RényiCL: Contrastive Representation Learning with Skew Rényi Divergence

Anonymous Author(s)
Affiliation
Address
email

Abstract

Contrastive representation learning seeks to acquire useful representations by estimating the shared information between multiple views of data. Here, the choice of data augmentation is sensitive to the quality of learned representations: as harder the data augmentations are applied, the views share more task-relevant information, but also task-irrelevant one that can hinder the generalization capability of representation. Motivated by this, we present a new robust contrastive learning scheme, coined RényiCL, which can effectively manage harder augmentations by utilizing Rényi divergence. Our method is built upon the variational lower bound of Rényi divergence, but a naïve usage of a variational method is impractical due to the large variance. To tackle this challenge, we propose a novel contrastive objective that conducts variational estimation of a skew Rényi divergence and provide a theoretical guarantee on how variational estimation of skew divergence leads to stable training. We show that Rényi contrastive learning objectives perform innate hard negative sampling and easy positive sampling simultaneously so that it can selectively learn useful features and ignore nuisance features. Through experiments on ImageNet, we show that Rényi contrastive learning with stronger augmentations outperforms other self-supervised methods without extra regularization or computational overhead. Moreover, we also validate our method on other domains such as graph and tabular, showing empirical gain over other contrastive methods.

1 Introduction

Recently, many AI studies are enamored by the power of contrastive methods in learning useful representations without any human supervision, e.g., image [1, 2, 3, 4, 5, 6], text [7, 8], audio [9, 10] and multimodal video [11, 12, 13]. The key components of contrastive learning can be summarized into two-fold: data augmentation and contrastive objectives. The data augmentation generates different views of data where the views share information that is relevant to the downstream tasks. The contrastive objective then enforces the representation to capture the shared information between the views by bringing the views from the same data together and pushing the views from different data away in the representation space. Therefore, the choice of data augmentation is critical in contrastive learning, cf., see [3, 14, 15]. If the views share insufficient information, the representations cannot learn sufficient features for downstream tasks. On the other hand, if there is too much information between the views, the views might share nuisance features that hamper the generalization.

In this paper, we propose a new contrastive learning objective using Rényi divergence [16], coined *RényiCL*, which can effectively handle the case when the views share excessive information. The Rényi divergence is a family of probabilistic divergence that is known for being "less forgiving"; they penalize more when two distributions have more differences. Thus by modeling contrastive learning with Rényi divergence, we can learn better representations under the harder augmentations.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Specifically, we show that RényiCL intrinsically gives more weight to easy positives and hard negatives, emphasizing the ability of RényiCL to focus more on informative views and less on nuisance views.

To implement RényiCL, one can consider variational lower bound of Rényi divergence [17]. However, we show that a variational lower bound of Rényi divergence does not suffice for contrastive learning due to their large variance. To tackle this challenge, we first made a key observation that the existing contrastive objectives are a variational form of a skew KL divergence, and show that variational estimation of skew divergence induces low variance property. Inspired by this, we consider a variational lower bound of a skew Rényi divergence and use it to implement RényiCL.

Through experiments under ImageNet [18], RényiCL (with harder data augmentations) achieves 75.3% linear evaluation accuracy with only 200 training epochs,[1] outperforming other self-supervised methods without using large batch size or extra computational overhead. Moreover, we also show that RényiCL achieves better performance on transfer learning and few-shot classification, compared to baselines. Finally, we demonstrate that RényiCL provides empirical gains over original contrastive learning methods on various domains including images, graphs, and tabular.

Related works. Since Chen et al. [3] emphasized that the usage of data augmentation is crucial for contrastive learning, a series of works proposed various contrastive learning frameworks that showed great empirical success on visual representation learning [2, 3, 4, 5, 6, 19]. Inspired by the empirical success, recent works focused on optimal view generation for both positive and negatives. Tian et al. [14] proposed the InfoMin principle for view generation, which states that the views should share minimal information that is relevant to the downstream tasks. Another line of research focuses on hard negative samples [20, 21, 22] to enhance the performance of contrastive learning. Especially, Robinson et al. [20] proposed to use importance sampling to learn with hard negative samples, which can expedite the contrastive learning. Meanwhile, others are interested in searching for different contrastive objectives by modeling with probabilistic divergences other than KL divergence. Hjelm et al. [23] proposed DeepInfoMax which used Jensen-Shannon divergence for variational estimation and maximization of mutual information. Zbontar et al. [24] proposed Barlow-Twins, which is shown to be a Hilbert-Schmidt Independence Criterion that approximates a Maximum Mean discrepancy measure between positives and negatives [25]. Our work is similar to that of Tsai et al. [26], where they proposed relative predictive coding that uses skew χ^2 divergence to model contrastive learning.

2 Preliminaries

2.1 Background on contrastive learning

Let $\mathcal{X}$ be a set of natural data (without any data augmentation) and denote p be a density of data distribution. To formalize data augmentation, let $v = t(x)$ be a view from a data $x \in \mathcal{X}$ by stochastic data transformation $t : \mathcal{X} \to \mathcal{X}$ with transition probability $\mathcal{A}(v|x)$. Let $V = t(X)$ be a random variable for views with distribution P_V, then the density of a random variable V can be written as $p(v) = \mathbb{E}_{p(x)}[\mathcal{A}(v|x)]$. To perform multiview contrastive learning, let V' be another random variable of views from data transformation t'. Then given a pair of views $(v, v') \sim V \times V'$, the joint density can be written by $p(v, v') = \mathbb{E}_{p(x)}[\mathcal{A}(v|x)\mathcal{A}(v'|x)]$. A pair of views (v, v') is called a *positive pair* if they are augmented from same data, i.e., $v \sim \mathcal{A}(\cdot|x)$ and $v' \sim \mathcal{A}(\cdot|x)$ for some $x \in \mathcal{X}$. If a pair of views (v, v') are augmented from different data, we call it a *negative pair* and we have

$$p(v, v') = \mathbb{E}_{p(x)}[\mathcal{A}(v|x)\mathcal{A}(v'|x)] = \mathbb{E}_{p(x)}[\mathcal{A}(v|x)]\mathbb{E}_{p(x)}[\mathcal{A}(v'|x)] = p(v)p(v').$$

Throughout the paper, we denote v^+ for positive counterpart and v^- for negative counterpart of v, i.e. $(v, v^+) \sim p(v, v^+)$ and $(v, v^-) \sim p(v)p(v^-)$. Also, denote $P_{V,V'}$ and $P_V \times P_{V'}$ be distribution of positive and negative pairs respectively.

