Improving Self-Supervised Learning by Characterizing Idealized Representations

Yann Dubois, Tatsunori Hashimoto, Stefano Ermon, Percy Liang
Stanford University
{yanndubs,thashim,ermon,pliang}@stanford.edu

Abstract

Despite the empirical successes of self-supervised learning (SSL) methods, it is unclear what characteristics of their representations lead to high downstream accuracies. In this work, we characterize properties that SSL representations should *ideally* satisfy. Specifically, we prove necessary and sufficient conditions such that for any task invariant to given data augmentations, desired probes (e.g., linear or MLP) trained on that representation attain perfect accuracy. These requirements lead to a unifying conceptual framework for improving existing SSL methods and deriving new ones. For contrastive learning, our framework prescribes simple but significant improvements to previous methods such as using asymmetric projection heads. For non-contrastive learning, we use our framework to derive a simple and novel objective. Our resulting SSL algorithms outperform baselines on standard benchmarks, including SwAV+multicrops on linear probing of ImageNet.

1 Introduction

We study self-supervised learning (SSL), where the goal is to learn representations from minimal supervision, such that simple probes trained on these representations achieve high downstream accuracy. Recently, there has been many different SSL methods achieving impressive empirical results (e.g. SimCLR [1], SwAV [2]) using label-preserving augmentations (e.g. cropping or color jittering) as supervision. We dub this setting *invariant SSL* (ISSL). Despite these empirical successes, it remains unclear how these various SSL methods relate to one another, how to improve them, and how to derive new ones. Our goal is to provide a simple conceptual framework to think about those questions.

To derive such a framework, we ask ourselves: *what are the ideal requirements that ISSL representations should aim to satisfy?* We prove necessary and sufficient requirements to ensure that probes from a specified family, e.g. linear or multi-layer perceptron (MLP), perfectly classify any task that is invariant to desired data augmentations. This complements theoretical work in ISSL [3–8], which analyze specific ISSL algorithms. Our work instead focuses on properties of representations that should serve as a goal for any ISSL algorithm. These ideal properties are: (i) desired predictors should be able to distinguish positive and negative examples from the representation; (ii) the dimensionality of the representation should be sufficiently large; (iii) augmented inputs should map to the same representations.

Previous ISSL methods can be seen as approximations of our ideal requirements. Using our requirements, we derive simple improvements to those approximations as well as new ISSL objectives. Our theory thus results in a unifying conceptual ISSL framework, with practical prescriptions including:

- improvements to existing methods, such as increasing the dimensionality of representations and using asymmetric projections heads, which lead to 5% point gains on TinyImageNet;
- a novel non-contrastive ISSL objective that outperforms all baselines, including SwAV+multicrops, by at least 1% point on linear classification of ImageNet;
- extensions of ISSL algorithms to learn representations that are better suited for non-linear probes.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

https://doi.org/10.52202/068431-0820

2 Problem statement

2.1 Invariant Self-Supervised Learning

SSL learns an encoder ϕ that maps an input x from a finite space $\mathcal{X}$ (e.g., 256×256 images) into a representation $\phi(x) \in \mathbb{R}^d$. Given the encoder and a dataset $\mathcal{D}_t$ drawn from some task of interest $p_t(X,Y)$, we fit a classifier f from a desired family of probes $\mathcal{F}$. Families of probes are sets of k-ary classifiers for any $k \geq 2$ such as linear or MLP probes. For clarity, we consider linear probes $\mathcal{F}$ until Sec. 5.

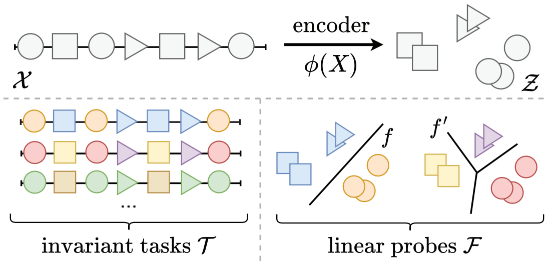


Figure 1: ISSL setting. (Top) 1D inputs, partitioned into 3 equivalence classes (shapes), are encoded by ϕ into a 2D representation. (Bot. left) 3 $\sim$-invariant tasks, where labels are the colors. (Bot. right) examples of probes for 2 of the invariant tasks.

Supervision for ISSL comes from unlabeled data p_X and label-preserving augmentations. Augmentations are ways of sampling "positive" x, x^+ examples that are equivalent for downstream classification tasks. We formalize this using an equivalence relation $x \sim x^+$ that partitions the inputs $\mathcal{X}$ into equivalent classes $[x] \in \mathcal{X}/\sim$, and we consider the following downstream tasks $\mathcal{T}$ whose labelings are deterministic and constant within these classes (we allow stochastic labeling in appendices).

Definition 1. The $\sim$-*invariant tasks* $\mathcal{T}$, is the set of all input-label distributions $p_t(X,Y)$ such that the labeling $p_t(Y|X)$ is deterministic and invariant to $\sim$, i.e.,

$$\text{for all } \; p_t \in \mathcal{T},\; x, x^+ \in \mathcal{X}: \qquad x \sim x^+ \implies \arg\max_{y\in\mathcal{Y}} p_t(y|x) = \arg\max_{y\in\mathcal{Y}} p_t(y|x^+). \tag{1}$$

As an illustration, consider the 3 classes (triangle, square, and circle) shown in Fig. 1. Then $\mathcal{T}$ consists of all tasks that are predictable from those shapes, e.g., recognizing shapes with vertices (blue/orange in Fig. 1) or recognizing the shape (yellow/red/purple in Fig. 1). Importantly, equivalence classes (here shapes) are different from—and essentially refinements of—downstream classes $\mathcal{Y}$ (here colors). Note that equivalence relations can model arbitrary transformations including cropping and adding Gaussian noise, which contrasts with typical restrictions to augmentations defined by group actions[9–11].

2.2 Idealized representations for ISSL

In this section, we define the optimal encoders, those that induce idealized representations that ISSL should be striving for. Although such encoders exist, they will likely not be learned in practice. Those idealized representations will nevertheless allow us to derive practical algorithms in Sec. 4. Note that our approach to defining ideal representations is to take into account how they will be used downstream, and thus depends on the desired family of probe $\mathcal{F}$ and potential invariant tasks $\mathcal{T}$.

The goal of ISSL is to learn representations from which (typically simple) probes $f \in \mathcal{F}$ classify well downstream tasks $p_t \in \mathcal{T}$, i.e., they achieve low 0-1 risk $\mathrm{R}_t(\phi, f) := \mathbb{E}_{p_t(X,Y)}[\mathbb{1}[Y \neq f(\phi(X))]]$. Ideally, for any task of interest $p_t \in \mathcal{T}$ there will be a probe that can classify it perfectly.

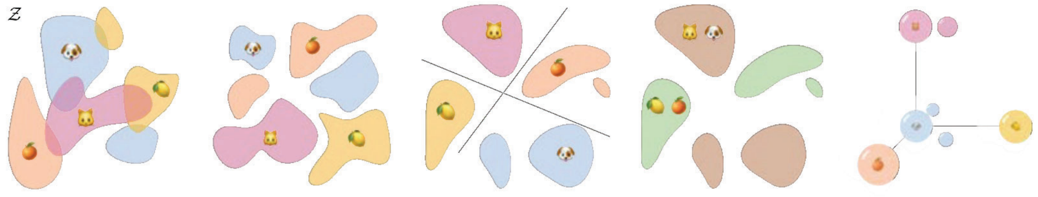
Definition 2. An encoder ϕ is *population optimal* for $\mathcal{T}, \mathcal{F}$, denoted as $\phi \in \Phi_{\text{pop}}$, iff predictors realize the Bayes error on any invariant task, i.e., for all $p_t \in \mathcal{T}$ we have $\inf_{f\in\mathcal{F}} \mathrm{R}_t(\phi, f) = 0$.

Population optimality ensures the existence of perfect downstream probes, which would be learned with infinite downstream data. In practice, however, predictors will be trained from finite datasets $\mathcal{D}_t$ of possibly small size n with empirical risk minimization. When n is small a fitted probe (ERM) $\hat{f} \in \widehat{\mathcal{F}}(\mathcal{D}_t, \phi) := \arg\min_{f\in\mathcal{F}} |\mathcal{D}_t|^{-1} \sum_{x,y\in\mathcal{D}_t} \mathbb{1}[y \neq f(\phi(x))]$ could be a terrible population predictor even when the underlying encoder is population optimal. Ideally, representations would thus also guarantee that *any* ERM performs as well as possible for *any* desired task and dataset size n. This suggests minimizing the following worst-case expected risk over tasks and ERMs.

$$\mathrm{W}_n(\phi, \mathcal{F}, \mathcal{T}) := \sup_{t\in\mathcal{T}} \mathbb{E}_{\mathcal{D}_t \overset{\text{iid}}{\sim} p_t^n(X,Y)} \left[\sup_{\hat{f}\in\widehat{\mathcal{F}}(\mathcal{D}_t,\phi)} \mathrm{R}_t(\phi, \hat{f}) \right]. \tag{2}$$

Definition 3. An encoder ϕ^* is *sample optimal* for $\mathcal{T}, \mathcal{F}$ iff if it is population optimal and minimizes the worst-case expected risk of ERMs for arbitrary sample sizes, i.e.,

$$\text{for all } \; n \geq 1: \qquad \phi^* \in \arg\min_{\phi\in\Phi_{\text{pop}}} \mathrm{W}_n(\phi, \mathcal{F}, \mathcal{T}). \tag{3}$$



(a) M not predictable (b) M non-lin. predictable (c) M linearly predictable (d) Tasks are not lin. pred. (e) Population optimal

Figure 2: Representations are *population optimal* for linear $\mathcal{F}$ iff their dimensionality is sufficiently large and equivalence classes $M(X)$ are linearly classifiable. The first 2 figures show representations from which $M(X)$ is (a) never or (b) only non-linearly classifiable. Although (c) ensures linear classification of $M(X)$, there exist invariant tasks, e.g. (d), that are not classifiable linearly. (e) Linear classifiability of $M(X)$ ensures population optimality iff the dimensionality is at least $|\mathcal{X}/{\sim}| - 1$.

