Self-Supervised Learning Disentangled Group Representation for Efficient Classification

Tan Wang[1] **Wanqi Yin**[2] **Jiaxin Qi**[1] **Jin Liu**[2] **Jayashree Karlekar**[2] **Hanwang Zhang**[1]
[1]Nanyang Technological University [2]Panasonic R&D Center Singapore
{tan317, jiaxin003}@e.ntu.edu.sg; hanwangzhang@ntu.edu.sg
{wanqi.yin, jin.liu, karlekar.jayashree}@sg.panasonic.com

Abstract

A good visual representation is an inference map from observations (images) to features (vectors) that faithfully reflects the structure and transformations of the underlying generative factors (semantics), which are invariant to environmental changes. In this technical report, we formulate the notion of "good" representation from a group-theoretic view using Higgins' definition of ***disentangled representation*** [35] to improve existing Self-Supervised Learning (SSL). Specifically, we applied an iterative SSL method in the first stage: Iterative Partition-based Invariant Risk Minimization (IP-IRM), which successfully grounds the abstract group actions into a concrete SSL optimization. At each iteration, IP-IRM first partitions the training samples into subsets. In particular, the partition reflects an entangled semantic group action. Then, it leverages IRM to learn subset-invariant sample similarities, where the invariance guarantees to disentangle the corresponding semantic. After the SSL pretraining, we performed the knowledge distillation to transfer the learned disentangled representations to the student model in the second stage for efficient image classification. Together with other tricks, we achieved competitive performance (71.6% test accuracy) in the Visual Inductive Priors for Data-Efficient Computer Vision (VIPriors) challenge image classification track.

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 [31, 15], 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 large-scale pre-training trend in NLP community [23, 8]. Most works only care how much stage-2 performance an SSL feature can improve, but overlook what features SSL is learning, why they can be learned, what cannot be learned, how far is SSL from supervised learning, and can SSL surpass it?

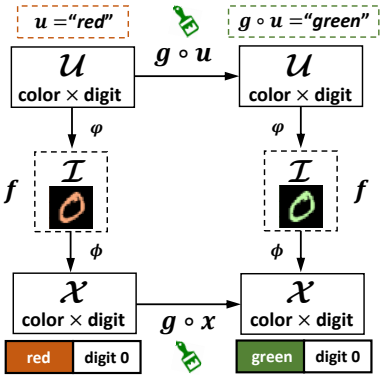

Figure 1: Disentangled representation is an equivariant map between the semantic space $\mathcal{U}$ and the vector space $\mathcal{X}$, which are decomposed into "color" and "digit".

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

Preprint. Under review.

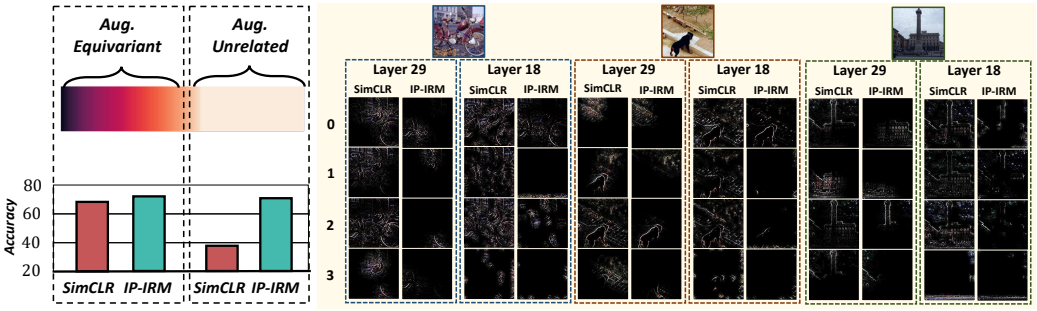

Figure 2: (a) The heat map visualizes feature dimensions equivariant to augmentations (aug. equivariant) and unrelated to augmentations (aug. unrelated), whose respective classification accuracy is shown in the bar chart below. Experiment was performed on STL10 [20] with representation learnt with SimCLR [15] and our IP-IRM. (b) Visualization of CNN activations [63] of 4 filters on layer 29 and 18 of VGG [62] trained on ImageNet100 [67]. 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.

model of visual generation and feature representation [1, 55] 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* [52] $\varphi : \mathcal{U} \to \mathcal{I}$, generating images from semantics, *e.g.*, writing a digit "0" when thinking of "0" [61]. 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. 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 propose to use Higgins' definition of **disentangled representation** [35] to define what is "good". Note that this definition is a re-formulation of the model in Figure 1 using group representation theory [30] (See Appendix for the group theory review).

