# Addressing Feature Suppression in Unsupervised Visual Representations

Tianhong Li[1,*] Lijie Fan[1,*] Yuan Yuan[1] Hao He[1] Yonglong Tian[1]
Rogerio Feris[2] Piotr Indyk[1] Dina Katabi[1]

[1]MIT CSAIL, [2]MIT-IBM Watson AI Lab

## Abstract

*Contrastive learning is one of the fastest growing research areas in machine learning due to its ability to learn useful representations without labeled data. However, contrastive learning is susceptible to feature suppression – i.e., it may discard important information relevant to the task of interest, and learn irrelevant features. Past work has addressed this limitation via handcrafted data augmentations that eliminate irrelevant information. This approach however does not work across all datasets and tasks. Further, data augmentations fail in addressing feature suppression in multi-attribute classification when one attribute can suppress features relevant to other attributes. In this paper, we analyze the objective function of contrastive learning and formally prove that it is vulnerable to feature suppression. We then present Predictive Contrastive Learning (PrCL), a framework for learning unsupervised representations that are robust to feature suppression. The key idea is to force the learned representation to predict the input, and hence prevent it from discarding important information. Extensive experiments verify that PrCL is robust to feature suppression and outperforms state-of-the-art contrastive learning methods on a variety of datasets and tasks.*

## 1. Introduction

The area of unsupervised or self-supervised representation learning is growing rapidly [12, 50, 27, 2, 53, 18, 35, 22, 33, 16, 14, 13, 15, 48, 49, 28, 34, 51, 52]. It refers to learning data representations that capture potential labels of interest, and doing so without human supervision. Contrastive learning is increasingly considered as a standard and highly competitive method for unsupervised representation learning. Features learned with this method have been shown to generalize well to downstream tasks, and in some cases surpass the performance of supervised models [37, 3, 43, 5, 6, 17, 8, 31].

*Indicates equal contribution. This work was supported by the MIT-IBM Watson Research Collaboration grant.

(a) Digit & Bkgd

(b) Face Attribute

Figure 1. (a) In Colorful-Moving-MNIST [42], the input has two types of information: digit and background object. But contrastive learning methods focus on the background object and ignore the digit. (b) Each image in FairFace [30] has multiple attributes such as age, gender, ethnicity, etc. Existing contrastive learning methods focus on ethnicity and partially ignore other attributes.

Contrastive learning learns representations by contrasting positive samples against negative samples. During training, a data sample is chosen as an anchor (e.g., an image); positive samples are chosen as different augmented versions of the anchor (e.g., randomly cropping and color distorting the image), whereas negative samples come from other samples in the dataset.

Yet contrastive learning is vulnerable to feature suppression [19, 40, 32] – i.e., if simple features are contrastive enough to separate positive samples from negative samples, contrastive learning might learn such simple (or simpler) features even if irrelevant to the tasks of interest, and other more relevant features are suppressed. For example, the authors of [5] show that color distribution can be used to distinguish patches cropped from the same image, from patches from different images; yet such feature is not useful for object classification. Past work addresses this problem by designing handcrafted data augmentations that eliminate the irrelevant features, so that the network may learn the relevant information [24, 5, 6, 8, 7].

However, in many scenarios it is hard to design augmentations to solve the problem of feature suppression. For example, the authors of [42] highlight the scenario in Figure 1 (a), where each image shows a digit (from MNIST) on a randomly chosen background object (from STL-10). They show that features related to background objects can create a shortcut that prevent contrastive learning from learning features related to digits. In this case, one cannot simply eliminate the background information since such a design, though would help digit classification, would harm the background classification task. A similar problem exists in the task of human face attribute classification, where each face image can be used in multiple downstream tasks including gender, age, and ethnicity classification (Figure 1 (b)), but the features learned by contrastive learning can be biased to only one of the attributes (e.g., ethnicity) and show poor performance on other attributes (gender and age) as shown in the experiments section. It is hard to come up with data augmentations that eliminate the dominant attribute without harming the corresponding classification task. Moreover, as machine learning keeps expanding to new modalities it becomes increasingly difficult to design handcrafted data augmentations because many new modalities are hard to directly interpret by humans (e.g., acceleration from wearable devices), or the interpretation requires domain experts (e.g., medical data).

In this paper, we first provide a theoretical analysis of contrastive learning and prove it is vulnerable to feature suppression. Our analysis shows that even with large feature dimensions, contrastive learning has many local minima that discard significant information about the input, and hence cause feature suppression. Furthermore, the value of the loss function at such local minima is very close to its value at the global minimum, making it hard to propel the model out of such local minima.

Second, we propose predictive contrastive learning (PrCL) as a training scheme that prevents feature suppression. PrCL learns representations using contrastive and predictive learning simultaneously. We use the term predictive learning to refer to tasks that force the representation to predict the input, such as inpainting, colorization, or autoencoding. Such tasks counter the effect of feature suppression because they force the learned features to retain the information in the input. More formally, if the contrastive loss (i.e., the InfoNCE loss) gets stuck in a local minimum that loses semantic information, the predictive loss naturally becomes very high, forcing the model to exit such local minima. An interesting feature of PrCL is that the predictive task is used only during training, and hence introduces no computation overhead during testing.

We evaluate PrCL and compare it with state-of-the-art contrastive learning baselines on four different datasets: ImageNet, MPII [1], Colorful-Moving-MNIST [42], and FairFace [30]. For all tasks, PrCL achieves superior performance and outperforms the state-of-the-art baselines by large margins, demonstrating robustness against feature suppression.

The paper makes the following contributions:

- It provides a theoretical analysis of contrastive learning that proves its vulnerability to feature suppression.
- It introduces PrCL, an unsupervised learning framework that automatically avoids feature suppression and provides a representation that learns all of the semantics in the input and can support different downstream tasks and multi-attribute classification.
- It empirically shows that SOTA contrastive learning baselines (e.g., SimCLR, MoCo, and BYOL) suffer from feature suppression, and that PrCL outperforms those baselines on several important tasks including object recognition, pose estimation, and face attribute classification.

## 2. Related Work

Early work on unsupervised representation learning has focused on designing pretext tasks and training the network to predict their pseudo labels. Such tasks include solving jigsaw puzzles [36], restoring a missing patch in the input [38], or predicting image rotation [20]. However, pretext tasks have to be handcrafted, and the generality of their representations is typically limited [5].

