Self-supervised learning with rotation-invariant kernels
Léon Zheng, Gilles Puy, Elisa Riccietti, Patrick Pérez, Rémi Gribonval

To cite this version:
Léon Zheng, Gilles Puy, Elisa Riccietti, Patrick Pérez, Rémi Gribonval. Self-supervised learning with rotation-invariant kernels. The Eleventh International Conference on Learning Representations, May 2023, Kigali, Rwanda. hal-03738466v4

HAL Id: hal-03738466
https://hal.science/hal-03738466v4
Submitted on 6 Mar 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
SELF-SUPERVISED LEARNING WITH ROTATION-INVARIANT KERNELS

Léon Zheng\textsuperscript{1,2}, Gilles Puy\textsuperscript{1}, Elisa Riccietti\textsuperscript{2}, Patrick Pérez\textsuperscript{1}, Rémi Gribonval\textsuperscript{2}

\textsuperscript{1}valeo.ai, Paris, France
\textsuperscript{2}Univ Lyon, EnsL, UCBL, CNRS, Inria, LIP, F-69342, LYON Cedex 07, France.

ABSTRACT

We introduce a regularization loss based on kernel mean embeddings with rotation-invariant kernels on the hypersphere (also known as dot-product kernels) for self-supervised learning of image representations. Besides being fully competitive with the state of the art, our method significantly reduces time and memory complexity for self-supervised training, making it implementable for very large embedding dimensions on existing devices and more easily adjustable than previous methods to settings with limited resources. Our work follows the major paradigm where the model learns to be invariant to some predefined image transformations (cropping, blurring, color jittering, etc.), while avoiding a degenerate solution by regularizing the embedding distribution. Our particular contribution is to propose a loss family promoting the embedding distribution to be close to the uniform distribution on the hypersphere, with respect to the maximum mean discrepancy pseudometric. We demonstrate that this family encompasses several regularizers of former methods, including uniformity-based and information-maximization methods, which are variants of our flexible regularization loss with different kernels. Beyond its practical consequences for state-of-the-art self-supervised learning with limited resources, the proposed generic regularization approach opens perspectives to leverage more widely the literature on kernel methods in order to improve self-supervised learning methods.

1 INTRODUCTION

Self-supervised learning is a promising approach for learning visual representations: recent methods (He et al., 2020; Grill et al., 2020; Caron et al., 2020; Gidaris et al., 2021) reach the performance of supervised pretraining in terms of quality for transfer learning in many downstream tasks, like classification, object detection, semantic segmentation, etc. These methods rely on some prior knowledge on images: the semantic of an image is invariant (Misra & Maaten, 2020) to some small transformations of the image, such as cropping, blurring, color jittering, etc. One way to design an objective function that encodes such an invariance property is to enforce two different augmentations of the same image to have a similar representation (or embedding) when they are encoded by the neural network. However, the main issue with this kind of objective function is to avoid an undesirable loss of information (Jing et al., 2022) where, e.g., the network learns to represent all images by the same constant representation. Hence, one of the main challenges in self-supervised learning is to propose an efficient way to regularize the embedding distribution in order to avoid such a collapse.

Our contribution is to propose a generic regularization loss promoting the embedding distribution to be close to the uniform distribution on the hypersphere, with respect to the maximum mean discrepancy (MMD), a distance on the space of probability measures based on the notion of embedding probabilities in a reproducing kernel Hilbert space (RKHS), using the so-called kernel mean embedding mapping. Inspired by high-dimensional statistical tests for uniformity that are rotation-invariant (García-Portugués & Verdebout, 2018), we choose to embed probability distributions using rotation-invariant kernels on the hypersphere (dot-product kernels), i.e., kernels for which the evaluation for two vectors depends only on their inner product (Smola et al., 2000). This paper shows that such an approach leads to important theoretical and practical consequences for self-supervised learning.

Code: \url{https://github.com/valeoai/sfrik}
We demonstrate that our regularization loss family parameterized by such rotation-invariant kernels encompasses several regularizers of former methods. As illustrated in Table 1, they are variants of our generic loss with different kernels: the quadratic kernel yields the general sample-contrastive criterion of Garrido et al. (2023) that englobes many contrastive learning methods like (HaoChen et al., 2021) (cf. Appendix A.2); the radial basis function (RBF) kernel yields the uniformity loss of Alignment & Uniformity on the Hypersphere (AUH) (Wang & Isola, 2020); the generalized distance kernel (cf. Example 2) yields one of the regularization used in PointContrast (Xie et al., 2020); and a linear combination of the linear kernel and the quadratic kernel yields a regularizer that promotes the covariance matrix of the embedding distribution to be proportional to the identity matrix, similarly to information-maximization methods like VICReg (Bardes et al., 2022). In other words, these former methods turn out to be particular ways of minimizing the MMD between the embedding distribution and the uniform distribution on the hypersphere, in the sense of the MMD associated to a rotation-invariant kernel \( \mathcal{K}(\mathbf{u}, \mathbf{v}) = \varphi(\mathbf{u}^\top \mathbf{v}) \) defined on the hypersphere.

Numerically, we show in a rigorous experimental setting with a separate validation set for hyperparameter tuning that our method yields fully competitive results compared to the state of the art, when parameter tuning that our method yields fully competitive results compared to the state of the art. As illustrated in Table 1, they are variants of our generic loss with different kernels: the quadratic kernel yields the general sample-contrastive criterion of Garrido et al. (2023) that englobes many contrastive learning methods like (HaoChen et al., 2021) (cf. Appendix A.2); the radial basis function (RBF) kernel yields the uniformity loss of Alignment & Uniformity on the Hypersphere (AUH) (Wang & Isola, 2020); the generalized distance kernel (cf. Example 2) yields one of the regularization used in PointContrast (Xie et al., 2020); and a linear combination of the linear kernel and the quadratic kernel yields a regularizer that promotes the covariance matrix of the embedding distribution to be proportional to the identity matrix, similarly to information-maximization methods like VICReg (Bardes et al., 2022). In other words, these former methods turn out to be particular ways of minimizing the MMD between the embedding distribution and the uniform distribution on the hypersphere during training, with various specific kernel choices. The proposed generic regularization approach opens perspectives to leverage more widely the literature on kernel methods in order to improve self-supervised learning.

Importantly, our method significantly reduces time and memory complexity for self-supervised training compared to information-maximization methods. Due to the kernel trick, the complexity of SFRIK’s loss is quadratic in the batch size and linear in the embedding dimension, instead of being quadratic as in VICReg. In practice, SFRIK’s pretraining time is up to 19% faster than VICReg for an embedding dimension 16384, and it can scale at dimension 32768, as opposed to VICReg whose memory requirement is too large at this dimension for a machine with 8 GPUs and 32GB of memory per GPU. Hence our work opens perspectives in self-supervised learning on embedded devices with limited memory like in (Xiao et al., 2022). We summarize our contributions as follows:

- **We introduce a generic regularization loss based on kernel mean embeddings with rotation-invariant kernels on the hypersphere for self-supervised learning of image representations.**
- **We show that our loss family encompasses several previous self-supervised learning methods, like uniformity-based and information-maximization methods.**
- **We numerically show that SFRIK significantly reduces time and memory complexity for self-supervised training, while remaining fully competitive with the state of the art.**
2 RELATED WORK

Instance discrimination methods typically rely on a contrastive loss that behaves asymptotically like an alignment and uniformity loss on the hypersphere in the limit of infinite samples. Our contribution is to formalize and generalize existing uniformity-based methods by using kernel mean embeddings. To the best of our knowledge, the proposed kernel framework establishes the first connection between uniformity-based methods and information-maximization methods like VICReg.

Instance discrimination One way of learning image representations that are invariant to predefined image transformations (Misra & Maaten, 2020) is to rely on an instance classification approach (Wu et al., 2018). Typically, contrastive learning (Oord et al., 2018; Hjelm et al., 2019; Chen et al., 2020a,b; He et al., 2020; Henaff, 2020) discriminates instances within a batch of sampled images using the noise contrastive estimator (Gutmann & Hyvärinen, 2010), by attracting embeddings of transformed images coming from the same image instance, and repulsing embeddings coming from different image instances. In practice, this estimator needs a large number of image representations in order to achieve good results, which requires a large batch size like SimCLR (Chen et al., 2020a) or a memory bank (Wu et al., 2018; He et al., 2020). In the limit of infinite samples, the contrastive loss is shown to behave asymptotically like the alignment and uniformity loss of AUH.

Uniformity on the hypersphere Existing uniformity-based methods avoid collapse by regularizing the embedding distribution to be somehow close to the uniform distribution on the hypersphere, which has a high entropy. Bojanowski & Joulin (2017) perform this kind of regularization by aligning the learned representations on a fixed number of vectors sampled uniformly at random on the hypersphere. AUH maximizes the average pairwise distance between embeddings using an RBF kernel, in the spirit of energy minimization methods that address the problem of scattering points evenly on the hypersphere (Hardin & Saff, 2005; Liu et al., 2018; Borodachov et al., 2019). Although alternative high-entropy prior distributions (e.g., the uniform distribution on the hypercube) can be used for regularization (Chen et al., 2021), encoding images into $\ell_2$-normalized representations helps to stabilize training (Schroff et al., 2015; Parkhi et al., 2015; Liu et al., 2017).

Kernel mean embedding As a contribution, our generic loss formalizes and generalizes these previous uniformity losses, by relying on kernel mean embeddings (cf. Appendix A.1) to measure the distance between probability distributions on high-dimensional spaces, using the MMD pseudometric (Gretton et al., 2012; Li et al., 2015; Dziugaite et al., 2015; Briol et al., 2019) with rotation-invariant kernels on the hypersphere (Smola et al., 2000; Pennington et al., 2015; Lyu, 2017; Dutordoir et al., 2020). These tools are adapted for high-dimensional problems on the hypersphere whose geometry is different from the one in small dimension, as illustrated by García-Portugués & Verdebout (2018): many statistical tests for uniformity on the hypersphere, i.e., tests for rejecting the null hypothesis where a batch of normalized vectors is sampled from the uniform distribution on the hypersphere, are in fact precisely estimators of the MMD between the embedding distribution and the uniform distribution, for different kernels. Our kernel method for self-supervision is complementary to (Li et al., 2021), in which the dependency between image instances and their embedding is maximized with respect to the Hilbert-Schmidt independence criterion (cf. Appendix A.3).

Information maximization Our generic regularization approach has the benefit of connecting uniformity-based and information-maximization methods (Zbontar et al., 2021; Ermolov et al., 2021; Bardes et al., 2022). The latter are alternatives to distillation methods (Grill et al., 2020; Gidaris et al., 2020; 2021; Chen & He, 2021; Caron et al., 2021) where a student network learns to predict the representations of a teacher network. In such methods, using various architecture tricks (like prediction head, stop-gradient, momentum encoder, batch normalization or centering) is shown empirically to be sufficient to avoid collapse without instance discrimination, even though it is not fully understood how these multiple factors induce a regularization during training (Richemond et al., 2020; Tian et al., 2021). Instead of using these tricks, information-maximization methods use a Siamese architecture and avoid collapse by maximizing the statistical information of a batch of embeddings, using a whitening operation (Ermolov et al., 2021), or an explicit regularization term making the covariance (Bardes et al., 2022) or the cross-correlation (Zbontar et al., 2021) matrix close to a scaled identity matrix. This paper shows that our generic regularization loss with an appropriate kernel also promotes the covariance matrix of the embedding distribution to be proportional to the identity matrix. But in contrast to VICReg which explicitly computes the covariance matrix, our method uses the kernel trick to significantly reduce complexity at large embedding dimensions.
3 Method description

Given an unlabeled dataset of images \(x_i \sim P, i \in [N] := \{1, \ldots, N\}\), sampled independently from a data distribution \(P\), the goal is to learn a backbone network \(f_\theta\) parameterized by \(\theta\) (e.g., a convolutional neural network) such that any new image \(x \sim P\) is encoded by a good representation \(f_\theta(x)\) whose quality is evaluated in several downstream tasks (see Section 4).