Contrastive representation learning. The goal of contrastive representation learning is to train an encoder $g : \mathcal{X} \to \mathbb{R}^d$ that is a useful representation for downstream tasks by enforcing g to discriminate between positive pairs and negative pairs. From our definition, the KL divergence between the positives and negatives coincides with the *mutual information* (MI) between the views:

$$\mathcal{I}(V; V') \coloneqq D_{\mathrm{KL}}(P_{V,V'} \| P_V \times P_{V'}).$$

[1] Due to our current limited computing resources, we could not try longer epochs, but we strongly believe that RényiCL can be improved further by training >200 epochs. We will report such results in the revised draft.

Then the contrastive representation learning follows InfoMax principle [27, 28, 23], which can be formulated into following optimization problem:

$$\max_g D_{\mathrm{KL}}(P_{g(V),g(V')} \| P_{g(V)} \times P_{g(V')}) = \max_g \mathcal{I}(g(V); g(V')) \leq \mathcal{I}(V; V'),$$

here the inequality always hold due to the data-processing inequality, and equality holds iff g is injective. Generally, we use a neural network for encoder g (such as ResNet50 [29]). If two views share sufficient task-relevant information, the encoder learns useful information that is invariant against nuisance features. Therefore, the choice of data augmentation for view generation greatly affects the performance of contrastive learning [14].

Contrastive learning objectives. Since the MI cannot be computed unless the densities are known, many works resort to the variational lower bounds of MI [30] associated with neural networks [31]. Among them, the *contrastive predictive coding* (CPC) objective (also as known as InfoNCE) [32] is a lower bound to the MI that is widely used for contrastive representation learning [2, 4, 3]. Given B batch of views $\{v_i\}_{i=1}^B$, assume we have a single positive v_i^+ (i.e., $v_i^+ \sim p(v_i^+|v_i)$) and K negatives $\{v_{ij}^-\}_{j=1}^K$ (i.e., $(v_i, v_{ij}^-) \sim p(v_i)p(v_{ij}^-)$) for each v_i. Then the CPC objective is given by following:

$$\mathcal{I}_{\mathrm{CPC}}(f) = \mathbb{E}\left[\frac{1}{B}\sum_{i=1}^B \log \frac{(K+1)\cdot e^{f(v_i,v_i^+)}}{e^{f(v_i,v_i^+)} + \sum_{j=1}^K e^{f(v_i,v_{ij}^-)}}\right].$$

Then we have $\mathcal{I}_{\mathrm{CPC}}(f) \leq \mathcal{I}(V; V')$ for any bounded measurable function $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ [32]. However, as $\mathcal{I}_{\mathrm{CPC}}(f) \leq \log(K+1)$, the CPC objective becomes high-bias MI estimator if the true MI is larger than $\log(K+1)$. To address this, Song et al. [33] proposed α-CPC which can control the bias by inserting $\alpha \in [0,1]$ into CPC as following:

$$\mathcal{I}_{\mathrm{CPC}}^{(\alpha)}(f) = \mathbb{E}\left[\frac{1}{B}\sum_{i=1}^B \log \frac{e^{f(v_i,v_i^+)}}{\alpha e^{f(v_i,v_i^+)} + \frac{1-\alpha}{K}\sum_{j=1}^K e^{f(v_i,v_{ij}^-)}}\right]. \tag{1}$$

The α-CPC also admits a variational lower bound to the MI for any $\alpha \in [0,1]$ (equivalent to the original CPC when $\alpha = \frac{1}{K+1}$), and we can have smaller bias by using small α. Furthermore, they introduced α-*Multi Label CPC* (MLCPC) which is a tighter lower bound to the MI, defined by

$$\mathcal{I}_{\mathrm{MLCPC}}^{(\alpha)}(f) = \mathbb{E}\left[\frac{1}{B}\sum_{i=1}^B \log \frac{e^{f(v_i,v_i^+)}}{\frac{\alpha}{B}\sum_{i=1}^B e^{f(v_i,v_i^+)} + \frac{1-\alpha}{BK}\sum_{i=1}^B\sum_{j=1}^K e^{f(v_i,v_{ij}^-)}}\right]. \tag{2}$$

One can easily see that $\forall \alpha \in [0,1], \mathcal{I}_{\mathrm{CPC}}^{(\alpha)}(f) \leq \mathcal{I}_{\mathrm{MLCPC}}^{(\alpha)}(f) \leq \mathcal{I}(V; V')$ by using Jensen's inequality.

2.2 Rényi divergence

Let P and Q be two probability distributions with densities p and q with respect to a base measure dx on domain $\mathcal{X}$. Suppose P is absolutely continuous with respect to Q, denoted by $P \ll Q$ (i.e., the Radon-Nikodym derivative dP/dQ exists). Then the *Rényi divergence* [16] of order $\gamma \in (0,1) \cup (1,\infty)$ between P and Q is defined as

$$R_\gamma(P\|Q) \coloneqq \frac{1}{\gamma(\gamma-1)} \log \mathbb{E}_P\left[\left(\frac{dP}{dQ}\right)^{\gamma-1}\right] = \frac{1}{\gamma(\gamma-1)} \log \int_{\mathcal{X}} p^\gamma(x) q^{1-\gamma}(x) dx.$$

Remark that Rényi divergence is intertwined with f-divergences, e.g., it becomes KL divergence as $\gamma \to 1$, and Rényi divergence with $\gamma = 2$ corresponds to χ^2 divergence. We refer to [34] for detailed characteristics of Rényi divergence.

