcal{T}]_{pq} \qquad (34)
\end{aligned}$$

Note that setting $\beta = 1$ yields the dynamics equation presented by [36]. However, in practice, β is orders of magnitude less that 1. For sake of simplicity, we will analyze the extreme case of $\beta = 0$, which will yield us insights into the weak-orthogonality constraint case. Therefore,

$$\frac{dW_{pq}}{dt} \approx -4\eta(C_{pp}-1)\,[W\mathcal{T}]_{pq} \qquad (35)$$

Let us denote the eigendecomposition of $\mathcal{T}$ be written as $\mathcal{T} = V\Lambda V^T$. Here, Λ is a diagonal matrix with singular values as the diagonal elements. Let us also denote the projection of the weight vectors onto the singular vectors of $\mathcal{T}$, i.e. V as z. So, $z = WV$.
Therefore, using these definitions, we can write the following:

$$\begin{aligned}
C_{pp} &= \left[W\mathcal{T}W^T\right]_{pp} = \left[Z\Lambda Z^T\right]_{pp} = \sum_i z_{pi}^2\lambda_i \\
W\mathcal{T} &= WV\Lambda V^T = Z\Lambda V^T
\end{aligned}$$

Now, writing the update equations Equation (35) in terms of z_{pi}:

$$\begin{aligned}
\frac{dz_{pi}}{dt} &= \sum_q \frac{dW_{pq}}{dt}V_{qi} \\
&= -4\eta\left(\sum_j z_{pj}^2\lambda_j - 1\right)\sum_k z_{pk}\lambda_k(\sum_q V_{qk}V_{qi}) \\
&= -4\eta\left(\sum_j z_{pj}^2\lambda_j - 1\right)z_{pi}\lambda_i \qquad (36)
\end{aligned}$$

Assuming small initialization of weights W, we can assume that $\mid z_{pi}(0) \mid << 1$, i.e. magnitude z_{pi} at time 0 is very small.
Let us define $h_p(t) = 1 - \sum_j z_{pj}(t)^2\lambda_j$. For small initialization, $h_p(t) > 0\ \forall t$. Therefore,

$$sign\left(\frac{dz_{pi}(t)}{dt}\frac{1}{z_{pi}}\right) = sign(\lambda_i) \qquad (37)$$

It is clear from Equation (37) that if $\lambda_i < 0$, $\lim_{t\to\infty} z_{pi}(t) = 0$. Similarly, if $\lambda_i = 0$, then $z_{pi}(t) = z_{pi}(0)\ \forall t$.
Therefore, akin to the conclusions of [36], the BarlowTwins loss recovers directions corresponding to positive singular values in the augmentation-defined covariance matrix, $\mathcal{T}$ and suppresses directions corresponding to negative singular values. Thus, the network outputs span the top singular vectors of $\mathcal{T}$.

https://doi.org/10.52202/079017-1258

It is worth noting from Equation (36) that the following holds:

$$\begin{aligned} \frac{1}{\lambda_i}\frac{dlog(z_{pi})}{dt} &= \frac{1}{\lambda_j}\frac{dlog(z_{pj})}{dt} \\ \implies \frac{1}{\lambda_i} log\left(\frac{z_{pi}(t)}{z_{pi}(0)}\right) &= \frac{1}{\lambda_j} log\left(\frac{z_{pj}(t)}{z_{pj}(0)}\right) \\ \implies \frac{z_{pi}(t)}{z_{pi}(0)} &= \left(\frac{z_{pj}(t)}{z_{pj}(0)}\right)^{\frac{\lambda_i}{\lambda_j}} \end{aligned} \tag{38}$$

□

Without loss of generality, if $\lambda_i << \lambda_j$, then $z_{pi}(t) \approx z_{pi}(0)$. Therefore, under small initialization, i.e. $z_{pi}(0)$ is small $\forall i$, gradient descent biases the p^{th} weight vector to be more strongly aligned to the eigenvector corresponding to the strongest eigenvalue, for all p's. Hence, under weak orthogonalization constraints, the BarlowTwins loss will over "represent" the strong singular vectors of the augmentation-defined cross-correlation matrix.
When using high-dimensional projectors, specifically when $m >> \sum_i \mathbf{1}_{\lambda_i > 0}$, wherein $\mathbf{1}_\zeta$ is the indicator function that is 1 when condition ζ is true and 0 otherwise, this problem might be ameliorated because there are multiple weight vectors that might be aligned with the top singular vectors of $\mathcal{T}$ at initialization. However, when using low-dimensional projectors, we do not have such a luxury and therefore, using a weak orthogonalization constraint leads to dimensionality collapse in the representation space.

Extending to deep non-linear networks. Similar to the analysis in [36], we can repeat the above analysis by replace X and X' by the corresponding kernel versions, where the kernel corresponds to the Neural Tangent Kernel (NTK) of the network. Therefore, the implicit bias of gradient descent to yield dimensionality collapse in the representation space when using weak orthogonalization constraints still remains.

Dimensionality collapse under noisy optimization. From the rest of this section, we have seen that the BarlowTwins loss is a Monte-Carlo estimate of the true data-augmentation defined covariance matrix. Moreover, stochastic gradient descent adds noise due to mini-batch sampling to the optimization process. Note that there exist symmetries in our linear network, i.e. an orthogonal rotation of the weight matrix yields the same loss function. As explained in [11], such symmetry-invariant sets are potential candidates for stochastic collapse when performing noisy gradient-based optimization. Therefore, the presence of noise in the data-augmentation covariance matrix, $\mathcal{T}$, as well as the batch noise can further worsen the dimensionality collapse problem where different weight vectors become parallel to each other due to noise in updates. One possible mitigation strategy is to obtain a better estimate of the true augmentation-defined covariance matrix (see Figure 7), which we discuss in the next section.