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

1. *Equivariant*: $\forall g \in \mathcal{G}, \forall u \in \mathcal{U}, f(g \circ u) = g \circ f(u)$, *e.g.*, changing the semantic "red" to "green" in $\mathcal{U}$ is equivalent to change the color vector in $\mathcal{X}$ from the value encoding "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 $\mathcal{G}_j, j \neq i$ and affected only by $\mathcal{G}_i$, *e.g.*, changing the color semantic in $\mathcal{U}$ does not affect the digit vector in $\mathcal{X}$.

The above definition indeed defines "good" features for classification in the common views: 1) *Robustness*: a good feature should be invariant to the change of environmental semantics, such as external interventions [39] or domain shifts [27]. By the above definition, a change is always retained in a subspace $\mathcal{X}_i$, while others are not affected. Hence, given the labelled data in a downstream task, the trained classifier will focus on the task-relevant features and ignore the ever-changing $\mathcal{X}_i$. 2) *Generalizability*: the structure ($\mathcal{G}$-module [30]) of the feature vector space $\mathcal{X}$ corresponds to the decomposition in $\mathcal{U}$ faithfully, and any change in $\mathcal{U}$ will be preserved in $\mathcal{X}$, even for unseen ones. So, the metrics of $\mathcal{X}$ trained by seen samples remains valid for zero-shot unseen samples [75, 43] and provides more generalizability in other tasks [76, 56, 6].

Are the existing SSL methods learning disentangled representations? No. We show that they can only learn representations that are disentangled according to the hand-crafted augmentation semantics, *e.g.*, color, jitter, and rotation. For example, in Figure 2 (a), we show that one part of the SSL features is equivariant with augmentations, *i.e.*, it captures the augmentation-related semantics, while the other part is not. Then, the classification accuracy of a standard SSL (SimCLR [15]) is mostly attributed first part, demonstrating that the other part is still useless as the features are still entangled. Moreover,

in Figure 2 (b) the visualization of CNN activation in each layer indeed reveals entangled semantics (*e.g.*, tyre, motor and background in the motorcycle image). In contrast, our approach IP-IRM, to be introduced below, learns features useful for classification beyond augmentations as shown in Figure 2 (a) and disentangle the semantics in Figure 2 (b).

In this paper, we propose Iterative Partition-based Invariant Risk Minimization (**IP-IRM**) that guarantees to learn disentangled representations in an SSL fashion. We present the algorithm in Section 3. In a nutshell, the core of IP-IRM is to partition training data into disjoint subsets. At each iteration, it first discovers a group $\mathcal{G}_i$ that is still entangled, whose action corresponds to identity mapping in each subset of the partition. Then, we adopt the **Invariant Risk Minimization (IRM)** [2] to implement a **partition-based** SSL, which disentangles the representation f *w.r.t.* $\mathcal{G}_i$. Iterating the above two steps eventually converges at a fully disentangled representation *w.r.t.* $\mathcal{G}_1 \times \ldots \times \mathcal{G}_m$. We further explore IP-IRM for the downstream data efficient classification task [9] and show promising results as an empirical support.

2 Related Work

Self-Supervised Learning. SSL aims to learn representations from unlabeled data with hand-crafted pretext tasks [25, 49, 28]. Recently, Contrastive learning [51, 48, 32, 66, 15] 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 [68, 3, 74, 36] of each instance and the negative ones are other instances. Along this direction, follow-up methods are mainly four-fold: 1) Memory-bank [73, 48, 31, 17]: 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 [29, 18, 69]. 3) Assigning clusters to samples to integrate inter-instance similarity into contrastive learning [11, 12, 13, 72, 44]. 4) Seeking hard negative samples with adversarial training or better sampling strategies [60, 19, 38, 40]. 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 [70, 65, 71, 46]. Several works aim to provide a more precise description [24, 26, 59, 24] by adopting an information-theoretic view [16, 24] and measuring the properties of a disentangled representation explicitly [26, 59]. We adopt the recent group-theoretic definition from Higgins *et al.* [35], which not only unifies the existing, but also resolves the previous controversial points [64, 47]. Although supervised learning of disentangled representation is a well-studied field [79, 37, 10, 54, 57, 41], unsupervised disentanglement based on GAN [16, 50, 45, 58] or VAE [34, 14, 78, 42] is still believed to be theoretically challenging [47]. 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 Approach

