Representation Learning via Non-Contrastive Mutual Information

Zhaohan Daniel Guo **Bernardo Avila Pires** **Khimya Khetarpal**

Dale Schuurmans **Bo Dai**

Google DeepMind

Abstract

Labeling data is often very time consuming and expensive, leaving us with a majority of unlabeled data. Self-supervised representation learning methods such as SimCLR (Chen et al., 2020) or BYOL (Grill et al., 2020) have been very successful at learning meaningful latent representations from unlabeled image data, resulting in much more general and transferable representations for downstream tasks. Broadly, self-supervised methods fall into two types: 1) Contrastive methods, such as SimCLR; and 2) Non-Contrastive methods, such as BYOL. Contrastive methods are generally trying to maximize mutual information between related data points, so they need to compare every data point to every other data point, resulting in high variance, and thus requiring large batch sizes to work well. Non-contrastive methods like BYOL have much lower variance as they do not need to make pairwise comparisons, but are much trickier to implement as they have the possibility of collapsing to a constant vector. In this paper, we aim to develop a self-supervised objective that combines the strength of both types. We start with a particular contrastive method called the Spectral Contrastive Loss (HaoChen et al., 2021; Lu et al., 2024), and we convert it into a more general non-contrastive form; this removes the pairwise comparisons resulting in lower variance, but keeps the mutual information formulation of the contrastive method preventing collapse. We call our new objective the Mutual Information Non-Contrastive (MINC) loss. We test MINC by learning image representations on ImageNet (similar to SimCLR and BYOL) and show that it consistently improves upon the Spectral Contrastive loss baseline.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

1 INTRODUCTION

In practice, it is often the case that we have a large dataset of unlabeled data and a smaller dataset of labeled data. This is because labeling data can be expensive due to time or cost, especially if we want high quality labels. In order to still take advantage of the large unlabeled dataset, self-supervised algorithms have been developed that first learn some kind of latent representation from the unlabeled data. The learned latent representation is then further used with the labeled data, either by training new layers on top, or fine-tuning. SimCLR (Chen et al., 2020) and BYOL (Grill et al., 2020) are examples of this approach for large image datasets, and they have been shown to learn great representations in ImageNet. In addition to not needing labels, another advantage of self-supervised methods is that they learn more general representations. Because they are trained using unlabeled data, they learn representations that can better capture intrinsic qualities of the data, which are then much more transferable (Chen et al., 2020); then with fine-tuning or just a couple of new layers, they can quickly adapt to a desired downstream task.

Most of the effective self-supervised algorithms can be broadly split into two categories: 1) contrastive; and 2) non-contrastive. Contrastive losses have the following general idea: Pull together the embeddings of similar images closer, and push apart the embeddings of images that are different. A strength of many of these losses, such as InfoNCE (Oord et al., 2018), SimCLR, Spectral Contrastive (HaoChen et al., 2021), f-MICL (Lu et al., 2024), is that they are theoretically based in maximizing the mutual information between related

data points. This gives a solid theoretical foundation for analyzing what kind of representation is learned; for example, the Spectral Contrastive loss learns a type of spectral decomposition of the data. However these contrastive losses have an inherent weakness; because they are concerned about comparing data points to one another, they require every data point to eventually be compared with every other data point, resulting in a quadratic dependence on the size of the dataset. This quadratic dependence means that the objective function has very high variance, which means large batch sizes are required to minimize these losses effectively.

Non-contrastive losses like BYOL and SimSiam do not have the quadratic dependence because they do not need to compare all data points to each other. Furthermore, they currently achieve state-of-the-art (when properly tuned), outperforming many contrastive algorithms. They only have a linear dependence, meaning much lower variance. However the tradeoff is that, theoretically and empirically, they are more brittle and prone to collapse, *i.e.*, a constant representation minimizes their losses. They require specific empirical tweaks such as additional layers of MLP, target networks, and are not great to use with weight decay for regularization, as it can increase the chance of collapse or partial collapse. Recent theoretical understanding has shown when they do not collapse but learn a meaningful representation, however the theoretical assumptions do not hold in practice, and so the issue of collapse still remains.

In this paper, we propose a new non-contrastive objective, the Mutual Information Non-Contrastive (MINC) loss, which retains the non-collapse guarantees enjoyed by contrastive losses, while removing the quadratic data dependence of contrastive methods. Instead of comparing every data point to every other data point, the MINC loss compares each data point to a summary of all other data points. We derive MINC starting with the Spectral Contrastive Loss, and reformulating it using the lens of power iteration in combination with insights from the Generalized Hebbian Algorithm (Sanger, 1989). The power iteration formulation also acts as a bridge to connect MINC with BYOL using a linear predictor.

To evaluate MINC, we follow the same setup as for SimCLR and BYOL and test in ImageNet. We first learn a representation from the unlabeled dataset. Then we train a linear classifier on top of the fixed representation to evaluate how effective the learned representation is for image classification. Our experiments show that MINC is consistently improving upon the Spectral Contrastive Loss baseline, and matches up to BYOL with a linear predictor. While it does not quite match up to the regular BYOL that uses a non-linear predictor, these results do show the promising direction of turning a contrastive loss into a non-contrastive loss to get the best of both worlds. Perhaps a better contrastive loss is needed as a foundation to connect to the non-linear BYOL.

2 PRELIMINARIES

We first introduce the generalization of InfoNCE-based contrastive learning (Lu et al., 2024) through f-divergences, which includes the SimCLR (Chen et al., 2020), InfoNCE (Oord et al., 2018) and Spectral Contrastive loss (HaoChen et al., 2021) as special cases. We further provide a brief introduction to BYOL (Grill et al., 2020), the representative non-contrastive representation learning method, which not only inspired our algorithm, but also whose connection to mutual information will be revealed during our algorithm derivation.