Empirical validation. We empirically validate our results on the learning dynamics on simplistic 2-dimensional settings. These results, demonstrating the difference in feature learning dynamics for weak vs strong orthogonalization, are presented as GIFs in the supplementary material, and can also be viewed at the project website.

C Multi-Augmentation Learning

C.1 Augmentation graph

We use the population augmentation graph formulation introduced in [22]. Briefly, we define a graph $\mathcal{G}(\mathcal{X}, \mathcal{W})$, where the vertex set $\mathcal{X}$ comprises of all augmentations from the dataset (could be infinite when continuous augmentation functions are used) and $\mathcal{W}$ denotes the adjacency matrix with edge weights as defined below:

$$w_{xx'} := \mathbb{E}_{\bar{x} \sim \mathcal{P}_{\bar{X}}} \left[\mathcal{A}(x|\bar{x})\mathcal{A}(x'|\bar{x})\right] \tag{39}$$

, i.e. the joint probability of generating 'patches' x, x' from the same image $\bar{x}$. Here $\mathcal{A}$ defines the set of augmentation functions used in the SSL pipeline. It is worth noting that the magnitude of $w_{xx'}$ captures the relative similarity between x and x'. A higher value of $w_{xx'}$ indicates that it is more

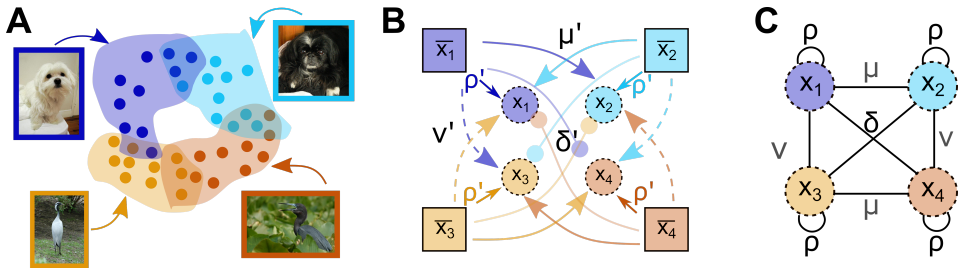

Figure 6: Schematic of augmentation graph. (A) Augmentations from each image span a region in the image space which could overlap with the augmentation span of other images. (B) An augmentation graph schematic that uses probabilities to characterize the interactions among augmentation spans of different instances.

likely that both patches came from the same image, and thereby are more similar. The marginal likelihood of each patch x can also be derived from this formulation:

$$w_x = \mathbb{E}_{x' \sim \mathcal{X}} \left[w_{xx'} \right] \tag{40}$$

C.2 Contrastive and non-contrastive losses suffer from the same issues

We will now show that the proposal of using multiple patches for the $\mathcal{L}_{invariance}$ is pertinent to both the contrastive and non-contrastive SSL. Following [22], we use the spectral contrastive loss formulation and incorporate the augmentation graph relations:

$$\begin{aligned} \mathcal{L}_c &= -\mathbb{E}_{x,x^+} \left[f(x)^T f(x^+) \right] + \beta \mathbb{E}_{x,x'} \left[\left(f(x)^T f(x') \right)^2 \right] \\ \mathcal{L}_c &\propto \| Z Z^T - D^{-\frac{1}{2}} \mathcal{W} D^{-\frac{1}{2}} \|_F^2 = \| Z Z^T - \bar{\mathcal{W}} \|_F^2 \end{aligned} \tag{41}$$

where $z := \sqrt{w_x} f(x)$, D is a $N \times N$ diagonal matrix with entries $\{w_x\}$ and $\bar{\mathcal{W}} = D^{-\frac{1}{2}} \mathcal{W} D^{-\frac{1}{2}}$.

We extend the duality results between contrastive and non-contrastive SSL loss, established by [19], to demonstrate how Equation (41) can be decomposed into the invariance and collapse-preventing loss terms.

$$\| Z Z^T - \bar{\mathcal{W}} \|_F^2 = \| Z^T Z - I_d \|_F^2 + 2 Tr \left[Z^T (I_N - \bar{\mathcal{W}}) Z \right] + \kappa \tag{42}$$

$$= \| Z^T Z - I_d \|_F^2 + 2 \sum_i \sum_x (1 - \bar{w}_x) z_i^2 - 2 \sum_i \sum_{x,x'} \bar{w}_{xx'} z_i z_i' + \kappa \tag{43}$$

where κ is some constant independent of Z. The first term in Equation (42) is the covariance regularization term in non-contrastive losses like BarlowTwins (implicit) or VIC-Reg (explicit), and the second term in Equation (43) is the variance regularization. Simplifying the third term in Equation (43) gives us:

$$\begin{aligned} \sum_i \sum_{x,x'} \bar{w}_{xx'} z_i z_i' &= \sum_i \sum_{x,x'} w_{xx'} f(x)_i f(x')_i = \sum_i \sum_{x,x'} \mathbb{E}_{\bar{x} \sim \mathcal{P}_{\bar{X}}} \left[\mathcal{A}(x|\bar{x}) \mathcal{A}(x'|\bar{x}) f(x)_i f(x')_i \right] \\ &= \sum_i \mathbb{E}_{\bar{x} \sim \mathcal{P}_{\bar{X}}} \left[\sum_x \mathcal{A}(x|\bar{x}) (f(x)_i \overline{f(x)}_i - f(x)_i^2) \right] \\ &= \mathbb{E}_{\bar{x} \sim \mathcal{P}_{\bar{X}}} \left[\sum_x \mathcal{A}(x|\bar{x}) \left(f(x)^T \overline{f(x)} - \| f(x) \|^2 \right) \right] \end{aligned} \tag{44}$$