3.1 Invariance and uniformity for self-supervision

Our self-supervised learning method (see Figure 1) follows the principle of the recent methods like SimCLR or VICReg. During self-supervised training, each image \(x_i\) is augmented using two different random transformations \(t^{(1)}\) and \(t^{(2)}\) sampled from a distribution \(T\), which yields two views \(x_i^{(1)} := t^{(1)}(x_i)\) and \(x_i^{(2)} := t^{(2)}(x_i)\) of the image \(x_i\). Two representations \(z_i^{(v)} (v = 1, 2)\) are obtained by encoding each \(x_i^{(v)}\) with the backbone \(f_\theta\) and \(\ell_2\)-normalizing the resulting feature vector. For a given subset of indices \(I \subseteq [N]\), we write \(Z_I^{(v)} := \{z_i^{(v)}\}_{i \in I}\). The backbone \(f_\theta\) is trained by minimizing the total objective function:

\[
\mathcal{L} = \mathbb{E}_{T^{(1)}, T^{(2)} \sim T} \mathbb{E}_{I \subseteq [N]} \ell(Z_I^{(1)}, Z_I^{(2)}),
\]

where batches \(I\) are drawn at random with a prescribed batch size, and the loss \(\ell\) is a weighted sum involving an alignment term \(\ell_a\) and a uniformity term \(\ell_u\), in the spirit of AUH:

\[
\ell(Z_I^{(1)}, Z_I^{(2)}) := \lambda \ell_a(Z_I^{(1)}, Z_I^{(2)}) + 0.5(\ell_u(Z_I^{(1)}) + \ell_u(Z_I^{(2)}));
\]

\(\lambda > 0\) is a hyperparameter that tunes the balance between the two terms. The loss \(\ell_a\) enforces the invariance property of the model, and is defined for a batch \(I \subseteq [N]\) of cardinality \(|I|\) as:

\[
\ell_a(Z_I^{(1)}, Z_I^{(2)}) := \frac{1}{|I|} \sum_{i \in I} \|z_i^{(1)} - z_i^{(2)}\|_2^2.
\]

Our main contribution is in the choice of the uniformity term \(\ell_u\), detailed in the rest of the section. Note that instead of applying the loss (2) to the output of \(f_\theta\) (called image representation), we add a projection head \(g_w\) (a multi-layer perceptron) parameterized by \(w\) to the output of \(f_\theta\) and apply (2) at the output of \(g_w\) (called image embedding). This common practice (Caron et al., 2020; Grill et al., 2020) improves the performance in the downstream tasks. Therefore, denoting \(S^{q-1}\) the unit hypersphere in \(\mathbb{R}^q\), the image embedding actually reads \(z_i^{(v)} := (g_w \circ f_\theta)(x_i^{(v)})/\|g_w \circ f_\theta)(x_i^{(v)})\|_2 \in S^{q-1}\). Both \(g_w\) and \(f_\theta\) are jointly trained without supervision by minimizing (1) using a stochastic mini-batch algorithm. After training, \(g_w\) is discarded and only \(f_\theta\) is kept for the downstream tasks.

3.2 Uniformity loss via MMD minimization

We continue by explaining our generic kernel formulation of \(\ell_u\) using the MMD pseudometric and rotation-invariant kernels. Then we provide examples of such kernels and describe our kernel choice.

3.2.1 MMD pseudometric and rotation-invariant kernels

Our uniformity loss relies on a divergence in the space of probability distributions based on a positive definite kernel \(\mathcal{K}\) defined on some space \(\mathcal{X}\). Denoting \(\mathcal{H}\) the corresponding RKHS with norm \(\|\cdot\|_{\mathcal{H}}\), the MMD between two probability distributions \(Q_1, Q_2\) on \(\mathcal{X}\) can be expressed as the distance in \(\|\cdot\|_{\mathcal{H}}\) between their kernel mean embeddings (Borgwardt et al., 2006; Muandet et al., 2017):

\[
\text{MMD}(Q_1, Q_2) = \left\| \int_{\mathcal{X}} \mathcal{K}(u, \cdot) dQ_1(u) - \int_{\mathcal{X}} \mathcal{K}(u, \cdot) dQ_2(u) \right\|_{\mathcal{H}}.
\]

We propose to use this pseudometric to measure the distance between the probability distributions of the embeddings \(z_i^{(v)} (v = 1, 2)\) and the uniform probability distribution on the hypersphere \(S^{q-1}\) defined by \(\mathbb{U} := \sigma_{q-1}/|S^{q-1}|\), where \(\sigma_{q-1}\) denotes the normalized Hausdorff surface measure on \(S^{q-1}\), and \(|S^{q-1}| := \int_{S^{q-1}} d\sigma_{q-1} = 2\pi^{q/2}/\Gamma(q/2)\) is the surface area of \(S^{q-1}\), with \(\Gamma\) denoting the Gamma function. Intuitively, a good choice of kernel for measuring the distance (4) should
distinguish any probability distribution from the uniform distribution. Inspired by statistical tests for uniformity that are rotation-invariant (García-Portugués & Verdebout, 2018), we propose to use rotation-invariant kernels on $\mathcal{X} := \mathcal{S}^{q-1}$ of the form $\mathcal{K}(u, v) := \varphi(u^\top v)$ with $\varphi$ a continuous function defined on $[-1, 1]$ (Smola et al., 2000). The following theorem characterizes the form of function $\varphi$ that ensures positive definiteness of $\mathcal{K}$, and thus that (4) is a valid pseudometric.

**Theorem 1 (Schoenberg (1942, Theorem 1))** The kernel $\mathcal{K}(u, v) := \varphi(u^\top v)$ on $\mathcal{X} := \mathcal{S}^{q-1}$ with $\varphi$ continuous is positive definite if, and only if, the function $\varphi$ admits an expansion:

$$\varphi(t) = \sum_{\ell=0}^{+\infty} b_\ell P_\ell(q; t), \quad \text{with} \quad b_\ell \geq 0, \tag{5}$$

where $P_\ell(q; t) := \ell! \Gamma\left(\frac{q-1}{2}\right) \sum_{k=0}^{\left\lfloor \frac{\ell}{2} \right\rfloor} \left(-\frac{1}{4}\right)^k \frac{(1-t^2)\ell^{\ell-2k}}{k!(\ell-2k)! \Gamma(k+\frac{q-1}{2})}$ is the Legendre (or Gegenbauer) polynomial of degree $\ell$ in dimension $q$ (Müller, 2012, (2.32)).

As we are interested in measuring the distance between the embedding distribution and the uniform distribution on the hypersphere $\mathbb{U}$, we compute the kernel mean embedding of $\mathbb{U}$ for a kernel satisfying the condition of Theorem 1 using the following known result used, e.g., implicitly in (Brauchart et al., 2014). As we could not locate a formal proof, we provide one in Appendix B.1.

**Lemma 2** Let $\mathcal{K}(u, v) := \varphi(u^\top v)$ be a rotation-invariant kernel on $\mathcal{X} := \mathcal{S}^{q-1}$ where $\varphi$ admits the expansion (5). The kernel mean embedding of the uniform distribution $\mathbb{U}$ on $\mathcal{S}^{q-1}$ is constant: $\int_{\mathcal{S}^{q-1}} \mathcal{K}(u, v) \, d\mathbb{U}(u) = b_0 \in \mathbb{R}$ for all $v \in \mathcal{S}^{q-1}$. The kernel mean embedding of any probability distribution $\mathcal{Q}$ defined on the hypersphere satisfies: $\int_{\mathcal{S}^{q-1}} \mathcal{K}(u, \cdot) \, d\mathcal{Q}(u) = b_0 + \int_{\mathcal{S}^{q-1}} \tilde{\mathcal{K}}(u, \cdot) \, d\mathcal{Q}(u)$, where $\tilde{\mathcal{K}}(u, v) := \varphi(u^\top v)$ for any $u, v \in \mathcal{S}^{q-1}$ with $\varphi := \sum_{\ell=1}^{+\infty} b_\ell P_\ell(q; \cdot)$.

Using Lemma 2 in (4) yields $\text{MMD}(\mathcal{Q}, \mathbb{U}) = \| \int_{\mathcal{S}^{q-1}} \tilde{\mathcal{K}}(u, \cdot) \, d\mathcal{Q}(u) \|_{\mathcal{H}}$ for any probability distribution $\mathcal{Q}$ on $\mathcal{S}^{q-1}$. Then, by the reproducing property in the RKHS $\mathcal{H}$, the squared MMD satisfies, for any rotation-invariant kernel $\mathcal{K}$ verifying the condition of Theorem 1:

$$\text{MMD}^2(\mathcal{Q}, \mathbb{U}) = \mathbb{E}_{z, z' \sim \mathcal{Q}} \left[ \tilde{\mathcal{K}}(z, z') \right], \quad \text{with} \ z, z' \text{ i.i.d.} \tag{6}$$

### 3.2.2 Estimator of the squared MMD and kernel choices

The proposed uniformity loss $\ell_u$ for self-supervision is a biased estimator (Gretton et al., 2012) of $\text{MMD}^2(\mathcal{Q}, \mathbb{U})$ in (6). Given a batch $Z_I := \{z_i\}_{i \in I}$ sampled from $\mathcal{Q}$, our uniformity loss is:

$$\ell_u(Z_I) = \frac{1}{|I|^2} \sum_{i \in I} \sum_{i' \in I} \tilde{\mathcal{K}}(z_i, z_{i'}) = \frac{1}{|I|^2} \sum_{i \in I} \sum_{i' \in I} \varphi(z_i^\top z_{i'}). \tag{7}$$

In our framework, any rotation-invariant kernel satisfying the condition of Theorem 1 can be used to compute (7) and train a self-supervised model by minimizing (1). The uniformity term (7) can be interpreted as an energy functional (Brauchart et al., 2014): minimizing the average pairwise energy quantified by $\tilde{\mathcal{K}}$ tends to scatter evenly the embeddings on the hypersphere. We now give examples of kernels that can be used for this uniformity term. This illustrates that our framework offers a unification of several strategies for self-supervision.

**Example 1: RBF kernel.** Using $\mathcal{K}(u, v) = e^{-t\|u-v\|^2_2}$ (with $t > 0$) in the uniformity term $\ell_u$ (7) yields the regularization term from AUH, with the only difference that AUH uses the logarithm of the energy functional as their uniformity loss.

**Example 2: Generalized distance kernel.** It is defined as $\mathcal{K}(u, v) := C - \|u - v\|_2^{2s+g+1}$ with $\frac{s-1}{2} < s < \frac{s+1}{2}$ and $C > 0$ sufficiently large (Brauchart et al., 2014). A variation of this kernel choice is, e.g., used in the hard-contrastive loss of PointContrast for self-supervision on point clouds.

**Example 3: Truncations of the Laplace-Fourier series.** A truncated kernel up to order $L$ (Brauchart et al., 2014) is a kernel $\mathcal{K}(u, v) := \sum_{\ell=0}^{L} b_\ell P_\ell(q; u^\top v)$, with $b_\ell \geq 0$ for $\ell = 0, \ldots, L$. It admits a closed-form expression given by the definition of Legendre polynomials $P_\ell(q; \cdot)$ in Theorem 1, e.g., $P_1(q, t) = t$, $P_2(q, t) = \frac{q^2 t - 1}{q-1}$, $P_3(q, t) = \frac{(q+2)q^4 t^3 - 3q t}{q^4 - 1}$. We explore numerically this kernel choice in Section 4, since it has never been considered in previous self-supervision methods.
The expansion of \( \varphi \) in Legendre polynomials (5) for the RBF (Example 1) and the generalized distance kernel (Example 2) verifies \( b_2 > 0 \) for each integer \( \ell \) (see Appendix B.2). By (Micchelli et al., 2006, Theorem 10), this is a necessary and sufficient condition for a rotation-invariant kernel to be universal, and universality is a sufficient condition for injectivity of the corresponding kernel mean embedding mapping, i.e., the kernel is characteristic (Fukumizu et al., 2004). The benefit of this property is to guarantee that the uniform distribution \( \mathbb{U} \) is the unique solution to the minimization problem: \( \min \{ \text{MMD}(Q, \mathbb{U}) \mid Q \text { is a probability distribution on } S^{q-1} \} \). In contrast, the truncated rotation-invariant kernel up to an order \( L \) (Example 3) is not universal. Yet, our experiments in Section 4 show that truncated kernels up to order \( L = 2, 3 \) provide better results than, e.g., AUH whose uniformity loss is based on the RBF kernel.

In summary, the uniformity loss in our method, called SFRIK, corresponds to (7) with a truncated kernel up to order \( L = 3 \) and satisfies:

\[
\ell_u(\{z_i\}_{i \in I}) = \frac{1}{|I|^2} \sum_{i \in I} \sum_{i' \in I} \left( b_1 z_i^T z_{i'} + b_2 \frac{q(z_i^T z_{i'})^2 - 1}{q - 1} + b_3 \frac{(q + 2)(z_i^T z_{i'})^3 - 3z_i^T z_{i'}}{q - 1} \right),
\]

(8)

where \( b_\ell \geq 0, \ell = 1, 2, 3 \), are hyperparameters, and \( q \) is the dimension of the image embedding \( z_i \).

### 3.3 Connection with information-maximization methods

We now show that choosing an appropriate kernel in the proposed uniformity term (7) leads to a regularization that maximizes a statistical measure of information analog to the one used in VICReg. To the best of our knowledge, this is the first connection made between uniformity-based and information-maximization methods. The regularization loss of VICReg is a weighted sum between two terms:

\[
u(\mathbf{Z}_I) := \frac{1}{q} \sum_{j=1}^q \max \left( 0, \gamma - \sqrt{\text{Var}(z_j^I)} + \varepsilon \right), \quad c(\mathbf{Z}_I) := \frac{1}{q} \sum_{1 \leq j \neq j' \leq q} |C(\mathbf{Z}_I)_{j,j'}|^2,
\]

(9)

for a batch of image embeddings \( \mathbf{Z}_I := \{z_i\}_{i \in I} \), where \( z_j^I \) denotes the \( j \)-th coordinate of a (random) vector \( z \) and \( \varepsilon \) is a fixed small scalar. The variance term \( \nu(\mathbf{Z}_I) \) enforces the empirical variance \( \text{Var}(z_j^I) \) of the random variable \( z_j^I \) to be above a certain threshold \( \gamma^2 > 0 \) (here \( \bar{z} \) is the empirical mean of \( \mathbf{Z}_I \)). The covariance term \( c(\mathbf{Z}_I) \) enforces the non-diagonal entries of the empirical covariance matrix \( C(\mathbf{Z}_I) := \frac{1}{|I|} \sum_{i \in I} (z_i - \bar{z})(z_i - \bar{z})^\top \) to be zero.