Hence, researchers have recently focused on contrastive learning, which emerged as a competitive and systematic method for learning effective representations without human supervision. The learned features generalize well to downstream tasks, outperform representations learned through pretext tasks, and even surpass the performance of supervised models on some tasks [5, 6, 8, 24]. Multiple successful contrastive learning frameworks have been proposed, which typically differ in the way they sample negative pairs. To name a few, SimCLR [5] uses a large batch size, and samples negative pairs within each batch. The momentum-contrastive approach (MoCo) [24] leverages a moving-average encoder and a queue to generate negative samples on the fly during training. Contrastive-Multiview-Coding [41] maintains a memory-bank to store features and generate negative samples. Some recent methods, like BYOL [21], do not rely on negative pairs [9, 21]. Instead, they use two neural networks that learn from each other to boost performance.

Past work has also reported problems with contrastive learning. It can focus on irrelevant features such as color distribution, and suppress more relevant features [5]. Past work addressed this problem by using color-distortion as a data augmentation. Also, the authors of [42] noted that when the data includes multiple types of semantics, contrastive learning may learn one type of semantics and fail to learn effective features of the other semantics (as in Figure 1(b) where the background object information can suppress features related to digits). They proposed a solution that learns contrastive views suitable for the desired downstream task. While they share our goal of supporting different downstream tasks, their method requires supervision since they learn their contrastive views from labeled data. In contrast, our approach is completely unsupervised.

Another related work is contrastive-predictive-coding (CPC) [37, 25]. CPC has some similarities with PrCL in that it has a predictive task that aims to reconstruct missing information. However, CPC aims to reconstruct the features of a future frame, while PrCL reconstructs the raw input data. As a result, the representation learned by CPC is not forced to contain necessary information to reconstruct the input, making it susceptible to feature suppression, just like other contrastive learning methods.

The family of auto-encoders provides a popular framework for unsupervised representation learning using a reconstructive loss [26, 39, 45]. It trains an encoder to generate low-dimensional latent codes that could reconstruct the entire high-dimensional inputs. There are many types of AEs, such as denoising auto-encoders [45], which corrupt the input and let the latent codes reconstruct it, and variational auto-encoders [39], which force the latent codes to follow a prior distribution. Recently, masked auto-encoders with transformer-based network architectures have demonstrated great performance on unsupervised representation learning [23, 4]. However, these works use architectures such as BERT and ViT, which requires much more computation resources than convolutional neural networks. PrCL can be viewed as a special variant of the denoising auto-encoder that forces the latent codes to have a 'contrastive' property regularized by a contrastive loss. As a result, the latent codes, are good not only for reconstructing the input, but also for downstream classification tasks.

Finally, several concurrent papers published on Arxiv also used a combination contrastive and reconstructive loss [11, 29]. However, none of them explore the potential of this combination to solve the feature suppression problem, or provides a theoretical analysis of feature suppression. This paper is the first to demonstrate that the combination of contrastive and predictive loss can be used to avoid feature suppression and learn general representations that support multiple downstream tasks.

## 3. Analysis of Feature Suppression

Before delving into formal proofs, we provide an informal description of our analysis as follows:

1. At low feature dimensions, contrastive learning loss (InfoNCE) global minimum loses semantic information. This is because with small feature dimensions, it is impossible to keep all information about the input.
2. InfoNCE global minima at low dimensions (which loses information from (1.) above), are local minima at higher dimensions **[Corollary 2]**. Thus, even for high dimension features, it will have many local minima that lose information about the input (i.e., feature suppression).
3. The value of infoNCE at such local minima (from (2.) above) can be very close to its global minimum **[Lemma 1 and Figure 2]**, making it hard to escape from such local minima.
4. The above three points mean that, even at high dimensions, contrastive learning is likely to get stuck in a local minimum that exhibits feature suppression. Adding a predictive loss allows the model to exit such local minimum and avoid feature suppression. This is because suppressed features lose information about the input causing the predictive loss to become large, and push the model out from such local minimum and away from feature suppression.

### 3.1. Formal Proof.

Let $X = \{x_i\}_{i=1}^n$ be the set of the data points. We use $\lambda_{ij}$ to indicate whether a data pair $x_i$ and $x_j$ is positive or negative. Specifically, $\lambda_{ij} = 1$ indicates a positive pair while $\lambda_{ij} = 0$ indicates a negative pair. Let $Z = \{z_i\}_{i=1}^n$, where $z_i = f(x_i) = (z_i^1, \cdots, z_i^d) \in \mathcal{S}^{d-1}$, denote the learned features on the hypersphere, generated by the neural network $f$. $t \in \mathbb{R}^+$ is a scalar temperature parameter. We consider the following empirical asymptotics of the infoNCE objective function introduced in [46].

**Definition 1** (Empirical infoNCE asymptotics)**.**

$$\mathcal{E}_{\texttt{limNCE}}(Z; X, t, d) \triangleq$$
$$-\frac{1}{tn^2}\sum_{ij}\lambda_{ij} z_i^\top z_j + \frac{1}{n}\sum_i \log\left(\frac{1}{n}\sum_j e^{z_i^\top z_j/t}\right)$$

We are going to connect the landscape of empirical infoNCE asymptotics in the low dimension to that in the high dimension. We start by defining a *lifting operator* that maps a low dimensional vector to a higher dimension.