This term encourages $f(x)$ to be similar to $\overline{f(x)}$, i.e. the mean representation across all augmentations of $\bar{x}$, thereby requiring to "sufficiently" sample $A(.|\bar{x})$. Given that both the contrastive and non-contrastive losses rely on learning invariance properties from data augmentations, we believe that our multi-patch proposal would improve the probability density estimation of $A(.|\bar{x})$ and yield better performance with few training epochs.

https://doi.org/10.52202/079017-1258

C.3 Explaining training dynamics in low patch sampling regime

We now turn to a simple form of the augmentation graph to understand how using low number of augmentations affects the evolution of ZZ^T. Minimizing Equation (41) implies that the spectral decomposition of Z would align with the top eigenvectors (and values) of $\overline{\mathcal{W}}$. We will demonstrate that in the low sampling regime (using few augmentations), the eigenvectors of the sampled augmentation graph $\tilde{\mathcal{W}}$ *may not* align with those of $\overline{\mathcal{W}}$.

Augmentation graph setup. We define an augmentation graph with only two instances from two different classes, similar to the one presented in [35]. Let us denote the four instances as $\bar{x}_i$ for $i \in 1, 2, 3, 4$, where $\bar{x}_1, \bar{x}_2$ belong to class 1 (i.e. $y_1, y_2 = 1$) and $\bar{x}_3, \bar{x}_4$ belong to class 2 (i.e. $y_3, y_4 = 4$). Let us further assume that $\bar{x}_1, \bar{x}_3$ have the highest pixel-level similarity among $(\bar{x}_1, \bar{x}_i) \forall i \in 2, 3, 4$, thereby making it more likely to have similar patches. We denote this relationship among input examples using $\mathcal{G}$ to indicate (pixel-wise) global similarity groups. So, $\mathcal{G}_1, \mathcal{G}_3 = 1$ and $\mathcal{G}_2, \mathcal{G}_4 = 2$. We can use the following probabilistic formulation to model our augmentation functions (see Figure 6B):

$$A(x_j|\bar{x}_i) = \begin{cases} \rho' & \text{if } j = i \\ \mu' & \text{if } j \neq i \text{ and } y_j = y_i \text{ and } \mathcal{G}_j \neq \mathcal{G}_i \\ \nu' & \text{if } j \neq i \text{ and } y_j \neq y_i \text{ and } \mathcal{G}_j = \mathcal{G}_i \\ \delta' & \text{if } j \neq i \text{ and } y_j \neq y_i \text{ and } \mathcal{G}_j \neq \mathcal{G}_i \end{cases} \tag{45}$$

In our setting, $\rho' + \mu' + \nu' + \delta' = 1$. The adjacency matrix of our augmentation graph (as shown in Figure 6C) is as follows:

$$\overline{\mathcal{W}} = \begin{bmatrix} \rho & \mu & \nu & \delta \\ \mu & \rho & \delta & \nu \\ \nu & \delta & \rho & \mu \\ \delta & \nu & \mu & \rho \end{bmatrix} \tag{46}$$

We defer the relations between $\rho', \mu', \nu'\delta'$ and ρ, μ, ν, δ to the appendix. The eigenvalues of this matrix are: ($\rho + \mu + \nu + \delta$, $\rho + \mu - \nu - \delta$,$\rho - \mu + \nu - \delta$, $\rho - \mu - \nu + \delta$). Corresponding eigenvectors are along $[1, 1, 1, 1]^T$, $[1, 1, -1, -1]^T$. $[1, -1, 1, -1]^T$, $[1, -1, -1, 1]^T$. Assuming that the augmentation functions induce semantically-relevant invariance properties that are relevant for identifying y_i from $f(x_i)$, we can say that $\rho' > max\{\mu', \nu'\}$ and $min\{\nu', \mu'\} > \delta'$. When we have sufficiently sampled the augmentations, any SSL loss will learn Z such that its singular values are span the top eigenvectors of the augmentation graph, and the eigenspectrum of ZZ^T would simply be the above eigenvalues. In practical settings, the augmentation graph would have significantly higher dimension that the feature/embedding dimension [2]. Therefore, singular vectors of Z would span the top eigenvectors of $\overline{\mathcal{W}}$ and the smaller eigenmodes are not learned. When we have accurately sampled the augmentation graph, $\mu > \nu$ and therefore, the class-information preserving information is preferred over pixel-level preserving information during learning. But what happens when we *do not sufficiently sample the augmentation space?*

Ansatz. Based on our empirical experience, we define *an ansatz* pertaining to the eigenvalues of a sampled augmentation graph and validate it in tractable toy settings, such as the one described above. Specifically, we claim that when the augmentation space is not sufficiently sampled, $\{|\mu - \nu|, \delta\} \to 0$. In other words, we claim that when only few augmentations per example are used, it is more likely to have an equal empirical likelihood for augmentations that preserve (pixel-level) global information and class/context information. Moreover, it is very unlikely to have augmentations that change both the class and global information. This is demonstrated in Figure 7.