f-InfoNCE. Consider a pair of random variables (X, X') both taking values in $\mathcal{X}$ with joint density $p(x, x')$ and marginal densities $p(x)$ and $p(x')$, respectively. The f-mutual information between X and X' is defined as

$$I_f\left(X, X'\right) := \int f\left(\frac{p(x,x')}{p(x)p(x')}\right) p(x)p(x')dxdx', \quad (1)$$

where $f : \mathbb{R}_+ \to \mathbb{R}$ is a convex function with $f(1) = 0$. In fact, the f-mutual information (MI) is the f-divergence between $p(x, x')$ and $p(x)p(x')$, which intuitively characterizes the entangle part between X and X' only in joint distribution. Letting $f^* : \mathbb{R} \to \mathbb{R}$ be the Fenchel duality (convex conjugate) of f (Nguyen et al., 2010; Nachum and Dai, 2020), we have

$$f(u) = \max_{t\in\mathbb{R}} \left(u \cdot t - f^*(t)\right),$$

which reformulates the f-MI as

$$I_f\left(X, X'\right) = \max_{t:\mathcal{X}\times\mathcal{X}\to\mathbb{R}} \mathbb{E}_{p(x,x')}\left[t(x, x')\right] - \mathbb{E}_{p(x)p(x')}\left[f^*\left(t(x, x')\right)\right]. \quad (2)$$

Under this formulation, the representation learning consists of parametrizing t on the right-hand side using of an embedding function ϕ, and finding a ϕ that maximizes the expression. Since we experiment with image representations, for ease of explanation, we will assume that the data x are pixel images, and that ϕ is a ResNet neural network that encodes the image into a latent vector. Plugging the parametrization of $t\left(x, x'\right)$ with representation of $\phi\left(x\right)$ and $\phi\left(x'\right)$ into (2), with different choices of f but also recover several existing contrastive representation learning losses (Lu et al., 2024), therefore, revealing the mutual information essence in representation learning.

Specifically, with $f(u) = (u-1)^2$ and the corresponding $f^*(t) = \frac{t^2}{4} + t$ in (2), we consider $t(x, x') = 2(\phi(x)^\top \phi(x') - 1)$, leading to the representation learning objective (ignoring constants)

$$\max_{\phi} \; 2\mathbb{E}_{p(x,x')}\left[\phi(x)^\top \phi(x')\right] - \mathbb{E}_{p(x)p(x')}\left[\left(\phi(x)^\top \phi(x')\right)^2\right], \tag{3}$$

which is the Spectral Contrastive loss introduced by HaoChen et al. (2021). This choice of f corresponds to the commonly known χ^2-divergence. Similarly, SimCLR (Chen et al., 2020) and InfoNCE (Oord et al., 2018) can be related to this framework using a variational form of (2) with mutual information from KL-divergence through $f(u) = u \log u$, as discussed by Lu et al. (2024).

BYOL. A drawback of contrastive losses for representation learning is that they require careful treatment of how to compare every sample to every other sample (Wu et al., 2017), and large batch sizes (Chen et al., 2020; Tian et al., 2020). Non-contrastive representation learning methods like BYOL (Grill et al., 2020) aim to circumvent these issues. Specifically, BYOL is proposed to learn the representation ϕ through direct prediction of self-generated targets:

$$\min_{\phi,\Lambda} \; \mathbb{E}_{p(x,x')}\left[\left\|\Lambda\left(\phi\left(x\right)\right) - \mathrm{sg}\left(\phi\left(x'\right)\right)\right\|^2\right], \tag{4}$$

where Λ is the predictor, and $\mathrm{sg}\left(\phi\left(x'\right)\right)$ are the prediction targets (with a stop-gradient) used for representation learning. Often in practice, the prediction targets also come from a slow moving target network. Although BYOL achieves state-of-the-art performance, the objective has a quirk that a trivial $\phi(\cdot) = 0$, often referred to as a collapsed representation, is optimal. In practice, BYOL has been shown to avoids collapse and learn meaningful representations, and there have been several attempts to investigate why. Tang et al. (2023) characterized the optimal solution of the BYOL learning dynamics with semi-gradient optimization, while Richemond et al. (2023) revealed the orthonormalization effects in the updates through spectral view and Riemannian gradients. These theoretical approaches showed that BYOL is related to some kind of spectral decomposition of the data. However, these analyses make many assumptions that are not present in practice, in particular they both analyze the case of a linear predictor Λ as opposed to an MLP predictor used in practice. Thus, while BYOL does not have issues related to negative examples, in practice it still requires careful execution to prevent collapse.

3 MUTUAL INFORMATION THROUGH NON-CONTRASTIVE LOSS

As discussed, contrastive losses like the Spectral Contrastive loss are founded in the f-mutual information framework and so have a solid theoretical basis. They are not prone to representation collapse, but are held back by the quadratic dependence on the number of samples. In this section, we introduce the *Mutual Information Non-Contrastive (*MINC*)* loss, which retains the non-collapse guarantees enjoyed by f-MI contrastive losses, while, at the same time, avoiding the quadratic dependence by being *non-contrastive*.

3.1 First Attempt at Non-Contrastive

We start to convert the Spectral Contrastive loss (3) to a non-contrastive loss with rearranging:

$$\max_{\phi} \; 2\mathbb{E}_{p(x,x')}\left[\phi\left(x\right)^\top \phi\left(x'\right)\right] - \mathbb{E}_{p(x')}\left[\phi\left(x'\right)^\top \mathbb{E}_{p(x)}\left[\phi\left(x\right)\phi^\top\left(x\right)\right]\phi\left(x'\right)\right],$$

which has separated the quadratic dependence on the dataset (in the second term of (3)) into an inner and outer expectation. Then we can introduce an auxiliary variable with a constraint to completely decouple the inner expectation:

$$\max_{\phi} \; 2\mathbb{E}_{p(x,x')}\left[\phi\left(x\right)^\top \phi\left(x'\right)\right] - \mathbb{E}_{p(x')}\left[\phi\left(x'\right)^\top \Lambda \phi\left(x'\right)\right] \quad \text{subject to} \quad \Lambda = \mathbb{E}_{p(x)}\left[\phi\left(x\right)\phi^\top\left(x\right)\right], \tag{5}$$

and this becomes a non-contrastive objective, because it removes the quadratic dependence (the double-integral weighted by $p(x)p(x')$) through the auxiliary matrix Λ. Essentially, Λ has become a *summary* of the statistics of $\phi(x)$, allowing $\phi(x')$ to be compared with this summary matrix instead of every other point.

To tackle this constrained objective, we will first re-contextualize the Spectral Contrastive loss in an eigen-decomposition view. Then we will use the simple and efficient power iteration method for eigen-decomposition to help derive a new non-contrastive objective.