Variational representation of Rényi divergence. To perform contrastive learning with Rényi divergence, we consider variational methods similar to CPC for MI. The following lemma introduces a variational representations of the Rényi divergence.

Lemma 2.1 ([17]). *Let P and Q be two probability distributions with densities p and q with respect to a base measure dx on domain $\mathcal{X}$. Let $\mathcal{F}$ be a set of bounded measurable functions on $\mathcal{X}$. Suppose $P \ll Q$, then the Rényi divergence of order $\gamma \in (0,1) \cup (1,\infty)$ admits following variational form:*

$$R_\gamma(P\|Q) = \sup_{f\in\mathcal{F}} \frac{1}{\gamma-1} \log \mathbb{E}_P[e^{(\gamma-1)f(x)}] - \frac{1}{\gamma} \log \mathbb{E}_Q[e^{\gamma f(x)}], \tag{3}$$

where the optimum $f^* \in \mathcal{F}$ satisfies $f^* \propto \log \frac{dP}{dQ}$.

Note that (3) is a generalization of Donsker-Varadhan (DV) [35] objective, which is a variational representation of KL divergence defined as following:

$$D_{\mathrm{KL}}(P\|Q) = \sup_{f\in\mathcal{F}} \mathbb{E}_P[f(x)] - \log \mathbb{E}_Q[e^{f(x)}]. \tag{4}$$

Note that the optimum $f^* \in \mathcal{F}$ of (4) satisfies $f^* \propto \log \frac{dP}{dQ}$ and (3) is equivalent to (4) if $\gamma \to 1$.

3 Rényi Contrastive Learning

3.1 Challenges in contrastive learning with Rényi divergence

Similar to the InfoMax principle, we use Rényi divergence to characterize the shared information between the views. Thus, the goal of Rényi contrastive learning is to solve following:

$$\max_g \; R_\gamma(P_{g(V),g(V')}\|P_{g(V)} \times P_{g(V')}) \le R_\gamma(P_{V,V'}\|P_V \times P_{V'}).$$

To that end, one may consider the variational lower bound of Rényi divergence in (3). However, the contrastive learning with (3) is impractical due to the large variance. That large-variance problem of (3) resembles that of the variational MI estimator observed in [30]. Song et al. [36] showed that the variational MI estimators such as DV objective, require a large number of samples proportional to the exponential of ground-truth MI to achieve low variance. In the following theorem, we provide a similar result for the Rényi variational objective in (3).

Theorem 3.1. *Assume $P \ll Q$, and $Var_Q[dP/dQ] < \infty$. Let P_m and Q_n be the empirical distributions of m i.i.d samples from P and n i.i.d samples from Q, respectively. Define*

$$\mathcal{I}_{\textit{Rényi}}(f) \coloneqq \frac{1}{\gamma-1}\log \mathbb{E}_{P_m}[e^{(\gamma-1)f}] - \frac{1}{\gamma}\log \mathbb{E}_{Q_n}[e^{\gamma f}],$$

and assume we have optimum $f^* \propto \log(dP/dQ)$. Then $\forall\gamma > 1, \forall m \in \mathbb{N}$, we have

$$\lim_{n\to\infty} n\mathit{Var}_{P,Q}[\mathcal{I}_{\textit{Rényi}}(f^*)] \ge \frac{e^{\gamma D_{KL}(P\|Q)} - \gamma}{e^{2\gamma(\gamma-1)R_\gamma(P\|Q)}}.$$

The proof is in the appendix. Theorem 3.1 indicates that even though we achieve the optimal function, the variance of (3) could explode exponentially with the ground-truth MI. In the appendix, we show that even for a simple synthetic Gaussian dataset, (3) is not able to estimate Rényi divergence between two correlated Gaussian distributions if the correlation coefficient is large. Note that our result generalizes that of [36] for KL divergence if we take $\gamma \to 1$.

3.2 Contrastive objectives are variational skew divergence estimators

In this section, we claim that the CPC and MLCPC objectives are variational lower bounds of a skew KL divergence, and show that variational estimation of skew divergence induces a low variance estimator. For P and Q with $P \ll Q$ and for any $\alpha \in [0,1]$, the α-skew KL divergence between P and Q is defined by the KL divergence between P and the mixture $\alpha P + (1-\alpha)Q$:

$$D^{(\alpha)}_{\mathrm{KL}}(P\|Q) \coloneqq D_{\mathrm{KL}}(P\|\alpha P + (1-\alpha)Q)$$

Then the DV objective (4) for α-skew KL divergence can be written as

$$D^{(\alpha)}_{\mathrm{KL}}(P\|Q) = \sup_f \; \mathbb{E}_P[f(x)] - \log\left(\alpha\mathbb{E}_P[e^{f(x)}] + (1-\alpha)\mathbb{E}_Q[e^{f(x)}]\right). \tag{5}$$

Note that the α-CPC objective in (1) can be written as following generalized form:

$$\mathcal{I}^{(\alpha)}_{\mathrm{CPC}}(f) = \mathbb{E}_{v,v^+}[f(v,v^+)] - \mathbb{E}_{p(v)}[\log(\alpha\mathbb{E}_{v^+}[e^{f(v,v^+)}] + (1-\alpha)\mathbb{E}_{v^-}[e^{f(v,v^-)}])]. \tag{6}$$

Then one can observe that (6) is equivalent to the expectation of the variational DV objective (5) of α-skew KL divergence between $p(v^+|v)$ and $p(v^-)$. Thus we have

$$\sup_f \; \mathcal{I}^{(\alpha)}_{\mathrm{CPC}}(f) = \mathbb{E}_{P_V}[D^{(\alpha)}_{\mathrm{KL}}(P_{V'|V}\|P_{V'})] = D^{(\alpha)}_{\mathrm{KL}}(P_{V',V}\|P_{V'} \times P_V),$$

showing that the α-CPC is a DV variational lower bound of α-skew KL divergence between the $P_{V',V}$ and $P_{V'} \times P_V$. Similarly, the generalized form of α-MLCPC objective (2) is given by

$$\mathcal{I}^{(\alpha)}_{\text{MLCPC}}(f) = \mathbb{E}_{v,v^+}[f(v,v^+)] - \log\left(\alpha\mathbb{E}_{v,v^+}[e^{f(v,v^+)}] + (1-\alpha)\mathbb{E}_{v,v^-}[e^{f(v,v^-)}]\right),$$

and in the appendix, we show that α-MLCPC also admits a variational form of α-skew KL divergence.