3 Theoretical framework for linear probes $\mathcal{F}$

3.1 Characterizing optimal encoders for linear ISSL

In this section, we characterize all sample-optimal encoders (Def. 3), with simple properties that give insights into the form of the idealized representation. The key for our theory is that any $\sim$-invariant function g can be written as a composition between some function c_g and a *maximal invariant* M [12], i.e., $g = c_g \circ M$ where $M : \mathcal{X} \to \{1, \ldots, |\mathcal{X}/{\sim}|\}$ indexes the equivalence class:

$$\text{for any } x, x^+ \in \mathcal{X} : \quad x \sim x^+ \iff M(x) = M(x^+). \tag{4}$$

To build intuition for the final characterization, let us first discuss population optimal encoders (Def. 2) for unconstrained probes, then linear probes, and finally sample optimality (Def. 3).

Unconstrained probes. By Def. 1, labels of downstream tasks $p_t \in \mathcal{T}$ are $\sim$-invariant. Labels can thus be written as some function c_t of M, i.e., $\arg\max_y p_t(y|X) = c_t(M(X))$. This shows that $M(X)$ contains all and only information about desired tasks $\mathcal{T}$. If probes are unconstrained, an encoder is population optimal iff $M(X)$ is predictable/classifiable, i.e., there exist an h_M such that $M(X) = h_M(\phi(X))$. Indeed, this ensures that the probe defined by $c_t \circ h_M$ can classify the task p_t.

Linear probes $\mathcal{F}$. The problem with constrained probes is that they might not be able to use the information about desired tasks. In particular, the previous probe might not be linear $c_t \circ h_M \notin \mathcal{F}$. As an illustration, consider 4 equivalence classes: cats, dogs, oranges, and lemons. Fig. 2b shows a representation from which $M(X)$ is predictable but invariant tasks are not linearly classifiable. In fact, even when $M(X)$ is linearly predictable, i.e., $h_M \in \mathcal{F}$ as in Fig. 2c, downstream labels might not be, i.e., $c_t \circ h_M \notin \mathcal{F}$ as shown in Fig. 2d. By standard VC dimension arguments [13, 14], such binary task is linearly predictable if the representation's dimensionality d is one less than the number of equivalence classes $|\mathcal{X}/{\sim}| - 1$, as in Fig. 2e. Building on this intuition we prove that population optimal encoders are essentially[1] those that induce $d \geq |\mathcal{X}/{\sim}| - 1$ dimensional representations from which $M(X)$ is linearly predictable.

Sample optimality. Although population optimality ensures the existence of a perfect linear probe, ERM probes trained on finite samples might still be bad due to generalization issues (Fig. 3 left). Intuitively, one can remove such bad ERMs by mapping equivalent examples to the same representation, i.e., by using invariant encoders. Indeed, this ensures that ERMs that correctly predict one example in an equivalence class also correctly predict all the other (Fig. 3 right). We prove that such invariance of population optimal encoders is necessary and sufficient for sample optimality.

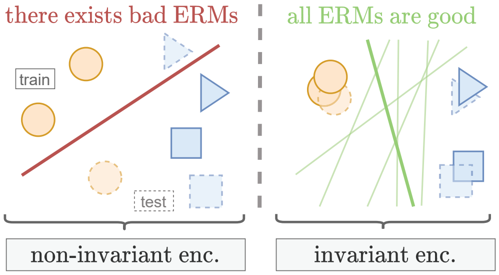


Figure 3: Invariance of population-optimal encoders is (a) necessary and (b) sufficient to ensure that there exists no bad ERM.

[1] The difference with learning theory is that instead of (binary) shatterability of all examples, we want k-ary shatterability of all equivalence classes from representations. The key is that both notions coincide for specific probes (e.g. linear) and invariant encoders, which are necessary for sample optimality.

Putting all together gives the following necessary and sufficient properties of sample-optimal encoders.

Theorem 1. An encoder ϕ^* is sample optimal for $\sim$-invariant tasks $\mathcal{T}$ and linear $\mathcal{F}$ if and only if

- **$\mathcal{F}$-predictability of M**: there exists a max. invariant M and an $f \in \mathcal{F}$ s.t. $M(X) = f(\phi^*(X))$;
- **Invariance**: ϕ^* is $\sim$-invariant, i.e., for any $x, x^+ \in \mathcal{X}$ we have $x \sim x^+ \implies \phi^*(x) = \phi^*(x^+)$;
- **Dimensionality**: the effective dimensionality of the representations is at least one less than the number of equivalence classes, i.e., $\dim(\operatorname{span}(\{\phi^*(x) | x \in \mathcal{X}\})) \geq |\mathcal{X}/{\sim}| - 1$.

3.2 The impact of augmentations on downstream performance

Let us compute the worst-case excess risk $\mathrm{W}_n(\phi, \mathcal{F}, \mathcal{T})$ of sample optimal encoders and show its dependence on the invariance structure. The key is that since sample optimal encoders are invariant (Theorem 1), ERMs only need to be trained on a single example per equivalence class to perfectly classify all other examples from that class. The risk then depends on the proportion of equivalence classes seen when training the probe, which can be computed in closed form as a function of the number of equivalence classes $|\mathcal{X}/{\sim}|$ and downstream samples n. For other similar results refer to Appx. B.3.

Proposition 2. The worst expected risk of a sample optimal ϕ^* for $\sim$-invariant tasks $\mathcal{T}$ and any $\mathcal{F}$ is

$$\mathrm{W}_n(\phi^*, \mathcal{F}, \mathcal{T}) = \left(1 - \frac{1}{|\mathcal{X}/{\sim}|}\right)^n \tag{5}$$

Prop. 2 shows that fewer equivalence classes, i.e., coarser $\sim$, leads to better downstream sample efficiency. Of course, the convergence rate will be slower for practical encoders than for sample optimal ones. The result nevertheless suggests that good augmentations should be as strong as possible (induce coarser $\sim$) while being label-preserving. Examples of coarse augmentations are those from CLIP [15], which map many images to similar sentences.

4 Practical ISSL objectives for linear probes $\mathcal{F}$

The main remaining question is how to practically enforce requirements from Theorem 1. In this section, we derive simple objectives that learn optimal encoders in *ideal settings* (infinite data, perfect optimizers, universal approximators). In the process, we shed light on why previously proposed ISSL methods work, and how to improve them in practice.

Our key insight is that we can learn sample-optimal encoders by jointly training an encoder and a logistic regression to predict $M(X)$ from the representations. Indeed, the resulting representations will be characterized by three properties, each of which implies a requirement from Theorem 1 ($\mathcal{F}$-predictability, invariance, effective dimensionality). First, the representations will allow linear predictability of $M(X)$ due to the linearity of logistic regression. Second, the variance of same-class representations will be minimized due to Jensen's inequality. Finally, the effective dimensionality will be maximal. Indeed, logistic regression favors maximal angles between representations of different classes to increase the confidence of the predicted class. Specifically, the representations learned by this joint procedure form a simplex equiangular tight frame (sETF)—as illustrated in Fig. 4 and discussed in the neural collapse literature [16–21]—and we prove that those are optimal.

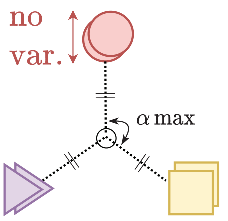


Figure 4: sETF are idealized representations that collapse same-class examples and maximize the angle across classes.

More formally, we prove in Prop. 3 that high dimension encoders trained to minimize the following multinomial logistic regression of $M(X)$, dubbed *$\sim$-ISSL log loss*, will be sample optimal,

$$\mathcal{L}_I(\phi; p_X) := \inf_{w \in \mathcal{W}_1} \mathbb{E}_{p_X}\left[-\log \frac{\exp\big(w(M(X))^\top \phi(X)\big)}{\sum_{m'=1}^{|\mathcal{X}/\sim|} \exp\big(w(m')^\top \phi(X)\big)}\right], \tag{6}$$

where w maps classes to weights, e.g., by indexing a weight matrix. Following previous work on neural collapse, we assume for this section that classes are equiprobable $p_X([x]) = {}^1\!/_{|\mathcal{X}/\sim|}$, and that classifier's weights and representations are unit-normalized, i.e., $w \in \mathcal{W}_1 := \{1, \ldots, |\mathcal{X}/{\sim}|\} \to \mathcal{S}$ and $\phi \in \Phi_1 := \mathcal{X} \to \mathcal{S}$ where $\mathcal{S}$ denotes the $(d{-}1)$-sphere (these assumptions can be relaxed e.g. [22, 23]). We then have the desired relation between ISSL log loss and sample-optimal encoders.