3.2 Spectral Contrastive as Eigen-Decomposition

Here we re-contextualize the Spectral Contrastive loss (3) through the eigen-decomposition perspective originally adopted by HaoChen et al. (2021); Ren et al. (2023, 2024). Given embeddings $\phi(x), \phi(x')$, and assuming a finite number of data point pairs N, let $F : \mathbb{R}^N \times \mathbb{R}^d$ be the matrix where we stack all the

scaled $\sqrt{p(x)}\phi(x)$ as rows (and similarly for F' and x'). Then we can formulate the following low-rank matrix decomposition problem:

$$\min_{\phi} \|M - FF'^{\top}\|_{\mathrm{F}}^2, \tag{6}$$

where $M_{ij} := \frac{p(x_i, x'_j)}{\sqrt{p(x_i)p(x'_j)}}$, is a quantity related to the mutual information of the dataset. The solution is then for F to share the eigen-space of M (more precisely, F will be the matrix of eigenvectors multiplied by the diagonal matrix of the square root of the eigenvalues). This means the solution satisfies (in pointwise notation):

$$\frac{p(x,x')}{\sqrt{p(x)p(x')}} = \left(\sqrt{p(x)}\phi\left(x\right)^{\top}\right)\left(\sqrt{p(x')}\phi\left(x'\right)\right), \tag{7}$$

Then in the more general case of (6) we have:

$$\begin{aligned}
&\min_{\phi} \int \left(\frac{p(x,x')}{\sqrt{p(x)p(x')}} - \sqrt{p(x)p(x')}\phi\left(x\right)^{\top}\phi\left(x'\right)\right)^2 dxdx' \\
&= \underbrace{\int \frac{p^2\left(x',x\right)}{p\left(x'\right)p\left(x\right)}dx'dx}_{\texttt{constant}} - 2\mathbb{E}_{p(x,x')}\left[\phi\left(x\right)^{\top}\phi\left(x'\right)\right] \\
&\quad + \mathbb{E}_{p(x)p(x')}\left[\left(\phi\left(x\right)^{\top}\phi\left(x'\right)\right)^2\right],
\end{aligned}$$

giving us back the Spectral Contrastive loss.

3.3 From Contrastive to Non-Contrastive Through Power Iteration

Given the eigen-decomposition view of the Spectral Contrastive loss in (7), we can exploit the rich numerical linear algebra literature for faster eigen-decomposition algorithms, among which power iteration (Golub and Van Loan, 2013; Trefethen and Bau, 2022) is a simple yet effective algorithm. It has been proved the convergence rate of subspace iteration is exponential, therefore motivating us to extend power iteration for representation learning.

Recall the properties of eigen-decomposition, first we have an orthogonality condition, which we can write as

$$\int \phi\left(x\right)\phi\left(x\right)^{\top}p(x)dx = \int \phi\left(x'\right)\phi\left(x'\right)^{\top}p(x')dx' = \Lambda, \tag{8}$$

where $\Lambda = \mathrm{diag}\left([\lambda_i]_{i=1}^d\right)$ as the diagonal matrix constructed by the eigenvalues. Note that we are reusing the symbol Λ that was for the auxiliary matrix in (5) because they will end up matching. Meanwhile, the eigen-function satisfies the following fixed point equation for all x':

$$\int \frac{p(x,x')}{\sqrt{p(x)p(x')}}\sqrt{p(x)}\phi\left(x\right)dx = \Lambda\sqrt{p(x')}\phi\left(x'\right). \tag{9}$$

Power iteration works by iterating a fixed point update through (9) while ensuring the orthogonality condition (8). Concretely we have the following update rules for each iteration t:

$$\Lambda_{t+1} = \mathbb{E}_{p(x)}\left[\phi_t\left(x\right)\phi_t\left(x\right)^{\top}\right], \tag{10a}$$

$$\Lambda_{t+1}\sqrt{p(x')}\psi_{t+1}\left(x'\right) = \int \frac{p(x,x')}{\sqrt{p(x')}}\phi_t\left(x\right)dx, \tag{10b}$$

$$\sqrt{p(x')}\phi_{t+1}\left(x'\right) = \mathbf{orth}\left(\sqrt{p(x')}\psi_{t+1}\left(x'\right)\right). \tag{10c}$$

Note how (10a) is actually the same constraint we introduce in (5); this means power iteration can give us an explicit and efficient approach that subsumes the constrained non-contrastive objective.

The exact integrals in the first two updates and the orthogonalization in the third update are intractable, however we will approximate them effectively to arrive at a practical power iteration procedure.

Moving Average for (10a)**.** Given ϕ_t, Λ_{t+1} is formulated as an expectation, therefore, it can be simply approximated by Monte Carlo estimation, *i.e.*,

$$\Lambda_{t+1} \approx \tfrac{1}{n}\textstyle\sum_{i=1}^n \phi_t(x_i)\phi_t(x_i)^{\top}, \quad \{x_i\}_{i=1}^n \sim p\left(\cdot\right).$$

However, this Monte Carlo ignores our iterative estimation procedure. We convert the estimation to an optimization, which leads to a moving average procedure, *i.e.*,

$$\Lambda_{t+1} = \operatorname*{argmin}_{\Lambda} \mathbb{E}_{p(x)}\left[\left\|\Lambda - \phi_t(x)\phi_t(x)^{\top}\right\|_2^2\right] + \eta\left\|\Lambda - \Lambda_t\right\|_{\mathrm{F}}^2, \tag{11}$$

inducing the exponential moving average (EMA) update

$$\Lambda_{t+1} = \beta\Lambda_t + (1-\beta)\mathbb{E}_{p(x)}\left[\phi_t(x)\phi_t(x)^{\top}\right] \tag{12}$$

with $\beta = \frac{\eta}{1+\eta}$.

