Singapore Management University

Institutional Knowledge at Singapore Management University

Research Collection School Of Computing and Information Systems | School of Computing and Information Systems

12-2021

Self-supervised learning disentangled group representation as feature

Tan WANG

Zhongqi YUE

Jianqiang HUANG

Qianru SUN
Singapore Management University, qianrusun@smu.edu.sg

Hanwang ZHANG

Follow this and additional works at: https://ink.library.smu.edu.sg/sis_research

 Part of the Databases and Information Systems Commons, and the Graphics and Human Computer Interfaces Commons

Citation

WANG, Tan; YUE, Zhongqi; HUANG, Jianqiang; SUN, Qianru; and ZHANG, Hanwang. Self-supervised learning disentangled group representation as feature. (2021). *Proceedings of the 35th Conference on Neural Information Processing Systems, Sydney, Australia, 2021 December 6-14*. 1-8.
Available at: https://ink.library.smu.edu.sg/sis_research/6227

This Conference Proceeding Article is brought to you for free and open access by the School of Computing and Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Computing and Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email cherylds@smu.edu.sg.

Self-Supervised Learning Disentangled Group Representation as Feature

Tan Wang[1] **Zhongqi Yue**[1,3] **Jianqiang Huang**[1,3] **Qianru Sun**[2] **Hanwang Zhang**[1]
[1]Nanyang Technological University [2]Singapore Management University [3]Damo Academy, Alibaba Group
{tan317,yuez0003,hanwangzhang}@ntu.edu.sg
jianqiang.jqh@gmail.com qianrusun@smu.edu.sg

Abstract

A good visual representation is an inference map from observations (images) to features (vectors) that faithfully reflects the hidden modularized generative factors (semantics). In this paper, we formulate the notion of "good" representation from a group-theoretic view using Higgins' definition of ***disentangled representation*** [38], and show that existing Self-Supervised Learning (SSL) only disentangles simple augmentation features such as rotation and colorization, thus unable to modularize the remaining semantics. To break the limitation, we propose an iterative SSL algorithm: Iterative Partition-based Invariant Risk Minimization (IP-IRM), which successfully grounds the abstract semantics and the group acting on them into concrete contrastive learning. At each iteration, IP-IRM first partitions the training samples into two subsets that correspond to an entangled group element. Then, it minimizes a subset-invariant contrastive loss, where the invariance guarantees to disentangle the group element. We prove that IP-IRM converges to a fully disentangled representation and show its effectiveness on various benchmarks.

1 Introduction

Deep learning is all about learning feature representations [5]. Compared to the conventional end-to-end supervised learning, Self-Supervised Learning (SSL) first learns a generic feature representation (*e.g.*, a network backbone) by training with unsupervised pretext tasks such as the prevailing contrastive objective [34, 16], and then the above stage-1 feature is expected to serve various stage-2 applications with proper fine-tuning. SSL for visual representation is so fascinating that it is the first time that we can obtain "good" visual features for free, just like the trending pre-training in NLP community [24, 8]. However, most SSL works only care how much stage-2 performance an SSL feature can improve, but overlook what feature SSL is learning, why it can be learned, what cannot be learned, what the gap between SSL and Supervised Learning (SL) is, and if SSL can even surpass SL?

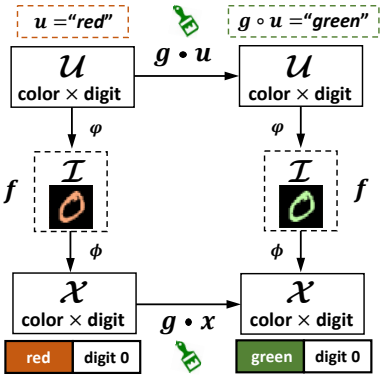


Figure 1: Disentangled representation is an equivariant map between the semantic space $\mathcal{U}$ and the vector space $\mathcal{X}$, which is decomposed into "color" and "digit".

The crux of answering those questions is to formally understand *what a feature representation is* and *what a good one is*. We postulate the classic world model of visual generation and feature representation [1, 65] as in Figure 1. Let $\mathcal{U}$ be a set of (unknown) *semantics*, *e.g.*, attributes such as "digit" and "color". There is a set of *independent and causal mechanisms* [62] $\varphi : \mathcal{U} \to \mathcal{I}$, generating images from semantics,

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

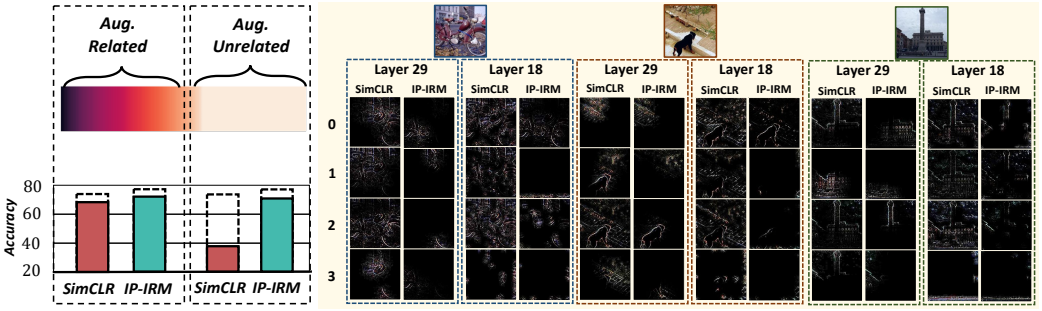


Figure 2: (a) The heat map visualizes feature dimensions related to augmentations (aug. related) and unrelated to augmentations (aug. unrelated), whose respective classification accuracy is shown in the bar chart below. Dashed bar denotes the accuracy using full feature dimensions. Experiment was performed on STL10 [22] with representation learnt with SimCLR [16] and our IP-IRM. (b) Visualization of CNN activations [72] of 4 filters on layer 29 and 18 of VGG [71] trained on ImageNet100 [76]. The filters were chosen by first clustering the aug. unrelated filters with k-means ($k = 4$) and then selecting the filters corresponding to the cluster centers.

e.g., writing a digit "0" when thinking of "0" [70]. A **visual representation** is the inference process $\phi : \mathcal{I} \to \mathcal{X}$ that maps image pixels to vector space features, *e.g.*, a neural network. We define **semantic representation** as the functional composition $f : \mathcal{U} \to \mathcal{I} \to \mathcal{X}$. In this paper, we are only interested in the parameterization of the inference process for feature extraction, but not the generation process, *i.e.*, we assume $\forall I \in \mathcal{I}, \exists u \in \mathcal{U}$, such that $I = \varphi(u)$ is fixed as the observation of each image sample. Therefore, we consider semantic and visual representations the same as **feature representation**, or simply **representation**, and we slightly abuse $\phi(I) := f\left(\varphi^{-1}(I)\right)$, *i.e.*, ϕ and f share the same trainable parameters. We call the vector $\mathbf{x} = \phi(I)$ as **feature**, where $\mathbf{x} \in \mathcal{X}$.

We propose to use Higgins' definition of **disentangled representation** [38] to define what is "good".