Proposition 3. Let p_X be a distribution with support $\mathcal{X}$ and equiprobable equivalence classes $p_X([x]) = {}^1\!/\!_{|\mathcal{X}/\sim|}$, $\forall x \in \mathcal{X}$. If $d \geq |\mathcal{X}/\sim| - 1$ then any unit-normalized encoder that minimizes the $\sim$-ISSL log loss $\phi^* \in \arg\min_{\phi\in\Phi_1} \mathcal{L}_I(\phi; p_X)$ is sample-optimal for $\sim$-invariant tasks and linear $\mathcal{F}$.

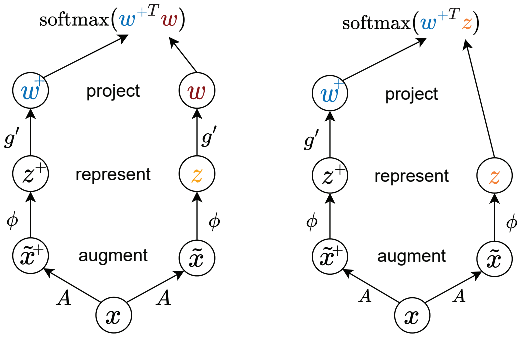
Prop. 3 shows that the ISSL log loss is a perfect pretext task in ideal settings. This suggests optimizing the ISSL log loss in practice and provides a formal relation between self-supervised and classical supervised learning. The challenge is that we typically neither have access to the maximal invariant $M(X)$ nor the number of equivalence classes $|\mathcal{X}/\sim|$ required to compute the denominator of Eq. (6). Instead, knowledge about the equivalence structure comes from data augmentations $\mathrm{A}(\tilde{X}|X)$ from which we can sample examples that are equivalent to the input $\tilde{X} \sim X$.

Having established our framework, we can re-interpret previous ISSL methods as practical approximations of the ISSL log loss using data augmentations (e.g., SimCLR, SwAV, DINO [24], SimSiam [25]). These approximations are nevertheless suboptimal: none of them learn sample- (nor population-) optimal encoders in idealized settings. By directly deriving ISSL methods from Eq. (6), we will prescribe improvements to previous ISSL objectives that ensure that sample-optimal encoders are learned in idealized settings. We broadly categorize prior methods into two families depending on whether they explicitly select the number of equivalence classes $|\mathcal{X}/\sim|$ or if it is implicitly inferred from augmentations. We call these approaches distillation and contrastive ISSL respectively. For derivations and Pytorch implementation see Appx. C.

4.1 Contrastive ISSL (CISSL)

Inspired by previous contrastive objectives [1, 26–28], we show how to optimize the ISSL log loss using data augmentations and negative samples to remove the need of knowing $M(X)$ and the number of classes $|\mathcal{X}/\sim|$. The resulting objective, dubbed *CISSL*, corresponds to SimCLR using only a projection head g on one branch. See Fig. 5. This asymmetry is necessary to learn population optimal encoders for linear $\mathcal{F}$.

CISSL bypasses the need for $M(X)$ by noticing that it only appears in the ISSL log loss through the class weights $w(M(X))$. CISSL thus learns a function g mapping equivalent (augmented) inputs $\tilde{X}$ to $w(M(\tilde{X}))$. Such mapping exists since M is invariant, e.g., $g := w \circ M$.



(a) SimCLR (b) CISSL

Figure 5: CISSL corresponds to SimCLR with a single (asymmetric) projection head $g'(z^+)^\top z$. This ensures linear predictability of downstream tasks, which will be computed by $W_t{}^\top z$.

When augmentations satisfy the Markov Chain $\tilde{X} - M(X) - X$, we show that we can replace the class weights $w(M(X))$ in the ISSL loss by $g(\tilde{X})$, where g is optimized over. Intuitively, this is because predicting $\tilde{X}$ contains only information about $M(X)$ due to the data processing inequality.

The only remaining challenge is computing the denominator of Eq. (6) without summing over the unknown classes $\mathcal{X}/\sim$. To do so we use ranking conditional noise contrastive estimation (NCE; [29–31]). NCE replaces classification of equivalence classes by classification of positives in a batch $\tilde{\mathbf{X}} := \{\tilde{X}^+, \tilde{X}_1^-, \ldots, \tilde{X}_k^-\}$ where the positive $\tilde{X}^+$ is sampled from the conditional $\mathrm{A}(\tilde{X} \mid X)$, while the k negatives $\tilde{X}_i^-$ come from the marginal $\mathrm{A}(\tilde{X}) = \mathbb{E}_{p_X}[\mathrm{A}(\tilde{X}|X)]$. Using Monte Carlo (MC) estimates with an unlabeled dataset $\mathcal{D} \overset{\text{i.i.d.}}{\sim} p_X$ we get our final empirical CISSL objective

$$\hat{\mathcal{L}}_C(\phi; \mathcal{D}) := \inf_{g\in\mathcal{S}^{\mathcal{X}}} \sum_{x\in\mathcal{D}} \mathbb{E}_{p(\tilde{\mathbf{X}}|x,\mathcal{D})}\left[-\log \frac{\exp g(\tilde{X}^+)^\top \phi(x)}{\sum_{\tilde{X}'\in\tilde{\mathbf{X}}} \exp g(\tilde{X}')^\top \phi(x)}\right]. \tag{7}$$

By Prop. 3 and NCE's consistency [31], encoders trained with CISSL $\phi^* \in \arg\min_{\phi\in\Phi_1} \hat{\mathcal{L}}_C(\phi; \mathcal{D})$ are sample optimal for linear $\mathcal{F}$ in our ideal setting assumption ($|\mathcal{D}| \to \infty$ and unconstrained g) and when $\tilde{X} - M(X) - X$ forms a Markov Chain. While consistency (and thus optimality) holds for $k \geq 1$, more negatives k improves statistical efficiency [31].

Typically $\tilde{X}, X$ take value in the same space (e.g. images). If so, we can tie parameters by encoding $\tilde{X}$, i.e., we can replace g by $g' \circ \phi$ where $g' : \mathbb{R}^d \to \mathcal{S}$ is called a projection head. The logits

https://doi.org/10.52202/068431-0820

inside of Eq. (7) are then computed by $g'(\phi(\tilde{X}))^\top \phi(x)$ as shown in Fig. 5b. This is very similar to SimCLR's objective $g'(\phi(\tilde{X})))^\top g'(\phi(x))$ shown in Fig. 5a. The difference is that, by projecting the current representation, *SimCLR learns encoders that are not even population optimal for linear $\mathcal{F}$*. In contrast, CISSL learns sample optimal encoders in ideal settings. Intuitively, this is because CISSL trains the representations in the same way as they will be used in downstream tasks p_t. Indeed, representations $\phi(X)$ will be dotted with the downstream tasks' weights ${W_t}^\top \phi(x)$ to compute logits. In ISSL, representations $\phi(x)$—rather than their projections $g'(\phi(x))$—should thus be used in the inner product with ISSL weights $w(M(X)) = g'(\phi(\tilde{X})))$ to compute logits. CISSL thus derives, from first principles, an asymmetric use of projection heads.

4.2 Distillation ISSL (DISSL)

In practice, CISSL requires contrasting many negatives. To avoid contrastive estimation, we can directly approximate $M(X)$—instead of the weight $w(M(X))$—so that the denominator of Eq. (6) can be computed exactly. The resulting method, dubbed *DISSL* (Fig. 6), is a simpler and theoretically-grounded version of non-contrastive losses (e.g. SwAV, DINO, SimSiam). The challenge and main difference between each of those methods is how they estimate $M(X)$.

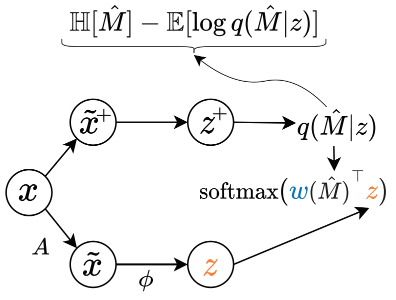

$\mathbb{H}[\hat{M}] - \mathbb{E}[\log q(\hat{M}|z)]$
$\tilde{x}^+$
z^+
$q(\hat{M}|z)$
x
softmax$(w(\hat{M})^\top z)$
A
$\tilde{x}$
ϕ
z

Figure 6: DISSL. The teacher (top branch) is trained with the top loss to ensure $\hat{M}$ is a maximal invariant r.v. The student (bottom branch) distills the teacher by predicting $\hat{M}$.

As $M(X)$ is discrete, we use a conditional categorical distribution $q(\hat{M} \,|\, X)$ to estimate $M(X)$ with a random variable $\hat{M}$. Replacing terms in the ISSL log loss then gives the following objective. Differences with Eq. (6) are in red.

$$\inf_{w \in \mathcal{W}_1} \mathbb{E}_{p_X q(\hat{M} \,|\, X)} \left[-\log s_{\phi,w}(\hat{M} \,|\, X) \right], \quad s_{\phi,w}(m \,|\, x) = \frac{\exp\big(w(m)^\top \phi(x)\big)}{\sum_{m=1}^{C} \exp\big(w(m)^\top \phi(x)\big)}, \tag{8}$$

To highlight similarities with previous methods [2, 24, 32] we refer to Eq. (8) as *distilling* the *teacher* $q(\hat{M} \,|\, X)$ into the student $s_{\phi,w}(\hat{M} \,|\, X)$. Importantly, Eq. (8) is exactly the ISSL log loss if the teacher outputs a maximal invariant random variable, i.e., $\hat{M} = M(X)$. By Prop. 3, distillation thus learns sample-optimal encoders when the teacher takes at least $C \geq |\mathcal{X}/{\sim}|$ values and satisfies the following requirements, which correspond to the definition of maximal invariance (Eq. (4)).