Low variance properties of CPC and MLCPC objectives. Now we show that the variational estimation of empirical skew divergence has low variance, explaining how CPC and MLCPC objectives become a low-variance estimator of MI. The following theorem shows that the variance of the variational estimator can be adjusted with the right choice of α.

Theorem 3.2. *Assume $P \ll Q$ and $Var_Q[dP/dQ] < \infty$. Let P_m and Q_n be empirical distributions of m i.i.d samples from P and n i.i.d samples from Q. Then define*

$$\mathcal{I}^{(\alpha)}_{KL}(f) = \mathbb{E}_{P_m}[f] - \log\left(\alpha\mathbb{E}_{P_m}[e^f] + (1-\alpha)\mathbb{E}_{Q_n}[e^f]\right),$$

and assume that there is $\widehat{f} \in \mathcal{F}$ that $|\mathcal{I}^{(\alpha)}_{KL}(\widehat{f}) - D^{(\alpha)}_{KL}(P_m\|Q_n)| < \varepsilon_f$ for some $\varepsilon_f > 0$. Then for $\forall\alpha < 1/8$, the variance of estimator satisfies

$$Var_{P,Q}[\mathcal{I}^{(\alpha)}_{KL}(\widehat{f})] \leq c_1\varepsilon_f + \frac{c_2}{\min\{n,m\}} + \frac{c_3\log^2(\alpha n)}{n} + \frac{c_4\log^2(c_5 m)}{\alpha^2 m},$$

for some constants $c_1, c_2, c_3, c_4, c_5 > 0$ independent of n and m.

The proof is in the appendix. Theorem. 3.2 demonstrates that if we can find a sufficiently close critic for empirical skew KL divergence, the variance of the estimator is asymptotically bounded unless we choose large enough $\alpha = \alpha_m = \omega(m^{-1/2})$. Also, the ε_f can be further formalized with universal approximation theorem [37], if we use a neural network for f.

3.3 Rényi contrastive learning objectives

Now we present Rényi contrastive learning (RényiCL) objectives based on variational estimation of skew Rényi divergence. As shown in Section 3.2, we consider the variational lower bound of skew Rényi divergence to guarantee stable optimization. Remark that the analogous version of Theorem 3.2 can be achieved for Rényi divergence, which we refer to in the appendix. For any $\alpha \in [0,1]$ and $\gamma \in (0,1) \cup (1,\infty)$, the α-skew Rényi divergence of order γ is defined as:

$$R^{(\alpha)}_{\gamma}(P\|Q) \coloneqq R_\gamma(P\|\alpha P + (1-\alpha)Q).$$

Then we define (α,γ)-*Rényi MLCPC* (RMLCPC) objective by using variational lower bound (3) for α-skew α-skew Rényi divergence between $P_{V,V'}$ and $P_V P_{V'}$:

$$\mathcal{I}^{(\alpha,\gamma)}_{\text{RMLCPC}}(f) \coloneqq \frac{1}{\gamma-1}\log\mathbb{E}_{v,v^+}[e^{(\gamma-1)f(v,v^+)}] - \frac{1}{\gamma}\log(\alpha\mathbb{E}_{v,v^+}[e^{\gamma f(v,v^+)}] + (1-\alpha)\mathbb{E}_{v,v^-}[e^{\gamma f(v,v^-)}]), \tag{7}$$

which satisfies $\sup_f \mathcal{I}^{(\alpha,\gamma)}_{\text{RMLCPC}}(f) = R^{(\alpha)}_{\gamma}(P_{V,V'}\|P_V \times P_{V'})$ for any $\alpha \in [0,1]$ and $\gamma \in (0,1)\cup(1,\infty)$. Note that the RMLCPC objective can be used for contrastive learning with multiple positive views. For example, when using multi-crops data augmentation [38], instead of computing contrastive objective for each crop, we gather all positive pairs and negative pairs and compute only once for the final loss. The PyTorch style pseudo-code for the RMLCPC objective can be found in the appendix.

Gradient analysis for Rényi contrastive learning. We provide an analysis of how RényiCL can manage harder views based on the gradient of contrastive objectives. Without loss of generality, we let $\alpha = 0$. Then note that the MLCPC objective becomes the DV objective in (4). Let f_θ be a neural network parameterized by θ, then the gradient of $\mathcal{I}_{\text{MLCPC}}(f_\theta)$ is

$$\begin{aligned}\nabla_\theta\mathcal{I}_{\text{MLCPC}}(f_\theta) &= \mathbb{E}_{v,v^+}[\nabla_\theta f_\theta(v,v^+)] - \frac{\mathbb{E}_{v,v^-}[e^{f_\theta(v,v^-)}\nabla_\theta f_\theta(v,v^-)]}{\mathbb{E}_{v,v^-}[e^{f_\theta(v,v^-)}]} \\ &= \mathbb{E}_{v,v^+}[\nabla_\theta f_\theta(v,v^+)] - \mathbb{E}_{q_\theta(v,v^-)}[\nabla_\theta f_\theta(v,v^-)],\end{aligned}$$