**Definition 2** (Lifting operator)**.** *A lifting operator $\mathcal{T}_\sigma$ parameterized by an indexing function $\sigma$ maps a $d_1$-dimensional vector to dimension $d_2$ ($d_2 > d_1$). Its parameter $\sigma$ is a permutation of length $d_2$. Given a $d_1$-dimensional vector $z$, the lifting operator maps it to a $d_2$-dimensional vector $\tilde{z} = \mathcal{T}_\sigma(z)$ by the following rules: $\tilde{z}^t = z^{\sigma(t)}$ if $\sigma(t) \leq d_1$, otherwise $\tilde{z}^t = 0$.*

With a slight abuse of notations, we allow the lifting operator to map a *set* of low dimensional vectors to higher dimension, i.e. $\mathcal{T}_\sigma(\{z_i\}) = \{\mathcal{T}_\sigma(z_i)\}$. We further allow the lifting operator to map a function $f$ of lower dimension to higher dimension, i.e., $\mathcal{T}_\sigma(f)(x) = \mathcal{T}_\sigma(f(x))$. Note that $\mathcal{T}_\sigma$ is a linear operator. We highlight several useful properties of $\mathcal{T}_\sigma$:

**Lemma 1** (Value Invariance)**.** *The value of the empirical infoNCE asymptotics is invariant under the lifting operation. Formally, consider any lifting operator $\mathcal{T}_\sigma$ from the* *dimension $d_1$ to the dimension $d_2$. We have*

$$\mathcal{E}_{\texttt{limNCE}}(\mathcal{T}_\sigma(Z); X, t, d_2) = \mathcal{E}_{\texttt{limNCE}}(Z; X, t, d_1)$$

*Proof.* Following the definition of $\mathcal{T}_\sigma$, $\forall z_i, z_j$, $z_i^\top z_j = \mathcal{T}_\sigma(z_i)^\top \mathcal{T}_\sigma(z_j)$. Therefore, $\mathcal{E}_{\texttt{limNCE}}(\mathcal{T}_\sigma(Z); X, t, d_2) = \mathcal{E}_{\texttt{limNCE}}(Z; X, t, d_1)$. □

**Lemma 2** (Gradient Equivariance)**.** *The gradient of the empirical infoNCE asymptotics is equivariant under the lifting operation. Formally, consider any lifting operator $\mathcal{T}_\sigma$ from the dimension $d_1$ to the dimension $d_2$. We have*

$$\nabla_{\tilde{z_k}} \mathcal{E}_{\texttt{limNCE}}(\mathcal{T}_\sigma(Z); X, t, d_2) = \mathcal{T}_\sigma\left(\nabla_{z_k} \mathcal{E}_{\texttt{limNCE}}(Z; X, t, d_1)\right)$$

*Proof.* The proof is in the supplemental material. □

**Corollary 1.** *For any lifting operator $\mathcal{T}_\sigma$, if $\hat{Z} = \{\hat{z}_i\}$ is a stationary point of $\mathcal{E}_{\texttt{limNCE}}(Z; X, t, d_1)$, then $\mathcal{T}_\sigma(\hat{Z})$ is a stationary point of $\mathcal{E}_{\texttt{limNCE}}(Z; X, t, d_2)$.*

*Proof.* $\hat{Z}$ is a stationary point of $\mathcal{E}_{\texttt{limNCE}}(Z; X, \tau, d_1)$ implies $\nabla_{z_i} \mathcal{E}_{\texttt{limNCE}}(\hat{Z}; X, \tau, d_1) = 0$. Therefore, by Lemma 2, $\nabla_{\tilde{z_i}} \mathcal{E}_{\texttt{limNCE}}(\mathcal{T}_\sigma(\hat{Z}); X, \tau, d_2) = \mathcal{T}_\sigma\left(\nabla_{z_i} \mathcal{E}_{\texttt{limNCE}}(\hat{Z}; X, \tau, d_1)\right) = 0$. □

**Corollary 2.** *For any lifting operator $\mathcal{T}_\sigma$, if $\hat{Z} = \{\hat{z}_i\}$ is a global minimum of $\mathcal{E}_{\texttt{limNCE}}(Z; X, t, d_1)$ with a positive definite Hessian matrix, then $\mathcal{T}_\sigma(\hat{Z})$ is a saddle point or a local minimum of $\mathcal{E}_{\texttt{limNCE}}(Z; X, t, d_2)$.*

*Proof.* From Corollary 1, $\mathcal{T}_\sigma(\hat{Z})$ is a stationary point of $\mathcal{E}_{\texttt{limNCE}}(Z; X, \tau, d_2)$. Since the Hessian matrix of $\mathcal{E}_{\texttt{limNCE}}(Z; X, \tau, d_1)$ at $\hat{Z}$ is positive definite, $\forall r > 0, \exists Z' \in B_r(\hat{Z})$ s.t. $\mathcal{E}_{\texttt{limNCE}}(Z'; X, \tau, d_1) > \mathcal{E}_{\texttt{limNCE}}(Z; X, \tau, d_2)$, where $B_r(Z) = \{Z' \in \mathcal{S}^{d-1} | \, ||Z - Z'||_2 < r\}$ is the neighborhood of $Z$ with radius $r$. Therefore, $\mathcal{E}_{\texttt{limNCE}}(\mathcal{T}_\sigma(Z'); X, \tau, d_1) > \mathcal{E}_{\texttt{limNCE}}(\mathcal{T}_\sigma(Z); X, \tau, d_2)$ (Lemma 1). Note that $Z' \in B_r(\hat{Z}) \rightarrow \mathcal{T}_\sigma(Z') \in B_r(\mathcal{T}_\sigma(\hat{Z}))$. Therefore, $\forall r > 0, \exists \mathcal{T}_\sigma(Z') \in B_r(\mathcal{T}_\sigma(\hat{Z}))$ s.t. $\mathcal{E}_{\texttt{limNCE}}(\mathcal{T}_\sigma(Z'); X, \tau, d_1) > \mathcal{E}_{\texttt{limNCE}}(\mathcal{T}_\sigma(Z); X, \tau, d_2)$. Therefore, $\mathcal{T}_\sigma(\hat{Z})$ is not a local maximum, so it can only be a local minimum or a saddle point of $\mathcal{E}_{\texttt{limNCE}}(Z; X, \tau, d_2)$. □

With Corollary 2, we can explain why contrastive learning can suffer from feature suppression. Suppose $f$ is a network that achieves the global minimum of $\mathcal{E}_{\texttt{limNCE}}(Z; X, t, d_1)$. When $d_1$ is relatively small (e.g., <100 for images), $f$ must lose some information about the input, i.e., suppress feature. From Corollary 2, $\mathcal{T}_\sigma(f)$ is a saddle point or a local minimum of $\mathcal{E}_{\texttt{limNCE}}(Z; X, t, d_2)$ where $d_2 > d_1$ and $\mathcal{T}_\sigma(f)$ carries no more information than $f$. Therefore, for any dimension $d > 1$, there exists saddle point/local minimum of $\mathcal{E}_{\texttt{limNCE}}(Z; X, t, d)$ which suppresses features.