Orthogonal Variational Iteration via GHA for (10b) **and** (10c)**.** With Λ_{t+1} being updated by an EMA, we can recast the iteration update for ψ as an optimization by matching LHS and RHS of (10) via Mahalanobis distance with Λ_{t+1}^{-1}, *i.e.*,

$$\begin{aligned}
&\min_{\psi} \\
&\int \left[\left\|\int \frac{p(x,x')}{\sqrt{p(x')}}\phi_t\left(x\right)dx - \Lambda_{t+1}\sqrt{p(x')}\psi\left(x'\right)\right\|_{\Lambda_{t+1}^{-1}}^2\right]dx' \\
&\propto \underbrace{-2\mathbb{E}_{p(x,x')}\left[\phi_t\left(x\right)^{\top}\psi\left(x'\right)\right] + \mathbb{E}_{p(x')}\left[\psi\left(x'\right)^{\top}\Lambda_{t+1}\psi\left(x'\right)\right]}_{\ell(\psi)}.
\end{aligned}$$

where its corresponding gradient update is

$$\frac{1}{2}\nabla_\psi \ell(\psi) = -\mathbb{E}_{p(x,x')}\left[\phi_t(x)^\top \nabla_\psi \psi(x')\right] + \mathbb{E}_{p(x')}\left[\psi(x')^\top \Lambda_{t+1}\nabla_\psi \psi(x')\right], \quad (13)$$

to approximate the update rule for ψ related to (10b).

To obtain ϕ_{t+1} via (10c) we need to orthogonalize ψ_{t+1}. One natural choice is Gram-Schmidt process, but is computationally expensive (or intractable). We consider to relax the strict orthogonal condition, only maintaining *asymptotic orthogonality*. This can be achieved through the generalized Hebbian algorithm (GHA; Sanger, 1989; Xie et al., 2015), which ensures the orthogonality in an asymptotic sense. By combining the GHA update rule and the gradient update of ψ from (13), we can have ψ_{t+1} asymptotically orthogonal by uptading it with suitably chosen step size κ_t as

$$\psi_{t+1} = \psi_t + \kappa_t\Big(\mathbb{E}_{p(x,x')}\left[\phi_t(x)^\top \nabla_\psi \psi_t(x')\right] - \mathbb{E}_{p(x')}\left[\psi_t(x')^\top \texttt{LT}\left[\Lambda_{t+1}\right]\nabla_\psi \psi_t(x')\right]\Big),$$

where $\texttt{LT}[\cdot]$ stands for making matrix lower triangular by setting all elements above the diagonal of its matrix argument to zero.

Because the GHA rule ensures asymptotic orthogonality, we skip the explicit orthogonalization in (10c), which gives $\phi_{t+1} = \psi_{t+1}$. Therefore, we can fuse (13) and (10c) into a direct update for ϕ:

$$\phi_{t+1} = \phi_t + \kappa_t\Big(\mathbb{E}_{p(x,x')}\left[\phi_t(x)^\top \nabla_\phi \phi_t(x')\right] - \mathbb{E}_{p(x')}\left[\phi_t(x')^\top \texttt{LT}\left[\Lambda_{t+1}\right]\nabla_\phi \phi_t(x')\right]\Big).$$

The combined updates for Λ and ϕ from (12) and (14), respectively, give us a tractable, non-contrastive power iteration method. This method is also an asymptotic non-contrastive approximation of gradient descent on the Spectral Contrastive loss. To see this, note that the gradient of the Spectral Contrastive loss (3) with respect to ϕ is,

$$-2\Big(\mathbb{E}_{p(x,x')}\left[\phi_t(x)^\top \nabla_\phi \phi(x')\right] + \mathbb{E}_{p(x')p(x)}\left[\phi(x')^\top \left(\phi(x)\phi(x)^\top\right)\nabla_\phi \phi(x')\right]\Big).$$

Comparing this with the update rules in (12) and (14), we can recover the gradient of the Spectral Contrastive loss with $\mathbb{E}_{p(x)}\left[\phi_t(x)\phi_t^\top(x)\right]$ approximated by $\texttt{LT}\left[\Lambda_{t+1}\right]$.

This asymptotic non-contrastive approximation not only holds for the χ^2-MI discussed so far, but it can also be extended to other f-MI objectives.

3.4 Generalization for f-MI Representation

The general f-MI objective is restated as follows with an embedding network ϕ, as well as a choice of a scalar encoding function $t : \mathbb{R}_+ \to \mathbb{R}$, *i.e.*,

$$I_f(X, X') = \max_\phi \mathbb{E}_{p(x,x')}\left[t\left(\phi(x)^\top\phi(x')\right)\right] - \mathbb{E}_{p(x)p(x')}\left[f^*\left(t\left(\phi(x)^\top\phi(x')\right)\right)\right]. \quad (14)$$

The first term of $\mathbb{E}_{p(x,x')}\left[t\left(\phi(x)^\top\phi(x')\right)\right]$ can be put aside, because it does not have the quadratic variance issue. So we focus on the second term $\mathbb{E}_{p(x)p(x')}\left[f^*\left(t\left(\phi(x)^\top\phi(x')\right)\right)\right]$. For the χ^2-divergence, we already know that we get a squared dot product term for this second term, which allows us to make the non-contrastive conversion. In the more general f-divergence case, it suffices then to pick f^* and t such that we get back to a squared dot product form again.

We consider a family of f-divergences, the α-divergences. The α-divergences generalize many common divergences through a real parameter α. For example, when $\alpha \in [0, 1, 2]$ we get the reverse KL, KL, and (Pearson) χ^2 divergences respectively. This means the Spectral Contrastive loss corresponds to $\alpha = 2$. We know f and f^* and we can pick t as follows:

$$f_\alpha(y) = \frac{(y^\alpha - 1) - \alpha(y-1)}{\alpha(\alpha-1)} \quad (15)$$

$$f_\alpha^\star(t) = \frac{1}{\alpha}\left|1 + (\alpha-1)t\right|^{\frac{\alpha}{\alpha-1}} - \frac{1}{\alpha} \quad (16)$$

$$t_\alpha(u) = \operatorname{sign}(u)\frac{1}{\alpha-1}\left|\sqrt{\frac{\alpha}{2}}u\right|^{\frac{2(\alpha-1)}{\alpha}} - \frac{1}{\alpha-1}. \quad (17)$$

Plugging these in leads to the following:

$$\max_\phi \mathbb{E}_{p(x,x')}\left[t_\alpha\left(\phi(x)^\top\phi(x')\right)\right] - \frac{1}{2}\mathbb{E}_{p(x)p(x')}\left[\left(\phi(x)^\top\phi(x')\right)^2\right] + \frac{1}{\alpha}. \quad (18)$$

Now the second term has the square dot product form as needed, which we can rewrite with the auxiliary matrix Λ, giving us the corresponding generalized objective (safely ignoring the constant $1/\alpha$):

$$\max_\phi \mathbb{E}_{p(x,x')}\left[t_\alpha\left(\phi(x)^\top\phi(x')\right)\right] - \frac{1}{2}\mathbb{E}_{p(x')}\left[\left(\phi(x')^\top \texttt{LT}\left[\Lambda\right]\phi(x')\right)\right]. \quad (19)$$

where Λ is updated through the EMA update rule (12).