In order to connect VICReg and SFRIK, let us consider for simplicity a truncated kernel \( \tilde{K}(u, v) = \sum_{\ell = 1}^L b_\ell P_\ell(q; u^\top v) \) of order \( L = 2 \) (the reasoning would be the same if the kernel was not truncated), and assume \( b_1, b_2 > 0 \). By the addition theorem (Müller, 2012, Theorem 2, §1), there exists a feature map \( \Phi : S^{q-1} \rightarrow \mathbb{R}^m \) involving an orthonormal basis of spherical harmonics (homogeneous harmonic polynomials restricted to the hypersphere) of order 1 and 2 such that \( \Phi(u)^\top \Phi(v) = \tilde{K}(u, v) \). Hence, the kernel mean embedding of a distribution in the associated RKHS contains its first and second-order moments (see Appendix B.3). Therefore, denoting \( N(q, \ell) \) the dimension of the space of spherical harmonics of order \( \ell \), dimension \( q \), and defining \( \Phi_\ell : S^{q-1} \rightarrow \mathbb{R}^{N(q, \ell)}, z \mapsto (Y_{\ell,k}(z))_{k=1}^{N(q, \ell)} \) for \( \ell \in \{1, 2\} \) with

\[
\{Y_{1,k}\}_{k=1}^{N(q,1)} := \left\{ u \mapsto u^j \mid 1 \leq j \leq q \right\}, \\
\{Y_{2,k}\}_{k=1}^{N(q,2)} := \left\{ u \mapsto u^j u^{j'} \mid 1 \leq j < j' \leq q \cup \left\{ u \mapsto (u^j)^2 - \frac{1}{q} \mid 2 \leq j \leq q \right\},
\]

(10)

it is possible to show (see Appendix B.3) that the squared MMD (6) can be written as

\[
\text{MMD}^2(Q, \mathbb{U}) = a_1 \|M_1 \mathbb{E}_{z \sim Q}[\Phi_1(z)]\|_2^2 + a_2 \|M_2 \mathbb{E}_{z \sim Q}[\Phi_2(z)]\|_2^2,
\]

(11)

where \( a_\ell := b_\ell |S^{q-1}|/N(q, \ell) \) for \( \ell \in \{1, 2\} \), and \( M_1, M_2 \) are two lower triangular matrices with nonzero diagonal entries. Hence, when \( Q \) plays the role of the embedding distribution during self-supervised training, minimizing \( \text{MMD}^2(Q, \mathbb{U}) \) in (11) as we propose for regularizing the embedding distribution promotes its expectation \( \mathbb{E}_{z \sim Q}[\Phi(z)] \) and its autocorrelation matrix \( \mathbb{E}_{z \sim Q}[\Phi(z)^\top \Phi(z)] \) to be close to 0 and \( q^{-1} I_q \), respectively, where \( I_q \) is the identity matrix. When \( \text{MMD}^2(Q, \mathbb{U}) = 0 \), the covariance matrix is equal to \( (\mathbb{E}[(z - \mathbb{E}[z])(z - \mathbb{E}[z])^\top] = \mathbb{E}[zz^\top] - \mathbb{E}[z]\mathbb{E}[z]^\top = \frac{1}{q} I_q \) because \( b_1, b_2 > 0 \) and the two terms on the right-hand side of (11) are null.
In conclusion, the regularization both in VICReg and SFRIK induces the embedding distribution to have a covariance matrix with zero non-diagonal entries. The diagonal entries of the covariance matrix are encouraged to be equal to $1/q$ in SFRIK, and greater than $\gamma^2$ in VICReg (we recall that the image embeddings $\{z_i\}_{i \in I}$ are not $\ell_2$-normalized in VICReg). However, one difference in terms of regularization behavior is that SFRIK encourages the expectation of the embedding distribution to be zero, as shown in the first term of (11). This is not the case for VICReg, as we can see in (9).

Finally, the memory and computational complexities for computing the uniformity term (8) in SFRIK are $O(|I|^2)$ and $O(q|I|^2)$, as opposed to $O(q^2)$ and $O(q^2|I|)$ for the variance and covariance terms (9) in VICReg. In the setting where SFRIK and VICReg work best, i.e., larger dimension $q$ and smaller batch size $|I|$, SFRIK has the lowest memory and computational complexities. This computational advantage is due to the kernel trick and is illustrated in Section 4.

4 Experiments

We first demonstrate numerically that the regularization loss (8) of SFRIK outperforms existing alternatives, in a rigorous experimental setting with a subset of ImageNet-1000 (Deng et al., 2009) for pretraining and a separate validation set for hyperparameter tuning. Then, we pretrain a ResNet-50 backbone (He et al., 2016) with SFRIK on the full ImageNet dataset and show competitive results compared to the state of the art, with significant computational benefits during pretraining.

4.1 Experimental setting

The backbone $f_a$ is either ResNet-18 or ResNet-50, depending on the experiment. Following Zbontar et al. (2021), the projection head $g_u$ is a three-layer MLP made of two hidden layers with ReLU activation and batch normalization (Ioffe & Szegedy, 2015), and a linear output layer. Unless otherwise specified, the size (number of neurons) of the two hidden layers is the same as the one, denoted $q$, of the output layer and the default value is $q = 8192$. The augmentations used for transforming images into views are the same as the ones used in VICReg. The backbone and the projection head are trained with a LARS optimizer (You et al., 2017). The weight decay is fixed at $10^{-6}$. The learning rate scheduling starts with 10 warm-up epochs (Goyal et al., 2017) with a linear increase from 0 to initial.lr = base.lr * bs/256, where base.lr is the base learning rate (Goyal et al., 2017) and bs is the batch size, followed by a cosine decay (Loshchilov & Hutter, 2017) with a final learning rate 1000 times smaller than initial.lr. For pretraining, we consider a 20% subset of ImageNet-1000 (denoted by IN20%), like in (Gidaris et al., 2021), and 100% of ImageNet-1000 (denoted by IN100%). In IN20%, we keep all the 1000 classes but only 260 images per class.

4.2 SFRIK’s regularizer outperforms existing alternative on IN20%

Many existing self-supervision methods are based on the Siamese architecture and have the same form of training loss $\lambda \ell_u(Z_i^{(1)}, Z_i^{(2)}) + \mu \ell_r(Z_i^{(1)}, Z_i^{(2)})$. This is the case of SimCLR, AUH and VICReg, for which Appendix B.4 gives the expression of the regularization loss $\ell_r$. For SFRIK, following (2), we have $\mu = 0.5$ and $\ell_r(Z_i^{(1)}, Z_i^{(2)}) = \ell_u(Z_i^{(1)}) + \ell_u(Z_i^{(2)})$ with $\ell_u$ given by (8).

Protocol. To isolate the impact of $\ell_r$ on the quality of the learned representations, we (re)implement all these four methods in the setting of Section 4.1, to get rid of the influence of other design choices, like image augmentations or projection head architecture. We fix the batch size at 2048, and tune the base learning rate and hyperparameters specific to each method’s loss. We also compare different embedding dimension $q \in \{1024, 2048, 4096, 8192\}$. In order to perform an extensive hyperparameter tuning by grid search of each method for fair comparisons, we choose a smaller backbone and a reduced dataset for pretraining, i.e., we pretrain a ResNet-18 on IN20% for 100 epochs with all methods. Pretrained backbones are then evaluated by linear probing trained on IN20% with labels.

Number of hyperparameters. Note that in total SFRIK with $L = 2$ has as many hyperparameters to tune as AUH or VICReg, and SFRIK with $L = 3$ has a single additional hyperparameter.

Rigorous hyperparameter tuning. In contrast to the common practice in the literature where hyperparameters are directly selected on the evaluation dataset, we choose to tune hyperparameters on
which the weights contrasts with the RBF and the generalized distance kernel for larger than $b$. Cated kernel performs well when the weights uniformity term (7). During tuning we observed that the truncated kernel performs better than the RBF kernel at order 2, 3 than on order 1 in the Legendre expansion (5). This suggests that it is important to focus more on order 2, 3 than on order 1 in the Legendre expansion (5).

Table 3 confirms empirically that a truncated kernel is better than the RBF\textsuperscript{†} or the generalized distance kernel for the uniformity term (7). During tuning we observed that the truncated kernel performs well when the weights $b_2, b_3$ in (8) are larger than $b_1$, e.g., $(b_1, b_2, b_3) = (1, 40, 40)$ for $q = 8192$. This contrasts with the RBF and the generalized distance kernel for which the weights $b_\ell$ decay polynomially with respect to $\ell$ (see Appendix B.2). This suggests that it is important to focus more on order 2, 3 than on order 1 in the Legendre expansion (5).

### 4.3 Results for ResNet-50 pretrained on IN100%

**Protocol.** We pretrain a ResNet-50 with SFRIK on IN100% under the setting of Section 4.1, with a batch size of 2048. We study the impact of a larger embedding dimension in SFRIK by considering a projection head with two hidden layers of size 8192, and an output layer of size $q \in \{8192, 16384, 32768\}$. Truncation order is either $L = 2$ or $L = 3$. For comparison, we also pretrain a ResNet-50 with VICReg under the same setting with $q = 8192$. Similarly to the original paper (Bardes et al., 2022), the alignment, variance and covariance weights are respectively 25, 25, 1, and the base learning rate is 0.2 for VICReg. All pretrained backbones are evaluated by: linear probing on IN100%; linear classification on Places205 and VOC2007 in order to measure how the learned representations generalize to an unseen dataset; and semi-supervised learning with few labels of IN100% (backbones are fine-tuned for classification using 1% or 10% of labeled images).

**Computational complexity.** We show under this protocol that SFRIK’s time and memory complexity during pretraining is significantly smaller than the one of VICReg for large dimensions. This allows us to scale SFRIK at dimension 16384 and even to 32768 for better results on downstream tasks.\textsuperscript{2} We measure the peak memory per GPU during pretraining on IN100% with a batch size of 2048 and the pretraining wall time of both methods on a 8× AMD Radeon Instinct MI50 32GB:

\textsuperscript{1}The performance gap between AUH and the RBF kernel is only due to the presence of the logarithm in AUH (cf. Example 1). Future work could clarify the role of this logarithm for regularization in self-supervision.

\textsuperscript{2}We recall that the time and memory complexity is identical for all methods on downstream tasks.
Table 4: Linear classification on IN100%, Places205, VOC2007, and semi-supervised learning with few labels of IN100% (accuracy or mean average precision). Methods are pretrained on IN100% with ResNet-50. We only include methods relying on a Siamese architecture with image augmentations limited to two views. The scores of methods marked with \( \ast \) are from Chen & He (2021). The score of VICReg\( \ast \) was obtained by retraining the model ourselves. For each downstream task, we highlight in bold the best score among all backbones pretrained on 200 epochs.

<table>
<thead>
<tr>
<th>Method</th>
<th>Epochs</th>
<th>IN100% Top-1</th>
<th>Places205 Top-1</th>
<th>VOC07 mAP</th>
<th>1% labels Top-1</th>
<th>1% labels Top-5</th>
<th>10% labels Top-1</th>
<th>10% labels Top-5</th>
</tr>
</thead>
<tbody>
<tr>
<td>SimCLR( \ast ) (Chen et al., 2020a)</td>
<td>200</td>
<td>68.3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SwAV( \ast ) (Caron et al., 2020) (no multi-crop)</td>
<td>200</td>
<td>69.1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SimSiam (Chen &amp; He, 2021)</td>
<td>200</td>
<td>70.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>VICReg( \ast ) (Bardes et al., 2022) ( q = 8192 )</td>
<td>200</td>
<td>70.0</td>
<td>89.3</td>
<td>54.1</td>
<td>83.4</td>
<td>84.9</td>
<td>49.4</td>
<td>75.1</td>
</tr>
<tr>
<td>SFRIK ( L = 2, q = 8192 )</td>
<td>200</td>
<td>70.1</td>
<td>89.3</td>
<td>53.8</td>
<td>83.0</td>
<td>85.1</td>
<td>46.6</td>
<td>73.3</td>
</tr>
<tr>
<td>SFRIK ( L = 3, q = 8192 )</td>
<td>200</td>
<td>70.2</td>
<td>89.6</td>
<td>54.5</td>
<td>83.9</td>
<td>84.6</td>
<td>46.9</td>
<td>73.6</td>
</tr>
<tr>
<td>SFRIK ( L = 2, q = 16384 )</td>
<td>200</td>
<td>70.3</td>
<td>89.6</td>
<td>54.3</td>
<td>83.4</td>
<td>85.2</td>
<td>46.0</td>
<td>73.0</td>
</tr>
<tr>
<td>SFRIK ( L = 2, q = 32768 )</td>
<td>200</td>
<td>70.3</td>
<td>89.6</td>
<td>54.1</td>
<td>83.0</td>
<td>85.0</td>
<td>46.1</td>
<td>73.0</td>
</tr>
<tr>
<td>SFRIK ( L = 3, q = 32768 )</td>
<td>200</td>
<td>70.3</td>
<td>89.7</td>
<td>54.4</td>
<td>83.2</td>
<td>85.1</td>
<td>46.6</td>
<td>73.0</td>
</tr>
<tr>
<td>SFRIK ( L = 2, q = 8192 )</td>
<td>400</td>
<td>70.8</td>
<td>89.9</td>
<td>54.4</td>
<td>83.5</td>
<td>85.7</td>
<td>47.8</td>
<td>74.3</td>
</tr>
</tbody>
</table>

- at \( q = 8192 \), SFRIK is 8% faster than VICReg and needs 3% less memory per GPU;
- at \( q = 16384 \), SFRIK is 19% faster than VICReg and needs 8% less memory per GPU;
- at \( q = 32768 \), SFRIK is still 2% faster than VICReg run in the lower dimension 16384. It only requires 30.9GB per GPU while VICReg at \( q = 32768 \) needs more than the available memory. Table 5 emphasizes this memory advantage at reduced batch sizes.