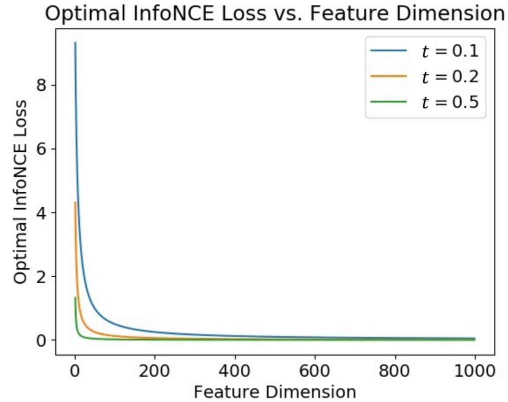



Figure 2. Optimal infoNCE loss vs. different output feature dimension $d$ and temperature $t$.

Furthermore, the value of the aforementioned saddle point/local minimum of $\mathcal{E}_{\texttt{limNCE}}(Z; X, t, d)$ is quite close to that of the global minimum. This is because the optimal value of $\mathcal{E}_{\texttt{limNCE}}(Z; X, t, d)$ converges quickly as $d$ increases. Figure 2 shows the curve of $\log_0 F_1(; d; \frac{1}{4t^2})$, which is the optimal value of the infoNCE loss [47]. As shown in the figure, the curve essentially converges when $d > 200$. Therefore, $\mathcal{T}_\sigma(f)$ can be a saddle point/local minimum of $\mathcal{E}_{\texttt{limNCE}}(Z; X, t, d_2)$, and its value can also be quite close to that of the global minimum, making it hard to escape from such local minimum. So effectively one can achieve a value pretty close to the global minimum by suppressing features, and stay at that saddle point being unable to escape. This motivates our solution, which adds a predictive loss to force the model out from such local minima that suppress features.

# 4. Predictive Contrastive Learning (PrCL)

Predictive contrastive learning (PrCL) is a framework for self-supervised representation learning. It aims to learn representations that are robust to feature suppression, and capable of supporting multiple diverse downstream tasks.

The idea underlying PrCL is as follows: feature suppression is harmful because the representation loses important information that was available in the input. Thus, to counter feature suppression, PrCL uses a prediction loss to ensure that the representation can restore the input, i.e., the features have the information available at the input. Yet, keeping all information in the features is not enough; the input already has all information. By adding a contrastive loss, PrCL reorganizes the information in the feature space to make it amenable to downstream classification, i.e., samples that have similar attributes/objects are closer to each other than samples that have different attributes/objects. Figure 3 shows the PrCL framework which has two branches: a contrastive branch and a predictive branch.

**(a) Contrastive Branch:** The contrastive branch is illustrated in the orange box in Figure 3. Here, we use SimCLR

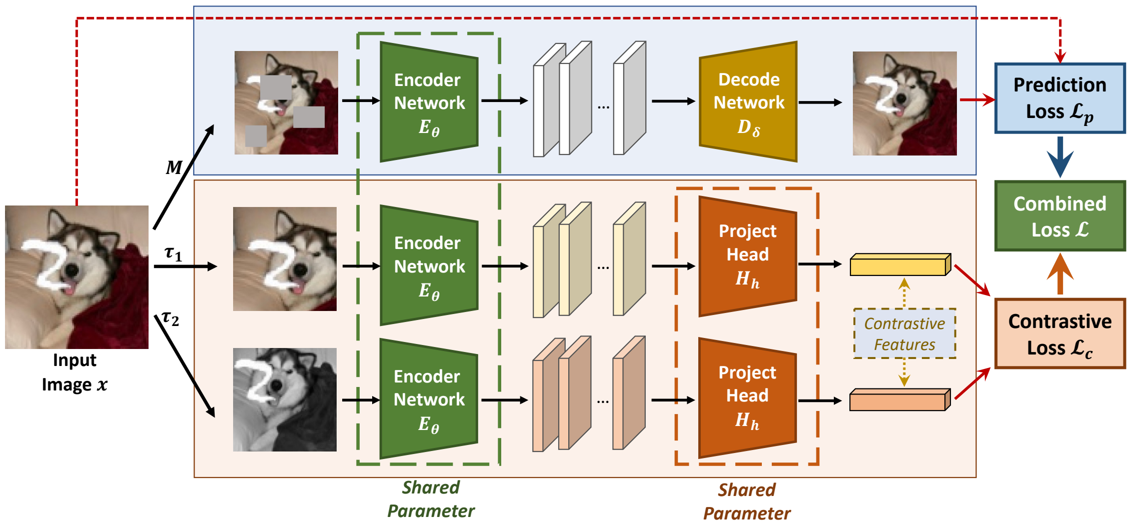



Figure 3. Illustration of the PrCL framework. PrCL has two branches: 1) a predictive branch, illustrated in the blue box, which ensures that the representation has enough information to restore missing patches in the input, and 2) a contrastive branch, illustrated in the orange box, which ensures that the representation keeps positive samples close to each other and away from negative samples.

as an example to demonstrate the basic idea. However, this contrastive branch can be easily adapted to any contrastive learning method such as CPC, MoCo, and BYOL. For each image, we first generate a pair of positive samples by using two random augmentations $\tau_1$ and $\tau_2$, then we forward the two augmented inputs separately to the encoder $E$, parameterized by $\theta$ and a multi-layer nonlinear projection head $H$ parameterized by $h$ to get the latent representations $z_1$ and $z_2$ for these two positive samples. We use the commonly used InfoNCE loss [5] as the contrastive loss $\mathcal{L}_c$. Namely, for a batch of $N$ different input images $x_i, i = 1, ..., N$,

$$\mathcal{L}_c = -\sum_{i=1}^{N} \log \sum \frac{\exp\left(\text{sim}(z_{2i}, z_{2i+1})/t\right)}{\sum_{k=1}^{2N} \mathbb{1}_{k\neq 2i} \exp\left(\text{sim}(z_{2i}, z_k)/t\right)},$$

where $\text{sim}(u, v) = u^T v/(\|u\|_2 \|v\|_2)$ denotes the dot product between the normalized $u$ and $v$ (i.e., cosine similarity), and $z_{2i}, z_{2i+1}$ are the encoded features of positive pairs generated from $x_i$, i.e., $z_{2i} = H_h(E_\theta(\tau_1(x_i)))$ and $z_{2i+1} = H_h(E_\theta(\tau_2(x_i)))$.