Consequences of the *Ansatz*. When only a few augmentations are sampled, learning can suppress the class information at the cost of preserving the pixel-level information, thereby leading to an increased smoothness in the learned feature space.

[2] Contrastive algorithms use a large batch size, thereby optimizing a high-dimensional ZZ^T whereas non-contrastive algorithms use a large embedding dimension, thereby optimizing a high-dimensional Z^TZ.

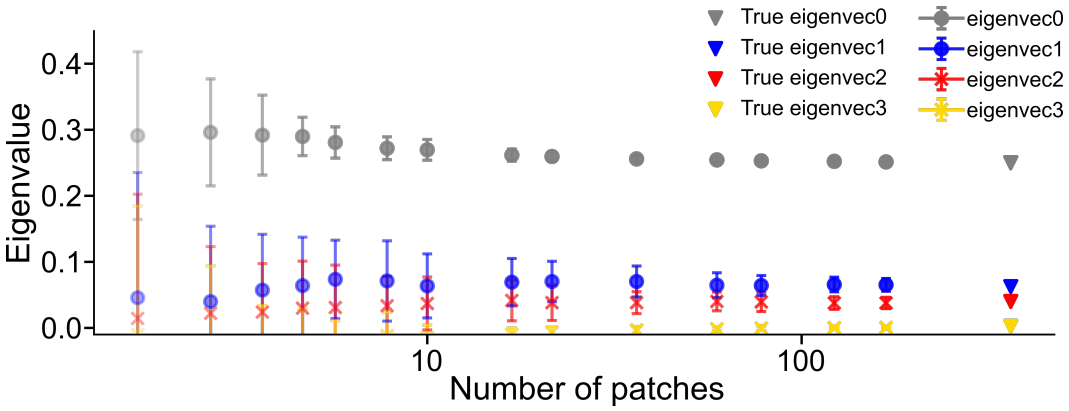

Figure 7: Empirical verification of the subsampling Ansatz.

D Implementation Details

Image Classification Datasets Across all experiments, our settings mainly follow [13]. In particular, Table 4a summarizes our pretraining settings on Cifar-10 [26], STL-10 [14] and Imagenet-100 [33]. The Imagenet-100 dataset was generated by sampling 100 classes from the original Imagenet-1k dataset, according to this list [38]. In Table 4b, we outline the corresponding linear evaluation settings for Resnet-50 (for CIFAR-10 and STL-10) and ResNet-18 (for Imagenet). Note that we add a linear classifier layer to the encoder's features and discard the projection layers for evaluation. Our code base is publicly available on github.

config	value
optimizer	Adam
learning rate	1e-3
batch size	128 (Imagnet), 256 (CIFAR, STL)
epochs	100
weight-decay	1e-6

(a) Pretraining

config	value
optimizer	Adam
learning rate	1e-3
batch size	512
epochs	200
weight-decay	1e-6
test-patches	16

(b) Linear Evaluation

Table 4: Experiment Protocol for comparing SSL algorithms

The key SSL loss functions that we use in this work are BarlowTwins [43] and VICReg [5]. Let us suppose that the embeddings of two augmentations of a batch of images are denoted as z and z'. The BarlowTwins loss function is as follows:

$$\mathcal{L}_{BT} = \sum_i (C_{ii} - 1)^2 + \beta \sum_i \sum_{j \neq i} C_{ij}^2 \tag{47}$$

$$\text{where} \quad C = \frac{1}{n-1} \sum_{k=1}^{n} (z_k - \bar{z})(z'_k - \bar{z'})^T$$

$$\text{and} \quad \bar{z} = \frac{1}{n} \sum_{k=1}^{n} z_k \quad , \quad \bar{z'} = \frac{1}{n} \sum_{k=1}^{n} z'_k \tag{48}$$

C_{ij} is the element of C at row i, column j and n is the batch size. For each projector dimensionality, d, we search for the hyperparameter, β, that yields the best downstream task performance.

https://doi.org/10.52202/079017-1258

The VICReg loss function is as follows:

$$\mathcal{L}_{VIC} = \frac{1}{n}\mu \sum_{k=1}^{n} \|z_k - z'_k\|^2 + \frac{1}{2}\mu\left[v(Z) + v(Z')\right] + \frac{1}{2}\left[c(Z) + c(Z')\right] \tag{49}$$

$$\text{where} \quad v(Z) = \frac{1}{d}\sum_{i=1}^{d} max(0, 1 - Stdev(z_{:,i}))$$

$$\text{and} \quad c(Z) = \frac{1}{d}\sum_{i}\sum_{j\neq i}[C(Z)_{ij}]^2 \quad , \quad C(Z) = \frac{1}{n-1}\sum_{k=1}^{n}(z_k - \bar{z})(z_k - \bar{z})^T$$

For each projector dimensionality, d, we search for the hyperparameter, μ, that yields the best downstream task performance.