Results. Table 4 compares methods that have the same Siamese architecture and use the same image augmentations described in our protocol. For completeness, this table is completed in Appendix D.3 by evaluation results of other existing methods such as BYOL (Grill et al., 2020), OBoW (Gidaris et al., 2020) and SwAV with multi-crop (Caron et al., 2020), which are not comparable to the methods of Table 4 as they use a teacher-student architecture with momentum encoder and/or image augmentations with multi-scale cropping, and are beyond the setting of a Siamese architecture with only two views. Incorporating such designs in SFRIK is possible, and is left as a future work.

Table 4 demonstrates the competitiveness of SFRIK: it has the best accuracy for linear probing on IN100% among SimCLR, SwAV with no multi-crop, SimSiam and VICReg, and it performs better than VICReg for linear classification on Places205, VOC2007, and semi-supervised-learning with 10% of labels. We observe that SFRIK and VICReg offer a different trade-off between performance on linear probing on IN100% and performance on semi-supervised learning with 1% of labels. But as shown in Appendix D.3, other methods like BYOL and SwAV with multi-crop similarly have a performance drop compared to VICReg on semi-supervised learning with 1% of labels, even though they perform better on linear probing. Future work will therefore involve understanding what specific ingredients of VICReg make it more robust for semi-supervised learning with few labels. Ideally we could combine these ingredients with our generic kernel framework to design an improved version of SFRIK that can still benefit from its computational advantages over VICReg.

5 Conclusion

We proposed a regularization loss family based on the MMD and rotation-invariant kernels. We demonstrated that several regularizers of former methods are indeed variants of our flexible loss with different kernels. This generic regularization approach allowed us to leverage degrees of freedom in rotation-invariant kernel design to improve self-supervision methods. In practice, using a truncated kernel, we derived from the proposed framework a fully competitive self-supervised pretraining method, SFRIK, which significantly reduces time and memory complexity during pretraining compared to information-maximization methods. Combining the approach with kernel approximation techniques such as quadrature rules or random feature expansions offers promising perspectives to further enhance the ability to perform self-supervised training with limited computational resources.
ETHICS STATEMENT

The authors are concerned by the carbon footprint of deep learning research. To raise awareness of this issue in the community, we report the energy used for the computational resources of this project, which is approximately 12500 kWh.

REPRODUCIBILITY STATEMENT

In the interest of reproducible research, we provide our code and our pretrained ResNet-50 backbones with SFRIK on IN100% at https://github.com/valeoai/sfrik (Zheng et al., 2022). All details about our experimental setting can be found in either Section 4 or Appendix C. For the experiments on IN20%, as detailed in Appendix C.4, hyperparameters are tuned on a separate validation set different from the one used for evaluation, in contrast to the common practice in the literature where hyperparameters are directly selected on the evaluation dataset. All hyperparameters that yield the evaluation results reported in the paper are given in Appendix C.

ACKNOWLEDGMENTS

This project was funded by the CIFRE fellowship N°2020/1643 and supported in part by the Alle-groAssai ANR project ANR-19-CHIA-0009. This work was granted access to the HPC resources of IDRIS under the allocation 2021-AD011012940 made by GENCI. Experiments presented in this paper were carried out using the Grid’5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000.fr).

REFERENCES


Sergiy V. Borodachov, Douglas P. Hardin, and Edward B. Saff. Discrete energy on rectifiable sets. Springer, 2019. 3


Ting Chen, Calvin Luo, and Lala Li. Intriguing properties of contrastive losses. In NeurIPS, 2021. 3


Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In CVPR, 2009. 7, 25

Vincent Dutordoir, Nicolas Durrande, and James Hensman. Sparse Gaussian processes with spherical harmonic features. In ICML, 2020. 3


Priya Goyal, Quentin Duval, Jeremy Reizenstein, Matthew Leavitt, Min Xu, Benjamin Lefaudeux, Mannat Singh, Vinicius Reis, Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Ishan Misra. VISSL. https://github.com/facebookresearch/vissl, 2021. 21, 25

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In ALT, 2005. 14


Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In CVPR, 2020. 1, 3

Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In ICML, 2020. 3

R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In ICLR, 2019. 3


Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In ICLR, 2022. 1


Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In CVPR, 2017. 3


Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, 2020. 2


Saining Xie, Jiatao Gu, Demi Guo, Charles R. Qi, Leonidas Guibas, and Or Litany. PointContrast: Unsupervised pre-training for 3D point cloud understanding. In *ECCV*, 2020. 2


A Extended related work

We further discuss some related works referenced in the main paper.

A.1 Reminders on kernel mean embeddings

In this appendix we provide a high-level introduction to the notion of kernel mean embedding. We refer the reader to (Muandet et al., 2017) for a complete survey.

The idea of kernel mean embedding is to encode a probability distribution in an RKHS $\mathcal{H}$. Denoting $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ the reproducing kernel of $\mathcal{H}$ defined on some space $\mathcal{X}$, the kernel mean embedding of a probability distribution $Q$ defined on $\mathcal{X}$ is

$$
\mu_Q := \int_{\mathcal{X}} K(u, \cdot) dQ(u) \in \mathcal{H}.
$$

In other words, the kernel mean embedding mapping $Q \mapsto \mu_Q$ transforms a probability distribution into an element in $\mathcal{H}$. As an application, this allows one to quantify the divergence between probabilities using the norm $\| \cdot \|_\mathcal{H}$ associated to $\mathcal{H}$. Given two probability distributions $Q_1, Q_2$ defined on $\mathcal{X}$, one can indeed quantify their divergence by

$$
\|\mu_{Q_1} - \mu_{Q_2}\|_{\mathcal{H}} = \left\| \int_{\mathcal{X}} K(u, \cdot) dQ_1(u) - \int_{\mathcal{X}} K(u, \cdot) dQ_2(u) \right\|_{\mathcal{H}},
$$

which is precisely the MMD between $Q_1$ and $Q_2$ defined in (4).

A.2 Sample-contrastive criterion

Given a batch of embeddings $Z_I := \{z_i\}_{i \in I}$ (that are not necessarily normalized), the general sample-contrastive criterion of Garrido et al. (2023) is defined by:

$$
\ell_c(Z_I) = \sum_{i,i' \in I, i \neq i'} (z_i^\top z_{i'})^2.
$$

Garrido et al. (2023) show that this criterion is minimized in many contrastive learning methods like (HaoChen et al., 2021). In the case where the embeddings are normalized, this sample-contrastive criterion can be derived from the proposed generic uniformity loss $\ell_u$ defined by (7) with the quadratic kernel $K(u, v) = (u^\top v)^2$ where $u, v \in S^{q-1}$, as claimed in Section 1. Indeed, since $\|z_i\|_2 = 1$ for all $i \in I$:

$$
\ell_c(Z_I) = \sum_{i,i' \in I} (z_i^\top z_{i'})^2 - \sum_{i \in I} (z_i^\top z_i)^2 = |I|^2 \ell_u(Z_I) - |I|,
$$

where $|I|$ is the batch size. Therefore, the sample-contrastive criterion in the normalized case is an estimator of the MMD associated to the quadratic kernel between the embedding distribution and the uniform distribution on the hypersphere.

A.3 Kernel dependence maximization

We further explain the positioning of our paper with respect to (Li et al., 2021), which proposes a self-supervised learning method based on kernel dependence maximization, using the Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al., 2005). The HSIC measures the dependence between two random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ using two RKHS $F$ on $\mathcal{X}$ with kernel $k$ and $G$ on $\mathcal{Y}$ with kernel $l$, in order to capture nonlinear correlations. It is defined as the squared MMD associated to the reproducing kernel of the tensor product space $F \otimes G$ between the joint probability distribution $\mathbb{P}_{X,Y}$ and the product $\mathbb{P}_X \mathbb{P}_Y$ of marginal probability distributions:

$$
\text{HSIC}(X, Y) := \| \mu_{\mathbb{P}_{X,Y}} - \mu_{\mathbb{P}_X \mathbb{P}_Y} \|^2_{F \otimes G},
$$

where $Q \mapsto \mu_Q$ is the kernel mean embedding mapping defined by (12). Then, the self-supervised learning loss in (Li et al., 2021) is defined as:

$$
\mathcal{L}_{\text{SSL-HSIC}} := -\text{HSIC}(Z, Y) + \gamma \sqrt{\text{HSIC}(Z, Z)},
$$

where $\mathbb{P}$ is the uniform distribution on the hypersphere.
where $Z$ encodes embeddings of transformed images, and $Y$ encodes image identity as the index of the original image (before transformation) in the training dataset. By maximizing $\text{HSIC}(Z, Y)$, the backbone learns image representations that are invariant to image transformations. To avoid collapse, high-variance representations are penalized by minimizing $\text{HSIC}(Z, Z)$. This is similar to previous information-maximization methods (Bardes et al., 2022; Zhontar et al., 2021), with the difference that they take into account nonlinear correlations using kernels.

Although both our approach and the one in (Li et al., 2021) view self-supervised learning as a kernel method, we highlight here a main distinction between the two works. Both approaches use the MMD, but they do not use it to measure the same quantity. As explained above, Li et al. (2021) use the MMD to measure dependency between random variables (like $Z$ and $Y$), while the regularization loss we propose uses the MMD to measure the divergence between the embedding distribution and the uniform distribution on the hypersphere. As explained in Sections 1 and 2, this kernel approach for self-supervised learning is new in the literature and allows for the unification of several previous self-supervised learning methods as illustrated in Table 1.

Note that when the image identity $Y$ is a one-hot encoding, (Li et al., 2021) shows that
\[
- \text{HSIC}(Z, Y) = C \left( -E_{(Z_1, Z_2)\sim \text{pos}}[k(Z_1, Z_2)] + E_{Z_3, Z_4}[k(Z_3, Z_4)] \right),
\]
where $C > 0$ is a constant, $(Z_1, Z_2)$ is a positive pair of embeddings, i.e., they are embeddings of two transformations of the same original image, and $(Z_3, Z_4)$ is a pair of independent embeddings. In other words, $\text{HSIC}(Z, Y)$ is proportional to the sum of an alignment term $-E_{(Z_1, Z_2)\sim \text{pos}}[k(Z, Z')]$ and an energy term $E_{Z_3, Z_4}[k(Z_3, Z_4)]$, similarly to (3) combined with (7), which yields the proposed loss (2). Our paper shows that, if $k(\cdot, \cdot)$ is a rotation-invariant kernel on the hypersphere, then the energy term $E_{Z_3, Z_4}[k(Z_3, Z_4)]$ is precisely the MMD between the embedding distribution and the uniform distribution on the hypersphere, cf. (6). However there are two differences between the maximization of $\text{HSIC}(Z, Y)$ and the minimization of the proposed loss (2). First, the alignment term and the energy term in (18) are quantified with the same kernel $k(\cdot, \cdot)$, which is not the case in (2) where the alignment term is quantified by the $\ell_2$-distance between embeddings (equivalent to the linear kernel when the embeddings are normalized), and the uniformity term (7) is quantified by another rotation-invariant kernel. Second, the loss (2) is a weighted sum between the alignment loss (3) and the uniformity loss (7) controlled by the hyperparameter $\lambda$ that tunes the balance between the two terms, which is not the case of (18).

### B Theoretical results

We provide proofs and more details about the theoretical results in the main text.