**(b) Predictive Branch:** To choose a proper predictive task, we need to consider two aspects: its ability to summarize and abstract the input, and its applicability to different datasets and tasks. In fact, many self-supervised learning tasks, such as Auto-encoder, Colorization and Inpainting, are predictive since they all aim to restore the input. But, those tasks do not have the same ability to both retain and abstract information. For example, inpainting is a stronger predictive task than autoencoding in terms of its ability to both abstract and retain information. Thus, although both of them would help in strengthening contrastive learning against feature suppression, inpainting is likely to provide more gains.

Another issue to consider is the applicability of the chosen task to various datasets. For example, colorization is applicable only to colorful RGB datasets, but not to greyscale datasets such as MNIST or medical image datasets. In contrast, a task like inpainting is easier to translate across different datasets.

Given the above considerations, we adopt inpainting as the default predictive task. In the supplemental material, we compare various tasks and show that while they all improve performance, inpainting delivers higher gains.

Figure 3 shows how PrCL uses the inpainting task, where given an input image $x$, we first randomly mask several patches to get the masked input $M(x)$. Then the masked input is passed through an encoder network $E$ with parameter $\theta$, and a decoder network $D$, with parameter $\delta$, to obtain the reconstruction result $D_\delta(E_\theta(M(x)))$. The prediction loss $\mathcal{L}_p$ is defined as the reconstruction error between the original input $x$ and the reconstructed one $D_\delta(E_\theta(M(x)))$:

$$\mathcal{L}_p = ||D_\delta(E_\theta(M(x))) - x||_2.$$

**(c) Training Procedure:** We have empirically found that it is better to train the model in two phases. In the first phase, only the predictive branch is trained. In the second phase, both branches are trained together. In this latter case, the overall training loss is the combination of the prediction loss and the contrastive loss, i.e., $\mathcal{L} = \mathcal{L}_c + \lambda \cdot \mathcal{L}_p$. We set $\lambda = 10$ for all experiments. We also include results with different $\lambda$ in the supplemental material.

**(d) PrCL Avoids Feature Suppression:** With a combination of the prediction loss and the contrastive loss, PrCL is capable of escaping the aforementioned local minimum/saddle points of infoNCE loss where only partial semantics are learned. This is because learning only part of the semantics can result in very high prediction loss. For example, if the network learns only semantics related to the background object but ignores the digit (Figure 3), all pixels related to the digit are likely to be predicted incorrectly, introducing large gradients that force the model out of the saddle point.

# 5. Experiments

**Baselines.** We use state-of-the-art contrastive learning methods as baselines, including SimCLR [5], MoCo [8], CPC [25] and BYOL [21]. The same network structure, batch size, and training epochs are used for all baselines and PrCL's contrastive branch. For the contrastive branch of PrCL, we apply the same training scheme as MoCo. PrCL uses the predictive branch only for training. During inference it uses only the encoder, which is shared with the contrastive branch. Thus, the evaluation of PrCL uses exactly the same number of parameters as the baselines.
**Datasets.** We experiment with the following datasets:

- **ImageNet:** ImageNet[10] (CC BY 2.0) is a widely used image classification benchmark which contains 1.28M images in 1000 different categories. It is a standard benchmark to evaluate self-supervised learning methods [8, 5, 21].
- **MPII:** MPII [1] (the Simplified BSD License) is one of the most common datasets for the task of human pose estimation. It contains images of everyday human activities.
- **FairFace:** FairFace [30] (CC BY 4.0.) is a face attribute classification dataset, where each image contains multiple semantics including gender, age, and ethnicity.
- **Colorful-Moving-MNIST:** This is a synthetic dataset used by [42] to highlight the feature suppression problem. It is constructed by assigning each digit from MNIST a background object image selected randomly from STL-10. It supports two downstream tasks: digit and background classification.

**Setups.** On ImageNet, as common in the literature, we evaluate the representations with the encoder fixed and only the linear classifier is trained. On all other datasets, we evaluate the representations under two different settings: fixed feature encoder setting and fine-tuning setting. In the fixed feature encoder setting, the ResNet encoder is fixed and only the classifier (FairFace, Colorful-Moving-MNIST) or the 4-layer decoder network (MPII) is trained; In the fine-tuning setting, the encoder is initialized with the pre-trained model and fine-tuned during training. Please refer to the Appendix for architectural details and hyper-parameters.

## 5.1. Results

We report the main results for all datasets. The experiment setup, training details and hyper-parameter settings are provided in the supplemental material along with additional results.

**ImageNet.** Table 1 compares PrCL with the contrastive learning baselines on the task of object classification under different data augmentations. Here, we compare PrCL with SimCLR and MoCo since they use the same set of data augmentations. The results show that with fewer data augmentations, the accuracy of the contrastive learning baselines drops quickly due to feature suppression. For example, removing the color distortion augmentation significantly degrades the performance of the baseline approaches, as color distribution is known to be able to suppress other features in contrastive learning. In contrast, PrCL is significantly more robust. For example, with only random cropping, PrCL's Top-1 accuracy drops by only 6.9 whereas the Top-1 accuracy of SimCLR drops by 27.6 and the Top-1 accuracy of MoCo drops by 12.1. We also compare PrCL with a predictive baseline [38]. For the predictive baseline, though the model is not sensitive to different augmentations, the best performance is not comparable to contrastive learning, indicating predictive learning alone is not enough to learn fine-grained representations from images.

**MPII.** We use PrCL and the contrastive learning baselines to learn representations from MPII, and evaluate them on the task of pose estimation. Table 2 shows that PrCL improves the average PCKh (the standard metric for pose estimation) over the strongest contrastive baseline by 3.7 and achieves even higher gains on important keypoints such as Head and Wrist. This is because contrastive learning is likely to focus on features irrelevant to the downstream task, such as clothes and appearances.

**FairFace.** Table 3 compares the contrastive learning baselines to PrCL on the task of face-attribute classification. The results show how contrastive learning struggles with multi-attribute classification. Specifically, the performance of the contrastive learning baselines on ethnicity classification is close to supervised learning of that attribute (62% vs. 69%). However, their results on age and gender classifications are significantly worse than supervised learning of those attributes (44% and 78% vs. 54% and 91%). This indicates that ethnicity suppresses other features in contrastive learning. This feature is partial since there are dependencies in how ethnicity manifests itself across age and gender. In contrast, PrCL is much more robust to such feature suppression problem, and its performance results on age and gender classifications are much closer to those of fully-supervised classification of those attributes.