D.1 Empirical results for low-dimensional projectors

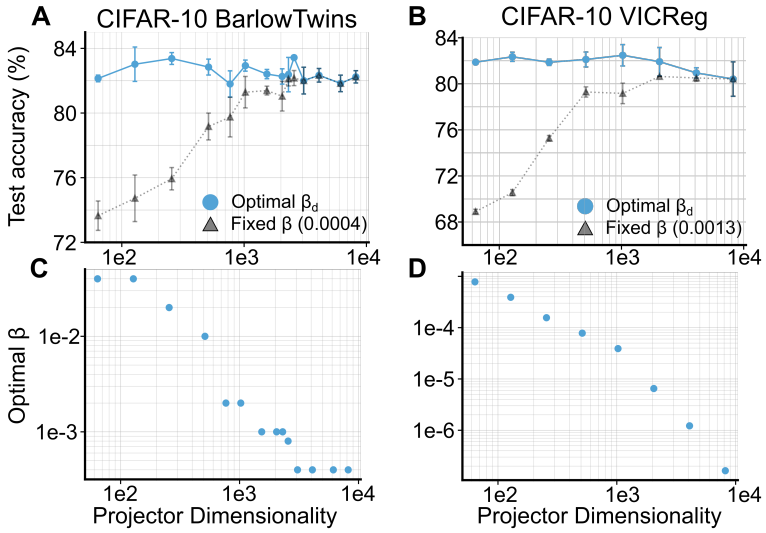

Figure 8: Low-dimensional projectors can yield good representations for both BarlowTwins and VICReg. We demonstrate that using a higher orthogonality constraint, β, for lower projector dimensionality can achieve similar performance over a wide range of projector dimensions (d). Note that for VICReg, we plot the ratio of the coefficient of the covariance loss to the coefficient of the invariance loss, i.e. $\beta = \frac{1}{d*\mu}$, where μ is the coefficient of the invariance loss. (See Equation (49) for details of the loss formulation.)

pdim	*Projector params (approx)*	**Barlow Twins**		**VICReg**	
		fixed β	optimal β^*	fixed β	optimal β^*
64	135k	73.6 ± 0.9	82.1 ± 0.2	68.9 ± 0.2	81.9 ± 0.1
128	278k	74.7 ± 1.4	83.0 ± 1.1	70.6 ± 0.3	82.3 ± 0.4
256	589k	75.9 ± 0.7	83.4 ± 0.4	75.3 ± 0.2	81.9 ± 0.3
512	1.3M	79.2 ± 0.8	82.8 ± 0.5	79.3 ± 0.4	82.1 ± 0.6
1024	3.1M	81.3 ± 1.0	82.9 ± 0.3	79.2 ± 0.9	82.5 ± 0.9
2048	8.3M	81.0 ± 0.9	82.3 ± 0.5	80.6 ± 0.0	81.9 ± 1.2
4096	25.2M	82.3 ± 0.4	82.3 ± 0.4	80.5 ± 0.3	81.0 ± 0.4
8192	83.9M	82.2 ± 0.4	82.2 ± 0.4	80.4 ± 1.5	80.4 ± 1.5

Table 5: Extended version of Table 1. Optimizing for orthogonality appropriately allows low-dimensional projectors to match the performance for BarlowTwins and VICReg (on CIFAR-10) of much higher-dimensional projectors.

D.2 Empirical results with multi-augmentations along with Time

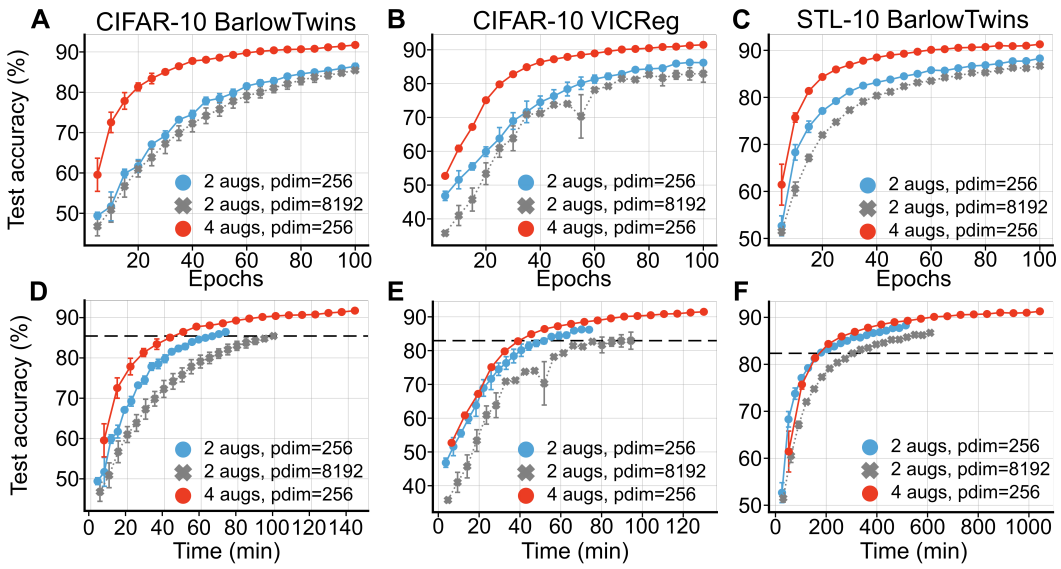

Figure 9: Using multiple augmentations improves representation learning performance and convergence. (A-C) Across BarlowTwins and VICReg for CIFAR-10 and STL-10 pretraining, using 4 augmentations instead of 2 helps improve performance. (D-F) Although the 4-augmentations take longer for each epoch, its performance still trumps the 2-augmentation version of the algorithm at the same wall clock time. Please see Appendix E.3 for more results.