#### B.1 Proof of Lemma 2

Consider a rotation-invariant kernel $\mathcal{K}(\mathbf{u}, \mathbf{v})$ defined on the hypersphere $S^{q-1}$ such that:
\[
\mathcal{K}(\mathbf{u}, \mathbf{v}) = \sum_{\ell=0}^{+\infty} b_{\ell} P_{\ell}(q; \mathbf{u}^\top \mathbf{v}), \quad \forall \mathbf{u}, \mathbf{v} \in S^{q-1},
\]
with weights $b_{\ell} \geq 0$ and $P_{\ell}(q; \cdot)$ the Legendre polynomial of order $\ell$ in dimension $q$. The proof of Lemma 2 relies on an orthonormal system of spherical harmonics. Let $(f, g)_{(q)} := \int_{S^{q-1}} \int_{S^{q-1}} f g d\sigma_{q-1}$ be the inner product in the space of continuous functions defined on $S^{q-1}$ and, for each $\ell \in \mathbb{N}$, consider $\{Y_{\ell, k} \mid k = 1, \ldots, N(q, \ell)\}$ an orthonormal basis of spherical harmonics of order $\ell$ in dimension $q$ (homogeneous harmonic polynomials in $q$ variables restricted to $S^{q-1}$, see e.g. (Müller, 2012) for more details), where $N(q, \ell)$ denotes the dimension of this space, which is by (Müller, 2012, Exercise 6, §3):
\[
N(q, \ell) = \begin{cases} 
q \frac{(2\ell+q-2)(\ell+q-3)!}{\ell!(q-2)!} & \text{for } \ell = 1, \\
\ell & \text{for } \ell \geq 2.
\end{cases}
\]

By the addition theorem (Müller, 2012, Theorem 2, §1):
\[
\sum_{k=1}^{N(q, \ell)} Y_{\ell, k}(\mathbf{u}) Y_{\ell, k}(\mathbf{v}) = \frac{N(q, \ell)}{|S^{q-1}|} P_{\ell}(q; \mathbf{u}^\top \mathbf{v}), \quad \mathbf{u}, \mathbf{v} \in S^{q-1}.
\]
Hence, the kernel $K(u, v)$ can be rewritten as:

$$K(u, v) = \sum_{\ell=0}^{+\infty} \sum_{k=1}^{N(q, \ell)} \frac{b_\ell |S_{\ell}^{-1}|}{N(q, \ell)} Y_{\ell,k}(u) Y_{\ell,k}(v).$$

(22)

Since $\{Y_{\ell,k} \mid \ell = 0, \ldots, +\infty, k = 1, \ldots, N(q, \ell)\}$ is an orthonormal system for the inner product $(\cdot, \cdot)_q$ and since $Y_{0,1}$ is constant on $S_{\ell}^{-1}$, we have for any integer $\ell$ and $k \in \{1, \ldots, N(q, \ell)\}$ that:

$$\int_{S_{\ell}^{-1}} Y_{\ell,k} \mathrm{d}\sigma_{\ell-1} = \frac{1}{Y_{0,1}} (Y_{\ell,k}, Y_{0,1})_q = \begin{cases} \frac{1}{Y_{0,1}} & \text{if } \ell = 0, k = 1 \\ 0 & \text{otherwise} \end{cases}.$$  

(23)

Moreover $Y_{0,1} = 1/\sqrt{|S_{\ell}^{-1}|}$, because $1 = (Y_{0,1}, Y_{0,1})_q = \int_{S_{\ell}^{-1}} Y_{0,1}^2 \mathrm{d}\sigma_{\ell-1} = Y_{0,1}^2 |S_{\ell}^{-1}|$. Therefore, the kernel mean embedding of the uniform distribution on the hypersphere $U := \sigma_{\ell-1}/|S_{\ell}^{-1}|$ associated to the kernel $K$ is:

$$\forall v \in S_{\ell}^{-1}, \int_{S_{\ell}^{-1}} K(u, v) \mathrm{d}\mu(u) = \frac{1}{Y_{0,1}} \int_{S_{\ell}^{-1}} \sum_{\ell=0}^{+\infty} \sum_{k=1}^{N(q, \ell)} \frac{b_\ell}{N(q, \ell)} Y_{\ell,k}(u) Y_{\ell,k}(v) \mathrm{d}\sigma_{\ell-1}(u)$$

$$= \sum_{\ell=0}^{+\infty} \int_{S_{\ell}^{-1}} \sum_{k=1}^{N(q, \ell)} \frac{b_\ell}{N(q, \ell)} Y_{\ell,k}(u) Y_{\ell,k}(v) \mathrm{d}\sigma_{\ell-1}(u)$$

$$= \sum_{\ell=0}^{+\infty} \frac{b_\ell}{N(q, \ell)} \sum_{k=1}^{N(q, \ell)} \int_{S_{\ell}^{-1}} Y_{\ell,k}(u) \mathrm{d}\sigma_{\ell-1}(u) Y_{\ell,k}(v)$$

(24)

$$= b_0 \frac{1}{Y_{1,0}} Y_{1,0} = b_0,$$

where we inverted series and integral in the second equation using the dominated convergence theorem: the series $\sum_{\ell=0}^{+\infty} \frac{b_\ell}{N(q, \ell)} Y_{\ell,k}(u) Y_{\ell,k}(v)$ converges for every $u, v$, and for any $L$, $|\sum_{\ell=0}^{L} b_\ell P_\ell(q; u^\top v)| \leq \sum_{\ell=0}^{+\infty} b_\ell = \sum_{\ell=0}^{+\infty} b_\ell P_\ell(q; 1)$, because $|P_\ell(q; \cdot)| \leq 1$ for all $\ell$ by (Müller, 2012, Lemma 2, §8), $P_\ell(q; 1) = 1$ for all $\ell$ by (Müller, 2012, §9), and $\sum_{\ell=0}^{+\infty} b_\ell P_\ell(q; 1) < +\infty$ is integrable on $S_{\ell}^{-1}$. This yields the first claim of Lemma 2.

Consider now any probability distribution $Q$ defined on the hypersphere. The kernel mean embedding of $Q$ is simply rewritten as:

$$\forall v \in S_{\ell}^{-1}, \int_{S_{\ell}^{-1}} K(u, v) \mathrm{d}Q(u) = \int_{S_{\ell}^{-1}} \sum_{\ell=0}^{+\infty} \sum_{k=1}^{N(q, \ell)} \frac{b_\ell}{N(q, \ell)} P_\ell(q; u^\top v) \mathrm{d}Q(u)$$

$$= \sum_{\ell=0}^{+\infty} \frac{b_\ell}{N(q, \ell)} \sum_{k=1}^{N(q, \ell)} \int_{S_{\ell}^{-1}} P_\ell(q; u^\top v) \mathrm{d}Q(u)$$

$$= b_0 + \int_{S_{\ell}^{-1}} \tilde{K}(u, v) \mathrm{d}Q(u),$$

(25)

because the Legendre polynomial of order 0 is the constant function equal to 1 (see the closed form expression of $P_0(q; \cdot)$ in Theorem 1 in the main text) and $\int_{S_{\ell}^{-1}} \mathrm{d}Q = 1$. This ends the proof of Lemma 2.

### B.2 LEGENDRE EXPANSION OF ROTATION-INvariant Kernels

We show that the kernel weights $b_\ell$ in the Legendre expansion (5) of the RBF kernel and the generalized distance kernel decay with a rate at least polynomial with respect to $\ell$.

**RBF kernel** The RBF kernel is defined as:

$$K_{\text{RBF}}(u, v) = e^{-\sigma |u - v|^2} = e^{-2\sigma(1-u^\top v)} \quad \text{for } u, v \in S_{\ell},$$

(26)
where $\sigma > 0$ is the scale of the RBF kernel. Denote $\varphi(t) := e^{-2\sigma(1-t)}$ for $t \in [-1, 1]$. Since the RBF kernel is positive definite and rotation-invariant, by Theorem 1, there exist weights $b_\ell \geq 0$, $\ell = 0, \ldots, +\infty$, such that:

$$\varphi(t) = e^{-2\sigma(1-t)} = \sum_{\ell=0}^{+\infty} b_\ell P_\ell(q; t) \quad \text{for } t \in [-1, 1]. \quad (27)$$

The Legendre polynomials $P_\ell(q; \cdot)$ are orthogonal on the interval $[-1, 1]$ with respect to the weight function $(1 - t^2)^{-\frac{\sigma}{2}}$, see e.g. (Müller, 2012):

$$\int_{-1}^{1} P_n(q; t) P_m(q; t)(1 - t^2)^{-\frac{\sigma}{2}} dt = 0 \quad \text{for } m \neq n. \quad (28)$$

Moreover, by (Müller, 2012, Exercise 3, §2):

$$\int_{-1}^{1} (P_n(q; t))^2 (1 - t^2)^{-\frac{\sigma}{2}} dt = \frac{|S^{q-1}|}{|S^{q-2}|} \frac{1}{N(q, n)} \quad \text{for any } n. \quad (29)$$

We multiply (27) by $P_\ell(q; t)(1 - t^2)^{-\frac{\sigma}{2}}$ and integrate the equation on $[-1, 1]$:}

$$\int_{-1}^{1} \varphi(t) P_\ell(q; t)(1 - t^2)^{-\frac{\sigma}{2}} dt = \int_{-1}^{1} \sum_{n=0}^{+\infty} b_n P_n(q; t) P_\ell(q; t)(1 - t^2)^{-\frac{\sigma}{2}} dt$$

$$= \sum_{n=0}^{+\infty} b_n \int_{-1}^{1} P_n(q; t) P_\ell(q; t)(1 - t^2)^{-\frac{\sigma}{2}} dt$$

$$= b_\ell \frac{|S^{q-1}|}{|S^{q-2}|} \frac{1}{N(q, \ell)}, \quad (30)$$

where the inversion between series and integral is justified by the dominated convergence theorem: the series $\sum_{n=0}^{+\infty} b_n P_n(q; t) P_\ell(q; t)(1 - t^2)^{-\frac{\sigma}{2}}$ converges for every $t$, and for any $N$, $|\sum_{n=0}^{N} b_n P_n(q; t) P_\ell(q; t)(1 - t^2)^{-\frac{\sigma}{2}}| \leq \sum_{n=0}^{+\infty} b_n (1 - t^2)^{-\frac{\sigma}{2}} := g(t)$ since $|P_n(q; \cdot)| \leq 1$ for any $n$ by (Müller, 2012, Lemma 2, §8), and $g$ is integrable on $[-1, 1]$ because $\sum_{n=0}^{+\infty} b_n < +\infty$. Hence:

$$b_\ell = N(q, \ell) \frac{|S^{q-2}|}{|S^{q-1}|} \int_{-1}^{1} \varphi(t) P_\ell(q; t)(1 - t^2)^{-\frac{\sigma}{2}} dt. \quad (31)$$

By the Rodrigues rule (Müller, 2012, Exercise 1, §2), since $\varphi$ has continuous derivatives of all orders on $[-1, 1]$, we have:

$$b_\ell = N(q, \ell) \frac{|S^{q-2}|}{|S^{q-1}|} \frac{\Gamma\left(\frac{q-1}{2}\right)}{2^\ell \Gamma\left(\ell + \frac{q-1}{2}\right)} \int_{-1}^{1} \varphi^{(\ell)}(t)(1 - t^2)^{\frac{\ell}{2} + \frac{\sigma-1}{2}} dt, \quad \ell \in \mathbb{N}, \quad (32)$$

where $\varphi^{(\ell)}$ is the $\ell$-th derivative of $\varphi$, which is $\varphi^{(\ell)}(t) = e^{-2\sigma}(2\sigma)^\ell e^{2\sigma t}$. We now show that the weights $b_\ell$ decay very fast with respect to $\ell$. We bound the integral:

$$\int_{-1}^{1} \varphi^{(\ell)}(t)(1 - t^2)^{\frac{\ell}{2} + \frac{\sigma-1}{2}} dt = \int_{-1}^{1} e^{-2\sigma (2\sigma)^\ell} e^{2\sigma t}(1 - t^2)^{\frac{\ell}{2} + \frac{\sigma-1}{2}} dt \leq \int_{-1}^{1} (2\sigma)^\ell dt = 2(2\sigma)^\ell. \quad (33)$$

Hence:

$$b_\ell \leq 2 N(q, \ell) \frac{|S^{q-2}|}{|S^{q-1}|} \frac{\Gamma\left(\frac{q-1}{2}\right)}{2^\ell \Gamma\left(\ell + \frac{q-1}{2}\right)} (2\sigma)^\ell. \quad (34)$$

Denote $(a)_n := \Gamma(n + a)/\Gamma(a)$ the Pochhammer symbol defined for any integer $n$ and any scalar $a$. By the Stirling approximation of the Gamma function (Spiegel et al., 2013, (25.15)), the asymptotic behavior of $(a)_n$ when $n$ goes to infinity is:

$$(a)_n \sim \frac{\sqrt{2\pi}}{\Gamma(a)} e^{-n} n^{a-n-1/2} \quad \text{as } n \to \infty. \quad (35)$$
Moreover, for a fixed dimension $q$, the asymptotic behavior of $N(q, \ell)$ defined by (20) when $\ell$ goes to infinity is:

$$N(q, \ell) \sim \frac{2}{(q - 2)!} \ell^{q-2} \quad \text{as } \ell \to \infty. \quad (36)$$

Therefore, the asymptotic behavior of $b_\ell$ as $\ell$ goes to infinity is:

$$b_\ell = O \left( \sigma^q \ell^{q/2 - 1 - \ell} \right) \quad \text{as } \ell \to +\infty. \quad (37)$$

**Generalized distance kernel** For $\frac{q - 1}{2} < s < \frac{q + 1}{2}$, the generalized distance kernel on the hypersphere $S^{q-1}$ is defined in (Brauchart et al., 2014, Section 5) as:

$$K^{(s)}_{gd}(u, v) := 2V_{q-1-2s}(S^{q-1}) - \|u - v\|_2^{2s+1} \quad \text{for } u, v \in S^{q-1}, \quad (38)$$

where

$$V_{q-1-2s}(S^{q-1}) := \int_{S^{q-1}} \int_{S^{q-1}} \|u - v\|_2^{2s+1} \, d\sigma_{q-1}(u) \, d\sigma_{q-1}(v) = 2^{q-1} \frac{\Gamma(q/2)\Gamma(s)}{\sqrt{\pi}\Gamma((q - 1)/2 + s)}. \quad (39)$$

Following Brauchart et al. (2014, Section 5), the Legendre expansion of the generalized distance kernel $K^{(s)}_{gd}$ is:

$$K^{(s)}_{gd}(u, v) = V_{q-1-2s}(S^{q-1}) + \sum_{\ell=1}^{+\infty} \alpha^{(s)}_\ell N(q, \ell) P_\ell(q; u^\top v), \quad (40)$$

$$\alpha^{(s)}_\ell := -V_{q-1-2s}(S^{q-1}) \frac{((q - 1)/2 - s)\ell}{((q - 1)/2 + s)}, \quad \ell \geq 1. \quad (41)$$

The kernel weights indeed decay polynomially with respect to $\ell$, because according to Brauchart et al. (2014, Section 5), the asymptotic behavior of the $\alpha^{(s)}_\ell$ is:

$$\alpha^{(s)}_\ell \sim 2^{q-1} \frac{\Gamma(q/2)\Gamma(s)}{\sqrt{\pi}\Gamma((q - 1)/2 + s)} \ell^{-2s} \quad \text{as } \ell \to +\infty. \quad (42)$$

**B.3 Connection between SFRIK and VICReg**

Consider a rotation-invariant kernel $\tilde{K}(u, v) := \sum_{\ell=1}^{+\infty} b_\ell P_\ell(q; u^\top v)$ defined on $S^{q-1}$ such that $b_\ell \geq 0$ for $\ell \in \{1, \ldots, +\infty\}$, with $b_1, b_2 > 0$. To show the connection between SFRIK and VICReg, we construct a high-dimensional feature map $\Phi : S^{q-1} \to \ell_2(\mathbb{N})$, where $\ell_2(\mathbb{N})$ denotes the space of square-summable sequences with its canonical inner product $\langle \cdot, \cdot \rangle_{\ell_2}$, such that $\langle \Phi(u), \Phi(v) \rangle_{\ell_2} = \tilde{K}(u, v)$ for any $u, v \in S^{q-1}$.