Definition 1 (Disentangled Representation). *Let $\mathcal{G}$ be the group acting on $\mathcal{U}$, i.e., $g \cdot u \in \mathcal{U} \times \mathcal{U}$ transforms $u \in \mathcal{U}$, e.g., a "turn green" group element changing the semantic from "red" to "green". Suppose there is a direct product decomposition*[1] *$\mathcal{G} = g_1 \times \ldots \times g_m$ and $\mathcal{U} = \mathcal{U}_1 \times \ldots \times \mathcal{U}_m$, where g_i acts on $\mathcal{U}_i$ respectively. A feature representation is disentangled if there exists a group $\mathcal{G}$ acting on $\mathcal{X}$ such that:*

1. *Equivariant*: $\forall g \in \mathcal{G}, \forall u \in \mathcal{U}, f(g \cdot u) = g \cdot f(u)$, *e.g., the feature of the changed semantic: "red" to "green" in $\mathcal{U}$, is equivalent to directly change the color vector in $\mathcal{X}$ from "red" to "green".*
2. *Decomposable*: *there is a decomposition $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_m$, such that each $\mathcal{X}_i$ is fixed by the action of all $\}_j, j \neq i$ and affected only by $\}_i$, e.g., changing the "color" semantic in $\mathcal{U}$ does not affect the "digit" vector in $\mathcal{X}$.*

Compared to the previous definition of feature representation which is a static mapping, the disentangled representation in Definition 1 is dynamic as it explicitly incorporate **group representation** [33], which is a homomorphism from group to group actions on a space, *e.g.*, $\mathcal{G} \to \mathcal{X} \times \mathcal{X}$, and it is common to use the feature space $\mathcal{X}$ as a shorthand—this is where our title stands.

Definition 1 defines "good" features in the common views: 1) *Robustness*: a good feature should be invariant to the change of environmental semantics, such as external interventions [43] or domain shifts [30]. By the above definition, a change is always retained in a subspace $\mathcal{X}_i$, while others are not affected. Hence, the subsequent classifier will focus on the invariant features and ignore the ever-changing $\mathcal{X}_i$. 2) *Zero-shot Generalization*: even if a new combination of semantics is unseen in training, each semantic has been learned as features. So, the metrics of each $\mathcal{X}_i$ trained by seen samples remain valid for unseen samples [88].

Are the existing SSL methods learning disentangled representations? No. We show in Section 4 that they can only disentangle representations according to the hand-crafted augmentations, *e.g.*, color, jitter, and rotation. For example, in Figure 2 (a), even if we only use the augmentation-related feature, the classification accuracy of a standard SSL (SimCLR [16]) does not lose much as compared to the

[1]Note that g_i can also denote a cyclic subgroup $\mathcal{G}_i$ such as rotation $[0^\circ : 1^\circ : 360^\circ]$, or a countable one but treated as cyclic such as translation $[(0, 0) : (1, 1) : (\text{width}, \text{height})]$ and color $[0 : 1 : 255]$.

full feature use. Figure 2 (b) visualizes that the CNN features in each layer are indeed entangled (*e.g.*, tyre, motor, and background in the motorcycle image). In contrast, our approach IP-IRM, to be introduced below, disentangles more useful features beyond augmentations.

In this paper, we propose Iterative Partition-based Invariant Risk Minimization (**IP-IRM** [ˌaiˈpəːm]) that guarantees to learn disentangled representations in an SSL fashion. We present the algorithm in Section 3, followed by the theoretical justifications in Section 4. In a nutshell, at each iteration, IP-IRM first partitions the training data into two disjoint subsets, each of which is an orbit of the already disentangled group, and the cross-orbit group corresponds to an entangled group element g_i. Then, we adopt the **Invariant Risk Minimization (IRM)** [2] to implement a **partition-based** SSL, which disentangles the representation $\mathcal{X}_i$ *w.r.t.* g_i. Iterating the above two steps eventually converges to a fully disentangled representation *w.r.t.* $\prod_{g\in\mathcal{G}} g$. In Section 5, we show promising experimental results on various feature disentanglement and SSL benchmarks.

2 Related Work

Self-Supervised Learning. SSL aims to learn representations from unlabeled data with hand-crafted pretext tasks [26, 59, 31]. Recently, Contrastive learning [61, 57, 36, 75, 16] prevails in most state-of-the-art methods. The key is to map positive samples closer, while pushing apart negative ones in the feature space. Specifically, the positive samples are from the augmented views [77, 3, 87, 40] of each instance and the negative ones are other instances. Along this direction, follow-up methods are mainly four-fold: 1) Memory-bank [83, 57, 34, 18]: storing the prototypes of all the instances computed previously into a memory bank to benefit from a large number of negative samples. 2) Using siamese network [7] to avoid representation collapse [32, 19, 78]. 3) Assigning clusters to samples to integrate inter-instance similarity into contrastive learning [11, 12, 13, 82, 52]. 4) Seeking hard negative samples with adversarial training or better sampling strategies [69, 20, 42, 45]. In contrast, our proposed IP-IRM jumps out of the above frame and introduces the *disentangled representation* into SSL with group theory to show the limitations of existing SSL and how to break through them.

Disentangled Representation. This notion dates back to [4], and henceforward becomes a high-level goal of separating the factors of variations in the data [79, 74, 81, 54]. Several works aim to provide a more precise description [25, 27, 68, 25] by adopting an information-theoretic view [17, 25] and measuring the properties of a disentangled representation explicitly [27, 68]. We adopt the recent group-theoretic definition from Higgins *et al.* [38], which not only unifies the existing, but also resolves the previous controversial points [73, 55]. Although supervised learning of disentangled representation is a well-studied field [92, 41, 10, 64, 66, 46], unsupervised disentanglement based on GAN [17, 60, 53, 67] or VAE [37, 15, 91, 47] is still believed to be theoretically challenging [55]. Thanks to the Higgins' definition, we prove that the proposed IP-IRM converges with full-semantic disentanglement using group representation theory. Notably, unlike all the existing unsupervised methods based on generative models, our IP-IRM is the first approach to learn an inference process, making it widely applicable even on large-scale datasets.

3 IP-IRM Algorithm

Notations. Our goal is to learn the feature extractor ϕ in a self-supervised fashion. We define a partition matrix $\mathbf{P} \in \{0,1\}^{N\times 2}$ that partitions N training images into 2 disjoint subsets. $P_{i,k} = 1$ if the i-th image belongs to the k-th subset and 0 otherwise. Suppose we have a pretext task loss function $\mathcal{L}(\phi, \theta = 1, k, \mathbf{P})$ defined on the samples in the k-th subset, where $\theta = 1$ is a "dummy" parameter used to evaluate the invariance of the SSL loss across the subsets (later discussed in Step 1). For example, $\mathcal{L}$ can be defined as:

$$\mathcal{L}(\phi, \theta = 1, k, \mathbf{P}) = \sum_{\mathbf{x}\in\mathcal{X}_k} -\log \frac{\exp\left(\mathbf{x}^T\mathbf{x}^* \cdot \theta\right)}{\sum_{\mathbf{x}'\in\mathcal{X}_k\cup\mathcal{X}^*\setminus\mathbf{x}} \exp\left(\mathbf{x}^T\mathbf{x}' \cdot \theta\right)}, \tag{1}$$

where $\mathcal{X}_k = \phi(\{I_i | P_{i,k} = 1\})$, and $\mathbf{x}^* \in \mathcal{X}^*$ is the augmented view feature of $\mathbf{x} \in \mathcal{X}_k$.