where $q_\theta(v, v^-) \propto p(v)p(v^-)e^{f_\theta(v,v^-)}$ is a importance weights. Thus the MLCPC objective is equivalent to following in terms of gradient:

$$\mathcal{I}_{\text{MLCPC}}(f_\theta) = \mathbb{E}_{p(v,v^+)}[f_\theta(v, v^+)] - \mathbb{E}_{q_\theta(v,v^-)}[f_\theta(v, v^-)],$$

which is same as contrastive divergence that is used for training energy-based models [39]. This shows that the original contrastive objective performs innate hard negative sampling with importance weight q_θ [20], as the gradient of negative pairs becomes larger as the value of f_θ becomes larger. The general argument for α-MLCPC and α-CPC is in the appendix. On the other hand, for RMLCPC objective, by letting $\alpha = 0$, and taking gradient with respect to θ, we have

$$\begin{aligned}\nabla_\theta \mathcal{I}^{(\gamma)}_{\text{RMLCPC}}(f_\theta) &= \frac{\mathbb{E}_{v,v^+}[e^{(\gamma-1)f_\theta(v,v^+)}\nabla_\theta f_\theta(v, v^+)]}{\mathbb{E}_{v,v^+}[e^{(\gamma-1)f_\theta(v,v^+)}]} - \frac{\mathbb{E}_{v,v^-}[e^{\gamma f_\theta(v,v^-)}\nabla_\theta f_\theta(v, v^-)]}{\mathbb{E}_{v,v^-}[e^{\gamma f_\theta(v,v^-)}]} \\ &= \mathbb{E}_{q_\theta(v,v^+;\gamma-1)}[\nabla_\theta f_\theta(v, v^+)] - \mathbb{E}_{q_\theta(v,v^-;\gamma)}[\nabla_\theta f_\theta(v, v^-)],\end{aligned}$$

where $q_\theta(v, v^+; \gamma - 1) \propto p(v, v^+)e^{(\gamma-1)f_\theta(v,v^+)}$ and $q_\theta(v, v^-; \gamma) \propto p(v)p(v^-)e^{\gamma f_\theta(v,v^-)}$. Thus the RMLCPC objective is equivalent to following in terms of gradient:

$$\mathcal{I}^{(\gamma)}_{\text{RMLCPC}}(f_\theta) = \mathbb{E}_{q_\theta(v,v^+;\gamma-1)}[f_\theta(v, v^+)] - \mathbb{E}_{q_\theta(v,v^-;\gamma)}[f_\theta(v, v^-)]$$

When comparing MLCPC and RMLCPC, it occurs two differences:

- **Hard negative sampling**: the gradient weights more on harder negatives, i.e., (v, v^-) with high value of $f_\theta(v, v^-)$, as γ increases [20].
- **Easy positive sampling**: the gradient weights more on easier positives, i.e., (v, v^+) with high value of $f_\theta(v, v^+)$ as $\gamma \in (1, \infty)$ increases.

Therefore, when there are positive views that share less information, the RMLCPC objective resists updating, and it regularizes the model to ignore task-irrelevant information. The hard negative sampling with importance weight was proposed in [20], but our analysis shows that CPC and MLCPC conduct intrinsic hard negative sampling. Also, the Rényi contrastive learning can control the level of hard negative sampling by choosing the appropriate γ.

3.4 MI estimation with contrastive objectives

By interpreting contrastive objectives with a variational estimator of skew KL divergence, we can explain the large bias of CPC and MLCPC. Note that $\forall \alpha \in (0, 1)$,

$$D^{(\alpha)}_{\text{KL}}(P\|Q) \le (1-\alpha)D_{\text{KL}}(P\|Q) < D_{\text{KL}}(P\|Q),$$

from the convexity of KL divergence. Therefore, CPC and MLCPC are intrinsically strictly-biased MI estimators, and similarly, the RMLCPC is a strictly-biased estimator of Rényi divergence. However, we provide a general method to recover the MI by using critic f learned from CPC, MLCPC, and RMLCPC objectives. From the optimal condition of (4) and (3), the optimal critic f^* of α-CPC, α-MLCPC and (α, γ)-RMLCPC of any γ satisfies

$$f^*(v, v') \propto \log \frac{p(v, v')}{\alpha p(v, v') + (1-\alpha)p(v)p(v')} = \log \frac{r(v, v')}{\alpha r(v, v') + 1 - \alpha},$$

where $r(v, v') := p(v, v')/(p(v)p(v'))$ is a density ratio. Thus, we have $\exp(f^*(v, v'))/Z = \frac{r(v,v')}{\alpha r(v,v')+1-\alpha}$ for some normalization constant $Z > 0$, and we use Monte-Carlo estimation $\widehat{Z}$ to approximate Z. Then the approximate density ratio $\hat{r}(v, v')$ is given by $\hat{r}(v, v') = \frac{(1-\alpha)e^{f^*(v,v')}}{\widehat{Z}-\alpha e^{f^*(v,v')}}$. Lastly, we estimate the MI by the empirical mean of the log-density ratio:

$$I(V; V') = \frac{1}{B}\sum_{i=1}^{B} \log \hat{r}(v_i, v_i^+). \tag{8}$$

In Section 4.3 and the appendix, we empirically show that (8) can effectively estimate the true MI with low variance and low bias.

Table 1: Linear evaluation on the ImageNet validation set. We report pre-training epochs and Top-1 classification accuracies (%).

Method	Epochs	Top-1
SimCLR [3]	800	70.4
SimSiam [40]	800	71.3
Barlow Twins [24]	800	73.2
BYOL [41]	800	74.3
MoCo v3 [6]	800	74.6
NNCLR [19]	800	74.9
SwAV [38]	800	**75.3**
RényiCL	**200**	**75.3**

Table 2: Semi-supervised learning results on ImageNet. We report Top-1 and Top-5 classification accuracies (%) by fine-tuning a pre-trained ResNet-50 with 1% and 10% ImageNet datasets.