3.5 The MINC Objective

There are two more additions that we make to the objective Equation (19) to further improve its empirical performance. The first is to normalize the embedding vector $\phi(x)$ so that it has Euclidean norm 1. This means the dot product is now a cosine similarity. Because this will constrain the magnitude of the dot product, it becomes important to allow the dot product to be scaled by a separate hyperparameter that we call the *inner scale* s.

$$\max_{\phi} \mathbb{E}_{p(x,x')}\left[t_\alpha\left(s\phi(x)^\top\phi(x')\right)\right] - \frac{1}{2}\mathbb{E}_{p(x')}\left[s^2\left(\phi(x')^\top \mathtt{LT}\left[\Lambda\right]\phi(x')\right)\right]. \tag{20}$$

This modification is also done in the Spectral Contrastive loss and SimCLR, though they use $1/s$ as the hyperparameter and denote it as the temperature.

The final addition we make is to use a target network ϕ_{target} for x and Λ, which gives the following objective and update rule (t_α is defined in equation (17)):

$$\max_{\phi} \mathbb{E}_{p(x,x')}\left[t_\alpha\left(s\phi_{\text{target}}(x)^\top\phi(x')\right)\right] - \frac{1}{2}\mathbb{E}_{p(x')}\left[s^2\left(\phi(x')^\top \mathtt{LT}\left[\Lambda\right]\phi(x')\right)\right] \tag{21}$$

$$\Lambda_{t+1} = \beta\Lambda_t + (1-\beta)\mathbb{E}_{p(x)}\left[\phi_{\text{target},t}(x)\phi_{\text{target},t}(x)^\top\right] \tag{22}$$

This is similar to BYOL in that ϕ_{target} is a slow, EMA updated version of ϕ. We found this target network to both improve performance and increase learning stability under large learning rates. Putting it together we get Algorithm 1.

Algorithm 1 MINC

Input: batch of data $\{(x, x')\}$, α-divergence α, inner scale s, auxiliary EMA β, neural network embedding ϕ, target network EMA γ
Initialize target network $\phi_{\text{target},0} = \phi$
Initialize auxiliary matrix $\Lambda_0 = 0$
for $i = 1$ **to** ... **do**
  Sample minibatch of data $B = \{(x_j, x'_j)\}$
  Update Λ_i through Equation (22) using B
  Update ϕ_i through gradient of Equation (21) using B
  Update target network $\phi_{\text{target},i}$ with EMA γ
end for

3.6 More Theoretical Connections

Remark (Choice of Matching Metric): In (14), we consider the Mahalanobis distance with Λ_{t+1}^{-1}. In fact, one can also select vanilla L_2 distance, *i.e.*,

$$\int\left[\left\|\int\frac{p(x,x')}{\sqrt{p(x')}}\phi_t(x)\,dx - \Lambda_{t+1}\sqrt{p(x')}\phi(x')\right\|^2\right]dx' \propto -2\mathbb{E}_{p(x,x')}\left[\phi_t(x)^\top\Lambda_{t+1}\phi(x')\right] + \mathbb{E}_{p(x')}\left[\phi(x')^\top\Lambda_{t+1}^\top\Lambda_{t+1}\phi(x')\right].$$

Comparing to (14), the L_2 distance leads to a similar objective, which is also stochastic gradient descent compatible, but with worse condition number. We can consider other metrics, which is out of the scope of the paper and we leave for future research.

Remark (Connection to BYOL and the variants): BYOL (Grill et al., 2020) and its variants (Richemond et al., 2023; Tian et al., 2021; Wang et al., 2021) can be recast as variants in implementing power iteration. The major difference between MINC and BYOL and its variants mainly lies in the update rule of Λ. We emphasize these update rules in BYOL (Grill et al., 2020) and its variants (Richemond et al., 2023; Tian et al., 2021; Wang et al., 2021) can also be derived by exploiting different properties of Λ. For example, as we show in (9), Λ plays as the eigenvalue matrix in stationary solution, therefore, it should satisfy the variational characteristic of eigenvalues, *i.e.*,

$$\Lambda = \operatorname*{argmin}_A \int\left[\left\|\int\frac{p(x,x')}{\sqrt{p(x')}}\phi(x)\,dx - A\sqrt{p(x')}\phi(x')\right\|^2\right]dx' \propto -2\mathbb{E}_{p(x,x')}\left[\phi_t(x)^\top A\phi(x')\right] + \mathbb{E}_{p(x')}\left[\phi(x')^\top A^\top A\phi(x')\right],$$

which leads to the update rule in BYOL with a linear predictor. Similarly, we can also derive other update rule for Λ through the properties of eigen-values, as shown in Richemond et al. (2023).

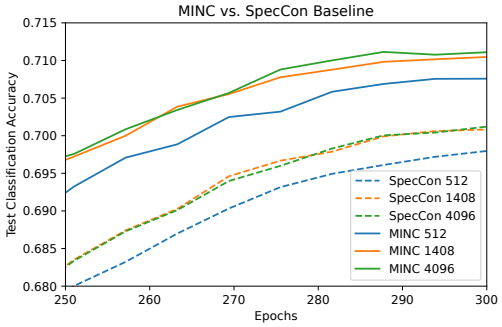
4 EXPERIMENTS

We follow the same experimental procedures as in SimCLR (Chen et al., 2020) and BYOL (Grill et al., 2020). We train latent image representations in ImageNet using a standard ResNet-50 network (He et al., 2016) and our objective, with standard random image augmentations and without labels, as in SimCLR and BYOL. On top of the ResNet, we have a projector MLP akin to SimCLR, with output sizes $(2048, 2048, 2048)$ i.e. two hidden layers with an output embedding size of 2048. The learned representations are evaluated by training a separate linear classifier on top of the frozen learned representations on a validation or test split. The representation that is evaluated is the immediate output of the ResNet, before the projector. For training we follow the same initial learning rate of 0.3 with 10 epochs of linear warmup, followed by cosine decay with the LARS optimizer and weight decay of 10^{-4}, identical to SimCLR. When using a target network, we use an