Input. N training images. Randomly initialized ϕ. A partition matrix $\mathbf{P}$ initialized such that the first column of $\mathbf{P}$ is 1, *i.e.*, all samples belong to the first subset. Set $\mathcal{P} = \{\mathbf{P}\}$.

Output. Disentangled feature extractor ϕ.

Step 1 [Update ϕ]. We update ϕ by:

$$\min_{\phi} \sum_{\mathbf{P}\in\mathcal{P}} \sum_{k=1}^{2} \left[\mathcal{L}(\phi, \theta = 1, k, \mathbf{P}) + \lambda_1 \left\| \nabla_{\theta=1} \mathcal{L}(\phi, \theta = 1, k, \mathbf{P}) \right\|^2 \right], \tag{2}$$

where λ_1 is a hyper-parameter. The second term delineates how far the contrast in one subset is from a constant baseline $\theta = 1$. The minimization of both of them encourages ϕ in different subsets close to the same baseline, *i.e.*, invariance across the subsets. See IRM [2] for more details. In particular, the first iteration corresponds to the standard SSL with $\mathcal{X}_1$ in Eq. (1) containing all training images.

Step 2 [Update $\mathbf{P}$]. We fix ϕ and find a new partition $\mathbf{P}^*$ by

$$\mathbf{P}^* = \arg\max_{\mathbf{P}} \sum_{k=1}^{2} \left[\mathcal{L}(\phi, \theta = 1, k, \mathbf{P}) + \lambda_2 \left\| \nabla_{\theta=1} \mathcal{L}(\phi, \theta = 1, k, \mathbf{P}) \right\|^2 \right], \tag{3}$$

where λ_2 is a hyper-parameter. In practice, we use a continuous partition matrix in $\mathbb{R}^{N\times 2}$ during optimization and then threshold it to $\{0, 1\}^{N\times 2}$.

We update $\mathcal{P} \leftarrow \mathcal{P} \cup \mathbf{P}^*$ and iterate the above two steps until convergence.

4 Justification

Recall that IP-IRM uses training sample **partitions** to learn the disentangled representations *w.r.t.* $\prod_{g\in\mathcal{G}} g$. As we have a $\mathcal{G}$-equivariant feature map between the sample space $\mathcal{I}$ and feature space $\mathcal{X}$ (the equivariance is later guaranteed by Lemma 1), we slightly abuse the notation by using $\mathcal{X}$ to denote both spaces. Also, we assume that $\mathcal{X}$ is a **homogeneous** space of $\mathcal{G}$, *i.e.*, any sample $\mathbf{x}' \in \mathcal{X}$ can be transited from another sample $\mathbf{x}$ by a group action $g \cdot \mathbf{x}$. Intuitively, $\mathcal{G}$ is all you need to describe the diversity of the training set.

We show that partition and group are tightly connected by the concept of **orbit**. Given a sample $\mathbf{x} \in \mathcal{X}$, its group orbit *w.r.t.* $\mathcal{G}$ is a sample set $\mathcal{G}(\mathbf{x}) = \{g \cdot \mathbf{x} \mid g \in \mathcal{G}\}$. As shown in Figure 3 (a), if $\mathcal{G}$ is a set of attributes shared by classes, *e.g.*, "color" and "pose ", the orbit is the sample set of the class of $\mathbf{x}$; in Figure 3 (b), if $\mathcal{G}$ denotes augmentations, the orbit is the set of augmented images. In particular, we can see that the disjoint orbits in Figure 3 naturally form a partition. Formally, we have the following definition:

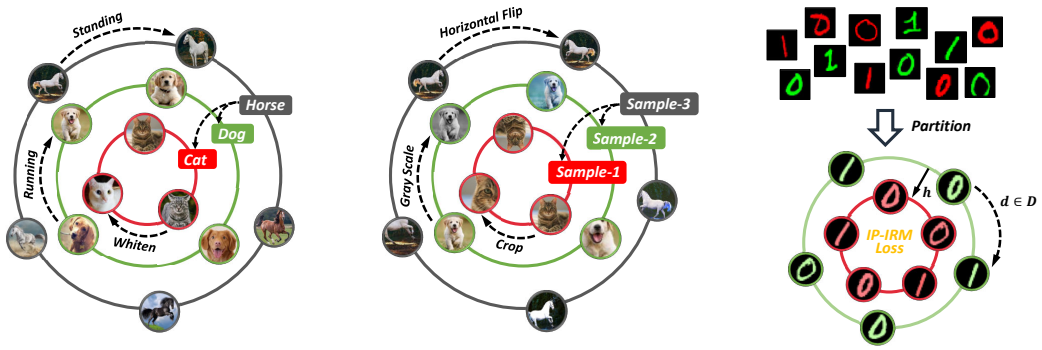
Definition 2 (Orbit & Partition [44]) *Given a subgroup $\mathcal{D} \subset \mathcal{G}$, it partitions $\mathcal{X}$ into the disjoint subsets: $\{\mathcal{D}(g_1 \cdot \mathbf{x}), ..., \mathcal{D}(g_k \cdot \mathbf{x})\}$, where k is the number of cosets $\{g_1\mathcal{D}, ..., g_k\mathcal{D}\}$, and $\{g_i\}_{i=1}^{k} = \mathcal{G}\backslash\mathcal{D}$. In particular, $g_i \cdot \mathbf{x}$ can be considered as a sample of the i-th class, transited from any sample $\mathbf{x} \in \mathcal{X}$.*

Interestingly, the partition offers a new perspective for the training data format in Supervised Learning (SL) and Self-Supervised Learning (SSL). In SL, as shown in Figure 3 (a), the data is labeled with k classes, each of which is an orbit with $\mathcal{D}(g_i \cdot \mathbf{x})$ training samples, whose variations are depicted by the class-sharing attribute group $\mathcal{D}$. The cross-orbit group action, *e.g.*, $g_{\text{dog}} \cdot \mathbf{x}$, can be read as "turn $\mathbf{x}$ into a dog" and such "turn" is always valid due to the assumption that $\mathcal{X}$ is a homogeneous space of $\mathcal{G}$. In SSL, as shown in Figure 3 (b), each training sample $\mathbf{x}$ is augmented by the group $\mathcal{D}$. So, $\mathcal{D}(g_i \cdot \mathbf{x})$ consists of all the augmentations of the i-th sample, where the cross-orbit group action $g_i \cdot \mathbf{x}$ can be read as "turn $\mathbf{x}$ into the i-th sample".