3.1 Stage-I: IP-IRM Algorithm

Notations. Our goal is to learn the feature extractor ϕ in an unsupervised way. We define a partition matrix $\mathbf{P} \in \{0,1\}^{N\times K}$ that partitions N training images into K 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 $\mathcal{L}(\phi, \theta = 1.0, k, \mathbf{P})$ defined on the samples in the k-th subset ($k \in \{1, \ldots, K\}$) of the partition $\mathbf{P}$, where $\theta = 1.0$ is a "dummy" parameter used to evaluate the invariance of the SSL loss across the subsets (later discussed in Step 1). For example, for contrastive learning [31, 15], $\mathcal{L}$ is given by:

$$\mathcal{L}(\phi, \theta = 1.0, k, \mathbf{P}) = -\log \frac{\exp\left(\mathbf{x}^T \mathbf{x}^* \cdot \theta\right)}{\sum_{\mathbf{x}' \in \mathcal{X}^- \cup \mathbf{x}^*} \exp\left(\mathbf{x}^T \mathbf{x}' \cdot \theta\right)} \tag{1}$$

where $\mathbf{x} = \phi(I), \mathbf{x}^* = \phi(I^*)$, with I, I^* being an image and its augmented view, and $\mathcal{X}^{-1}$ is the feature set of the negative samples $\mathcal{I}^- = \{I_i | i \in \{1, \ldots, N\} \wedge P_{i,k} = 1\}$ with $I \notin \mathcal{I}^-$.

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. The set $\mathcal{P} = \{\mathbf{P}\}$.

Output. Disentangled feature extractor ϕ.

Step 1 [Update ϕ]. We update ϕ with the objective:

$$\min_{\phi} \sum_{\mathbf{P} \in \mathcal{P}} \sum_{k=1}^{K} \left[\mathcal{L}(\phi, \theta = 1.0, k, \mathbf{P}) + \lambda_1 \left\| \nabla_{\theta=1} \mathcal{L}(\phi, \theta = 1.0, k, \mathbf{P}) \right\|^2 \right], \tag{2}$$

where λ_1 is a hyper-parameter. In particular, the scalar $\|\nabla_\theta \mathcal{L}\|^2$ delineates how far the ϕ-induced similarity is from a fixed baseline $\theta = 1$. Therefore, this scalar regularizes ϕ to be invariant across the subsets in a partition, as the baseline is constant. See IRM [2] for more details. We use SGD for Eq. (2) using samples from the entire dataset for one epoch.

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

$$\mathbf{P}^* = \arg\max_{\mathbf{P}} \sum_{k=1}^{K} \left[\mathcal{L}(\phi, \theta = 1.0, k, \mathbf{P}) + \lambda_2 \left\| \nabla_{\theta=1} \mathcal{L}(\phi, \theta = 1.0, k, \mathbf{P}) \right\|^2 \right], \tag{3}$$

where λ_2 is a hyper-parameter. To practically optimize $\mathbf{P}$ by back-propagation, we use a continuous partition matrix in $\mathbb{R}^{N \times K}$ during optimization and then threshold it to $\{0, 1\}^{N \times K}$.