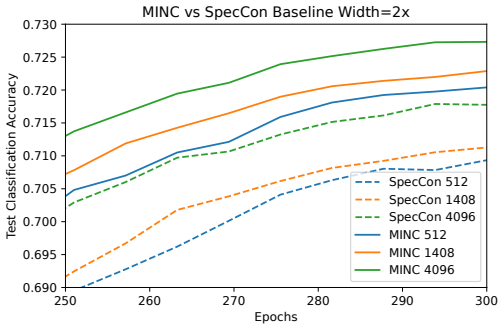
exponential moving average update with decay rate of $\gamma = 0.996$, same as for BYOL, except we use a fixed decay of 0.996 with no schedule. Unless otherwise specified, we use an auxiliary EMA $\beta = 0.8$. More details in the appendix. We train for 300 epochs and 1 seed due to computational cost.

4.1 Main Result

For our main results, we pick the $\alpha = 2$ (χ^2-divergence), as it was the best-performing and simplest compared to other nearby values of α (see the ablations in Section 4.2 and Figure 2 for more details).



(a) MINC vs the SpecCon Baseline. MINC improves consistently across all batch sizes.



(b) MINC vs the SpecCon Baseline but doubling the width of the ResNet-50. MINC still improves across the board.

Our main result is shown in Table 1, with curves in Figures 1a and 1b, where we compare our MINC vs. a SpecCon baseline across three different batch sizes of 512, 1408 and 4096. These results are evaluated using the ImageNet test split, and trained on both the train and validation splits. We also include reported results from SimCLR, MoCo v2 (He et al., 2020), SimSiam, Linear BYOL (the predictor is linear and trained), Linear NS^2 BYOL (Richemond et al., 2023) and BYOL (Grill et al., 2020), for reference. MINC consistently outperforms the Spectral Contrastive baseline across all batch sizes and for both network sizes (standard and 2x). This shows that MINC is a robust improvement over the baseline Spectral Contrastive loss.

Table 1: Main Result: Test Classification (Top-1) Accuracy

Method	ResNet50	ResNet50 2x
SpecCon 512	0.698	0.709
MINC 512	0.708	0.720
SpecCon 1408	0.701	0.711
MINC 1408	0.710	0.723
SpecCon 4096	0.701	0.718
MINC 4096	0.711	0.727
SimCLR	0.693	0.742
MoCo v2	0.711	—
SimSiam	0.713	—
Linear BYOL	0.715	—
Linear NS^2 BYOL	0.722	—
BYOL	0.743	0.774

Compared with other methods, it outperforms SimCLR, another contrastive method, and almost matches some non-contrastive methods such as SimSiam and Linear BYOL. Performing similarly to Linear BYOL suggests that the power iteration framing and connection (Section 3.6) of MINC to Linear BYOL may be very illuminating and useful for further development. However MINC is still not able to catch up with the regular BYOL that uses a non-linear predictor, which shows that perhaps the Spectral Contrastive loss (and other contrastive losses) may still be limited in representational power. This leaves the door open for better contrastive methods to be used and subsequently converted to improved non-contrastive objectives.

4.2 Ablations

MINC is made up of three main components: 1) the MINC objective with the auxiliary Λ; 2) The GHA update rule with the lower triangular transformation; 3) Usage of a target network. In this section we have ablations to show the contribution of each. The ablations are run with the same setup as the main result, using a batch size of 1408 unless otherwise specified, but trained only on the train split and evaluated on the validation split.

In Figure 2, we first sweep over various α-divergences. For this ablation, the inner scale s is not fixed, but actually learned through a separate SGD+Momentum optimizer with learning rate 0.1 and momentum 0.9. Surprisingly, $\alpha = 2$ (χ^2-divergence), the simplest Spectral Contrastive version, performs the best. With the larger $\alpha = 2.5$, raining becomes noticeably unstable and the performance is dramatically worse. This is because once $\alpha > 2$ there is higher pressure to push some of the embeddings much farther apart from each other, making it easier to diverge. As α gets smaller

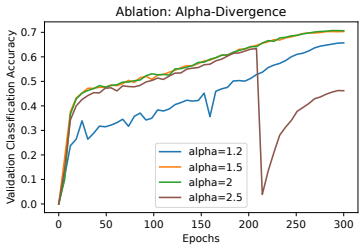


Figure 2: Ablation for the α-divergence with auxiliary EMA $= 0$. $\alpha = 2$ is the best performing one. $\alpha = 1.5$ is close, but smaller α or $\alpha > 2$ are much worse.

than 2 and closer to 1 ($\alpha = 1$ is the KL-divergence version), performance also gets worse. Owing to this result, we fixed $\alpha = 2$ (Spectral Contrastive) for our main results and other ablations. This result is surprising at first, due to the results for f-MICL (Lu et al., 2024), which showed much smaller differences between f-divergences. On closer inspection, we can see that the objectives of MINC and f-MICL are completely different because we rewrite the contrastive objective using a scalar transformation t_α that changes the second term (the quadratic variance term) into a squared dot product form. This makes MINC behave very differently than f-MICL, and thus perhaps the different α-divergences have a much stronger impact on the learned representation.

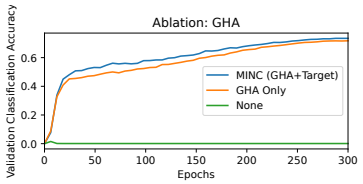


Figure 3: Ablation for GHA when auxiliary EMA $= 0.8$. Without GHA, the representation collapses. With GHA, the representation learns a meaningful representation. Combining GHA and the target network results in even better performance.

In Figure 3, we investigate the importance of the GHA-derived lower triangular transformation $\texttt{LT}[\cdot]$. We show how just directly using the MINC objective with an auxiliary EMA of 0.8 but without a target network nor the GHA transformation completely falls flat; the representation essentially collapses to a constant. Then by adding in the GHA transformation, we see that it no longer collapses and actually learns something. GHA is successfully preventing the collapse that could happen. Finally the complete MINC with both GHA and target network further improves the performance, showing the further benefit of the target network.