Thanks to the orbit and partition view of training data, we are ready to revisit model **generalization** in a group-theoretic view by using **invariance** and **equivariance**—the two sides of the coin, whose name is **disentanglement**. For SL, we expect that a good feature is disentangled into a class-agnostic part and a class-specific part: the former (latter) is invariant (equivariant) to $\mathcal{G}\backslash\mathcal{D}$—cross-orbit traverse, but equivariant (invariant) to $\mathcal{D}$—in-orbit traverse. By using such feature, a model can generalize to diverse testing samples (limited to $|\mathcal{D}|$ variations) by only keeping the class-specific feature. Formally, we prove that we can achieve such disentanglement by contrastive learning:

Lemma 1. (Disentanglement by Contrastive Learning) *Training loss* $-\log \frac{\exp(\mathbf{x}_i^T \mathbf{x}_j)}{\sum_{\mathbf{x}\in\mathcal{X}} \exp(\mathbf{x}_j^T \mathbf{x})}$ *disentangles $\mathcal{X}$ w.r.t. $(\mathcal{G}\backslash\mathcal{D}) \times \mathcal{D}$, where $\mathbf{x}_i$ and $\mathbf{x}_j$ are from the same orbit.*

We can draw the following interesting corollaries from Lemma 1 (details in Appendix):



(a) Supervised Learning (b) Self-Supervised Learning (c) Our IP-IRM

Figure 3: (a) Orbit: the training samples of a class, $\mathcal{D}$ in-orbit actions: intra-class variations, $\mathcal{G}\backslash\mathcal{D}$ cross-orbit actions: inter-class variations. (b) Orbit: a sample and its augmented samples, $\mathcal{D}$ in-orbit actions: augmentations, $\mathcal{G}\backslash\mathcal{D}$ cross-orbit actions: inter-sample variations. (c) Step 2 in IP-IRM discovers 2 orbits, where the cross-orbit action corresponds to a group action "green to red" or "red to green", which is yet disentangled.

1. If we use all the samples in the denominator of the loss, we can approximate to $\mathcal{G}$-equivariant features given limited training samples. This is because the loss minimization guarantees $\forall(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{X} \times \mathcal{X}, i \neq j \rightarrow \mathbf{x}_i \neq \mathbf{x}_j$, *i.e.*, any pair corresponds to a group action.
2. Conventional cross-entropy loss in SL is a special case, if we define $\mathbf{x}_i \in \mathcal{X} = \{\mathbf{x}_1, ..., \mathbf{x}_k\}$ as k classifier weights. So, SL does not guarantee the disentanglement of $\mathcal{G}\backslash\mathcal{D}$, which causes generalization error if the class domain of downstream task is different from SL pre-training, *e.g.*, a subset of $\mathcal{G}\backslash\mathcal{D}$.
3. In contrastive learning based SSL, $\mathcal{D}$ = "augmentations" (recall Figure 2), and the number of augmentations $|\mathcal{D}_{\text{aug}}|$ is generally much smaller compared to the class-wise sample diversity $|\mathcal{D}_{\text{SL}}|$ in SL. This enables the SL model to generalize to more diverse testing samples ($|\mathcal{D}_{\text{SL}}|$) by filtering out the class-agnostic features (*e.g.*, background) and focusing on the class-specific ones (*e.g.*, foreground), which explains why SSL is worse than SL in downstream classification.
4. In SL, if the number of training samples per orbit is not enough, *i.e.*, smaller than $|\mathcal{D}(g_i \cdot \mathbf{x})|$, the disentanglement between $\mathcal{D}$ and $\mathcal{G}\backslash\mathcal{D}$ cannot be guaranteed, such as the challenges in few-shot learning [89]. Fortunately, in SSL, the number is enough as we always include all the augmented samples in training. Moreover, we conjecture that $\mathcal{D}_{\text{aug}}$ only contains simple cyclic group elements such as rotation and colorization, which are easier for representation learning.

Note that Lemma 1 does not guarantee the decomposability of each $d \in \mathcal{D}$, and an improved contrastive objective is required to fully disentangle $\prod_{d\in\mathcal{D}} d$ [85]. However, the downstream model can still generalize by keeping the class-specific features affected by $\mathcal{G}\backslash\mathcal{D}$. Therefore, the key to fill the gap or even let SSL surpass SL is to achieve the full disentanglement of $\mathcal{G}\backslash\mathcal{D}_{\text{aug}}$.

Theorem 1. *The representation is fully disentangled w.r.t. $\mathcal{G}\backslash\mathcal{D}_{\text{aug}}$ if and only if $\forall g_i \in \mathcal{G}\backslash\mathcal{D}_{\text{aug}}$, the contrastive loss in Eq.* (1) *is invariant to the 2 orbits of partition $\{\mathcal{G}'(g_i \cdot \mathbf{x}), \mathcal{G}'(g_i^{-1} \cdot \mathbf{x})\}$, where $\mathcal{G}' = \mathcal{G}\backslash g_i = \mathcal{D}_{\text{aug}} \times g_1 \times \ldots \times g_{i-1} \times g_{i+1} \times \ldots \times g_k$.*

The maximization in **Step 2** is based on the contra-position of the sufficient condition of Theorem 1. Denote the currently disentangled group as $\mathcal{D}$ (initially as $\mathcal{D}_{\text{aug}}$). If we can find a partition $\mathbf{P}^*$ to maximize the loss in Eq. (3), *i.e.*, SSL loss is variant across the orbits, then $\exists h \in \mathcal{G}\backslash\mathcal{D}$ such that the representation of h is entangled, *i.e.*, $\mathbf{P}^* = \{\mathcal{G}'(h \cdot \mathbf{x}), \mathcal{G}'(h^{-1} \cdot \mathbf{x})\}$ with $\mathcal{G}' = \mathcal{G}\backslash h$. Figure 3 (c) illustrates a discovered partition about color. The minimization in **Step 1** is based on the necessary condition of Theorem 1. Based on the discovered $\mathbf{P}^*$, if we minimize Eq. (2), we can disentangle *w.r.t.* $\mathcal{D} \times h$ and update $\mathcal{D} \leftarrow \mathcal{D} \cup h$. Therefore, IP-IRM converges as $\mathcal{G}$ is finite.