Deterministic the teacher is a deterministic distribution $\max_{m \in \{1,\dots,C\}} q(m \,|\, X) = 1$;
Invariant the teacher maps positives together $x \sim x^+ \implies q(\hat{M} \,|\, x) = q(\hat{M} \,|\, x)$;
Maximal the teacher maps negatives separately $x \not\sim x^- \implies q(\hat{M} \,|\, x) \neq q(\hat{M} \,|\, x^-)$.

Intuitively, these requirements ensure that the teacher clusters examples by equivalence classes, which will then be classified by the student. There are many ways of enforcing such requirements. In the following, we use information-theoretic quantities (entropies and KL divergences) to jointly train the teacher (online clustering) and student. Such quantities have the advantage of being interpretable and computable in closed form for the categorical distributions given by our teacher and student.

Determinism and Invariance. Both properties hold if and only if for any equivalent example $x \sim x^+$ the cross-entropy of the teacher's outputs is minimized $\mathbb{E}_{q(\hat{M} \,|\, x)}[-\log q(\hat{M} \,|\, x^+)] = 0$. Indeed, zero cross-entropy simultaneously minimizes the conditional entropy $\mathbb{H}[\hat{M} \,|\, x] = 0$ (equivalent to determinism) and the KL divergence $\mathbb{D}_{\text{KL}}[q(\hat{M} \,|\, x) \| q(\hat{M} \,|\, x^+)] = 0$ (equivalent to invariance).

Maximality. A natural way of enforcing maximality is to train the teacher to predict differently negatives $x \not\sim x^-$. This requires accessing batches of negatives, which we wanted to avoid with DISSL. Instead, assume that we know (or have a prior on) the true distribution of equivalence classes $p(M(X))$. Then a deterministic and invariant teacher is maximal if and only if the KL divergence between its marginal $q(\hat{M}) = \mathbb{E}_{p_X}[q(\hat{M}|X)]$ and the true one is minimized $\mathbb{D}_{\text{KL}}[q(\hat{M}) \| p(M(X))] = 0$. We can thus avoid contrasting negatives using (a prior on) the distribution of equivalence classes.

Using a Lagrangian relaxation (with multipliers λ, β) of the teacher's constraints with an MC estimate of the distillation loss (Eq. (8)) we get the following empirical DISSL loss $\hat{\mathcal{L}}_D(\phi; \mathcal{D}) :=$

$$\inf_{q,w} \lambda \underbrace{\mathbb{D}_{\text{KL}}\left[q(\hat{M}) \middle\| p(M(X)) \right]}_{\text{Maximality}} - \sum_{x \in \mathcal{D}} \mathbb{E}_{\mathrm{A}(\tilde{X}|x) q(\hat{M} \,|\, x)} \Big[\underbrace{\beta \log q(\hat{M} \,|\, \tilde{X})}_{\text{Det. and Inv.}} + \underbrace{\log s_{\phi,w}(\hat{M} \,|\, x)}_{\text{Distillation}} \Big]. \tag{9}$$

In ideal settings ($|\mathcal{D}| \to \infty$, unconstrained q, known $p(M(X))$) and for $\lambda, \beta \to \infty$ we show that encoders trained with DISSL are sample optimal for linear $\mathcal{F}$ and tasks that are invariant to the equivalences defined by the connected components of the augmentation graph [8]. If the marginal is unknown, we follow the MaxEnt [33] and use a uniform prior $p(M(X)) = \text{Unif}(1, C)$. The KL in Eq. (9) then corresponds to maximizing entropy $\mathbb{H}[\hat{M}]$.

As with CISSL, when $\tilde{X}, X$ are in the same space, we tie parameters by encoding $\tilde{X}$ before the teacher, i.e., $q(\hat{M} \mid X) = \text{softmax}(g'(\phi(X)))$ for a head $g' : \mathbb{R}^d \to \mathbb{R}^C$. Algorithm 1 and Fig. 6 illustrate DISSL with a uniform prior, and a marginal $q(\hat{M})$ estimated from batches.

Algorithm 1 Batched DISSL

```
Require: head $g$, weights $W$, enc. $\phi$, batched inputs $X$, aug. A, hyp. $\beta$, $\lambda$.
1: $\tilde{X}, \tilde{X}^+ \leftarrow \text{sample}(\text{A}(\tilde{X} \mid X), 2)$
2: $q(\hat{M}|\tilde{X}) = \text{softmax}(g(\phi(\tilde{X}))$
3: $q(\hat{M}|\tilde{X}^+) = \text{softmax}(g(\phi(\tilde{X}^+))$
4: $s(\hat{M}|\tilde{X}^+) = \text{softmax}(W^T\phi(\tilde{X}^+))$
5: $q(M) = \text{batch\_avg}(q(M|\tilde{X}))$
6: $\text{mxml} = \sum_m q(m) \log q(m)$
7: $\text{det\_inv} = \sum_m q(m|\tilde{X}) \log q(m|\tilde{X}^+)$
8: $\text{dstl} = \sum_m q(m|\tilde{X}) \log s(m|\tilde{X}^+)$
9: return $\lambda * \text{mxml} - \beta * \text{det\_inv} - \text{dstl}$
```

DISSL is one of the many distillation methods that can be derived from our teacher's requirements. These requirements generally lead to a framework for deriving, comparing, and analyzing distillation objectives. In Appx. D we provide a taxonomy of 12 previous SSL methods from this perspective—none of which recover sample optimal encoders. Typically, previous methods favor: (i) determinism by "sharpening" the teacher with a temperature parameter; (ii) invariance by making matching the teacher and student on equivalent inputs $q(\hat{M} \mid x^+) \approx s_{\phi,w}(\hat{M} \mid x)$; (iii) maximality through optimization tricks (e.g. stop gradients or momentum encoder) to avoid a constant teacher —referred to as "collapsing" [25, 34–36]. Note that avoiding constant collapse $q(\hat{M} \mid x') = q(\hat{M} \mid x)$ for any $x, x' \in \mathcal{X}$ is insufficient, e.g., mapping half the inputs to a constant is also undesirable. Maximality formalizes the desired requirement: non-equivalent examples should not be mapped together.

5 ISSL for non-linear predictors $\mathcal{F}^+$

Linear probes $\mathcal{F}$ are standard for evaluating representations [27, 37, 38]. However, if the goal is to maximize performance it is natural to consider non-linear predictors $\mathcal{F}^+$. The question then becomes how should we learn optimal representations for any $\mathcal{F}^+$? In Appx. B.5 we extend our framework to certain non-linear probes, such as neural networks, that separate into a hidden component h and linear layer, i.e., $f(\cdot) = W^T h(\cdot)$. We informally summarize the theory and implications below.

Theory. There are two main differences between our characterization of optimal encoders for linear and non-linear probes. First, encoders should ensure predictability of $M(X)$ for the desired $\mathcal{F}^+$. Second, the dimensionality requirement decreases with the complexity of $\mathcal{F}^+$, e.g., for universal $\mathcal{F}^+$ it is $d \geq 1$. More generally, the necessary dimension and the sufficient dimension for optimality do not coincide and respectively depend on the predictors' VC dimension [13] and memory capacity [39].

Implications. Both theoretical differences lead to direct implications for non-linear ISSL. First, we can decrease the dimensionality when performing ISSL for more complex non-linear probes. Second, we should use a projection head that ensures the predictability of the maximal invariant with using the desired probes $\mathcal{F}^+$. Similarly to the linear case, we should match how the representation is used in the ISSL log loss and downstream tasks. We should thus apply the non-linear predictor h in a similar asymmetric way. For CISSL we should thus compute the right hand side of Fig. 5 as $g(\phi(\tilde{X}))h(\phi(x))$. For DISSL, we should change the line 4 of Algorithm 1 to $s(\hat{M}|\tilde{X}^+) = \text{softmax}(W^T h(\phi(\tilde{X}^+)))$.

6 Summary of insights and relation to previous work

Let summarize our framework's main insights and their relation to previous work. Details in Appx. D.

Dimensionality. Theorem 1 shows large representation's dimensionality is needed to ensure probes can classify all invariant tasks (note: this is different from the projection's dimensionality analyzed in [1, 40]). This suggests: (i) increasing the dimensionality of the representation; and (ii) ensuring that representations do not live in a lower dimensional subspace. Although the first has surprisingly not been investigated in SSL, there have been recent empirical analyses about the second [41, 42].

Projection heads. Secs. 4 and 5 theoretically show how to choose projection heads, namely, one should be as large as possible while the other should have the architecture of downstream probes.

To our knowledge, we are the first to relate the architecture of the probing family and the projection head. Furthermore, our theory suggests why projection heads empirically improve performances [1, 25, 43] in general SSL, i.e., beyond avoiding collapsing in non-contrastive learning [25, 36].