	1% ImageNet		10% ImageNet	
Method	Top-1	Top-5	Top-1	Top-5
Supervised [3]	25.4	48.4	56.4	80.4
SimCLR [3]	48.3	75.5	65.6	87.8
BYOL [41]	53.2	78.4	68.8	89.0
SwAV [38]	53.9	78.5	70.2	89.9
Barlow Twins [24]	55.0	79.2	69.7	89.3
NNCLR [19]	**56.4**	**80.7**	69.8	89.3
RényiCL	55.9	80.3	**70.3**	**90.0**

Table 3: Transfer learning performance on object classification datasets. For Aircraft and Flowers, we report mean per-class accuracy (%), and for VOC2007, we report 11-point mAP. Otherwise, we report Top-1 classification accuracies (%).

Method	CIFAR10	CIFAR100	Food101	Flowers	Cars	Aircraft	DTD	SUN397	VOC2007
Supervised [3]	93.6	78.3	72.3	94.7	66.7	61.0	74.9	61.9	82.8
SimCLR [3]	90.6	71.6	68.4	91.2	50.3	50.3	74.5	58.8	81.4
BYOL [41]	91.3	78.4	75.3	96.1	**67.8**	60.6	75.5	62.2	82.5
NNCLR [19]	93.7	**79.0**	76.7	95.1	67.1	**64.1**	75.5	62.5	83.0
RényiCL	**94.4**	**79.0**	**77.5**	**96.2**	66.4	61.1	**76.3**	**65.9**	**88.2**

4 Experiments

4.1 ImageNet results

Setup. For ImageNet [18] experiments, we use ResNet-50 [29] for encoder g. We use MLP projection head with momentum encoder [4], which is updated by EMA, and we use predictor, following the practice of [41, 6]. Then the critic f is implemented by the cosine similarity between the output of momentum encoder and base encoder with predictor, divided by temperature $\tau = 0.5$. We maximize RMLCPC objective defined in (7) with $\gamma = 2.0$. For optimization, we use LARS [42] optimizer with a batch size of 512, and train for 200 epochs.

Data augmentation. Given the popular data augmentation baseline [41, 3], we further use RandAugment [43] and multi-crop augmentation [38] to implement harder data augmentation. When using multi-crop, we gather all positives and negatives, then directly compute the RMLCPC objective, which differs from the original method in [38] (see the appendix for details).

Linear evaluation. We follow the linear evaluation protocol, where we report the Top-1 ImageNet validation accuracy (%) of a linear classifier trained on the top of frozen features. Table 1 compares the performance of different self-supervised methods using linear evaluation protocol. First, RényiCL achieves better performance than other self-supervised methods with only 200 epochs of training. Meanwhile, RényiCL enjoys better computational efficiency in that it does not require a large batch size [3, 41, 38, 6] or extra memory queue [19].

Semi-supervised learning. We evaluate the usefulness of the learned feature in a semi-supervised setting with 1% and 10% subsets of ImageNet dataset [3, 38]. The results are in Table 2. We observe that RényiCL achieves the best performance comparable with SwAV [38] on fine-tuning with 10% subset, and slightly less than NNCLR [19] on fine-tuning with 1% subset, validating that our method has decent generalization capacity in low-shot learning scenarios.

Transfer learning. We evaluate the learned representation through the performance of transfer learning on fine-grained object classification datasets by using linear evaluation. We use CIFAR10/100 [44], Food101 [45], Flowers [46], Cars [47], Aircraft [48], DTD [49], SUN397 [50], and VOC2007 [51]. In Table 3, we compare it with other self-supervised methods. Note that RényiCL achieves the best performance in 7 out of 9 datasets, especially showing superior performance on SUN397 (3.4%) and VOC2007 (5.2%) datasets.

Table 4: Comparison on contrastive self-supervised methods with harder augmentations. We compare Top-1 linear evaluation accuracy (%) on ImageNet, Pets, Caltech101 datasets and few-shot classification accuracy (%) over 2000 episodes on FC100, CUB200, and Plant disease datasets. † denotes trained without multi-crops. Here, we consider RényiCL without multi-crops as other baselines do, i.e., for fair comparison.

		Linear evaluation			5-way 1-shot			5-way 5-shot		
Method	Epochs	ImageNet	Pets	Caltech101	FC100	CUB200	Plant	FC100	CUB200	Plant
InfoMin [14]	800	**73.0**	86.37	88.48	31.80	53.81	66.11	45.09	72.20	84.12
CLSA [52]	800	72.2	85.18	91.21	34.46	50.83	67.39	49.16	67.93	86.15
RényiCL†	200	72.6	**88.43**	**94.04**	**36.31**	**59.73**	**80.68**	**53.39**	**82.12**	**94.29**

Table 5: Ablation on contrastive objectives for unsupervised representation learning on image (ImageNet-100, CIFAR-100, CIFAR-10) and tabular (CovType, Higgs-100K) datasets. We report Top-1 accuracy (%) with linear evaluation. Average over 5 runs.

	ImageNet-100		CIFAR-100		CIFAR-10		CovType		Higgs-100K	
Method	Base	Hard	Base	Hard	Base	Hard	Base	Hard	Base	Hard
CPC	79.7	81.1(+1.4)	65.4	67.1(+1.7)	91.7	91.9(+0.2)	71.6	74.3(+2.7)	64.7	71.3(+6.6)
MLCPC	79.5	81.2(+1.7)	65.6	66.6(+1.0)	91.9	92.1(+0.2)	71.7	74.1(+2.4)	64.9	71.5(+6.6)
RMLCPC	78.9	**81.6(+2.7)**	64.5	**68.5(+4.0)**	90.7	**92.5(+1.8)**	72.1	**74.9(+2.8)**	64.5	**72.4(+7.9)**

Comparison among self-supervised methods using hard augmentations. To demonstrate the effectiveness of RényiCL on learning with harder augmentations, we compare with InfoMin [14] and CLSA [52] where they also use harder augmentations for contrastive representation learning. Table 4 demonstrates the performance of representations learned with different self-supervised methods by linear evaulation on ImageNet [18], Pets [53], Caltech101 [54], and few-shot classification on FC100 [55], Caltech-UCSD Birds (CUB200) [43], and Plant Disease [56] datasets. While RényiCL achieves slightly lower performance than InfoMin on ImageNet validation accuracy, it outperforms InfoMin and CLSA on other transfer learning tasks with a large margin, which indicates the superiority of RényiCL on learning better representations under the hard data augmentations.