One way to construct such a feature map is to consider an orthonormal system of spherical harmonics. For any integer $\ell$, denote $\{Y_{\ell,k}\}_{\ell=1}^{N(q,\ell)}$ an orthonormal basis of spherical harmonics of order $\ell$ in dimension $q$. By the addition theorem (Müller, 2012, Theorem 2, §1) recalled in (21), the kernel $\tilde{K}(u, v)$ admits the decomposition:

$$\tilde{K}(u, v) = \sum_{\ell=1}^{+\infty} b_\ell P_\ell(q; u^\top v) = \sum_{\ell=1}^{+\infty} \sum_{k=1}^{N(q,\ell)} b_\ell |S^{q-1}|^{-1} N(q, \ell) Y_{\ell,k}(u) Y_{\ell,k}(v) \quad \text{as } \ell \to +\infty, \quad (43)$$

where

$$\Phi := \left( \sqrt{\frac{b_\ell |S^{q-1}|^{-1}}{N(q, \ell)}} \Phi_\ell \right)_{\ell=1}^{+\infty} \quad \text{with} \quad \Phi_\ell : \left\{ S^{q-1} \to \mathbb{R}^{N(q,\ell)} \quad \text{for } \ell \in \{1, \ldots, +\infty\}. \quad (44)$$

Then, the MMD in (6) between any probability distribution $\mathbb{Q}$ defined on the hypersphere and the uniform distribution $\mathbb{U}$ on the hypersphere can be written as the norm in $\ell_2(\mathbb{N})$ of the generalized
moment of $Q$ with the mapping $\Phi$: 
\[
\text{MMD}^2(Q, U) = \mathbb{E}_{z, z' \sim Q}[\mathcal{K}(z, z')]
= \mathbb{E}_{z, z' \sim Q}[\langle\Phi(z), \Phi(z')\rangle_{\ell_2}]
= \langle \mathbb{E}_{z \sim Q}[\Phi(z)], \mathbb{E}_{z' \sim Q}[\Phi(z')]\rangle_{\ell_2}
= \|\mathbb{E}_{z \sim Q}[\Phi(z)]\|^2_{\ell_2}
= \sum_{\ell=1}^{+\infty} \frac{b_{\ell}|S^{q-1}|}{N(q, \ell)} \|\mathbb{E}_{z \sim Q}[\Phi_{\ell}(z)]\|^2_{\ell_2}.
\tag{45}
\]

We now explain how to construct explicitly an orthonormal basis of spherical harmonics $\{Y_{\ell, k} \mid \ell = 1, \ldots, +\infty; k = 1, \ldots, N(q, \ell)\}$, based on the following theorem.

**Theorem 3 (Axler et al. (2013, Theorem 5.25))** For any order $\ell \in \mathbb{N}$ and any dimension $q \geq 3$, the family 
\[
\{Y_{\ell, k}^{N(q, \ell)}\}_{k=1}^{N(q, \ell)} := \left\{ u \mapsto \partial_{\ell_1}^{\alpha_1} \partial_{\ell_2}^{\alpha_2} \cdots \partial_{\ell_q}^{\alpha_q} \|u\|_{2, q}^{2-q} \mid \alpha_1 + \alpha_2 + \ldots + \alpha_q = \ell \text{ and } \alpha_j \leq 1 \right\}
\tag{46}
\]
is a (non-orthonormal) basis of the space of spherical harmonics of order $\ell$ in dimension $q$, where $\alpha_j$ ($j = 1, \ldots, q$) are nonnegative integers, and $\partial_{\ell_j}$ denotes the $\ell_j$-th partial derivative with respect to the $j$-th coordinate.

Typically, we construct the orthonormal basis $\{Y_{\ell, k}^{N(q, \ell)}\}_{k=1}^{N(q, \ell)}$ by orthonormalizing the basis $\{Y_{\ell, k}^{1N(q, \ell)}\}_{k=1}^{N(q, \ell)}$ of Theorem 3 using, e.g., the Gram-Schmidt procedure. For $\ell = 1, \ldots, +\infty$, denote:
\[
\Phi_{\ell} : S^{q-1} \rightarrow \mathbb{R}^{N(q, \ell)}, \ u \mapsto \left( Y_{\ell, k}(u) \right)_{k=1}^{N(q, \ell)}.
\tag{47}
\]

Then, for each $\ell = 1, \ldots, +\infty$, there exists a lower triangular matrix $M_{\ell}$ such that:
\[
\Phi_{\ell}(u) = M_{\ell} \Phi'_{\ell}(u), \quad \text{for all } u \in S^{q-1}.
\tag{48}
\]

Remark that it is possible to compute explicitly the entries of the matrices $M_{\ell}$, $\ell = 1, \ldots, +\infty$, because there exists a closed-form expression for the inner product $Q_{\ell,k}$ for any $\ell, k, k'$; indeed, the function $Y_{\ell, k}$ for any $\ell, k$ is a polynomial defined on the hypersphere, and the integral of any monomial with respect to the measure $\sigma_{q-1}$ on the hypersphere $S^{q-1}$ admits a closed-form expression given by Weyl (1939, Section 3).

By injecting (48) in (45), we obtain:
\[
\text{MMD}^2(Q, U) = \sum_{\ell=1}^{+\infty} \frac{b_{\ell}|S^{q-1}|}{N(q, \ell)} \|M_{\ell}\|_{\ell_2}^2.
\tag{49}
\]

This yields the claim of Section 3.3 by remarking with Theorem 3 that the families 
\[
\left\{Y_{1, k}^{N(q, 1)} \mid 1 \leq j \leq q \right\},
\left\{Y_{2, k}^{N(q, 2)} \mid 1 \leq j < j' \leq q \right\} \cup \left\{u \mapsto (u^j)^2 - \frac{1}{q} \mid 2 \leq j \leq q \right\}
\tag{50}
\]
are bases of the space of spherical harmonics of order 1 and 2 in dimension $q$.

### B.4 Regularization Loss of SimCLR, AUH and VICReg

In SimCLR, AUH, VICReg and SFRIK, each image $x_i$ in a batch $\{x_i\}_{i \in I}$ is augmented into two different views $x_i^{(1)}$ and $x_i^{(2)}$, which are encoded into two embeddings $z_i^{(1)}$ and $z_i^{(2)}$. These embeddings are normalized in SimCLR, AUH and SFRIK, but not in VICReg. This yields two batches of embeddings $Z_i^{(u)} := \{z_i^{(u)}\}_{i \in I}$ ($u = 1, 2$). The four methods share the same form of loss function:
\[
\ell(Z_i^{(1)}, Z_i^{(2)}) := \lambda \ell_{\text{a}}(Z_i^{(1)}, Z_i^{(2)}) + \mu \ell_r(Z_i^{(1)}, Z_i^{(2)}),
\tag{51}
\]
for some scalars $\lambda, \mu > 0$, where $\ell_{\text{a}}$ is the alignment loss defined by (3) (which is the same for all the four methods), and $\ell_r$ is the regularization loss specific to each method.
SimCLR

The regularization loss in SimCLR is:

\[
\ell_r(Z^{(1)}_I, Z^{(2)}_I) = \frac{1}{2|\mathcal{I}|^2} \sum_{v=1}^{2} \sum_{i \in \mathcal{I}} \log \left( \sum_{i' \in \mathcal{I}} \exp \left( -t \|z^{(v)}_i - z^{(v')(i')} \|^2_2 \right) \right),
\]

where \( t > 0 \) is a hyperparameter called the scale of the RBF kernel. The scalar \( \lambda \) is tuned as a hyperparameter and \( \mu \) is fixed at \( \mu = 1 \).

Alignment & Uniformity

The regularization loss in AUH is:

\[
\ell_r(Z^{(1)}_I, Z^{(2)}_I) = \frac{1}{2|\mathcal{I}|^2} \sum_{v=1}^{2} \sum_{i \in \mathcal{I}} \log \left( \sum_{i' \in \mathcal{I}} \exp \left( -t \|z^{(v)}_i - z^{(v')(i')} \|^2_2 \right) \right),
\]

where \( t > 0 \) is a hyperparameter called the scale of the RBF kernel. The scalar \( \lambda \) is tuned as a hyperparameter and \( \mu \) is fixed at \( \mu = 1 \).

VICReg

As introduced in Section 3.3, the regularization loss in VICReg is:

\[
\ell_r(Z^{(1)}_I, Z^{(2)}_I) = \frac{1}{2} \left[ v(Z^{(1)}_I) + v(Z^{(2)}_I) \right] + \frac{1}{2\mu} \left[ c(Z^{(1)}_I) + c(Z^{(2)}_I) \right],
\]

where \( \mu \) is the scalar from (51). Here, \( v(\cdot) \) and \( c(\cdot) \) respectively the variance and covariance terms defined by (9). Both \( \lambda \) and \( \mu \) are tuned as hyperparameters.

C EXPERIMENTAL SETTING

In the interest of reproducible research, we give more details about the setting of our experiments presented in Section 4.

C.1 IN20% DATASET DESCRIPTION

The datasets used in our experiments include a subset of 20% of ImageNet-1000 as in (Gidaris et al., 2021). This reduced dataset, denoted IN20%, contains all the 1000 classes of ImageNet, but we keep only 260 images per class. The 260 images extracted are the same as those extracted in the official implementation of OBoW (https://github.com/valeoai/obow). In Section 4.2, we also use another 20% subset of the ImageNet train set as a separate validation set for hyperparameter tuning (see Appendix C.4 below). The construction of this validation set is based on the code of OBoW, and will be exactly detailed in our code that will be released at publication.

C.2 IMAGE AUGMENTATIONS

We follow the same image augmentation pipeline as in (Bardes et al., 2022). Our experiments include the following image augmentations implemented by PyTorch (torchvision.transforms):

- RandomResizedCrop(224, scale=(0.08, 1.0)): crop a random area of the image between 8% and 100% of the total area, and resize it to an image of size 224 × 224;
- RandomHorizontalFlip(): flip horizontally an image;
- ColorJitter(brightness=0.4, contrast=0.4, saturation=0.2, hue=0.1): randomly change brightness, contrast, saturation and hue of an image by a factor randomly sampled in respectively [0.6, 1.4], [0.6, 1.4], [0.8, 1.2] and [−0.1, 0.1];
- RandomGrayscale(): convert an image into grayscale.

We also use image augmentations implemented by PIL, as in VICReg’s code available at https://github.com/facebookresearch/vicreg:

- GaussianBlur(): blur an image using a Gaussian kernel with standard deviation uniformly sampled in [0.1, 2.0];
- Solarization(): randomly invert all pixel values above a threshold, which is 130.
In our experiments, the first image view is obtained by composing the following random augmentations: random cropping resized to $224 \times 224$, random horizontal flip applied with probability 0.5, random color jittering applied with probability 0.8, random grayscale conversion applied with probability 0.2, random Gaussian blur applied with probability 0.1, and random solarization applied with probability 0.2. The second view is obtained by composing the same random augmentations as the first view, except that Gaussian blur is applied every time (probability 1), and solarization is never applied (probability 0).

**C.3 Evaluation protocol**

We describe the downstream tasks on which self-supervision methods are evaluated in our experiments of Section 4.