Augmentations. Prop. 2 shows the benefit of using coarse label-preserving augmentations, by proving the exact relation between optimal sample efficiency and the number of equivalence classes. This gives a new theoretical perspective on the use of augmentations that remove a lot of information of the input, which have been suggested to be useful for many different reasons [12, 44–48].

DISSL. In Sec. 4 we derive DISSL to learn optimal encoders in ideal settings, contrary to prior work. DISSL follows recent methods [49–51], in particular SwAV and DINO, that learn representations by jointly predicting and learning clusters of examples. DISSL/DINO/SWAV each distill a categorical teacher but differ in how they enforce its maximality. DINO does it implicitly through optimization tricks, e.g., exponentially moving averages and stop-gradients. SwAV explicitly enforces maximality by equiprobably clustering a queue of examples using Sinkhorn's algorithm [52]. DISSL is also explicit but uses a more efficient max entropy regularization (no queue/stop-gradient/internal algorithm). Furthermore, DISSL's student uses a linear projection head to ensure good linear probing (Sec. 4.1).

Theoretical SSL. Theorem 1 characterizes optimal encoders for ISSL. This contrasts with standard theoretical work in SSL that typically analyze specific SSL algorithms [3, 5, 6, 8, 53–55] and / or focus on upper-bounding downstream performance [4]. The advantage of our theory is that it provides a simple unifying framework from which we can derive a variety of SSL methods and actionable insights. The downside is that by distancing itself from specific algorithms, the framework does not provide guarantees outside of ideal settings. We thus see our theory as complementary to prior work.

7 Experiments

The experiments in the main paper focus on evaluating our frameworks' prescriptive implications. For more results see Appx. G. For experimental details see Appx. F and our GitHub repository.

In summary, our experimental results show that: (i) CISSL/DISSL outperform their respective baselines on standard benchmarks as suggested by Sec. 4; (ii) increasing dimensionality improves downstream probing as suggested by Theorem 1; (iii) coarser augmentations improve sample efficiency as in Prop. 2; (iv) smaller ISSL log loss improves downstream performance as suggested by Prop. 3; (v) projection heads should be related to probing family as discussed in Secs. 4 and 5.

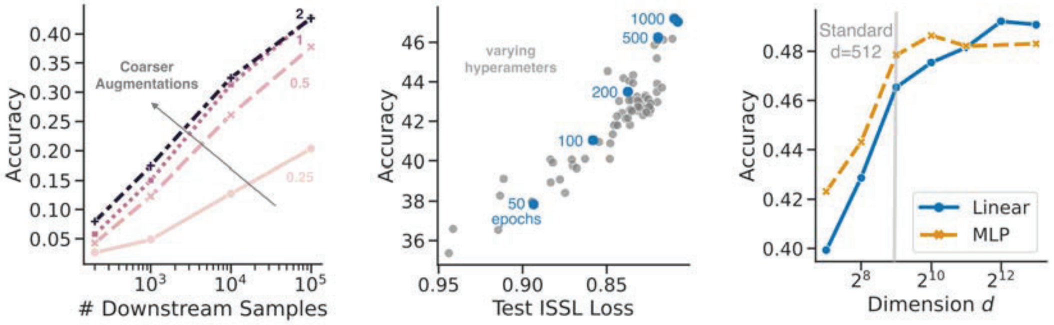
For the first experiments, we use TinyImageNet [56], 300 pretraining epochs, and ResNet18s. Note that for ResNet18s the dimensionality of the representation (`res5avg`) is $d = 512$. For contrastive baselines we use SimCLR. For distilling baselines we use the best over DINO, SwAV, SimSiam. We use standard TinyImageNet augmentations [34] (color jittering, grayscaling, cropping) with a parameter controlling the probability and strength of augmentations to study the effect of coarsening.

CISSL/DISSL improve linear probing Table 1 shows that each of our prescriptions significantly improve distillation and contrastive ISSL. "Ours" vs "Base" shows that removing one non-linear projection head (as in Fig. 5) gives a 1% gain in contrastive learning, while our new distillation objective achieves 2% gains compared to the best distillation baseline. "dim. $\uparrow$" shows that increasing the dimensionality of representation $512 \rightarrow 2048$ further improves linear probing by 2%. "Epochs $\uparrow$" and "coarse aug." show that training for longer ($300 \rightarrow 1000$ epochs) and using coarser augmentations both further improve performance by 1–2%. Altogether our framework's insights significantly improve linear probing accuracy: 6% gains for contrastive learning and 7% for distillation.

Table 1: Our prescriptive implications improves ISSL for linear probes on TinyImageNet with ResNet18 encoders. 3 seeds.

	CONTR.	DISTIL.
BASE. (SIMCLR | DINO)	44.9 ±0.2	43.5 ±0.2
OURS (CISSL | DISSL)	45.8 ±0.0	45.5 ±0.1
+ DIM. ↑	47.6 ±0.1	47.9 ±0.1
+ EPOCHS ↑	48.7 ±0.1	49.2 ±0.1
+ COARSE AUG.	51.0 ±0.1	50.6 ±0.2

Coarse augmentations improve sample efficiency. Prop. 2 suggests that coarser label-preserving augmentations improve downstream sample efficiency. We test that by training four DISSL encoders with augmentations of varying strengths ($[25\%, 50\%, 100\%, 200\%]$ relative to standard augmentations) and evaluating them at different downstream samples sizes. As suggested by the theory, Fig. 7a shows that stronger label-preserving augmentations improve sample efficiency.



(a) Augmentations

(b) ISSL log loss

(c) Dimensionality

Figure 7: As suggested by our theory: (a) coarsening augmentations improves sample efficiency; (b) decreasing ISSL log loss improves representations, which can be achieved by longer training; and (c) increasing the representation's dimensionality improves performance for linear $\mathcal{F}$ but less so for MLPs $\mathcal{F}^+$. Y-axis is TinyImageNet probing performance. Each point is obtained by sweeping: (a) augmentation strengths for DISSL and downstream sample sizes; (b) CISSL's optimization hyperparameters including epochs; and (c) the dimensionality of CISSL's representations.

ISSL log loss correlates with performance. Previous work [57] suggested contrastive losses may not predict probing performance. On the contrary, Prop. 3 shows that minimal ISSL log loss implies the optimality of encoders if d is large. This suggests that, for fixed augmentations, ISSL log loss is highly related to performance. To test this relation, we trained 80 CISSL models with various hyperparameters, while fixing augmentations and negatives k. Fig. 7b shows that the test ISSL loss indeed correlates with probing accuracy (gray points). Appx. G.4 shows similar results for the train ISSL loss.

Longer training decreases ISSL log loss. An efficient way of decreasing ISSL log loss is longer training (blue points in Fig. 7b only vary the number ISSL epochs). Prop. 3 thus provides an explanation of the well-known performance gains of longer ISSL training [1, 58]. Namely, longer training decreases ISSL log loss, which results in representations that are closer to optimality.

Increasing dimensionality improves performance. Theorem 1 shows that linear ISSL requires a dimensionality of $d = |\mathcal{X}/{\sim}| - 1$ to ensure predictability of all invariant tasks. In practice, the number of equivalence classes is most likely very large. Although such dimensionality is impractical, it does suggest increasing the dimensionality beyond the $d = 512$ of ResNet18s. To test this, we train CISSL with a larger dimensionality by increasing the output channels before the average pooling layer. To keep the number of parameters comparable, we use bottlenecks or low-rank linear layers before and after the representations layer as detailed in Appx. F. Fig. 7c shows that increasing the dimensionality has a significant effect on downstream performance for linear $\mathcal{F}$ (blue).

Previous results concern linear $\mathcal{F}$ which are standard for ISSL evaluation. In practice, one would likely use more complex probes $\mathcal{F}^+$ to improve results. In Sec. 5 we discussed how to perform ISSL for $\mathcal{F}^+$. In the following, we evaluate ISSL for MLPs $\mathcal{F}^+$ with two hidden layers of 2048 units.

Performing ISSL for MLPs $\mathcal{F}^+$. Table 2 shows that MLP probes ("$\mathcal{F}_{\text{EVAL}}$: MLP") outperforms linear ones ("$\mathcal{F}_{\text{EVAL}}$: lin.") even in few-shot regimes (10 shot). Furthermore, when predicting with MLP probes it is desirable to train the ISSL encoders for MLPs ("$\mathcal{F}_{\text{ISSL}}$: MLP") instead of linear probes (row 4 vs 3). As discussed in Sec. 5, the only difference between DISSL for different probes is that the student's representation is projected using a head with the same architecture as downstream probes.

Table 2: MLP probes outperform linear ones on TinyImageNet when using MLP ISSL.

	$\mathcal{F}_{\text{ISSL}}$	$\mathcal{F}_{\text{EVAL}}$	ALL	10 SHOT
DINO	-	MLP	44.7 ±0.3	22.0 ±0.2
DISSL	LIN.	LIN.	45.9 ±0.0	25.3 ±0.0
	LIN.	MLP	46.4 ±0.1	25.6 ±0.0
	MLP	MLP	47.5 ±0.2	26.1 ±0.1

Dimensionality affects less MLPs $\mathcal{F}^+$. In Sec. 5 we saw that the optimal dimensionality d decreases for more complex predictors $\mathcal{F}^+$. Fig. 7c indeed shows that larger dimensions result in smaller gains when the training and evaluation probe are MLPs $\mathcal{F}^+$ (orange) compared to linear ones $\mathcal{F}$ (blue).