5 Experiments

5.1 Unsupervised Disentanglement

Datasets. We used two datasets. **CMNIST** [2] has 60,000 digit images with semantic labels of digits (0-9) and colors (red and green). These images differ in other semantics (*e.g.*, slant and font) that are

	Method	DCI	IRS	MOD	EXP	LR	GBT	Average
CMNIST	VAE [48]	**0.952**	-	0.842	0.972	0.829	**0.952**	0.909
	β-VAE [39]	0.947	-	0.790	0.970	0.825	0.947	0.896
	β-AnnealVAE [9]	0.914	-	0.866	0.974	0.839	0.914	0.901
	β-TCVAE [15]	0.924	-	**0.982**	0.958	0.782	0.924	0.914
	Factor-VAE [47]	0.921	-	0.97	0.959	0.788	0.922	0.912
	SimCLR [16]	0.895	-	0.773	0.986	0.900	0.894	0.890
	IP-IRM (Ours)	0.921	-	0.830	**0.991**	**0.927**	0.920	**0.918**
Shapes3D	VAE [48]	0.322	0.295	0.807	0.764	0.368	0.321	0.480
	β-VAE [39]	0.347	0.303	**0.820**	0.824	0.468	0.353	0.519
	β-AnnealVAE [9]	0.351	**0.446**	0.718	0.636	0.269	0.352	0.462
	β-TCVAE [15]	0.410	0.287	0.806	0.742	0.304	0.413	0.494
	Factor-VAE [47]	0.369	0.333	**0.820**	0.786	0.371	0.371	0.508
	SimCLR [16]	0.512	0.402	0.594	0.942	0.639	0.514	0.601
	IP-IRM (Ours)	**0.533**	0.417	0.791	**0.953**	**0.734**	**0.532**	**0.660**

Table 1: Results on disentanglement metrics of existing unsupervised disentanglement methods, standard SSL (SimCLR [16]) and IP-IRM using CMNIST [2] and Shapes3D [47]. Note that IRS is based on intervening the semantics which requires access to the labels of all the semantics, and hence not applicable for CMNIST dataset.

not labeled. Moreover, there is a strong correlation between digits and colors (most 0-4 in red and 5-9 in green), increasing the difficulty to disentangle them. **Shapes3D** [47] contains 480,000 images with 6 labelled semantics, *i.e.*, size, type, azimuth, as well as floor, wall and object color. Note that we only considered the first three semantics for evaluation, as the standard augmentations in SSL will contaminate any color-related semantics.

Settings. We adopted 6 representative disentanglement metrics: *Disentangle Metric for Informativeness* (DCI) [27], *Interventional Robustness Score* (IRS) [74], *Explicitness Score* (EXP) [68], *Modularity Score* (MOD) [68] and the accuracy of predicting the ground-truth semantic labels by two classification models called *logistic regression* (LR) and *gradient boosted trees* (GBT) [55]. Specifically, DCI and EXP measure the explicitness, *i.e.*, the values of semantics can be decoded from the feature using a linear transformation. MOD and IRS measure the modularity, *i.e.*, whether each feature dimension is equivariant to the shift of a single semantic. See Appendix for more detailed formula of the metrics. In evaluation, we trained CNN-based feature extractor backbones with comparable number of parameters for all the baselines and our IP-IRM, and the full implementation details are in Appendix.

Results. In Table 1, we compared the proposed IP-IRM to the standard SSL method SimCLR [16] as well as several generative disentanglement methods [48, 39, 9, 15, 47]. On both CMNIST and Shapes3D dataset, IP-IRM significantly outperforms SimCLR regrading all metrics where the most relative gain is 19.7% for MOD. For this MOD, we notice that β-VAE and Factor-VAE perform better than our IP-IRM by 3 points, *i.e.*, 0.82 v.s. 0.79 for Shapes3D. This is because VAE explicitly pursues a high modularity score through regularizing the dimension-wise independence in the feature space. However, this regularization is adversarial to discriminative objectives [14, 88]. Indeed, we can observe from the column of LR (i.e., the performance of downstream linear classification) that VAE methods have clearly poor performance especially on the more challenging dataset Shapes3D. We can draw the same conclusion from the results of GBT. Different from VAE methods, our IP-IRM is optimized towards disentanglement without such regularization, and is thus able to outperform the others in downstream tasks while obtaining a competitive value of modularity.

What does IP-IRM feature look like? Figure 4 visualizes the features learned by SimCLR and our IP-IRM on two datasets: CMNIST in Figure 4 (a) and STL10 dataset in Figure 4 (b). In the following, we use Figure 4 (a) as the example, and can easily draw the similar conclusions from Figure 4 (b). On the left-hand side of Figure 4 (a), it is obvious that there is no clear boundary to distinguish the semantic of color in the SimCLR feature space. Besides, the features of the same digit semantic are scattered in two regions. On the right-hand side of (a), we have 3 observations for IP-IRM. 1) The features are well clustered and each cluster corresponds to a specific semantic of either digit or color. This validates the *equivariant* property of IP-IRM representation that it responds to any changes of the existing semantics, *e.g.*, digit and color on this dataset. 2) The feature space has the symmetrical structure for each individual semantic, validating the *decomposable* property of IP-IRM representation. More specifically, i) mirroring a feature (*w.r.t.* "*" in the figure center) indicates the change on the only semantic of color, regardless the other semantic (digit); and ii) a

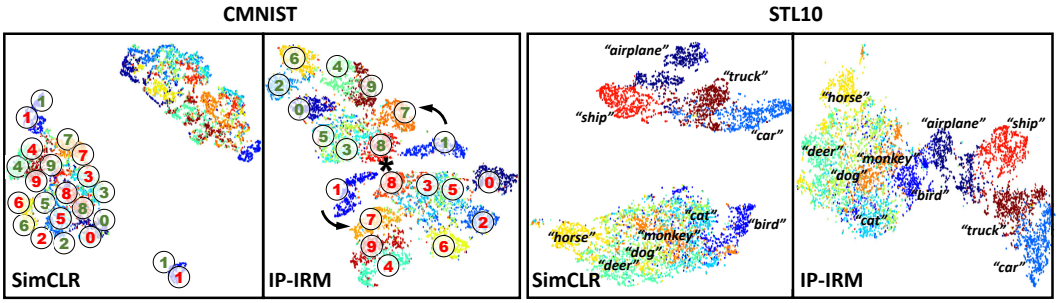


Figure 4: The t-SNE [80] visualizations of learned feature spaces using SimCLR [16] and IP-IRM on CMNIST [2] and STL10 [22]. For CMIST in (a), we annotate the digit and color near each cluster. We annotate only half of the feature points for SimCLR to avoid clutter. For STL10 in (b), we show the labels of the classes.

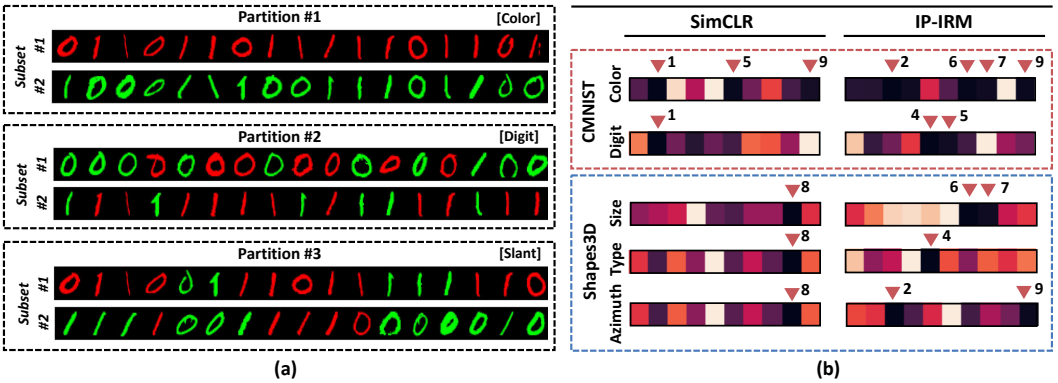


Figure 5: (a) Visualization of the obtained partitions $\mathbf{P}^*$ during training. Each partition has two subset and the displayed images are *randomly* sampled from each subset. (b) Visualization of the variance of each feature dimension when perturbing the semantic indicated on the left. The most equivariant dimensions are indicated by triangles and their corresponding indices.

counterclockwise rotation (denoted by black arrows from same-colored 1 to 7) indicates the change on the only semantic of digit. 3) IP-IRM reveals the true distribution (similarity) of different classes. For example, digits 3, 5, 8 sharing sub-parts (curved bottoms and turnings) have closer feature points in the IP-IRM feature space.