The next ablation in Figure 4 is a sweep over the auxiliary EMA β. An auxiliary EMA of 0 means that the auxiliary Λ is only computed from the current mini-batch, making it very similar to the Spectral Contrastive loss but with GHA and a target network, which performs comparatively the worst, but is still learning a meaningful representation. But increasing the auxiliary EMA further improves the performance, with 0.8 being the best. Going larger than 0.8 starts to result in worse performance. The fact that 0.8 is a sweetspot suggests there is a bias-variance tradeoff due to the non-stationarity of the embeddings ϕ. A larger EMA means we are accumulating Λ over many past mini-batches, which is more biased because the past mini-batches have stale representations. A smaller EMA accumulates over fewer mini-batches, resulting in a less stable but higher variance estimate.

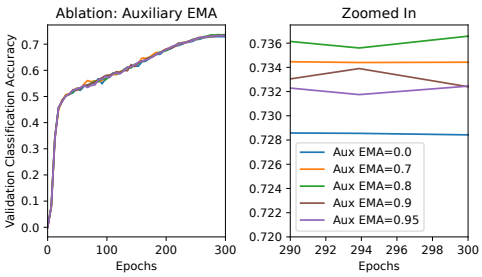


Figure 4: Ablation for auxiliary EMA of MINC. Any amount of auxiliary EMA is better than 0, with 0.8 being the sweetspot.

5 RELATED WORK

See Appendix for related work section.

6 CONCLUSION

Our MINC algorithm successfully improves upon the Spectral Contrastive loss by turning it from a contrastive loss into a non-contrastive loss. Theoretically, we provide a power iteration perspective that can relate contrastive and non-contrastive losses, connecting the Spectral Contrastive loss and the Linear BYOL method together. In ImageNet, MINC consistently shows an improvement over the Spectral Contrastive loss, and reduces the gap between contrastive methods and BYOL, the (non-contrastive) state-of-the-art. Interestingly, we found that MINC performs on par with Linear BYOL, which suggests that the non-linear predictor in BYOL can inspire further improvements to contrastive and non-contrastive spectral losses. More broadly, beyond ImageNet, our work does show the feasibility of combining the strengths of contrastive and non-contrastive objectives together, and could pave the

way for more powerful contrastive to non-contrastive conversions in the future. On the theory side, the power iteration perspective may also prove a useful tool to strengthen existing results and weaken some of the limiting assumptions.

References

S. Arora, H. Khandeparkar, M. Khodak, O. Plevrakis, and N. Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019.

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

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

X. Chen and K. He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.

G. H. Golub and C. F. Van Loan. *Matrix computations*. JHU press, 2013.

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

J. Z. HaoChen, C. Wei, A. Gaidon, and T. Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 5000–5011. Curran Associates, Inc., 2021.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

J. D. Lee, Q. Lei, N. Saunshi, and J. Zhuo. Predicting what you already know helps: Provable self-supervised learning. *Advances in Neural Information Processing Systems*, 34:309–323, 2021.

K.-J. Liu, M. Suganuma, and T. Okatani. Bridging the gap from asymmetry tricks to decorrelation principles in non-contrastive self-supervised learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 19824–19835. Curran Associates, Inc., 2022.

Y. Lu, G. Zhang, S. Sun, H. Guo, and Y. Yu. f-MICL: Understanding and generalizing infonce-based contrastive learning. *arXiv preprint arXiv:2402.10150*, 2024.

O. Nachum and B. Dai. Reinforcement learning via fenchel-rockafellar duality. *arXiv preprint arXiv:2001.01866*, 2020.

X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

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

T. Ren, T. Zhang, L. Lee, J. E. Gonzalez, D. Schuurmans, and B. Dai. Spectral decomposition representation for reinforcement learning. In *Proceedings of the International Conference on Learning Representations*, 2023.

T. Ren, H. Sun, A. Moulin, A. Gretton, and B. Dai. Spectral representation for causal estimation with hidden confounders. *arXiv preprint arXiv:2407.10448*, 2024.

P. H. Richemond, A. Tam, Y. Tang, F. Strub, B. Piot, and F. Hill. The edge of orthogonality: A simple view of what makes BYOL tick. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29063–29081. PMLR, 23–29 Jul 2023.

T. D. Sanger. Optimal unsupervised learning in a single-layer linear. *Neural Networks*, 1989.

Y. Tang, Z. D. Guo, P. H. Richemond, B. Á. Pires, Y. Chandak, R. Munos, M. Rowland, M. G. Azar, C. L. Lan, C. Lyle, et al. Understanding self-predictive learning for reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2023.

Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020.

Y. Tian, X. Chen, and S. Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10268–10278. PMLR, 18–24 Jul 2021.

C. Tosh, A. Krishnamurthy, and D. Hsu. Contrastive learning, multi-view redundancy, and linear models. In *Algorithmic Learning Theory*, pages 1179–1206. PMLR, 2021.

L. N. Trefethen and D. Bau. *Numerical linear algebra*. SIAM, 2022.

X. Wang, X. Chen, S. S. Du, and Y. Tian. Towards demystifying representation learning with non-contrastive self-supervision. *arXiv preprint arXiv:2110.04947*, 2021.

C.-Y. Wu, R. Manmatha, A. J. Smola, and P. Krahenbuhl. Sampling matters in deep embedding learning. In *Proceedings of the IEEE international conference on computer vision*, pages 2840–2848, 2017.

B. Xie, Y. Liang, and L. Song. Scale up nonlinear component analysis with doubly stochastic gradients. *Advances in Neural Information Processing Systems*, 28, 2015.

Y. You, I. Gitman, and B. Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

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

Checklist

The checklist follows the references. For each question, choose your answer from the three possible options: Yes, No, Not Applicable. You are encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description (1-2 sentences). Please do not modify the questions. Note that the Checklist section does not count towards the page limit. Not including the checklist in the first submission won't result in desk rejection, although in such case we will ask you to upload it during the author response period and include it in camera ready (if accepted).