Given encouraging TinyImageNet results, we investigated whether our objectives can be used as a replacement for standard objectives on ImageNet [59]. We specifically replaced SimCLR with CISSL and SwAV with DISSL in `VISSL`'s codebase [60] without modifying hyperparameters.

Table 3: Our models outperform baselines on ImageNet. All models use ResNet50, 100 epochs, 2560 batch size. For our models, we increase the dimensionality of the representations (2048 $\rightarrow$ 8192).

SimCLR	SwAV	Barlow T.	**CISSL**	**DISSL**
65.1	64.6	66.1	67.7	**68.9**

Our models outperform baselines on ImageNet. Table 3 shows that both CISSL and our novel DISSL objective significantly outperform all considered baselines, including SwAV by a 4%, when using standard augmentations. For CISSL we found (contrary to Table 1) that gains compared to SimCLR were mostly due to increasing the dimensionality of the representation.

Given encouraging results on standard augmentations, we compared the performance of DISSL with a near SOTA model: SwAV trained with their special multi-crop augmentations.

DISSL is competitive with SOTA models at scale. Table 4 shows that DISSL outperforms SwAV. Combined with Table 3 this suggests that DISSL works well in different settings. This is encouraging given the lack of tuning and the simplicity of DISSL's out-of-the-box objective. In contrast, SwAV requires stopping gradients, storing a queue, running Sinkhorn's algorithm, and freezing certain parameters during initial training steps.

Table 4: Our DISSL outperforms SwAV using $2 \times 160 + 4 \times 96$ multi-crops on ImageNet. 2560 batch size, ResNet50.

	100 epochs	400 epochs
SwAV	69.5	73.5
DISSL	**70.7**	**74.0**

Table 5: DISSL is competitive on transfer tasks. Same models as in Table 4 evaluated by linear probes.

	Food	CIFAR10	CIFAR100	Cars	Aircrafts	DTD	Pets	Caltech	Flowers
SwAV	75.5	92.0	76.2	58.2	**49.1**	72.6	86.9	**92.0**	94.7
DISSL	**77.9**	**93.6**	**77.6**	**62.2**	48.1	**73.9**	**88.0**	91.5	**95.3**

DISSL is competitive on transfer. Table 5 shows that DISSL generally outperforms SwAV on the standard transfer benchmarks from [61] even though our theory does not discuss transfer.

8 Summary and Outlook

We presented a simple conceptual framework to understand, compare and improve ISSL methods. Inspired by recent work on optimal representations for supervised [62] and robust [63] learning, we derived such a framework by studying algorithmic-agnostic goals for ISSL. In particular, Theorem 1 provides the minimal and sufficient requirements for optimal encoders. On the algorithmic side, Sec. 4 uses our framework to derive simpler and theoretically-motivated variants of prior SSL objectives. Altogether our framework provides actionable insights for ISSL such as how to choose: SSL algorithms, the dimensionality of the representations, projection heads, and augmentations. On the empirical side, Sec. 7 shows that our prescriptions leads to significant gains on linear probing benchmarks.

There are many limitations that should be addressed for a more realistic prescriptive framework for ISSL. See Appx. E for more details. First, our theory is binary, e.g., augmentations are either label preserving or not, and encoders are optimal or not. Although our framework likely captures the right intuition, nuance would be more realistic. Second, we consider unconstrained encoders ϕ, dimensionality, and infinite unlabeled data. As a result, our framework cannot currently improve the computational or data efficiency of the pretraining phase. Finally, we consider all invariant tasks $\mathcal{T}$ even though only a subset of those are meaningful. Despite those limitations, we showed that our framework's insights lead to substantial improvements in ISSL performance.

Acknowledgments and Disclosure of Funding

We thank Ananya Kumar, Shengjia Zhao, Mo Tiwari, Michael Xie, Niladri Chatterji, Shibani Santurkar, Jiaming Song, Ali Malik, Colin Wei, Yangjun Ruan and Chris Maddison for helpful feedback. YD is supported by a Knights-Hennessy Scholarship. The work is partially supported by an Open Philanthropy Project Award, Sloan Fellowship, and ARO (W911NF-21-1-0125).

References

[1] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *International Conference on Machine Learning (ICML)*, 2020. (Cited on 1, 5, 7, 8, 9, 43, 44, 45, 49)

[2] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, “Unsupervised learning of visual features by contrasting cluster assignments,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. (Cited on 1, 6, 43, 44, 45)

[3] N. Saunshi, O. Plevrakis, S. Arora, M. Khodak, and H. Khandeparkar, “A theoretical analysis of contrastive unsupervised representation learning,” in *International Conference on Machine Learning (ICML)*, 2019. (Cited on 1, 8, 37, 42, 43)

[4] Y. Bansal, G. Kaplun, and B. Barak, “For self-supervised learning, rationality implies generalization, provably,” in *International Conference on Learning Representations (ICLR)*, 2021. (Cited on 8, 42)

[5] J. D. Lee, Q. Lei, N. Saunshi, and J. Zhuo, “Predicting what you already know helps: Provable self-supervised learning,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. (Cited on 8, 42)

[6] C. Tosh, A. Krishnamurthy, and D. Hsu, “Contrastive learning, multi-view redundancy, and linear models,” in *Conference on Algorithmic Learning Theory (ALT)*, 2021. (Cited on 8, 42)

[7] Z. Wen and Y. Li, “Toward understanding the feature learning process of self-supervised contrastive learning,” in *International Conference on Machine Learning (ICML)*, 2021. (Cited on 42, 45)

[8] J. Z. HaoChen, C. Wei, A. Gaidon, and T. Ma, “Provable guarantees for self-supervised deep learning with spectral contrastive loss,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. (Cited on 1, 7, 8, 39, 42, 43)

[9] T. Leen, “From data distributions to regularization in invariant learning,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 1994. (Cited on 2)

[10] S. Chen, E. Dobriban, and J. H. Lee, “A group-theoretic framework for data augmentation,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. (Cited on 43)

[11] C. Lyle, M. van der Wilk, M. Kwiatkowska, Y. Gal, and B. Bloem-Reddy, “On the benefits of invariance in neural networks,” *arXiv preprint arXiv:2005.00178*, 2020. (Cited on 2, 43)

[12] Y. Dubois, B. Bloem-Reddy, K. Ullrich, and C. J. Maddison, “Lossy compression for lossless prediction,” *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. (Cited on 3, 8, 21, 22, 25, 37, 42, 43, 45)

[13] V. N. Vapnik and A. Y. Chervonenkis, “On uniform convergence of the frequencies of events to their probabilities,” *Teoriya Veroyatnostei i ee Primeneniya*, vol. 16, no. 2, pp. 264–279, 1971. (Cited on 3, 7, 32, 34, 35)

[14] C. J. C. Burges, “A tutorial on support vector machines for pattern recognition,” *Data mining and knowledge discovery*, vol. 2, no. 2, pp. 121–167, 1998. (Cited on 3, 25, 32)

[15] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning transferable visual models from natural language supervision,” in *International Conference on Machine Learning (ICML)*, 2021. (Cited on 4, 21, 43)

[16] V. Papyan, X. Y. Han, and D. L. Donoho, “Prevalence of neural collapse during the terminal phase of deep learning training,” *Proceedings of the National Academy of Sciences*, vol. 117, pp. 24 652 – 24 663, 2020. (Cited on 4, 43)

[17] T. Ergen and M. Pilanci, “Revealing the structure of deep neural networks via convex duality,” in *International Conference on Machine Learning (ICML)*, 2021.

[18] J. Lu and S. Steinerberger, “Neural collapse under cross-entropy loss,” *Applied and Computational Harmonic Analysis*, vol. 59, pp. 224–241, 2022. (Cited on 34, 43)

[19] F. Graf, C. Hofer, M. Niethammer, and R. Kwitt, “Dissecting supervised contrastive learning,” in *International Conference on Machine Learning (ICML)*, 2021.

https://doi.org/10.52202/068431-0820

[20] J. Zarka, F. Guth, and S. Mallat, "Separation and concentration in deep networks," in *International Conference on Learning Representations (ICLR)*, 2021.