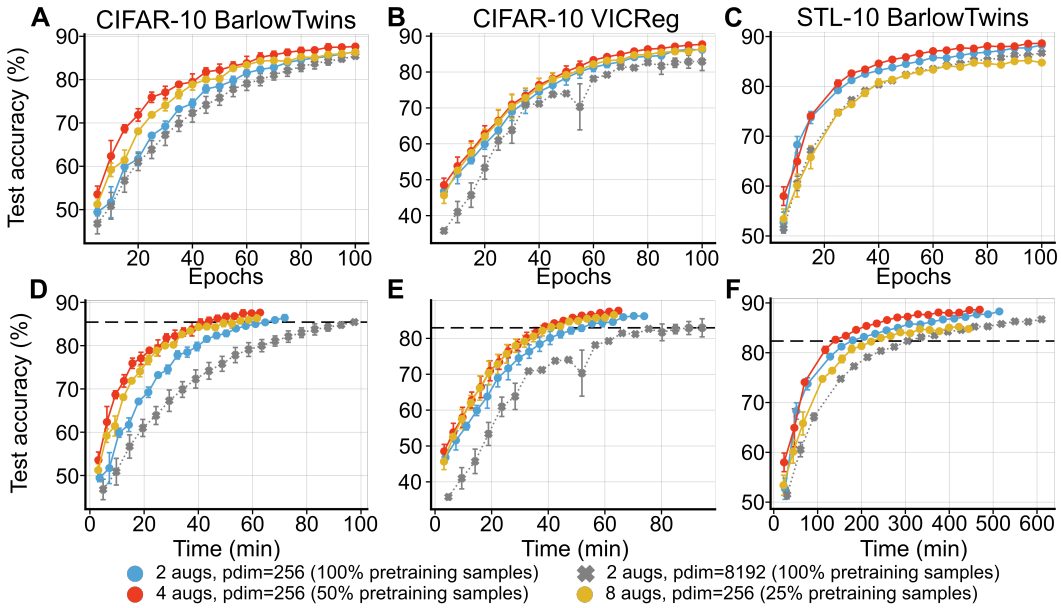

Figure 10: Multi-augmentation improves sample efficiency, recovering similar performance with significantly fewer unique samples in the pretraining dataset. Across BarlowTwins and VICReg pretraining on CIFAR-10 and STL-10, for the same effective dataset size ($\#augs \times \#unique\_samples$), using more patches improves performance at the same epoch (A-C) or wall clock time (D-F). However, a tradeoff exists wherein more data augmentations fail to improve performance in the scarce data regime.

https://doi.org/10.52202/079017-1258

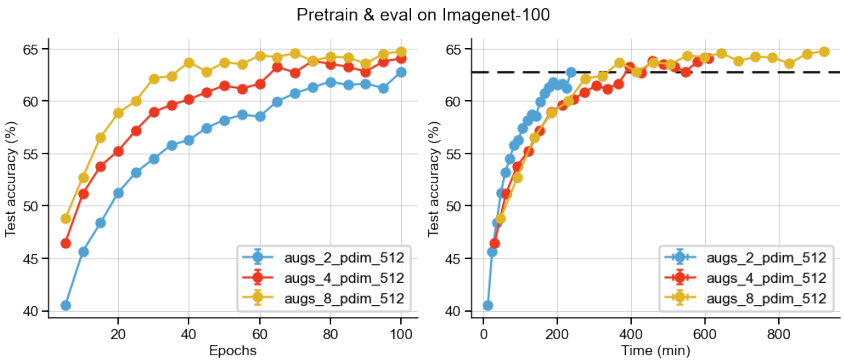

Figure 11: BarlowTwins pretraining on full Imagenet-100 dataset with 2, 4 and 8 augmentations.

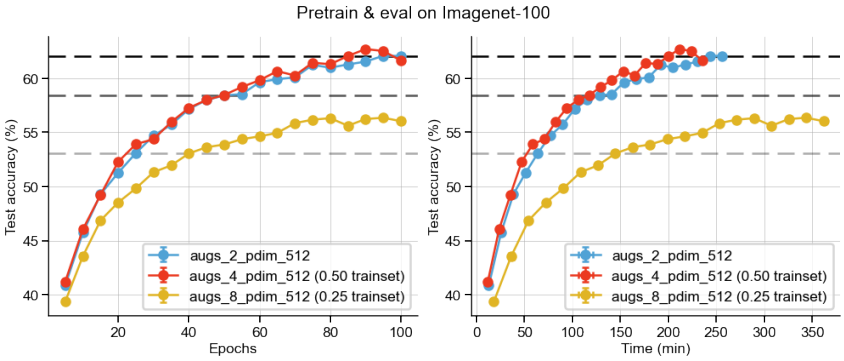

Figure 12: BarlowTwins pretraining on fraction of Imagenet-100 dataset with 2, 4 and 8 augmentations.

https://doi.org/10.52202/079017-1258

D.3 Empirical results on transfer learning

In this section, we present extended version of results presented in Figure 3, Figure 4 but pretraining on CIFAR-10 (or STL-10) and evaluating on STL-10 (or CIFAR-10). These results, coupled with the ones in Figure 3 Figure 4, present a strong case for the advantage of using the proposed multi-augmentation loss for better convergence as well as downstream accuracy.

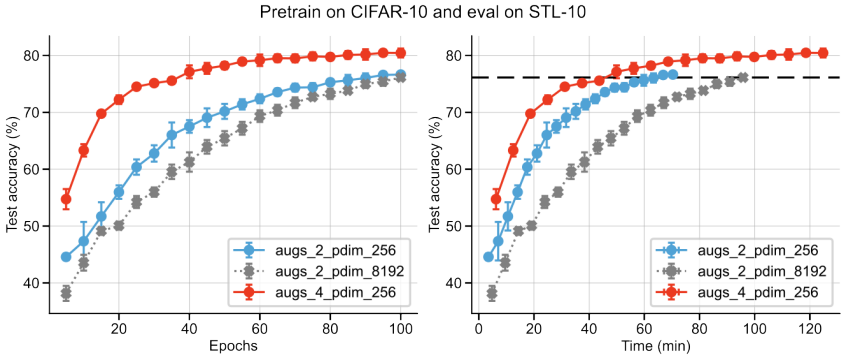

Figure 13: BarlowTwins pretraining on CIFAR-10, linear evaluation on STL-10 labelled set.