**Linear probing on IN20\% and IN100\%**  
Following, e.g., (Bardes et al., 2022), the weights of the backbone (ResNet-18 or ResNet-50) are frozen and a linear layer followed by a softmax on top of the backbone is trained in a supervised setting on a training set. Then the model is evaluated on a test set. The training set is either IN20\% or IN100\%, but with labels. The test set is the validation set of ImageNet. The linear layer is trained using an SGD optimizer with momentum parameter equal to 0.9 during 100 epochs. We apply a weight decay of $10^{-6}$. The learning rate follows a cosine decay scheduling. The batch size is fixed at 256. Training images are augmented by composing a random cropping of an area between 8\% and 100\% of the total area resized to $224 \times 224$, and a random horizontal flip of probability 0.5. Images at test time are resized to $256 \times 256$, and cropped at the center with a size $224 \times 224$. The initial learning rate is tuned as a hyperparameter, and we report the top-1 accuracy on the validation set of ImageNet obtained after the last training epoch, along with the corresponding top-5 accuracy. The code that we use for linear probing on IN20\% or IN100\% is adapted from (Bardes et al., 2022) available at [https://github.com/facebookresearch/vicreg](https://github.com/facebookresearch/vicreg).

**Linear probing on Places205**  
We use the code of Gidaris et al. (2021), available at [https://github.com/valeoai/obow](https://github.com/valeoai/obow), for the evaluation by linear probing on Places205. The weights of the backbone (ResNet-50) pretrained on IN100\% are frozen and a linear prediction layer is trained for the classification task on Places205. We note that a batch normalization layer with non-learnable scale and bias parameters is added at the output of the backbone in (Gidaris et al., 2021). The linear prediction layer is trained with an SGD optimizer with a 0.9 momentum parameter during 28 epochs. The weight decay is $10^{-4}$. The batch size is 256. The learning rate decreases by a factor of 10 at epoch 10 and epoch 20. We use the same image augmentation pipeline for training and testing as in linear probing on IN100\%. The initial learning rate is tuned as a hyperparameter, and we report the top-1 accuracy on the validation set of Places205 obtained after the last training epoch, along with the corresponding top-5 accuracy.

**Linear classification on VOC2007**  
After pretraining a ResNet backbone, we use the VISSL library (Goyal et al., 2021) to extract features of VOC2007 images resized to $224 \times 224$ by taking the output of the last average pooling layer of the pretrained ResNet backbone. We then learn a linear SVM with LIBLINEAR (Fan et al., 2008) on top of these features to predict the presence or the absence of a given class in the test images. An average precision score is then computed for each class after a 3-fold cross-validation, and we report the mean score over all classes as the mean average precision (mAP).

**Semi-supervised learning**  
After pretraining a ResNet backbone by self-supervision, we fine-tune this backbone and the linear classifier on the ImageNet classification task with only 1\% or 10\% of the labeled data. The labeled images that are considered in these subsets are the ones used in the official code of SimCLR available at [https://github.com/google-research/simclr](https://github.com/google-research/simclr). We use an SGD optimizer with momentum parameter equal to 0.9 during 20 epochs, without weight decay. The batch size is fixed at 256. The learning rates of the backbone and the linear classifier follow a cosine decay scheduling with different initial learning rates. These initial learning rates are tuned as hyperparameters. We report the top-1 accuracy on the validation set of ImageNet obtained after the last training epoch, along with the corresponding top-5 accuracy. We use the same image augmentation pipeline for training and testing as in linear probing on IN100\%. The code that we use for semi-supervised learning with few labels of IN100\% is the one of Bardes et al. (2022) available at [https://github.com/facebookresearch/vicreg](https://github.com/facebookresearch/vicreg).
Weighted kNN classification  We follow the usual protocol of Wu et al. (2018); Caron et al. (2021). We compute the normalized representations \( f_\theta(x_i) \) of the images \( x_i, i \in [N] \), in the training set. The label of an image \( x_{\text{test}} \) in the test set is predicted by a weighted vote of its \( k \) nearest neighbors \( N_k \) in the representation space: the class \( c \) gets a score of \( w_c := \sum_{i \in N_k} \exp(f_\theta(x_i)^T f_\theta(x_{\text{test}})/0.07) \mathbb{1}_{[c_i = c]} \) where \( f_\theta(x_{\text{test}}) \) is normalized, \( c_i \) is the class of \( x_i \), and \( \mathbb{1}_{[c_i = c]} \) is equal to 1 if \( c_i = c \), and 0 otherwise. We report the kNN classification top-1 accuracy for \( k = 20 \). The image augmentation pipeline for both training and testing is the following one: images are resized to \( 256 \times 256 \), and cropped at the center with a size \( 224 \times 224 \). The code that we use for kNN classification is the one of Caron et al. (2021) available at https://github.com/facebookresearch/dino.

C.4 HYPERPARAMETERS FOR EXPERIMENTS ON IN20%

We describe in detail our hyperparameter tuning protocol for the experiments in Section 4.2 on IN20%. For a rigorous tuning, it is important that the dataset used for the final evaluation remains unseen during pretraining and hyperparameter tuning. For each pretraining method, we pretrain on the IN20% training set (blue subset in Figure 2) a backbone for each choice of hyperparameters. These backbones are then evaluated by weighted kNN classification on a separate validation set, which is another 20% subset of the ImageNet train set (purple subset in Figure 2), and we select the hyperparameters yielding the highest top-1 accuracy on this kNN evaluation. Then, we tune the learning rate for the linear probing evaluation, again on our separate validation set (purple subset in Figure 2). Finally, we use the model trained with the best learning rate discovered for linear probing evaluation on the usual ImageNet validation set (red subset in Figure 2), which has never been seen during hyperparameter tuning.

We report the values of the optimal hyperparameters found after our hyperparameter tuning on a separate validation set, for each pretraining experiment on IN20% with a ResNet-18 presented in Section 4.2. These hyperparameters yield the evaluation results reported in Section 4.2 for linear probing on the usual ImageNet validation set. We recall that the hyperparameters specific to each self-supervision method was introduced in Appendix B.4.

SimCLR. For each embedding dimension, we fix the batch size at 2048, and tune the temperature \( \tau \) and the base learning rate \( \text{base}\_lr \) for pretraining with SimCLR. Then, we tune the initial learning rate \( \text{lr}\_\text{head} \) for linear probing on IN20%. The optimal hyperparameters are shown in Table 6.

AUH. For each embedding dimension, we fix the batch size at 2048, and tune the alignment weight \( \lambda \), the scale of the RBF kernel \( t \), and the base learning rate \( \text{base}\_lr \) for pretraining with AUH. Then we tune the initial learning rate \( \text{lr}\_\text{head} \) for linear probing on IN20%. The optimal hyperparameters are shown in Table 7.

Table 6: Hyperparameter choice for SimCLR pretrained on IN20% with a ResNet-18 during 100 epochs, evaluated by linear probing on IN20%.

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Temperature</th>
<th>\text{base}_lr</th>
<th>\text{lr}_\text{head}</th>
</tr>
</thead>
<tbody>
<tr>
<td>( q = 1024 )</td>
<td>0.15</td>
<td>1.0</td>
<td>0.2</td>
</tr>
<tr>
<td>( q = 2048 )</td>
<td>0.15</td>
<td>1.0</td>
<td>0.2</td>
</tr>
<tr>
<td>( q = 4096 )</td>
<td>0.15</td>
<td>1.0</td>
<td>0.2</td>
</tr>
<tr>
<td>( q = 8192 )</td>
<td>0.15</td>
<td>0.8</td>
<td>0.2</td>
</tr>
</tbody>
</table>
Table 7: Hyperparameter choice for AUH pretrained on IN20% with a ResNet-18 during 100 epochs, evaluated by linear probing on IN20%.

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Alignment weight λ</th>
<th>Scale t</th>
<th>base_lr</th>
<th>lr_head</th>
</tr>
</thead>
<tbody>
<tr>
<td>q = 1024</td>
<td>400</td>
<td>2.5</td>
<td>1.0</td>
<td>10.0</td>
</tr>
<tr>
<td>q = 2048</td>
<td>1000</td>
<td>2.5</td>
<td>1.0</td>
<td>4.0</td>
</tr>
<tr>
<td>q = 4096</td>
<td>2000</td>
<td>2.5</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>q = 8192</td>
<td>3000</td>
<td>2.5</td>
<td>1.0</td>
<td>2.0</td>
</tr>
</tbody>
</table>

Table 8: Hyperparameter choice for VICReg pretrained on IN20% with a ResNet-18 during 100 epochs, evaluated by linear probing on IN20%.

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Alignment weight λ</th>
<th>Variance weight µ</th>
<th>base_lr</th>
<th>lr_head</th>
</tr>
</thead>
<tbody>
<tr>
<td>q = 1024</td>
<td>4</td>
<td>10</td>
<td>0.4</td>
<td>0.2</td>
</tr>
<tr>
<td>q = 2048</td>
<td>4</td>
<td>4</td>
<td>0.7</td>
<td>1.0</td>
</tr>
<tr>
<td>q = 4096</td>
<td>10</td>
<td>10</td>
<td>0.6</td>
<td>0.2</td>
</tr>
<tr>
<td>q = 8192</td>
<td>10</td>
<td>10</td>
<td>0.7</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Table 9: Hyperparameter choice for SFRIK pretrained on IN20% with a ResNet-18 during 100 epochs, evaluated by linear probing on IN20%.

<table>
<thead>
<tr>
<th>Order</th>
<th>Dimension</th>
<th>Alignment weight λ</th>
<th>Kernel weights (b₁, b₂, b₃)</th>
<th>base_lr</th>
<th>lr_head</th>
</tr>
</thead>
<tbody>
<tr>
<td>L = 1</td>
<td>q = 8192</td>
<td>10000</td>
<td>(1, 0, 0)</td>
<td>0.4</td>
<td>10.0</td>
</tr>
<tr>
<td>L = 2</td>
<td>q = 1024</td>
<td>400</td>
<td>(1, 40, 0)</td>
<td>1.0</td>
<td>2.0</td>
</tr>
<tr>
<td></td>
<td>q = 2048</td>
<td>400</td>
<td>(1, 40, 0)</td>
<td>1.0</td>
<td>4.0</td>
</tr>
<tr>
<td></td>
<td>q = 4096</td>
<td>1000</td>
<td>(1, 40, 0)</td>
<td>1.0</td>
<td>4.0</td>
</tr>
<tr>
<td></td>
<td>q = 8192</td>
<td>2000</td>
<td>(1, 20, 0)</td>
<td>1.0</td>
<td>10.0</td>
</tr>
<tr>
<td>L = 3</td>
<td>q = 8192</td>
<td>4000</td>
<td>(1, 40, 40)</td>
<td>1.2</td>
<td>1.0</td>
</tr>
</tbody>
</table>

**VICReg.** For each embedding dimension, we fix the batch size at 2048, and we tune the alignment weight $\lambda$, the variance weight $\mu$, and the base learning rate $\text{base}_\text{lr}$ for pretraining with VICReg. Then, we tune the initial learning rate $\text{lr}_\text{head}$ for linear probing on IN20%. The optimal hyperparameters are shown in Table 8.

**SFRIK, batch size 2048.** For each embedding dimension, we fix the batch size at 2048, and tune the alignment weight $\lambda$ in (2), the kernel weights $b_\ell$ ($\ell \in \{2, 3\}$) in (8), and the base learning rate $\text{base}_\text{lr}$ for pretraining with SFRIK. Without loss of generality, the first kernel weight $b_1$ in (8) is fixed at $b_1 = 1$. Then, we tune the initial learning rate $\text{lr}_\text{head}$ for linear probing on IN20%. The optimal hyperparameters are shown in Table 9.

**SFRIK, batch size 4096.** We fix the dimension at $q = 8192$, and the batch size at 4096, and tune the alignment weight $\lambda$ in (2), the kernel weights $b_\ell$ ($\ell \in \{2, 3\}$) in (8), and the base learning rate $\text{base}_\text{lr}$ for pretraining with SFRIK. Without loss of generality, the first kernel weight $b_1$ in (8) is fixed at $b_1 = 1$. Then, we tune the initial learning rate $\text{lr}_\text{head}$ for linear probing on IN20%. The optimal hyperparameters are: $\lambda = 4000$, $(b_1, b_2, b_3) = (1, 20, 0)$, $\text{base}_\text{lr} = 0.8$, $\text{lr}_\text{head} = 1.0$. This yields a top-1 accuracy of 46.3% for linear probing on IN20%, meaning that it is not necessary to use a batch size larger than 2048 in SFRIK to obtain a better performance.

**C.5 Hyperparameters for the Ablation on the Kernel Choice**

We detail the experimental setting of our ablation study on the kernel choice for the generic uniformity term (7) in Table 3 of Section 4.2. We follow the protocol of Section 4.2, with an embedding dimension of $q = 8192$. The training loss is $\ell_u(Z^{(1)}_I, Z^{(2)}_I) + 0.5(\ell_u(Z^{(1)}_I) + \ell_u(Z^{(2)}_I))$, where $\ell_u$ is the loss (7) with an RBF or a general distance kernel.
**RBF kernel.** The kernel is \( K(u, v) = e^{-t \|u - v\|^2} \). We fix the batch size at 2048, and tune the alignment weight \( \lambda \), the scale of the RBF kernel \( t \), and the base learning rate \( \text{base}\.lr \) for pretraining. Then, we tune the initial learning rate \( lr\.head \) for linear probing on IN20%. The optimal hyperparameters are: \( \lambda = 100, t = 2, \text{base}\.lr = 1.0, lr\.head = 10 \).