4.2 RényiCL for representation learning on various domains

Setup. We perform ablation studies of RényiCL on various domains such as images, tabular, and graph datasets. For images, we consider CIFAR-10/100 [44] and ImageNet-100 [57], which is a 100-class subset of ImageNet [18]. For the base data augmentation, we use a popular benchmark from [3], and for the harder augmentation, we further apply RandAugment [43] and RandomErasing [58]. For ImageNet-100, we use the default settings of [6] with ResNet-50 backbone with only change in learning objective. For CIFAR-10 and CIFAR-100, we follow the settings in [3] with ResNet-18 backbone adjusted for the CIFAR dataset. We all use same α for CPC, MLCPC, RMLCPC, and use $\gamma = 2.0$ for RMLCPC.

For tabular experiments, we use Forest Cover Type (CovType) and Higgs Boson (Higgs) [59] dataset from UCI repository [60]. Due to its massive size, we consider a subset of 100K for the Higgs experiments. Since the tabular dataset has limited domain knowledge, we do not use any data augmentation for the baseline and when using data augmentation, we use random masking noise [61] for the Higgs dataset and random feature corruption [62] for the CovType dataset. We follow the settings of MoCo v3 [6] except that a 5-layer MLP is used for the backbone. We use $\gamma = 1.1$ for CovType and $\gamma = 1.2$ for Higgs. For evaluation, we use linear evaluation for CovType and we compute the linear regression layer by pseudo-inverse between the feature matrix and label matrix.

Lastly, we examine RényiCL on graph TUDataset [63], which contains numerous graph machine learning benchmarks of different domains such as bioinformatics, molecules, and social network. We follow the experimental setup of [15]: we use graph isomorphism network [64] backbone and use node dropout, edge perturbation, attribute masking, and subgraph sampling for data augmentations. For the graph experiments, we do not change the strength of data augmentation nor add new data augmentation. The detailed hyper-parameters for training and evaluation for the whole dataset can be found in the appendix.

Table 6: Comparison of different contrastive learning methods on graph TUDataset. Average over 5 runs.

Method	NCI1	PROTEINS	DD	MUTAG	COLLAB	RDT-B	RDT-M5K	IMDB-B
InfoGraph [65]	76.20	74.44	72.85	89.01	70.65	82.50	53.46	**73.03**
GraphCL [15]	77.87	74.39	78.62	86.80	71.36	89.53	55.99	71.14
JOAO [66]	78.07	74.55	77.32	87.35	69.50	85.29	55.74	70.21
JOAOv2 [66]	78.36	74.07	77.40	87.67	69.33	86.42	56.03	70.83
RényiCL	**78.60**	**75.11**	**78.98**	**90.22**	**71.88**	**90.92**	**56.18**	72.38

Results. Table 5 compares the Top-1 linear evaluation accuracy (%) of representations with different contrastive learning objectives. We observe that RMLCPC obtains the best performance when using harder data augmentations on all images and tabular datasets. Also, RMLCPC has the largest gain on using harder augmentation, showing that the effectiveness of RényiCL is amplified when we use stronger augmentation. Table 6 compare the Top-1 linear evaluation accuracy (%) of graph RényiCL with other graph contrastive learning methods. Even without harder data augmentation, RényiCL achieves the best performance in 7 out of 8 benchmarks. Especially, RényiCL outperforms JOAO [66], where the views are learned to boost the performance of graph contrastive learning.

4.3 MI estimation and view selection

MI estimation on synthetic Gaussian. We first consider estimating mutual information between two correlated synthetic Gaussian distributions following the procedure in [30]. In the appendix, we show that by using the critique learned from each α-CPC, α-MLCPC, and (α, γ)-RMLCPC objective, we can estimate the true MI with low variance and low bias by using (8).

InfoMin principle and RényiCL. By using (8), we plot the MI versus linear evaluation accuracy (%) by applying different data augmentations on the ImageNet-100 dataset. In Figure 1, we observe that for harder data augmentations, the RMLCPC induces lower MI than that of CPC and MLCPC, while showing better downstream performances. This is because of the effect of easy positive and hard negative sampling explained in Section 3.3. Note that InfoMin [14] principle argues that the optimal views must share sufficient and minimal information about the downstream tasks. Here, one can observe that RényiCL intrinsically follows the InfoMin principle as it extracts minimal and sufficient information between the views.

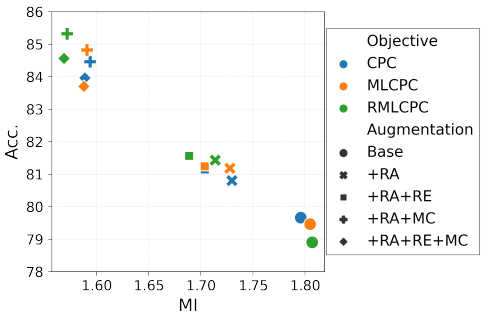

Figure 1: MI between the views and validation accuracy on ImageNet-100. RA: RandAugment [43], RE: RandomErasing [58], MC: multi-crops [38].

5 Conclusion and Discussion

We propose Rényi Contrastive Learning (RényiCL) which utilizes Rényi divergence to deal with stronger data augmentations. Since the variational lower bound of Rényi divergence is insufficient for contrastive learning due to large variance, we introduce a variational lower bound of skew Rényi divergence, namely Rényi-MLCPC. Indeed, we show that CPC and MLCPC are variational forms of skew KL divergence, and provide theoretical analysis on how they achieve low variance. Through experiments, we validate the effectiveness of RényiCL by using harder data augmentations.