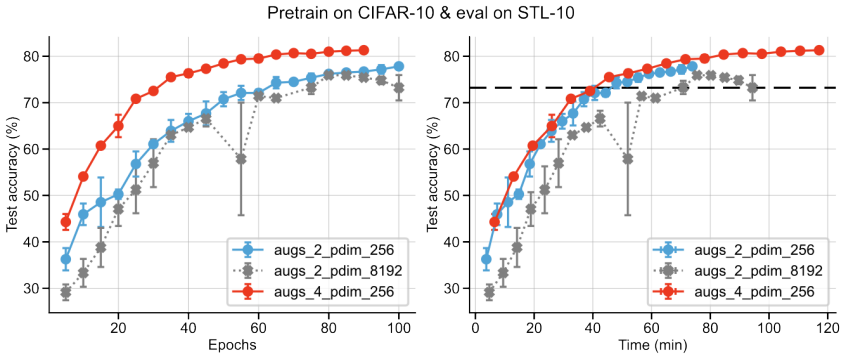

Figure 14: VICReg pretraining on CIFAR-10, linear evaluation on STL-10 labelled set.

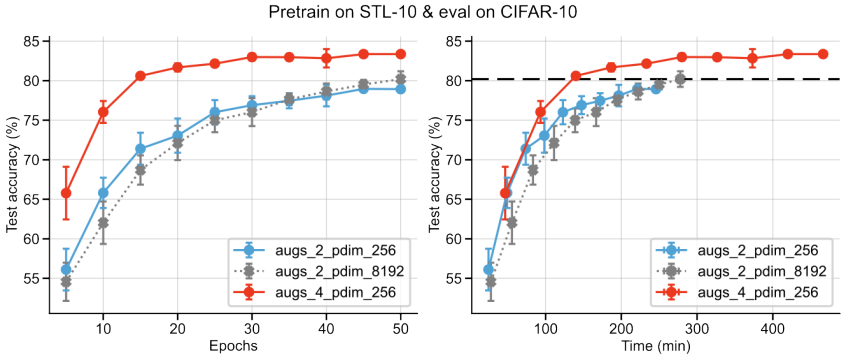

Figure 15: BarlowTwins pretraining on STL-10, linear evaluation on CIFAR-10 labelled set.

https://doi.org/10.52202/079017-1258

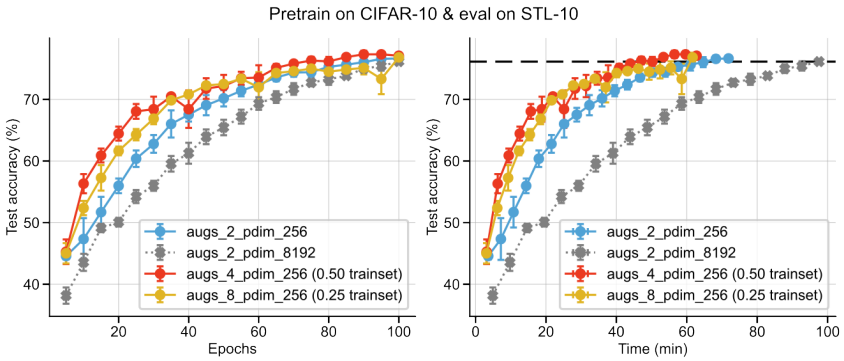

Figure 16: BarlowTwins pretraining on fraction of CIFAR-10 trainset, linear evaluation on STL-10 labelled set.

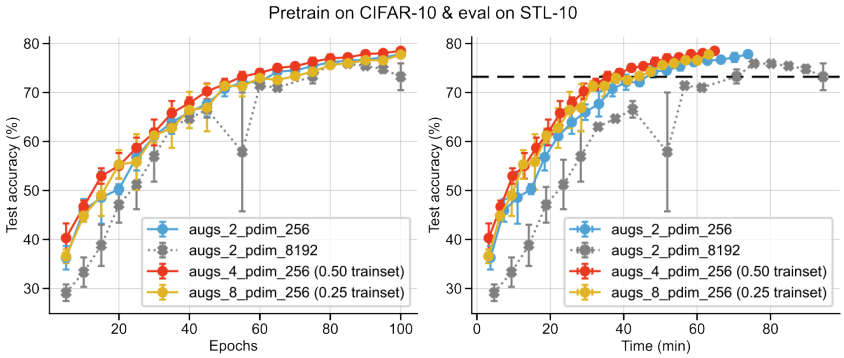

Figure 17: VICReg loss pretraining on fraction of CIFAR-10 trainset, linear evaluation on STL-10 labelled set.