How does IP-IRM disentangle features? *1) Discovered* $\mathbf{P}^*$: To visualize the discovered partitions $\mathbf{P}^*$ at each maximization step, we performed an experiment on a binary CMNIST (digit 0 and 1 in color red and green), and show the results in Figure 5 (a). Please kindly refer to Appendix for the full results on CMNIST. First, each partition tells apart a specific semantic into two subsets, *e.g.*, in Partition #1, red and green digits are separated. Second, besides the obvious semantics—digit and color (labelled on the dataset), we can discover new semantics, *e.g.*, the digit slant shown in Partition #3. *2) Disentangled Representation*: In Figure 5 (b), we aim to visualize how equivariant each feature dimension is to the change of each semantic, *i.e.*, a darker color shows that a dimension is more equivariant *w.r.t.* the semantic indicated on the left. We can see that SimCLR fails to learn the decomposable representation, *e.g.*, the 8-th dimension captures azimuth, type and size in Shapes3D. In contrast, our IP-IRM achieves disentanglement by representing the semantics into interpretable dimensions, *e.g.*, the 6-th and 7-th dimensions captures the size, the 4-th for type and the 2-nd and 9-th for azimuth on the Shapes3D. Overall, the results support the justification in Section 4, *i.e.*, we discover a new semantic (affected by h) through the partition $\mathbf{P}^*$ at each iteration and IP-IRM eventually converges with a disentangled representation.

5.2 Self-Supervised Learning

Datasets and Settings. We conducted the SSL evaluations on 2 standard benchmarks following [82, 20, 45]. **Cifar100** [51] contains 60,000 images in 100 classes and **STL10** [22] has 113,000 images in 10 classes. We used SimCLR [16], DCL [20] and HCL [45] as baselines, and learned the

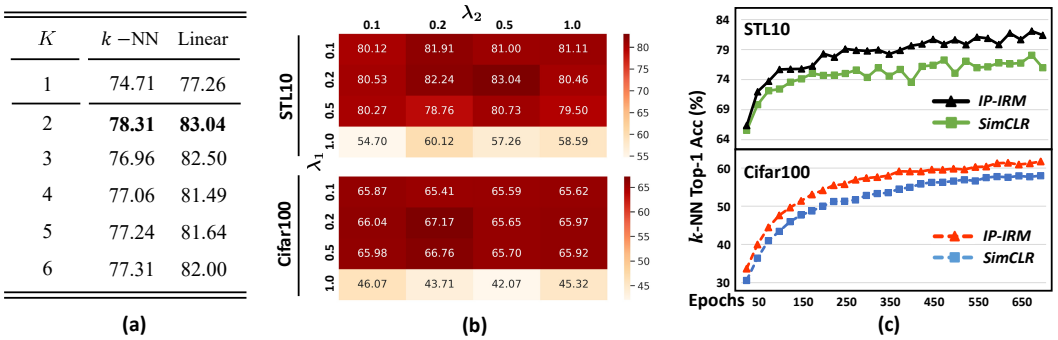


Figure 6: Our ablation study on the STL10 and Cifar100 datasets. (a) The effect of K, *i.e.*, the number of the subsets in $\mathbf{P}$, on the STL10 dataset. (b) The Top-1 accuracy (%) of linear classifiers using different values of λ_1 and λ_2 (in Eq. (2) and Eq. (3)), by training for 200 epochs on two datasets. (c) The Top-1 accuracy (%) of k-NN classifiers on two datasets, for which we trained the models for 700 epochs and updated $\mathbf{P}$ every 50 epochs.

representations for 400 epochs. We evaluated both linear and k-NN ($k = 200$) accuracies for the downstream classification task.

Results. We demonstrate our results and compare with baselines in Table 2. Incorporating IP-IRM to the 3 baselines brings consistent performance boosts to downstream classification models in all settings, *e.g.*, improving the linear models by 5.55% on STL10 and 2.92% on Cifar100. In particular, we observe that IP-IRM brings huge performance gain with k-NN classifiers, *e.g.*, 4.23% using HCL+IP-IRM on STL10, *i.e.*, the distance metrics in the IP-IRM feature space more faithfully reflects the class semantic differences. This validates that our algorithm further disentangles compared to the standard SSL Still, the quality of disentanglement cannot be fully evaluated when the training and test samples are identically distributed—while the improved accuracy demonstrates that IP-IRM representation is more equivariant to class semantics, it does not reveal if the representation is decomposable. Hence we present an out-of-distribution (OOD) setting in Section 5.3 to further show this property.

Method	STL10		Cifar100	
	k-NN	Linear	k-NN	Linear
SimCLR [16]	73.60	78.89	54.94	66.63
DCL [20]	78.82	82.56	57.29	68.59
HCL [45]	80.06	87.60	59.61	69.22
SimCLR+IP-IRM	79.66	84.44	59.10	69.55
DCL+IP-IRM	81.51	85.36	58.37	68.76
HCL+IP-IRM	**84.29**	**87.81**	**60.05**	**69.95**

Table 2: Accuracy (%) of k-NN and linear classifiers on STL10 [22] and Cifar100 [51] using the representations of SimCLR [16], DCL [20], HCL [45] and those after incorporating our IP-IRM.

Is IP-IRM sensitive to the values of hyper-parameters? 1) *The number of subsets* K. In Figure 6 (a), we find that increasing K causes a decrease of the model performance, as it reduces the difficulty of pretext tasks, impairing the model learning (see Appendix for details). 2) λ_1 *and* λ_2 *in Eq.* (2) *and Eq.* (3). In Figure 6 (b), we observe that the best performance is achieved with λ_1 and λ_2 taking values from 0.2 to 0.5 on both datasets. All accuracies drop sharply if using $\lambda_1 = 1.0$. The reason is that a higher λ_1 forces the model to push the ϕ-induced similarity to fixed baseline $\theta = 1$, rather than decrease the loss $\mathcal{L}$ on the pretext task, leading to poor convergence. 3) *The number of epochs*. In Figure 6 (c), we plot the Top-1 accuracies of using k-NN classifiers along the 700-epoch training of two kinds of SSL representations—SimCLR and IP-IRM. It is obvious that IP-IRM converges faster and achieves a higher accuracy than SimCLR. It is worth to highlight that on the STL10, the accuracy of SimCLR starts to oscillate and grow slowly after the 150-th epoch, while ours keeps on improving. This is an empirical evidence that IP-IRM keeps on disentangling more and more semantics in the feature space, and has the potential of improvement through long-term training.