[21] W. E and S. Wojtowytsch, "On the emergence of simplex symmetry in the final and penultimate layers of neural network classifiers," in *Mathematical and Scientific Machine Learning Conference (MSML)*, 2021. (Cited on 4, 34)

[22] Z. Zhu, T. Ding, J. Zhou, X. Li, C. You, J. Sulam, and Q. Qu, "A geometric analysis of neural collapse with unconstrained features," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. (Cited on 4, 34, 42, 43)

[23] W. Ji, Y. Lu, Y. Zhang, Z. Deng, and W. J. Su, "An unconstrained layer-peeled perspective on neural collapse," in *International Conference on Learning Representations (ICLR)*, 2022. (Cited on 4, 34, 42, 43)

[24] M. Caron, H. Touvron, I. Misra, H. J'egou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *International Conference on Computer Vision (ICCV)*, 2021. (Cited on 5, 6, 43, 44, 45)

[25] X. Chen and K. He, "Exploring simple siamese representation learning," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. (Cited on 5, 7, 8, 43, 44, 45)

[26] A. Mnih and K. Kavukcuoglu, "Learning word embeddings efficiently with noise-contrastive estimation," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2013. (Cited on 5)

[27] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:2110.02796*, 2019. (Cited on 7, 43, 45)

[28] I. Misra and L. van der Maaten, "Self-supervised learning of pretext-invariant representations," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. (Cited on 5)

[29] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Artificial Intelligence and Statistics (AISTATS)*, 2010. (Cited on 5, 37)

[30] R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu, "Exploring the limits of language modeling," *arXiv preprint arXiv:1602.02410*, 2016. (Cited on 37)

[31] Z. Ma and M. Collins, "Noise contrastive estimation and negative sampling for conditional models: Consistency and statistical efficiency," in *Empirical Methods in Natural Language Processing*, 2018. (Cited on 5, 37)

[32] Z. Fang, J. Wang, L. Wang, L. Zhang, Y. Yang, and Z. Liu, "SEED: Self-supervised distillation for visual representation," in *International Conference on Learning Representations (ICLR)*, 2021. (Cited on 6)

[33] E. T. Jaynes, "Information theory and statistical mechanics," *Physical Review*, vol. 106, no. 4, pp. 620–630, 1957. (Cited on 7)

[34] A. Ermolov, A. Siarohin, E. Sangineto, and N. Sebe, "Whitening for self-supervised representation learning," in *International Conference on Machine Learning (ICML)*, 2021. (Cited on 7, 8, 44, 45, 48)

[35] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. (Cited on 44)

[36] J. B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap Your Own Latent - a new approach to self-supervised learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. (Cited on 7, 8, 44, 45)

[37] Y. Bengio, A. C. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013. (Cited on 7, 43)

[38] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *European Conference on Computer Vision (ECCV)*, 2016. (Cited on 7, 43)

https://doi.org/10.52202/068431-0820

[39] T. M. Cover, “Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition,” *IEEE Transactions on Electronic Computers*, vol. EC-14, no. 3, pp. 326–334, Jun. 1965. (Cited on 7, 34, 35)

[40] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, “Barlow Twins: Self-supervised learning via redundancy reduction,” in *International Conference on Machine Learning (ICML)*, 2021. (Cited on 7, 44)

[41] T. Hua, W. Wang, Z. Xue, S. Ren, Y. Wang, and H. Zhao, “On feature decorrelation in self-supervised learning,” in *International Conference on Computer Vision (ICCV)*, 2021. (Cited on 7, 42)

[42] L. Jing, P. Vincent, Y. LeCun, and Y. Tian, “Understanding dimensional collapse in contrastive self-supervised learning,” in *International Conference on Learning Representations (ICLR)*, 2022. (Cited on 7, 42)

[43] T. Chen, C. Luo, and L. Li, “Intriguing properties of contrastive losses,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. (Cited on 8, 43, 45)

[44] Y. H. Tsai, Y. Wu, R. R. Salakhutdinov, and L. Morency, “Self-supervised learning from a multi-view perspective,” in *International Conference on Learning Representations (ICLR)*, 2021. (Cited on 8, 42, 45)

[45] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola, “What makes for good views for contrastive learning?” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[46] M. Federici, A. Dutta, P. Forré, N. Kushman, and Z. Akata, “Learning robust representations via multi-view information bottleneck,” in *International Conference on Learning Representations (ICLR)*, 2020. (Cited on 45)

[47] J. Mitrovic, B. McWilliams, J. Walker, L. Buesing, and C. Blundell, “Representation learning via invariant causal mechanisms,” in *International Conference on Learning Representations (ICLR)*, 2021. (Cited on 43)

[48] M. Wu, C. Zhuang, M. Mosse, D. L. K. Yamins, and N. D. Goodman, “On mutual information in contrastive learning for visual representations,” *arXiv preprint arXiv:2005.13149*, 2021. (Cited on 8, 42)

[49] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, “Deep clustering for unsupervised learning of visual features,” in *European Conference on Computer Vision (ECCV)*, 2018. (Cited on 8, 43, 44)

[50] Y. M. Asano, C. Rupprecht, and A. Vedaldi, “Self-labelling via simultaneous clustering and representation learning,” in *International Conference on Learning Representations (ICLR)*, 2020. (Cited on 43, 44)

[51] J. Yang, D. Parikh, and D. Batra, “Joint unsupervised learning of deep representations and image clusters,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. (Cited on 8)

[52] R. Sinkhorn and P. Knopp, “Concerning nonnegative matrices and doubly stochastic matrices,” *Pacific Journal of Mathematics*, vol. 21, no. 2, pp. 343–348, 1967. (Cited on 8, 44)

[53] C. Tosh, A. Krishnamurthy, and D. Hsu, “Contrastive estimation reveals topic posterior information to linear models,” *Journal of Machine Learning Research (JMLR)*, vol. 22, pp. 281:1–281:31, 2021. (Cited on 8)

[54] K. Nozawa and I. Sato, “Understanding negative samples in instance discriminative self-supervised representation learning,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. (Cited on 42)

[55] Y. Tian, X. Chen, and S. Ganguli, “Understanding self-supervised learning dynamics without contrastive pairs,” in *International Conference on Machine Learning (ICML)*, 2021. (Cited on 8, 43, 45)

[56] cs231n, “Tinyimagenet,” 2015. (Cited on 8, 47)

[57] N. Saunshi, J. T. Ash, S. Goel, D. Misra, C. Zhang, S. Arora, S. M. Kakade, and A. Krishnamurthy, “Understanding contrastive learning requires incorporating inductive biases,” in *International Conference on Machine Learning (ICML)*, 2022. (Cited on 9, 42)

https://doi.org/10.52202/068431-0820

[58] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, “Unsupervised feature learning via non-parametric instance discrimination,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. (Cited on 9)

[59] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. (Cited on 9, 47)

[60] P. Goyal, Q. Duval, J. Reizenstein, M. Leavitt, M. Xu, B. Lefaudeux, M. Singh, V. Reis, M. Caron, P. Bojanowski, A. Joulin, and I. Misra, “VISSL,” 2021. (Cited on 9, 48)

[61] S. Kornblith, T. Chen, H. Lee, and M. Norouzi, “Why do better loss functions lead to less transferable features?” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. (Cited on 10, 49)

[62] Y. Dubois, D. Kiela, D. J. Schwab, and R. Vedantam, “Learning optimal representations with the decodable information bottleneck,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. (Cited on 10, 42, 45)

[63] Y. Ruan, Y. Dubois, and C. J. Maddison, “Optimal representations for covariate shift,” in *International Conference on Learning Representations (ICLR)*, 2022. (Cited on 10, 42)

[64] M. L. Eaton, “Group invariance applications in statistics,” *Regional Conference Series in Probability and Statistics*, vol. 1, pp. i–133, 1989. (Cited on 22)

[65] E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses, Third Edition*, 3rd ed., ser. Springer texts in statistics. Springer, 2008. (Cited on 22)

[66] C. Zalinescu, *Convex Analysis in General Vector Spaces*. World Scientific, 2002. (Cited on 25)

[67] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[68] L. Narici and E. Beckenstein, *Topological Vector Spaces*, 2nd ed. Chapman and Hall/CRC, 2010. (Cited on 25)

[69] A. Brøndsted, *An introduction to convex polytopes*. Springer Science & Business Media, 1983, no. 90. (Cited on 29)

[70] G. M. Ziegler, *Lectures on polytopes*. Springer Science & Business Media, 2012, vol. 152.

[71] B. Grünbaum, V. Klee, M. A. Perles, and G. C. Shephard, *Convex polytopes*. Springer, 1967, vol. 16. (Cited on 29)

[72] S. Straszewicz, “Over exposed points of closed point sets,” *Fundamenta Mathematicae*, vol. 24, pp. 139–143, 1935. (Cited on 30)

[73] R. T. Rockafellar, *Convex analysis*. Princeton university press, 1970. (Cited on 30)

[74] S. Shalev-Shwartz and T. Zhang, “Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization,” *Mathematical Programming*, pp. 1–41, 2014. (Cited on 31, 32)

[75] S. Dutta and A. Goswami, “Mode estimation for discrete distributions,” *Mathematical Methods of Statistics*, vol. 19, no. 4, pp. 374–384, 2010. (Cited on 33)

[76] P. Flajolet, D. Gardy, and L. Thimonier, “Birthday paradox, coupon collectors, caching algorithms and self-organizing search,” *Discrete Applied Mathematics*, vol. 39, no. 3, pp. 207–229, 1992. (Cited on 33)

[77] P. Berenbrink and T. Sauerwald, “The weighted coupon collector’s problem and applications,” in *International Computing and Combinatorics Conference (COCOON)*, 2009. (Cited on 33)

[78] W. Feller, *An introduction to probability theory and its applications*. John Wiley & Sons, 2008. (Cited on 33)

[79] B. O’Neill, “The classical occupancy distribution: Computation and approximation,” *The American Statistician*, vol. 75, no. 4, pp. 364–375, 2021. (Cited on 33)

[80] C. Fang, H. He, Q. Long, and W. J. Su, “Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training,” *Proceedings of the National Academy of Sciences*, vol. 118, 2021. (Cited on 34, 42)

[81] D. J. C. MacKay, *Information theory, inference, and learning algorithms*. Cambridge University Press, 2003. (Cited on 35)

[82] A. Kowalczyk, "Dense shattering and teaching dimensions for differentiable families (extended abstract)," in *Conference on Learning Theory (COLT)*, 1997. (Cited on 35, 36)

[83] E. D. Sontag, "Remarks on interpolation and recognition using neural nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, 1990. (Cited on 35)

[84] ——, "Shattering all sets of k points in 'general position' requires (k - 1)/2 parameters," *Neural Computation*, vol. 9, no. 2, pp. 337–348, 1997. (Cited on 35, 36)

[85] E. B. Baum, "On the capabilities of multilayer perceptrons," *Journal of Complexity*, vol. 4, no. 3, pp. 193–215, 1988.