**Generalized distance kernel.** The kernel is \( K(u, v) = C - \|u - v\|^2_p \), where we fixed \( C = 0 \) because the value of this constant does not change the gradients of the optimization problem, and \( p = 2 \) because we verified empirically that choosing \( p<2 \) yields poor results. We fix the batch size at 2048, and tune the alignment weight \( \lambda \), and the base learning rate \( \text{base}\.lr \) for pretraining. Then, we tune the initial learning rate \( lr\.head \) for linear probing on IN20%. The optimal hyperparameters are: \( \lambda = 10000, \text{base}\.lr = 0.6, lr\.head = 10 \).

As shown in Table 3, a truncated kernel is empirically a better choice than the RBF or the generalized distance kernel for the uniformity term (7).

### C.6 Hyperparameters for experiments on IN100%

Table 10 reports the selected hyperparameters for pretraining ResNet-50 on IN100% with SFRIK in Section 4.3. The tuning protocol is as follows. Since hyperparameter tuning is costly on IN100%, we pretrain several ResNet-50 for different values of kernel weights \( b_\ell \), alignment weight \( \lambda \) and base learning rate \( \text{base}\.lr \), and pause the pretraining after 50 epochs. We evaluate the obtained backbones on kNN classification (top-1 accuracy, \( k = 20 \)), and select the best performing backbones. Then we continue pretraining these selected backbones until reaching epoch 200 or 400. Finally we select the hyperparameters that yield the highest top-1 accuracy for linear probing on IN100% after 200 or 400 epochs of pretraining. Because of the conclusions of Appendix D.1, our hyperparameter tuning follows the common practice in the self-supervised learning literature where hyperparameters are selected by measuring the performance on the validation set of ImageNet. Note that we verified experimentally a posteriori that the hyperparameters obtained by our tuning protocol are similar to the ones obtained from a tuning on a smaller dataset like STL-10 (Coates et al., 2011). This means that an alternative is to tune the hyperparameters on STL-10, and generalize these hyperparameters to pretrain SFRIK on IN100%.

**Table 10: Hyperparameter choice for SFRIK pretrained on IN100% with a ResNet-50 during 200 or 400 epochs.**

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Order</th>
<th>Epoch</th>
<th>Alignment weight ( \lambda )</th>
<th>Kernel weights ((b_1, b_2, b_3))</th>
<th>( \text{base}.lr )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( q = 8192 )</td>
<td>( L = 2 )</td>
<td>200</td>
<td>4000</td>
<td>((1, 20, 0))</td>
<td>0.4</td>
</tr>
<tr>
<td>( L = 3 )</td>
<td>200</td>
<td>4000</td>
<td>((1, 40, 40))</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>( q = 16384 )</td>
<td>( L = 2 )</td>
<td>200</td>
<td>20000</td>
<td>((1, 40, 0))</td>
<td>0.4</td>
</tr>
<tr>
<td>( L = 3 )</td>
<td>200</td>
<td>40000</td>
<td>((1, 40, 40))</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>( q = 8192 )</td>
<td>( L = 2 )</td>
<td>400</td>
<td>4000</td>
<td>((1, 20, 0))</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Tables 11 and 12 give the optimal hyperparameters found for linear probing on IN100%, linear probing on Places205, and semi-supervised learning with limited labels of IN100% when evaluating pretrained ResNet-50 backbones with SFRIK and VICReg. The hyperparameters that are tuned for evaluation are: the initial learning rate \( lr\.head \) of the linear layer in linear probing; and the initial learning rate \( lr\.backbone \) and \( lr\.head \) for respectively the backbone and the linear layer in semi-supervised learning. The reported hyperparameters in the two tables yield the evaluation results reported in Section 4.3.

### C.7 Computational resources

Pretrainings of ResNet-18 on IN20% with a batch size of 2048 (respectively 4096) are performed with 4 (respectively 8) NVIDIA Tesla V100 GPUs with 32GB of memory each. Pretrainings of ResNet-50 on IN100% are performed with 8 NVIDIA Tesla V100 GPUs with 32GB of memory each. The total amount of compute used for this work is around 50000 GPU hours.
Table 11: Hyperparameter tuning for linear probing on IN100% and Places205 for SFRIK and VICReg pretrained on IN100% with a ResNet-50.

<table>
<thead>
<tr>
<th>Method</th>
<th>Epochs</th>
<th>lr_head</th>
<th>IN100%</th>
<th>Places205</th>
</tr>
</thead>
<tbody>
<tr>
<td>VICReg(^\d) (q = 8192)</td>
<td>200</td>
<td>0.02</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>SFRIK (L = 2, q = 8192)</td>
<td>200</td>
<td>1.0</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>SFRIK (L = 3, q = 8192)</td>
<td>200</td>
<td>2.0</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>SFRIK (L = 2, q = 16384)</td>
<td>200</td>
<td>0.3</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>SFRIK (L = 2, q = 32768)</td>
<td>200</td>
<td>1.0</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>SFRIK (L = 3, q = 32768)</td>
<td>200</td>
<td>0.4</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>SFRIK (L = 2, q = 8192)</td>
<td>400</td>
<td>2.0</td>
<td>0.01</td>
<td></td>
</tr>
</tbody>
</table>

Table 12: Hyperparameter tuning for semi-supervised learning for SFRIK and VICReg pretrained on IN100% with a ResNet-50.

<table>
<thead>
<tr>
<th>Method</th>
<th>Epochs</th>
<th>lr_backbone</th>
<th>lr_head</th>
<th>lr_backbone</th>
<th>lr_head</th>
</tr>
</thead>
<tbody>
<tr>
<td>VICReg(^\d) (q = 8192)</td>
<td>200</td>
<td>0.02</td>
<td>0.2</td>
<td>0.2</td>
<td>0.04</td>
</tr>
<tr>
<td>SFRIK (L = 2, q = 8192)</td>
<td>200</td>
<td>0.004</td>
<td>1.6</td>
<td>0.02</td>
<td>0.4</td>
</tr>
<tr>
<td>SFRIK (L = 3, q = 8192)</td>
<td>200</td>
<td>0.002</td>
<td>1.0</td>
<td>0.01</td>
<td>0.2</td>
</tr>
<tr>
<td>SFRIK (L = 2, q = 16384)</td>
<td>200</td>
<td>0.004</td>
<td>1.4</td>
<td>0.04</td>
<td>0.2</td>
</tr>
<tr>
<td>SFRIK (L = 2, q = 32768)</td>
<td>200</td>
<td>0.004</td>
<td>1.0</td>
<td>0.04</td>
<td>0.1</td>
</tr>
<tr>
<td>SFRIK (L = 3, q = 32768)</td>
<td>200</td>
<td>0.004</td>
<td>1.0</td>
<td>0.02</td>
<td>0.1</td>
</tr>
<tr>
<td>SFRIK (L = 2, q = 8192)</td>
<td>400</td>
<td>0.004</td>
<td>1.4</td>
<td>0.02</td>
<td>0.2</td>
</tr>
</tbody>
</table>

C.8 Public Resources

We acknowledge the use of the following public resources, during the course of the experimental work of this paper:

- VICReg official code (Bardes et al., 2022) ........................................ MIT License
- DINO official code (Caron et al., 2021) ........................................ Apache License 2.0
- OBoW official code (Gidaris et al., 2021) ...................................... Apache License 2.0
- SwAV official code (Caron et al., 2020) ...................................... CC BY-NC 4.0
- SimCLR official code (Chen et al., 2020a) .................................. Apache License 2.0
- VISSL code (Goyal et al., 2021) ..................................................... MIT License
- ImageNet dataset (Deng et al., 2009) .............................................
- Places 205 dataset (Zhou et al., 2014) ........................................ Attribution CC BY
- VOC2007 dataset (Everingham et al., 2010) ...................................
D.1 HYPERPARAMETER TUNING WITHOUT A SEPARATE VALIDATION SET

A common practice in the self-supervised learning literature, e.g., (Bardes et al., 2022; Chen et al., 2020a), is to select the hyperparameters by measuring the performance on the validation set of ImageNet (red dataset in Figure 2) instead of a separate validation dataset (purple dataset in Figure 2). In this paragraph, we verify whether this less rigorous practice changes the conclusion of the experiments in Section 4.2. In Table 13, we report the evaluation of the different backbones pretrained on IN20% after tuning each method directly on the validation set of ImageNet, which is the same dataset used for evaluation in linear probing. By comparison with Table 2, which follows the more rigorous hyperparameter tuning protocol described in Appendix C.4, we observe that although the absolute figures of merit slightly vary if we use the less rigorous protocol instead of the more rigorous one, the conclusion of the experiments in Section 4.2 does not change. This gives an empirical justification to this common practice.

Table 13: Linear probing on IN20% (top-1 accuracy) at different embedding dimensions $q$. All methods were pretrained on IN20% with a ResNet-18 for 100 epochs. Hyperparameters specific to each method and the learning rate are tuned on the same dataset as the one used for evaluation in linear probing, which is less rigorous than tuning the hyperparameters on a separate validation set as described in Appendix C.4. Symbol † indicates models that we retrained ourselves.

<table>
<thead>
<tr>
<th>$q$</th>
<th>SimCLR†</th>
<th>AUH†</th>
<th>VICReg†</th>
<th>SFRIK</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L = 1$</td>
<td>$L = 2$</td>
<td>$L = 3$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1024</td>
<td>45.2</td>
<td>45.2</td>
<td>40.8</td>
<td>-</td>
</tr>
<tr>
<td>2048</td>
<td>45.8</td>
<td>45.6</td>
<td>44.1</td>
<td>-</td>
</tr>
<tr>
<td>4096</td>
<td>46.3</td>
<td>46.8</td>
<td>44.9</td>
<td>-</td>
</tr>
<tr>
<td>8192</td>
<td>46.2</td>
<td>46.8</td>
<td>46.0</td>
<td>27.5</td>
</tr>
</tbody>
</table>

D.2 ADDITIONAL EVALUATION OF SFRIK PRETRAINED ON IN20%

In complement to the results on linear probing on IN20% presented in Table 2, we evaluate SFRIK by linear probing on Places205 (Zhou et al., 2014) and linear SVM on VOC2007 (Everingham et al., 2010), following the protocol described in Appendix C.3. The learning rate for linear probing on Places205 is fixed at 0.01. Table 14 compares SFRIK, AUH and VICReg pretrained with ResNet-18 on IN20% under the setting of Section 4.2.

Table 14: Linear classification on Places205 and VOC2007 (accuracy and mean average precision). All methods were pretrained on IN20% with a ResNet-18 for 100 epochs. We tuned all hyperparameters specific to each pretraining method and the learning rate. The symbol † indicates models that we retrained ourselves.

<table>
<thead>
<tr>
<th>Linear classification</th>
<th>Places205</th>
<th>VOC07</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>Top-1</td>
<td>Top-5</td>
</tr>
<tr>
<td>VICReg† ($q = 8192$)</td>
<td>41.6</td>
<td>71.7</td>
</tr>
<tr>
<td>AUH† ($q = 8192$)</td>
<td>42.3</td>
<td>72.8</td>
</tr>
<tr>
<td>SFRIK ($L = 3$, $q = 8192$)</td>
<td>42.7</td>
<td>72.9</td>
</tr>
</tbody>
</table>

We conclude that, under the rigorous protocol of Section 4.2 and Appendix C.4, SFRIK performs better than AUH on various downstream tasks, while having the same computational saving offered by the kernel trick. Compared to VICReg, SFRIK performs better on these tasks with a reduced complexity.
D.3 Results of other pretraining methods on IN100% with ResNet-50

In complement to Table 4, Table 15 reports evaluation results for linear probing on IN100% and semi-supervised learning with few labels of IN100% of different ResNet-50 pretrained on IN100% with other state-of-the-art methods than the ones presented in Table 4. As mentioned in Section 4.3, we observe that, similarly to SFRIK, both BYOL and SwAV with multi-crop have a performance drop compared to VICReg on semi-supervised learning with 1% of labels, even though they perform better on linear probing on IN100%.

Table 15: Linear probing on IN100%, semi-supervised learning with few labels of IN100% (top-1 and top-5 accuracy). All methods have been pretrained on IN100% with a ResNet-50 during the reported number of epochs.

<table>
<thead>
<tr>
<th>Method</th>
<th>Epochs</th>
<th>Linear probing IN100%</th>
<th>Semi-supervised 1% labels</th>
<th>Semi-supervised 10% labels</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Top-1</td>
<td>Top-5</td>
<td>Top-1</td>
</tr>
<tr>
<td>SimCLR (Chen et al., 2020a)</td>
<td>1000</td>
<td>68.3</td>
<td>89.0</td>
<td>48.3</td>
</tr>
<tr>
<td>OBoW (Gidaris et al., 2021)</td>
<td>200</td>
<td>73.8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BYOL (Grill et al., 2020)</td>
<td>1000</td>
<td>74.3</td>
<td>91.6</td>
<td>53.2</td>
</tr>
<tr>
<td>SwAV (with multi-crop) (Caron et al., 2020)</td>
<td>800</td>
<td>75.3</td>
<td>-</td>
<td>53.9</td>
</tr>
<tr>
<td>Barlow Twins (Zbontar et al., 2021)</td>
<td>1000</td>
<td>73.2</td>
<td>91.0</td>
<td>55.0</td>
</tr>
<tr>
<td>VICReg (Bardes et al., 2022)</td>
<td>1000</td>
<td>73.2</td>
<td>91.1</td>
<td>54.8</td>
</tr>
</tbody>
</table>