Limitations and future works. Even though RényiCL is beneficial when using harder data augmentations, it cannot identify which data augmentation is actually beneficial. We think that the proposed scheme based on Rényi divergence could be useful for other tasks, e.g., information bottleneck [67], which is an interesting direction to explore in the future.

Negative societal impacts. Since contrastive learning often requires long epochs of training, it raises environmental concerns, e.g. carbon generation. Nevertheless, RényiCL is shown to be effective even under a smaller number of epochs, compared to existing schemes.

References

[1] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pages 776–794. Springer, 2020.

[2] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.

[3] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[4] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[5] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[6] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021.

[7] Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations. *arXiv preprint arXiv:1803.02893*, 2018.

[8] John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. Declutr: Deep contrastive learning for unsupervised textual representations. *arXiv preprint arXiv:2006.03659*, 2020.

[9] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

[10] Aaqib Saeed, David Grangier, and Neil Zeghidour. Contrastive learning of general-purpose audio representations. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3875–3879. IEEE, 2021.

[11] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6964–6974, 2021.

[12] Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. Active contrastive learning of audio-visual video representations. *arXiv preprint arXiv:2009.09805*, 2020.

[13] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12475–12486, 2021.

[14] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems*, 33:6827–6839, 2020.

[15] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. *Advances in Neural Information Processing Systems*, 33:5812–5823, 2020.

[16] Alfréd Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 4, pages 547–562. University of California Press, 1961.

[17] Jeremiah Birrell, Paul Dupuis, Markos A Katsoulakis, Luc Rey-Bellet, and Jie Wang. Variational representations and neural network estimation of rényi divergences. *SIAM Journal on Mathematics of Data Science*, 3(4):1093–1116, 2021.

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[19] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9588–9597, 2021.

[20] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020.

[21] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. *Advances in Neural Information Processing Systems*, 33:21798–21809, 2020.

[22] Tri Huynh, Simon Kornblith, Matthew R Walter, Michael Maire, and Maryam Khademi. Boosting contrastive self-supervised learning with false negative cancellation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2785–2795, 2022.

[23] 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. *arXiv preprint arXiv:1808.06670*, 2018.

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

[25] Yao-Hung Hubert Tsai, Shaojie Bai, Louis-Philippe Morency, and Ruslan Salakhutdinov. A note on connecting barlow twins with negative-sample-free contrastive learning. *arXiv preprint arXiv:2104.13712*, 2021.

[26] Yao-Hung Hubert Tsai, Martin Q Ma, Muqiao Yang, Han Zhao, Louis-Philippe Morency, and Ruslan Salakhutdinov. Self-supervised representation learning with relative predictive coding. *arXiv preprint arXiv:2103.11275*, 2021.

[27] Ralph Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988.

[28] Anthony J Bell and Terrence J Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[30] Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning*, pages 5171–5180. PMLR, 2019.

[31] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International conference on machine learning*, pages 531–540. PMLR, 2018.

[32] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[33] Jiaming Song and Stefano Ermon. Multi-label contrastive predictive coding. *arXiv preprint arXiv:2007.09852*, 2020.

[34] Tim Van Erven and Peter Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

[35] Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time, i. *Communications on Pure and Applied Mathematics*, 28(1):1–47, 1975.

[36] Jiaming Song and Stefano Ermon. Understanding the limitations of variational mutual information estimators. *arXiv preprint arXiv:1910.06222*, 2019.

[37] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[38] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020.

[39] Yang Song and Diederik P Kingma. How to train your energy-based models. *arXiv preprint arXiv:2101.03288*, 2021.

[40] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.

[41] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

[42] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

[43] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020.

[44] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[45] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer, 2014.

[46] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008.

[47] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013.

[48] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[49] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014.

[50] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010.

[51] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[52] Xiao Wang and Guo-Jun Qi. Contrastive learning with stronger augmentations. *arXiv preprint arXiv:2104.07713*, 2021.

[53] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012.

[54] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004.

[55] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. *Advances in neural information processing systems*, 31, 2018.

[56] Sharada P Mohanty, David P Hughes, and Marcel Salathé. Using deep learning for image-based plant disease detection. *Frontiers in plant science*, 7:1419, 2016.

[57] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020.

[58] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008, 2020.

[59] Pierre Baldi, Peter Sadowski, and Daniel Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature communications*, 5(1):1–9, 2014.

[60] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[61] Kibok Lee, Yian Zhu, Kihyuk Sohn, Chun-Liang Li, Jinwoo Shin, and Honglak Lee. i-mix: A domain-agnostic strategy for contrastive representation learning. In *ICLR*, 2021.

[62] Dara Bahri, Heinrich Jiang, Yi Tay, and Donald Metzler. Scarf: Self-supervised contrastive learning using random feature corruption. *arXiv preprint arXiv:2106.15147*, 2021.

[63] Christopher Morris, Nils M Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. Tudataset: A collection of benchmark datasets for learning with graphs. *arXiv preprint arXiv:2007.08663*, 2020.

[64] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

[65] Fan-Yun Sun, Jordan Hoffmann, Vikas Verma, and Jian Tang. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. *arXiv preprint arXiv:1908.01000*, 2019.

[66] Yuning You, Tianlong Chen, Yang Shen, and Zhangyang Wang. Graph contrastive learning automated. In *International Conference on Machine Learning*, pages 12121–12132. PMLR, 2021.

[67] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pages 1–5. IEEE, 2015.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all 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]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (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]
   (b) Did you include complete proofs of all theoretical results? [Yes] The proofs of theoretical results are in the appendix.
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] We provide code and instructions in the supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] They are in the appendix.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] They are in the appendix.
   (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)? [Yes] They are in the appendix.
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? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (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? [N/A]
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]