5.3 Potential on Large-Scale Data

Datasets. We evaluated on the standard benchmark of supervised learning **ImageNet ILSVRC-2012** [23] which has in total 1,331,167 images in 1,000 classes. To further reveal if a representation is decomposable, we used **NICO** [35], which is a real-world image dataset designed for OOD eval-

Method	ImageNet	NICO
InsDis [83]	56.5	41.6
PCL [52]	61.5	61.2
PIRL [57]	63.6	46.5
MoCo-v1 [34]	60.6	45.1
MoCo-v2 [18]	67.5	64.3
IP-IRM (Ours)	**69.1**	**76.8**

Table 3: ImageNet and NICO Top-1 Accuracy (%) of linear classifiers trained on the representations learnt with different SSL methods.

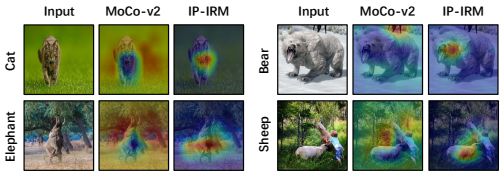


Figure 7: Visualization of CAM [90] on images from NICO [35] dataset using representations of the baseline MoCo-v2 [18] and our IP-IRM.

Method	Aircraft	Caltech	Cars	Cifar10	Cifar100	DTD	Flowers	Food	Pets	SUN	Average
InsDis [83]	35.07	75.97	37.49	51.49	57.61	69.38	77.35	50.01	66.38	74.97	59.57
PCL [52]	**36.86**	90.72	39.68	59.26	60.78	69.53	67.50	57.06	**88.31**	84.51	65.42
PIRL [57]	36.70	78.63	39.21	49.85	55.23	70.43	78.37	51.61	69.40	76.64	60.61
MoCo-v1 [34]	35.31	79.60	36.35	46.96	51.62	68.76	75.42	49.77	68.32	74.77	58.69
MoCo-v2 [18]	31.98	92.32	41.47	56.50	63.33	78.00	80.05	57.25	83.23	88.10	67.22
IP-IRM (Ours)	32.98	**93.16**	**42.87**	**60.73**	**68.54**	**79.30**	**82.68**	**59.61**	85.23	**89.38**	**69.44**

Table 4: Accuracy (%) of 5-way-5-shot few-shot evaluation using the image representation learned on ImageNet [23]. More detailed results are given in Appendix.

uations. It contains 25,000 images in 19 classes, with a strong correlation between the foreground and background in the train split (*e.g.*, most dogs on grass). We also studied the transferability of the learned representation following [28, 49]: FGVC Aircraft (**Aircraft**) [56], Caltech-101 (**Caltech**) [29], Stanford Cars (**Cars**) [86], **Cifar10** [50], **Cifar100** [50], **DTD** [21], Oxford 102 Flowers (**Flowers**) [58], Food-101 (**Food**) [6], Oxford-IIIT Pets (**Pets**) [63] and SUN397 (**SUN**) [84]. These datasets include coarse- to fine-grained classification tasks, and vary in the amount of training data (2,000-75,000 images) and classes (10-397 classes), representing a wide range of transfer learning settings.

Settings. For the ImageNet, all the representations were trained for 200 epochs due to limited computing resources. We followed the common setting [75, 34], using a linear classifier, and report Top-1 classification accuracies. For NICO, we fixed the ImageNet pre-trained ResNet-50 backbone and fine-tuned the classifier. See appendix for more training details. For the transfer learning, we followed [28, 49] to report the classification accuracies on Cars, Cifar-10, Cifar-100, DTD, Food, SUN and the average per-class accuracies on Aircraft, Caltech, Flowers, Pets. We call them uniformly as Accuracy. We used the few-shot n-way-k-shot setting for model evaluation. Specifically, we randomly sampled 2,000 episodes from the *test* splits of above datasets. An episode contains n classes, each with k training samples and 15 testing samples, where we fine-tuned the linear classifier (backbone weights frozen) for 100 epochs on the training samples, and evaluated the classifier on the testing samples. We evaluated with $n = k = 5$ (results of $n = 5, k = 20$ in Appendix).

ImageNet and NICO. In Table 3 ImageNet accuracy, our IP-IRM achieves the best performance over all baseline models. Yet we believe that this does not show the full potential of IP-IRM, because ImageNet is a larger-scale dataset with many semantics, and it is hard to achieve a full disentanglement of all semantics within the limited 200 epochs. To evaluate the feature decomposability of IP-IRM, we compared the performance on NICO with various SSL baselines in Table 3, where our approach significantly outperforms the best baseline by 2.9%. This validates IP-IRM feature is more decomposable—if each semantic feature (*e.g.*, background) is decomposed in some fixed dimensions and some classes vary with such semantic, then the classifier will recognize this as a non-discriminative variant feature and hence focus on other more discriminative features (*i.e.*, foreground). In this way, even though some classes are confounded by those non-discriminative features (*e.g.*, most of the "dog" images are with "grass" background), the fixed dimensions still help classifiers neglect those non-discriminative ones. We further visualized the CAM [90] on NICO in Figure 7, which indeed shows that IP-IRM helps the classifier focus on the foreground regions.

Few-Shot Tasks. As shown in Table 4, our IP-IRM significantly improves the performance of 5-way-5-shot setting, *e.g.*, we outperform the baseline MoCo-v2 by 2.2%. This is in line with recent

works [81] showing that a disentangled representation is especially beneficial in low-shot scenarios, and further demonstrates the importance of disentanglement in downstream tasks.

6 Conclusion

We presented an unsupervised disentangled representation learning method called Iterative Partition-based Invariant Risk Minimization (IP-IRM), based on Self-Supervised Learning (SSL). IP-IRM iteratively partitions the dataset into semantic-related subsets, and learns a representation invariant across the subsets using SSL with an IRM loss. We show that with theoretical guarantee, IP-IRM converges with a disentangled representation under the group-theoretical view, which fundamentally surpasses the capabilities of existing SSL and fully-supervised learning. Our proposed theory is backed by strong empirical results in disentanglement metrics, SSL classification accuracy and transfer performance. IP-IRM achieves disentanglement without using generative models, making it widely applicable on large-scale visual tasks. As future directions, we will continue to explore the application of group theory in representation learning and seek additional forms of inductive bias for faster convergence.

References

[1] Philip W Anderson. More is different. *Science*, 1972.

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

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

[4] Yoshua Bengio. *Learning deep architectures for AI*. Now Publishers Inc, 2009.

[5] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

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

[7] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*, 6:737–744, 1993.

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020.

[9] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae. *arXiv preprint arXiv:1804.03599*, 2018.

[10] Ruichu Cai, Zijian Li, Pengfei Wei, Jie Qiao, Kun Zhang, and Zhifeng Hao. Learning disentangled semantic representation for domain adaptation. In *IJCAI: proceedings of the conference*, 2019.

[11] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018.

[12] Mathilde Caron, Piotr Bojanowski, Julien Mairal, and Armand Joulin. Unsupervised pre-training of image features on non-curated data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2959–2968, 2019.