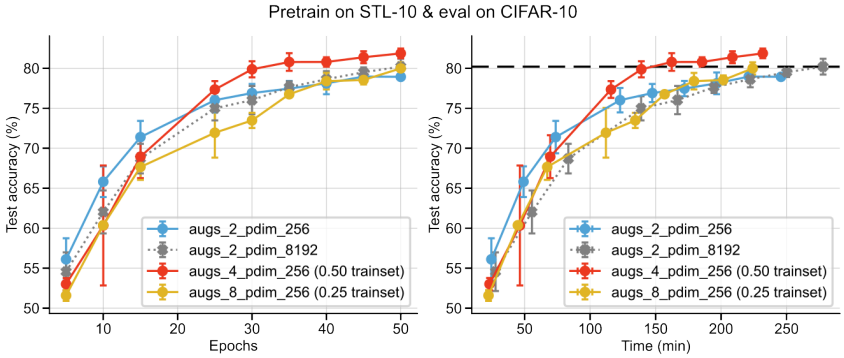

Figure 18: BarlowTwins loss pretraining on fraction of STL-10 unlabelled set, linear evaluation on CIFAR-10 train set.

E Additional Experiments probing multi-augmentation learning

E.1 Longer Pretraining to determine early stopping

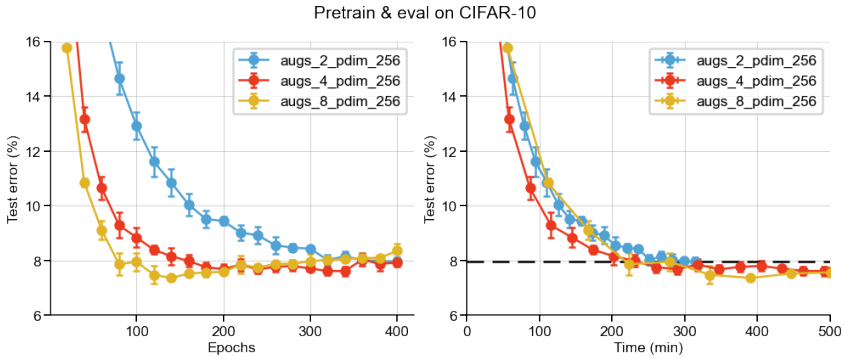

Figure 19: BarlowTwins pretraining on full CIFAR-10 dataset for 400 epochs.

Algorithm	Best accuracy	Best accuracy @ epoch
Barlow-Twins (2-augs) w/ pdim=256	92.04 +/- 0.16	400
Barlow-Twins (4-augs) w/ pdim=256	92.39 +/- 0.17	340
Barlow-Twins (8-augs) w/ pdim=256	92.64 +/- 0.10	140

Table 6: BarlowTwins pretraining on full CIFAR-10 dataset at 400 epochs (with early stopping)

E.2 SwAV-like augmentations for compute efficient multi-augmentation framework

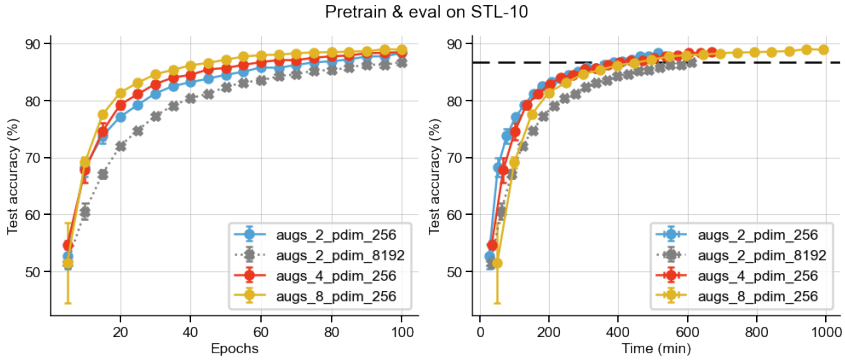

Figure 20: BarlowTwins pretraining on full STL-10 dataset for 100 epochs using SwAV-like augmentations. Specifically, the 2-augmentations setting uses two views that are 64×64, whereas the 4 (or 8) augmentation setting uses additional two (or six) augmentations that are 32×32.

https://doi.org/10.52202/079017-1258

E.3 Training with full dataset with 4/8 augmentations

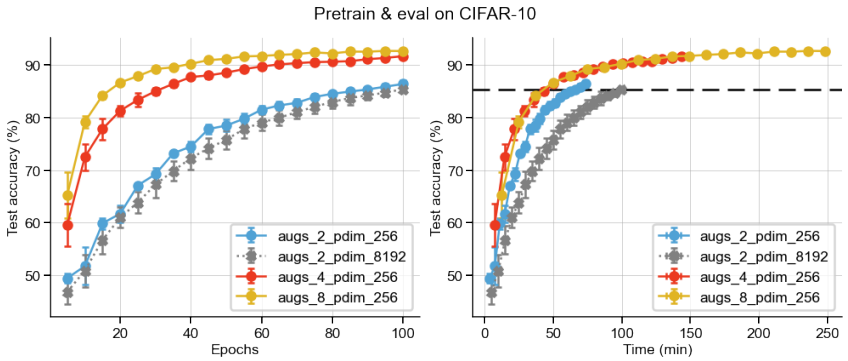

Figure 21: BarlowTwins pretraining on full CIFAR-10 dataset with 2, 4 and 8 augmentations.

Algorithm	**#augs=2**	**#augs=4**	**#augs=8**
Barlow-Twins w/ pdim=256	86.43 +/- 0.72	91.73 +/- 0.16	92.71 +/- 0.19
Barlow-Twins w/ pdim=8192	85.44 +/- 0.54	91.40 +/- 0.32	92.40 +/- 0.13

Table 7: BarlowTwins pretraining on full CIFAR-10 dataset at 100 epochs