[86] A. Sakurai, "n-h-1 networks store no less n*h+1 examples, but sometimes no more," in *International Joint Conference on Neural Networks (IJCNN)*, vol. 3, 1992.

[87] G. J. Mitchison and R. M. Durbin, "Bounds on the learning capacity of some multi-layer networks," *Biological Cybernetics*, vol. 60, no. 5, pp. 345–365, 1989. (Cited on 35)

[88] T. M. Cover, "Capacity problems for linear machines," *Pattern recognition*, pp. 283–289, 1968. (Cited on 35, 36)

[89] P. L. Bartlett, N. Harvey, C. Liaw, and A. Mehrabian, "Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks," *JMLR*, vol. 20, pp. 63:1–63:17, 2019. (Cited on 36)

[90] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014. (Cited on 36)

[91] A. Kowalczyk, "Estimates of storage capacity of multilayer perceptron with threshold logic hidden units," *Neural Networks*, vol. 10, no. 8, pp. 1417–1433, 1997. (Cited on 36)

[92] C. Yun, S. Sra, and A. Jadbabaie, "Small relu networks are powerful memorizers: a tight analysis of memorization capacity," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[93] P. Baldi and R. Vershynin, "The capacity of feedforward neural networks," *Neural Networks*, vol. 116, pp. 288–311, 2019. (Cited on 36)

[94] A. Xu and M. Raginsky, "Minimum excess risk in bayesian learning," *IEEE Transactions on Information Theory*, 2022. (Cited on 37)

[95] O. Kallenberg, *Foundations of modern probability*. Springer, 1997. (Cited on 37)

[96] H. Bao, Y. Nagano, and K. Nozawa, "Sharp learning bounds for contrastive unsupervised representation learning," *arXiv preprint arXiv:2110.02501*, 2021. (Cited on 42)

[97] K. Nozawa, P. Germain, and B. Guedj, "Pac-bayesian contrastive unsupervised representation learning," in *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2020.

[98] J. T. Ash, S. Goel, A. Krishnamurthy, and D. Misra, "Investigating the role of negatives in contrastive representation learning," in *Artificial Intelligence and Statistics (AISTATS)*, 2022.

[99] P. Awasthi, N. Dikkala, and P. Kamath, "Do more negative samples necessarily hurt in contrastive learning?" in *International Conference on Machine Learning (ICML)*, 2022. (Cited on 43)

[100] Y. Wang, Y. Zhang, Y. Wang, J. Yang, and Z. Lin, "Chaos is a ladder: A new theoretical understanding of contrastive learning via augmentation overlap," in *International Conference on Learning Representations (ICLR)*, 2022. (Cited on 42)

[101] Y. Tian, L. Yu, X. Chen, and S. Ganguli, "Understanding self-supervised learning with dual deep networks," *arXiv preprint arXiv:2010.00578*, 2021. (Cited on 42, 43, 45)

[102] D. G. Mixon, H. Parshall, and J. Pi, "Neural collapse with unconstrained features," *Sampling Theory, Signal Processing, and Data Analysis*, vol. 20, no. 2, pp. 1–13, 2022. (Cited on 42)

[103] T. Wang and P. Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere," in *International Conference on Machine Learning (ICML)*, 2020. (Cited on 42)

https://doi.org/10.52202/068431-0820

[104] F. Wang and H. Liu, "Understanding the behaviour of contrastive loss," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. (Cited on 42)

[105] P. H. Richemond, J.-B. Grill, F. Altché, C. Tallec, F. Strub, A. Brock, S. L. Smith, S. De, R. Pascanu, B. Piot, and M. Valko, "BYOL works even without batch statistics," *arXiv preprint arXiv:2010.10241*, 2020. (Cited on 43, 45)

[106] C. Zhang, K. Zhang, C. Zhang, T. X. Pham, C. D. Yoo, and I. S. Kweon, "How does simsiam avoid collapse without negative samples? a unified understanding with self-supervised contrastive learning," in *ICLR*, 2022. (Cited on 43, 45)

[107] L. Ericsson, H. Gouk, and T. M. Hospedales, "Why do self-supervised models transfer? investigating the impact of invariance on downstream tasks," *arXiv preprint arXiv:abs/2111.11398*, 2021. (Cited on 43)

[108] A. Foster, R. Pukdee, and T. Rainforth, "Improving transformation invariance in contrastive representation learning," in *International Conference on Learning Representations (ICLR)*, 2021. (Cited on 43)

[109] J. von Kügelgen, Y. Sharma, L. Gresele, W. Brendel, B. Schölkopf, M. Besserve, and F. Locatello, "Self-supervised learning with data augmentations provably isolates content from style," in *NeurIPS*, 2021. (Cited on 43)

[110] T. Galanti, A. György, and M. Hutter, "On the role of neural collapse in transfer learning," in *International Conference on Learning Representations (ICLR)*, 2022. (Cited on 43)

[111] L. Hui, M. Belkin, and P. Nakkiran, "Limitations of neural collapse for understanding generalization in deep learning," *arXiv preprint arXiv:2202.08384*, 2022. (Cited on 43, 56)

[112] A. Bardes, J. Ponce, and Y. LeCun, "VICReg: Variance-invariance-covariance regularization for self-supervised learning," in *International Conference on Learning Representations (ICLR)*, 2022. (Cited on 44)

[113] X. Yan, I. Misra, A. Gupta, D. Ghadiyaram, and D. Mahajan, "ClusterFit: Improving generalization of visual representations," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. (Cited on 44)

[114] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, "Learning deep representations by mutual information estimation and maximization," in *International Conference on Learning Representations (ICLR)*, 2019. (Cited on 45)

[115] P. Bachman, R. D. Hjelm, and W. Buchwalter, "Learning representations by maximizing mutual information across views," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[116] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *European Conference on Computer Vision (ECCV)*, 2020. (Cited on 45)

[117] M. Tschannen, J. Djolonga, P. K. Rubenstein, S. Gelly, and M. Lucic, "On mutual information maximization for representation learning," in *International Conference on Learning Representations (ICLR)*, 2020. (Cited on 45)

[118] Y. Xu, S. Zhao, J. Song, R. Stewart, and S. Ermon, "A theory of usable information under computational constraints," in *International Conference on Learning Representations (ICLR)*, 2020. (Cited on 45)

[119] A. Pokle, J. Tian, Y. Li, and A. Risteski, "Contrasting the landscape of contrastive and non-contrastive learning," in *Artificial Intelligence and Statistics (AISTATS)*, 2022. (Cited on 45)

[120] L. Bossard, M. Guillaumin, and L. V. Gool, "Food-101 – mining discriminative components with random forests," in *European Conference on Computer Vision (ECCV)*, 2014. (Cited on 47)

[121] A. Krizhevsky, "Learning multiple layers of features from tiny images," University of Toronto, Tech. Rep., 2009. (Cited on 47, 50)

[122] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *ICCV Wokshop on 3D Representation and Recognition*, 2013. (Cited on 47)

[123] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," *arXiv preprint arXiv:1306.5151*, 2013. (Cited on 47)

https://doi.org/10.52202/068431-0820

[124] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi, “Describing textures in the wild,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. (Cited on 47)

[125] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar, “Cats and dogs,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. (Cited on 47)

[126] F.-F. Li, M. Andreeto, M. A. Ranzato, and P. Perona, “Caltech 101,” 2022. (Cited on 47)

[127] M.-E. Nilsback and A. Zisserman, “Automated flower classification over a large number of classes,” in *Indian Conference on Computer Vision, Graphics & Image Processing (ICVGIP)*, 2008. (Cited on 47)

[128] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. (Cited on 47)

[129] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *International Conference on Machine Learning (ICML)*, 2015, pp. 448–456. (Cited on 47)

[130] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *International Conference on Learning Representations (ICLR)*, 2015. (Cited on 47)

[131] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research (JMLR)*, vol. 12, 2011. (Cited on 49)

[132] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, pp. 273–297, 1995. (Cited on 49)

[133] L. McInnes, J. Healy, N. Saul, and L. Großberger, “UMAP: uniform manifold approximation and projection,” *Journal of Open Source Software*, vol. 3, no. 29, p. 861, 2018. (Cited on 56)

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] in Sec. 8 expanded in Appx. E .
   (c) Did you discuss any potential negative societal impacts of your work? [No] .
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] in the paper the main paper we mention the important ones succintly, in Appx. A.2 we discuss each in detail.
   (b) Did you include complete proofs of all theoretical results? [Yes] in Appx. B.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] all the code to reproduce our results can be found at `github.com/YannDubs/Invariant-Self-Supervised-Learning`.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] in Appx. F.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] for the TinyImageNet table. We did not for ImageNet due to computational constraints.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] .
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] .
   (b) Did you mention the license of the assets? [No] .
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A] .
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] .
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] .
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] .
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] .
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] .

https://doi.org/10.52202/068431-0820