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

3.2 Stage-II: Downstream Knowledge Distillation

The self-supervised trained checkpoint from Stage-I is then used to initialize the teacher and student for fine-tuning on the whole dataset with labels. The distillation process can be seen as a regulation to prevent the student from overfitting the small train dataset and give the student a more diversed representation for classification. Specifically, the distillation loss can be formulated as follows:

$$q_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}, \quad \mathcal{L}_d = KL(p_i, q_i), \tag{4}$$

$$\mathcal{L} = \alpha \mathcal{L}_{ce} + \beta \mathcal{L}_d, \tag{5}$$

where z_i and p_i are the predicted logits of teacher and student model respectively, T is the temperature and $\mathcal{L}_d$ denotes the KL loss. The final loss function $\mathcal{L}$ of the student model is the addition of the cross entropy loss $\mathcal{L}_{ce}$ and the distillation loss. α and β are hyper-parameters.

4 Experiments

4.1 Dataset

Only the subset of the ImageNet dataset provided by the VIPrior challenge was used in our experiments. No external data or pre-trained checkpoint was used. The VIPrior challenge dataset contains 1,000 classes which are the same as the original ImageNet, and are split into train, val, and test splits, each of the splits has 50 images for each class, resulting in a total of 150,000 images. For comparison, we use the train split to train the model and test the model on the validation split. And for final submission, it is allowed to use both the train and validation split for training.

4.2 Implementation Details

Stage-I: IP-IRM

We built our IP-IRM on both SimCLR [15] and SimSiam [18]. Specifically, we used Adam as the optimizer with the learning rate set to 0.001. The temperature was set to 0.5 and the dimension of the latent vector is 128. All the models were trained for 800 epochs. We set $\lambda_1 = 0.2$ and $\lambda_2 = 0.5$. Partition $\mathbf{P}$ was updated every 50 epochs.

Model	Training Resolution	Accuracy (Stage-II)	Accuracy (TTA)
ResNest101	330	66.80	68.08
	448	65.00	65.63
ResNest269	300	62.00	63.37
ResNest200	320	66.40	67.94
	400	67.90	69.17
ResNet152	330	65.00	65.53
Wide_ResNet101	330	65.00	65.57
SEResNext101	330	66.40	67.80

Table 1: Test accuracy of various single models after knowledge distillation and TTA.

Number of models	Backbone	Policy	Accuracy (Ensemble)
2	Resnest101, SEResNext101	Mean	70.12
5	ResNest101, ResNest200, SEResNext101	T-sharpen	71.55
5	ResNest101, ResNest200, SEResNext101	Weighted	71.62

Table 2: Test accuracy of model ensemble.

Stage-II: Knowledge Distillation

Feature models trained in Stage-I with IP-IRM were used to initialize the weight of student and teacher model and the weight of teacher model was set frozen in the Stage-II training. Due to the small amount of training data, augmentation methods such as RandAugment [22] and AutoAugment [21] were applied. We also found label smoothing [53] and random erasing [77] helpful in boosting the performance as well.

We performed knowledge distillation on 8$\times$Tesla V100 GPU with parameters set as follows:

- Label smoothing: 0.1
- Resolution: 300-448
- Learning rate: $\frac{batch\ size}{160} \times 0.1$
- Warm-up epoch: 10
- Total epoch: 100

The models were evaluated by the top-1 accuracy on the test set, and the top-1 accuracy of various single models after knowledge distillation can be found in Table 1.

Other Tricks

Test time augmentation (TTA) and model ensemble were applied to further boost the model performance and final score during the inference on the test set.

HorizontalFlip, FiveCrop, or TenCrop was applied, with t-sharpening or geometric mean as the merging policy among the augmentation channels. With a proper policy and cropping size selected for TTA, the accuracy of every single model can be improved by around 1% as shown in Table 1.

Our final submission is the ensemble of 5 models with the ResNest101, ResNest200, and SEResNext101 backbone trained with various input sizes. The t-sharpening and weighted ensemble was used as our ensemble policy and the final result is shown in Table 2. With our assembled model, we achieved an accuracy of 71.62% on the test set.

4.3 Ablation Results

Effect of the proposed IP-IRM algorithm

In this subsection, we compare our proposed IP-IRM with vanilla SSL methods to evaluate its effectiveness. Specifically, after the stage-I SSL pretraining using ResNet50, we fixed the feature backbone and trained a linear classifier on the given ImageNet-50 dataset. As shown in Table 3, the model achieves better results in Stage-II finetuning with our IP-IRM algorithm. This indicates that

ResNet50	Stage-I Epoch	Finetune Epoch	Val Acc
Supervised Training	-	100	28.1
SimCLR	800	100	34.6
SimSiam	800	100	35.8
SimCLR+IP-IRM	800	100	36.1
SimSiam+IP-IRM	800	100	36.3

Table 3: Training and pretraining the model on the train split and evaluate the performance on the validation split.