**Colorful-Moving-MNIST.** We use this dataset to further investigate how contrastive learning performs on multi-attribute classification. Recall that each image in this dataset contains a digit from MNIST on a randomly selected background object from the STL-10. We investigate whether the learned representation supports both digit and background

Table 1. Performance on ImageNet with progressive removal of data augmentations for different self-supervised learning techniques. The baseline corresponds to the original set of augmentations used in SimCLR and MoCo: random flip, random resized crop, color distortion, and random Gaussian blur.

(a) ImageNet TOP-1 accuracy and its DROP w.r.t. inclusion of all augmentations.

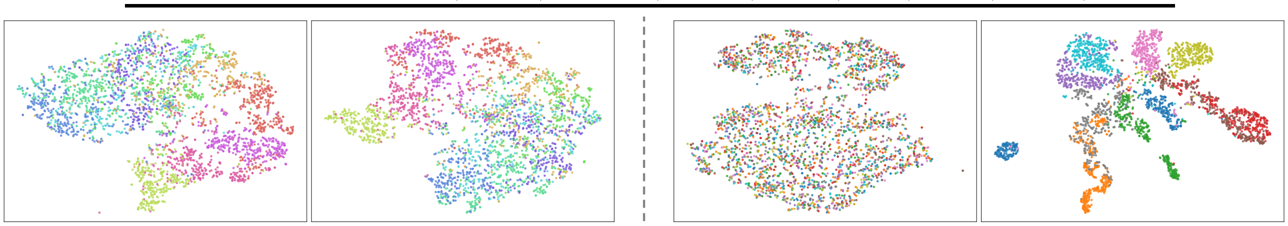
| Method | Inpainting | | SimCLR | | MoCo | | **PrCL(ours)** | | **IMPROVE** |
|---|---|---|---|---|---|---|---|---|---|
| **METRIC** | **TOP-1** | **DROP** | **TOP-1** | **DROP** | **TOP-1** | **DROP** | **TOP-1** | **DROP** | |
| Baseline | 43.7 | / | 67.9 | / | **71.1** | / | 71.0 | / | **-0.1** |
| Remove flip | 43.4 | -0.3 | 67.3 | -0.6 | 70.6 | -0.5 | **70.8** | -0.2 | **+0.2** |
| Remove blur | 43.6 | -0.1 | 65.2 | -2.7 | 69.7 | -1.4 | **70.6** | -0.4 | **+0.9** |
| Crop color only | 43.2 | -0.5 | 64.2 | -3.7 | 69.5 | -1.6 | **70.1** | -0.9 | **+0.6** |
| Remove color distort | 43.5 | -0.2 | 45.7 | -22.2 | 60.4 | -10.7 | **65.9** | -5.1 | **+5.5** |
| Crop blur only | 42.8 | -0.9 | 41.7 | -26.2 | 59.8 | -11.3 | **65.1** | -5.9 | **+5.3** |
| Crop flip only | 43.3 | -0.4 | 40.2 | -27.7 | 59.4 | -11.7 | **64.6** | -6.4 | **+5.2** |
| Crop only | 42.7 | -1.0 | 40.3 | -27.6 | 59.0 | -12.1 | **64.1** | -6.9 | **+5.1** |

(b) ImageNet TOP-5 accuracy and its DROP w.r.t. inclusion of all augmentations.

| Method | Inpainting | | SimCLR | | MoCo | | **PrCL(ours)** | | **IMPROVE** |
|---|---|---|---|---|---|---|---|---|---|
| **METRIC** | **TOP-5** | **DROP** | **TOP-5** | **DROP** | **TOP-5** | **DROP** | **TOP-5** | **DROP** | |
| Baseline | 68.3 | / | 88.5 | / | **90.1** | / | 90.0 | / | **-0.1** |
| Remove flip | 67.9 | -0.4 | 88.2 | -0.3 | 89.9 | -0.2 | **89.9** | -0.1 | **+0.0** |
| Remove blur | 68.1 | -0.2 | 86.6 | -1.9 | 89.7 | -0.4 | **89.8** | -0.2 | **+0.1** |
| Crop color only | 67.8 | -0.5 | 86.2 | -2.3 | 89.6 | -0.5 | **89.7** | -0.3 | **+0.1** |
| Remove color distort | 68.0 | -0.3 | 70.6 | -17.9 | 84.2 | -5.9 | **88.3** | -1.7 | **+4.1** |
| Crop blur only | 67.4 | -0.9 | 66.4 | -22.1 | 83.1 | -7.0 | **88.0** | -2.0 | **+4.9** |
| Crop flip only | 67.7 | -0.6 | 64.8 | -23.7 | 82.0 | -8.1 | **87.7** | -2.3 | **+5.7** |
| Crop only | 67.4 | -0.9 | 64.8 | -23.7 | 81.6 | -8.5 | **87.6** | -2.4 | **+6.0** |

Table 2. Performance on MPII for the downstream task of human pose estimation. ↑ indicates the larger the value, the better the performance.