In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes, see Section 3]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Not Applicable]
   (b) Complete proofs of all theoretical results. [Not Applicable]
   (c) Clear explanations of any assumptions. [Not Applicable]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [No code]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [No]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

APPENDIX

A RELATED WORK

Our algorithm MINC starts from the same theoretical framework as the Spectral Contrastive Loss (HaoChen et al., 2021) and f-MICL (Lu et al., 2024), which seek to maximize a notion of mutual information between data points. Other contrastive losses such as InfoNCE (Oord et al., 2018) and SimCLR (Chen et al., 2020), and more (Arora et al., 2019; Lee et al., 2021; Tosh et al., 2021) do not fall into the exact same framework, but are still similar in that the goal is still to maximize some notion of mutual information. Because contrastive methods need to compare every data point to every other data point, their objective have a quadratic dependence on the size of the dataset, which results in high variance. Our work builds on this mutual information framework, and goes on to derive a non-contrastive version based on the theoretical ideas of power iteration in order to reduce this variance dependence.

On the non-contrastive side, BYOL (Grill et al., 2020) amongst many others (Chen and He, 2021; Bardes et al., 2021; Zbontar et al., 2021; Liu et al., 2022; Tian et al., 2021; Wang et al., 2021) have been exploring ways to remedy the quadratic dependence on the dataset. Our MINC algorithm, is a non-contrastive version of the Spectral Contrastive loss that does not have the quadratic dependence. Interestingly, through a power iteration viewpoint, MINC can in fact be related to a linear predictor version of BYOL (Section 3.6). This connection brings non-contrastive methods and contrastive methods closer together, in that they all can be seen to be ultimately related to maximizing mutual information, just in somewhat different ways.

B Experiment Details

We follow the training setup as in Chen et al. (2020). This means we use the same image augmentations (random crops, color distortions, and Gaussian blurs). We use the same LARS optimizer (You et al., 2017), but instead of a linear scaling with batch size, we use the suggested square root scaling (learning rate $= 0.075 \times \sqrt{\text{batch size}}$) that was tested in the appendix, which was better performing for smaller batch sizes and fewer epochs.

Table 2: Optimization Hyperparameters

Optimizer	LARS
Learning Rate	$0.075 \times \sqrt{\text{batch size}}$
Learning Rate Linear Warmup	10 Epochs
Learning Rate Decay	Cosine
eta	10^{-3}
momentum	0.9
weight decay	10^{-4}

We use almost the same architecture as in SimCLR, with a ResNet-50 (He et al., 2016) base and an MLP projector on top. Our MLP has two hidden layers instead of the one in SimCLR. After the MLP, we normalize the embeddings to have Euclidean norm 1 as follows:

$$\frac{1}{\sqrt{\phi(x)^\top \phi(x) + 10^{-5}}}\phi(x)$$

where we add a small positive constant to avoid gradient divergence issues.

Below are the default algorithm hyperparameters that we use for the main results. For the inner scale hyperparameter, we found that doing an initial linear warmup starting from 1 (similar to the learning rate) increase the initial training stability.

Table 3: Network Architecture

ResNet-50
MLP [2048, relu, 2048, relu, 2048]
Normalize to norm 1

Table 4: Algorithm Hyperparameters

α-divergence	$\alpha = 2$
inner scale	$s = 100$
inner scale linear warmup	10 Epochs
auxiliary EMA	$\beta = 0.8$
target network EMA	$\gamma = 0.996$

For numerical stability, when we implement the MINC loss:

$$\max_{\phi} \mathbb{E}_{p(x,x')}\left[t_\alpha\left(s\phi_{\text{target}}(x)^\top \phi(x')\right)\right] - \frac{1}{2}\mathbb{E}_{p(x')}\left[s^2\left(\phi(x')^\top \texttt{LT}\left[\Lambda\right]\phi(x')\right)\right]$$

where

$$t_\alpha(u) = \text{sign}(u)\frac{1}{\alpha-1}\left|\sqrt{\frac{\alpha}{2}}u\right|^{\frac{2(\alpha-1)}{\alpha}} - \frac{1}{\alpha-1}$$

the u passed to $t_\alpha(u)$ can span both positive, zero, and negative values. Hence to avoid issues with infinite gradients, we ensure u cannot be zero by clipping as follows:

$$t_\alpha(u) = \text{sign}(u)\frac{1}{\alpha-1}\left(\max\left(\left|\sqrt{\frac{\alpha}{2}}u\right|, 10^{-6}\right)\right)^{\frac{2(\alpha-1)}{\alpha}} - \frac{1}{\alpha-1}$$

C Additional Ablation

Another ablation in Figure 5 shows a sweep for the inner scale (inverse temperature) hyperparameter for the contrastive baseline. An inner scale of 100 performs the best, and so we also used the same inner scale of 100 for the MINC objective to be consistent. Explicitly picking an inner scale does turn out to be a little better than letting the inner scale be trained through SGD+Momentum, as letting it be trained likely results in a little bit more overfitting.

In the ablation of α-divergence Figure 2 and inner scale Figure 5 where we trained the inner scale parameter s, we used a separate SGD+Momentum optimizer with learning rate 0.1 and momentum 0.9 for s. We found LARS to be not as suitable for training scalars, and a simple SGD+Momentum was very effective.

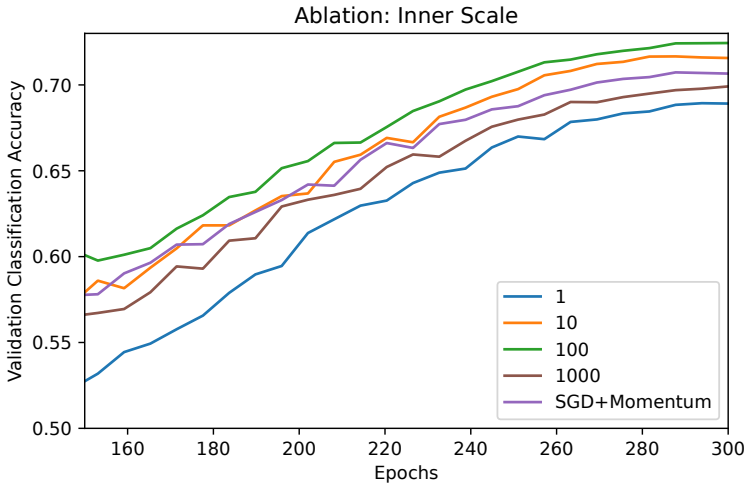


Figure 5: Ablation for inner scale hyperparameter for Spectral Contrastive Baseline, showing that 100 is best.