[13] 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.

[14] Long Chen, Hanwang Zhang, Jun Xiao, Wei Liu, and Shih-Fu Chang. Zero-shot visual recognition using semantics-preserving adversarial embedding networks. In *CVPR*, 2018.

[15] Ricky TQ Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in neural information processing systems*, 2018.

[16] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[17] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, 2016.

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

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

[20] Ching-Yao Chuang, Joshua Robinson, Lin Yen-Chen, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *arXiv preprint arXiv:2007.00224*, 2020.

[21] 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*, 2014.

[22] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

[23] 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.

[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019.

[25] Kien Do and Truyen Tran. Theory and evaluation metrics for learning disentangled representations. In *International conference on learning representations*, 2020.

[26] 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.

[27] Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. In *International conference on learning representations*, 2018.

[28] Linus Ericsson, Henry Gouk, and Timothy M. Hospedales. How Well Do Self-Supervised Models Transfer? In *CVPR*, 2021.

[29] 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*, 2004.

[30] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 2016.

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

[32] 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.

[33] W.F.J. Harris, W. Fulton, and J. Harris. *Representation Theory: A First Course*. 1991.

[34] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.

[35] Yue He, Zheyan Shen, and Peng Cui. Towards non-iid image classification: A dataset and baselines. *Pattern Recognition*, 110:107383, 2021.

[36] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pages 4182–4192. PMLR, 2020.

[37] I. Higgins, Loïc Matthey, A. Pal, C. Burgess, Xavier Glorot, M. Botvinick, S. Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

[38] Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

[39] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *International conference on learning representations*, 2017.

[40] 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.

[41] Jun-Ting Hsieh, Bingbin Liu, De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. Learning to decompose and disentangle representations for video prediction. In *Advances in neural information processing systems*, 2018.

[42] Qianjiang Hu, Xiao Wang, Wei Hu, and Guo-Jun Qi. Adco: Adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries. *arXiv preprint arXiv:2011.08435*, 2020.

[43] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, 2019.

[44] Thomas W. Judson. *Abstract Algebra: Theory and Applications (The Prindle, Weber & Schmidt Series in Advanced Mathematics)*. Prindle Weber & Schmidt, 1994.

[45] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. *arXiv preprint arXiv:2010.01028*, 2020.

[46] Theofanis Karaletsos, Serge Belongie, and Gunnar Rätsch. Bayesian representation learning with oracle constraints. *International conference on learning representations*, 2015.

[47] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, 2018.

[48] Diederik Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[49] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[50] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2012.

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

[52] Junnan Li, Pan Zhou, Caiming Xiong, Richard Socher, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020.

[53] Zinan Lin, Kiran Thekumparampil, Giulia Fanti, and Sewoong Oh. Infogan-cr and modelcentrality: Self-supervised model training and selection for disentangling gans. In *International Conference on Machine Learning*, 2020.

[54] F. Locatello, M. Tschannen, S. Bauer, G. Rätsch, B. Schölkopf, and O. Bachem. Disentangling factors of variations using few labels. In *8th International Conference on Learning Representations (ICLR)*, 2020.

[55] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, 2019.

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

[57] 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.

[58] 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*, 2008.

[59] 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.

[60] Utkarsh Ojha, Krishna Kumar Singh, Cho-Jui Hsieh, and Yong Jae Lee. Elastic-infogan: Unsupervised disentangled representation learning in class-imbalanced data. In *Advances in neural information processing systems*, 2020.

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

[62] Giambattista Parascandolo, Niki Kilbertus, Mateo Rojas-Carulla, and Bernhard Schölkopf. Learning independent causal mechanisms. In *Proceedings of the 35th International Conference on Machine Learning*, pages 4036–4044, 2018.

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

[64] Robin Quessard, Thomas Barrett, and William Clements. Learning disentangled representations and group structure of dynamical environments. *Advances in Neural Information Processing Systems*, 2020.

[65] Rajesh PN Rao and Daniel L Ruderman. Learning lie groups for invariant visual perception. *Advances in neural information processing systems*, 1999.

[66] Scott Reed, Kihyuk Sohn, Yuting Zhang, and Honglak Lee. Learning to disentangle factors of variation with manifold interaction. In *International conference on machine learning*, 2014.

[67] Xuanchi Ren, Tao Yang, Yuwang Wang, and Wenjun Zeng. Do generative models know disentanglement? contrastive learning is all you need. *arXiv preprint arXiv:2102.10543*, 2021.

[68] Karl Ridgeway and Michael C Mozer. Learning deep disentangled embeddings with the f-statistic loss. In *Advances in neural information processing systems*, 2018.

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

[70] Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

[71] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations*, 2015.

[72] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[73] Raphael Suter, Djordje Miladinovic, Stefan Bauer, and Bernhard Schölkopf. Interventional robustness of deep latent variable models. *arXiv*, 2018.

[74] Raphael Suter, Djordje Miladinovic, Bernhard Schölkopf, and Stefan Bauer. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In *International Conference on Machine Learning*, 2019.

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

[76] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *European conference on computer vision*, 2020.

[77] 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.

[78] Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. *arXiv preprint arXiv:2102.06810*, 2021.

[79] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

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

[81] Sjoerd Van Steenkiste, Francesco Locatello, Jürgen Schmidhuber, and Olivier Bachem. Are disentangled representations helpful for abstract visual reasoning? In *Advances in neural information processing systems*, 2019.

[82] 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.

[83] 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.

[84] 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*, 2010.

[85] Tete Xiao, Xiaolong Wang, Alexei A Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. In *International Conference on Learning Representations*, 2021.

[86] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[87] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 6210–6219, 2019.

[88] Zhongqi Yue, Tan Wang, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. Counterfactual zero-shot and open-set visual recognition. In *CVPR*, 2021.

[89] Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. Interventional few-shot learning. In *NeurIPS*, 2020.

[90] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[91] Yizhe Zhu, Martin Renqiang Min, Asim Kadav, and Hans Peter Graf. S3vae: Self-supervised sequential vae for representation disentanglement and data generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[92] Zhenyao Zhu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Multi-view perceptron: a deep model for learning face identity and view representations. *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, 2014.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] The theoretical justification can be found in Section 4 and the experiment results are included in Section 5.
   (b) Did you describe the limitations of your work? [Yes] See Section 6
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] We propose a practical self-supervised approach for representation disentanglement, which will help downstream tasks to remove bias, hence promoting fairness and transparency.
   (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] See Section 4 with details in Appendix.
   (b) Did you include complete proofs of all theoretical results? [Yes] See 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] See supplementary materials.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See supplementary materials.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We reported the standard deviation for few-shot experiments in Appendix following the standard setting.
   (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] See 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? [Yes]
   (b) Did you mention the license of the assets? [Yes] See Appendix.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We included our code in the supplementary materials.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] We only used publicly available assets. Details in Appendix.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See Appendix.

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] No crowdsourcing or human subjects involved.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] No crowdsourcing or human subjects involved.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] No crowdsourcing or human subjects involved.