| METRIC | | Head↑ | Shoulder↑ | Elbow↑ | Wrist↑ | Hip↑ | Knee↑ | Ankle↑ | PCKh↑ |
|---|---|---|---|---|---|---|---|---|---|
| FIXED FEATURE EXTRACTOR | SimCLR | 78.4 | 74.6 | 56.7 | 45.2 | 61.8 | 51.3 | 47.1 | 60.8 |
| | MoCo | 79.2 | 75.1 | 57.4 | 45.9 | 62.4 | 52.0 | 47.6 | 61.4 |
| | CPC | 78.0 | 74.3 | 56.0 | 44.8 | 61.2 | 51.4 | 46.5 | 60.3 |
| | BYOL | 79.1 | 75.0 | 57.1 | 46.0 | 62.4 | 52.2 | 47.7 | 61.4 |
| | **PrCL (ours)** | **85.7** | **78.8** | **61.7** | **51.3** | **64.4** | **55.6** | **49.2** | **65.1** |
| | **IMPROVEMENTS** | **+6.5** | **+3.7** | **+4.3** | **+5.3** | **+2.0** | **+3.4** | **+1.5** | **+3.7** |
| FINE-TUNING | SimCLR | 96.2 | 94.7 | 87.3 | 81.2 | 87.5 | 81.0 | 77.2 | 87.1 |
| | MoCo | 95.9 | 94.7 | 87.5 | 81.6 | 87.4 | 81.7 | 76.9 | 87.2 |
| | CPC | 96.0 | 94.5 | 87.0 | 81.1 | 87.3 | 80.8 | 77.0 | 87.0 |
| | BYOL | 96.2 | 94.8 | 87.5 | 81.4 | 87.6 | 81.5 | 77.0 | 87.2 |
| | **PrCL (ours)** | **96.3** | **94.9** | **88.1** | **82.3** | **87.9** | **82.8** | **77.8** | **87.8** |
| | **IMPROVEMENTS** | **+0.1** | **+0.1** | **+0.6** | **+0.7** | **+0.3** | **+1.1** | **+0.6** | **+0.6** |
| SUPERVISED | | 96.3 | 95.1 | 87.9 | 82.2 | 87.8 | 82.7 | 77.8 | 87.7 |



(a) SimCLR (Background) (b) PrCL (Background) (c) SimCLR (Digit) (d) PrCL (Digit)

Figure 4. Visualization of latent features learned using different approaches on Colorful-Moving-MNIST dataset. The color of the left two figures corresponds to background object labels, and the color of the right two figures corresponds to the digit label.

classifications. Table 4 shows that the contrastive learning baselines learn only the task of background classification, and fail to learn a representation relevant to digit classification. This shows that information related to the background

Table 3. Performance on FairFace with different unsupervised learning methods. The models are evaluated on downstream tasks of age, gender and ethnicity classification.

| Metric | | Age cls Acc. (%) | Gender cls Acc. (%) | Ethn. cls Acc. (%) |
|---|---|---|---|---|
| Fixed feature extractor | SimCLR | 43.9 | 78.1 | 61.7 |
| | MoCo | 44.5 | 78.6 | 61.9 |
| | CPC | 43.5 | 76.2 | 61.0 |
| | BYOL | 44.3 | 78.6 | **62.3** |
| | **PrCL (ours)** | **50.0** | **87.2** | 61.2 |
| | **Improvement** | **+5.7** | **+8.6** | **-1.1** |
| Fine-tuning | SimCLR | 54.3 | 91.1 | 69.1 |
| | MoCo | 54.7 | 91.3 | 69.2 |
| | CPC | 54.2 | 91.0 | 68.8 |
| | BYOL | 54.6 | 91.5 | **69.3** |
| | **PrCL (ours)** | **55.3** | **92.3** | 69.0 |
| | **Improvement** | **+0.6** | **+0.8** | **-0.3** |
| Supervised on Age | | 55.5 | 78.8 | 45.1 |
| Supervised on gender | | 43.3 | 92.5 | 45.4 |
| Supervised on Ethn. | | 42.1 | 76.8 | 69.4 |
| Supervised on All | | 54.8 | 91.9 | 68.8 |

prevents contrastive learning from capturing digit-relevant features. Note that the performance gap on digit classification between contrastive learning and supervised learning is very large (the accuracy is 15% vs. 93%). This is much larger than the gap we saw on FairFace because the information related to digit and background are totally independent, whereas features related to ethnicity, age, and gender have a significant overlap. In contrast, the representation learned by PrCL achieves very good accuracy on both background and digit classifications.

Figure 4 provides a t-SNE visualization [44] of the learned features for SimCLR and PrCL. For a clear visualization, when generating t-SNE for background, we choose samples from the same digit class, and when generating t-SNE for digits we choose samples from the same background class. This is done for both SimCLR and our method. The figure shows how predictive learning complement contrastive learning. Comparing Figures 4(c) and 4(d) reveals that PrCL's predictive branch allows it to capture information about digits that is lost in SimCLR.

Finally, we run SimCLR and PrCL on Colorful-Moving-MNIST with different feature dimensions of 512 and 1024, as shown in Table 5. These results show that the performance of SimCLR does not change with larger dimensions. In fact, the same result can be seen from our theoretical analysis, which proves that when increasing the feature dimensions, contrastive learning experiences many local minima that correspond to all of the global minima of the lower dimensions, which tend to suppress features, while PrCL can escape from those local minima.

Table 4. Performance on Colorful-Moving-MNIST under different unsupervised methods. The models are evaluated on the downstream tasks of digit classification and background object classification.

| Metric | | Digit cls Acc. (%) | Bkgd cls Acc. (%) |
|---|---|---|---|
| Fixed feature extractor | SimCLR | 14.9 | 47.3 |
| | MoCo | 15.7 | 48.5 |
| | CPC | 15.8 | 35.2 |
| | BYOL | 15.5 | **49.0** |
| | **PrCL (ours)** | **88.3** | 46.5 |
| | **Improvement** | **+72.5** | **-2.5** |
| Fine-tuning | SimCLR | 92.4 | 54.8 |
| | MoCo | 92.7 | 54.9 |
| | CPC | 92.3 | 54.7 |
| | BYOL | 92.7 | **54.9** |
| | **PrCL (ours)** | **93.3** | 54.7 |
| | **Improvement** | **+0.6** | **-0.2** |
| Supervised on digit | | 96.1 | 11.4 |
| Supervised on bkgd | | 12.9 | 56.7 |
| Supervised on digit & bkgd | | 93.0 | 54.5 |

Table 5. Performance on Colorful-Moving-MNIST with different feature dimensions under different unsupervised methods.

| Method | Feature Dimension | Digit cls Acc. (%) | Bkgd cls Acc. (%) |
|---|---|---|---|
| SimCLR | 512 | 16.0 | 48.4 |
| | 1024 | 15.8 | 48.6 |
| PrCL | 512 | 88.1 | 46.3 |
| | 1024 | 88.2 | 46.5 |

## 6. Conclusion & Limitations

In this paper, we introduce predictive contrastive learning (PrCL), a novel framework for making unsupervised contrastive learning more robust and allow it to preserve useful information in the presence of feature suppression. We theoretically analyze the reason why contrastive learning is vulnerable to feature suppression, and show that the predictive loss can help avoid feature suppression and preserve useful information. Extensive empirical results on a variety of datasets and tasks show that PrCL is effective at addressing the feature suppression problem.