Distillation method	Accuracy (Stage-II)	Remarks
OFD	61.31	Multiple feature layers
OFD	59.31	Last feature layer only
KL loss	60.00	Logits; Resolution 224
KL loss	58.70	Feature; Resolution 224
CRD	59.81	with KL loss
CRD	59.92	without KL loss

Table 4: Test accuracy of ResNet50 with various Stage-II knowledge distillation methods.

the feature learned by our IP-IRM contains more disentangled information and can generalize well on the downstream classification task.

Effect of the downstream knowledge distillation

Various knowledge distillation methods were tested in the Stage-II training, including overhaul-feature-distillation (OFD) [33] of multiple intermediate feature layers and last feature layer, contrastive representation distillation (CRD) [67] and the simple knowledge distillation with KL loss.

With the experiments on ResNet50 shown in Table 4, OFD was found the most effective distillation method. However, when it comes to more complex backbone networks, the performance heavily relies on the selection of feature layers. Without distillation of multiple intermediate feature layers and only distill the last feature layer with OFD works with various backbones, while simple distillation with KL loss led to better performance even with smaller input resolution. CRD did not bring impressive improvement with this dataset. Thus, for the complex backbone networks that are not compatible with OFD, we use KL loss instead for the Stage-II knowledge distillation.

5 Conclusion

In this competition, we applied the two-state strategy, where the first stage was an unsupervised disentangled representation learning method called Iterative Partition-based Invariant Risk Minimization (IP-IRM), based on Self-Supervised Learning (SSL), and the second stage was the supervised knowledge distillation.

With the provided small dataset, IP-IRM iteratively partitions the dataset into semantic-related subsets and learns a representation invariant across the subsets using SSL with an IRM loss. This effectively achieved feature disentanglement in Stage-I. Supervised knowledge distillation in Stage-II utilized the learned feature in Stage-I. By adding the distillation loss as the regulation, Stage-II fine-tuned the feature and trained a classifier without letting the model overfitting to the training set.

In addition to the two-stage strategy, with a proper data augmentation policy, inference tricks, and model ensemble, our approach was proved to be able to train the model in a data-efficient manner. The final submission to Visual Inductive Priors for Data-Efficient Computer Vision (VIPriors) challenge image classification track achieved an accuracy of 71.62% on the test set.

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] Lorenzo Brigato, Björn Barz, Luca Iocchi, and Joachim Denzler. Tune it or don't use it: Benchmarking data-efficient image classification, 2021.

[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] Robert-Jan Bruintjes, Attila Lengyel, Marcos Baptista Rios, Osman Semih Kayhan, and Jan van Gemert. Vipriors 1: Visual inductive priors for data-efficient deep learning challenges, 2021.

[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] 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.

[15] 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.

[16] 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.

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

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

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

[20] 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.

[21] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation policies from data, 2019.

[22] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space, 2019.

[23] 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.

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

[25] 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.

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

[27] 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.

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

[29] 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.

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

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

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

[33] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation, 2019.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

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

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

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

[43] Attila Lengyel, Sourav Garg, Michael Milford, and Jan C. van Gemert. Zero-shot domain adaptation with a physics prior, 2021.

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

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

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

[52] 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.

[53] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017.

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

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

[56] Carolina Redondo-Cabrera, Marcos Baptista-Ríos, and Roberto J. López-Sastre. Learning to exploit the prior network knowledge for weakly supervised semantic segmentation. *IEEE Transactions on Image Processing*, 28(7):3649–3661, 2019.

[57] 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.

[58] 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.

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

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

[61] 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.

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

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

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

[65] 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.

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

[67] 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.

[68] 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.

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

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

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

[76] Bingchen Zhao and Xin Wen. Distilling visual priors from self-supervised learning. In *European Conference on Computer Vision*, pages 422–429. Springer, 2020.

[77] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation, 2017.

[78] 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.

[79] 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.