The problem of feature suppression is complex; and, while PrCL provides an important improvement over the current SOTA, it has some limitations. First, PrCL sees some performance drop with fewer augmentations. The drop however is much better than the contrastive baselines. Second, PrCL tries to abstract and preserve the information in the input, but some of this information may be unnecessary or irrelevant to the downstream tasks of interest. Yet, despite these limitations, we believe that PrCL provides an important step forward toward making self-supervised learning more robust and providing richer self-supervised representations that support multi-attribute classifications and generalize well across diverse tasks.

## References


[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014.

[2] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *arXiv preprint arXiv:1906.00910*, 2019.

[3] 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.

[4] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11315–11325, 2022.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[6] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in Neural Information Processing Systems*, 33, 2020.

[7] Ting Chen and Lala Li. Intriguing properties of contrastive losses. *arXiv preprint arXiv:2011.02803*, 2020.

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

[9] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *arXiv preprint arXiv:2011.10566*, 2020.

[10] 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.

[11] Jonas Dippel, Steffen Vogler, and Johannes Höhne. Towards fine-grained visual representations by combining contrastive learning with image reconstruction and attention-weighted pooling. *arXiv preprint arXiv:2104.04323*, 2021.

[12] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015.

[13] Lijie Fan, Wenbing Huang, Chuang Gan, Stefano Ermon, Boqing Gong, and Junzhou Huang. End-to-end learning of motion representation for video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6016–6025, 2018.

[14] Lijie Fan, Wenbing Huang, Chuang Gan, Junzhou Huang, and Boqing Gong. Controllable image-to-video translation: A case study on facial expression generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3510–3517, 2019.

[15] Lijie Fan, Tianhong Li, Rongyao Fang, Rumen Hristov, Yuan Yuan, and Dina Katabi. Learning longterm representations for person re-identification using radio signals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10699–10709, 2020.

[16] Lijie Fan, Tianhong Li, Yuan Yuan, and Dina Katabi. In-home daily-life captioning using radio signals. *arXiv preprint arXiv:2008.10966*, 2020.

[17] Lijie Fan, Sijia Liu, Pin-Yu Chen, Gaoyuan Zhang, and Chuang Gan. When does contrastive learning preserve adversarial robustness from pretraining to finetuning? *Advances in Neural Information Processing Systems*, 34:21480–21492, 2021.

[18] Lijie Fan, Shengjia Zhao, and Stefano Ermon. Adversarial localization network. In *Learning with limited labeled data: weak supervision and beyond, NIPS Workshop*, volume 2, 2017.

[19] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[20] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

[21] 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.

[22] Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation learning. *arXiv preprint arXiv:2008.01065*, 2020.

[23] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[24] 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.

[25] Olivier J Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, SM Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*, 2019.

[26] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[27] 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.

[28] Wenbing Huang, Lijie Fan, Mehrtash Harandi, Lin Ma, Huaping Liu, Wei Liu, and Chuang Gan. Toward efficient action recognition: Principal backpropagation for training two-stream networks. *IEEE Transactions on Image Processing*, 28(4):1773–1782, 2018.

[29] Dongwei Jiang, Wubo Li, Miao Cao, Ruixiong Zhang, Wei Zou, Kun Han, and Xiangang Li. Speech simclr: Combining contrastive and reconstruction objective for self-supervised speech representation learning. *arXiv preprint arXiv:2010.13991*, 2020.

[30] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv preprint arXiv:1908.04913*, 2019.

[31] Tianhong Li, Peng Cao, Yuan Yuan, Lijie Fan, Yuzhe Yang, Rogerio S Feris, Piotr Indyk, and Dina Katabi. Targeted supervised contrastive learning for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6918–6928, 2022.

[32] Tianhong Li, Lijie Fan, Yuan Yuan, and Dina Katabi. Unsupervised learning for human sensing using radio signals. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3288–3297, January 2022.

[33] Tianhong Li, Lijie Fan, Mingmin Zhao, Yingcheng Liu, and Dina Katabi. Making the invisible visible: Action recognition through walls and occlusions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 872–881, 2019.

[34] Tianhong Li, Jianguo Li, Zhuang Liu, and Changshui Zhang. Few sample knowledge distillation for efficient network compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14639–14647, 2020.

[35] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020.

[36] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016.

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

[38] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[39] Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin. Variational autoencoder for deep learning of images, labels and captions. *arXiv preprint arXiv:1609.08976*, 2016.

[40] Joshua Robinson, Li Sun, Ke Yu, Kayhan Batmanghelich, Stefanie Jegelka, and Suvrit Sra. Can contrastive learning avoid shortcut solutions? *Advances in neural information processing systems*, 34:4974–4986, 2021.

[41] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

[42] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. *arXiv preprint arXiv:2005.10243*, 2020.

[43] Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. *arXiv preprint arXiv:1907.13625*, 2019.

[44] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[45] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

[46] 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.

[47] Xudong Wang, Ziwei Liu, and Stella X Yu. Unsupervised feature learning by cross-level discrimination between instances and groups. *arXiv preprint arXiv:2008.03813*, 2020.

[48] Yangtao Wang, Xi Shen, Shell Xu Hu, Yuan Yuan, James L Crowley, and Dominique Vaufreydaz. Self-supervised transformers for unsupervised object discovery using normalized cut. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14543–14553, 2022.

[49] Yangtao Wang, Xi Shen, Yuan Yuan, Yuming Du, Maomao Li, Shell Xu Hu, James L Crowley, and Dominique Vaufreydaz. Tokencut: Segmenting objects in images and videos with self-supervised transformer and normalized cut. *arXiv preprint arXiv:2209.00383*, 2022.

[50] Mang Ye, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pages 6210–6219, 2019.

[51] Yuan Yuan, Xiaodan Liang, Xiaolong Wang, Dit-Yan Yeung, and Abhinav Gupta. Temporal dynamic graph lstm for action-driven video object detection. In *ICCV*, 2017.

[52] Yuan Yuan, Yueming Lyu, Xi Shen, Ivor W Tsang, and Dit-Yan Yeung. Marginalized average attentional network for weakly-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2019.

[53] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6002–6012, 2019.