Distortion-Disentangled Contrastive Learning

Jinfeng Wang[a,b*], Sifan Song[a,b*], Jionglong Su[a†], S. Kevin Zhou[b,c†]
[a] School of AIAC, Xi'an Jiaotong-liverpool University, Suzhou, China
[b] School of BME & Suzhou Institute for Advanced Research,
Center for Medical Imaging, Robotics, Analytic Computing & Learning (MIRACLE),
University of Science and Technology of China, Suzhou, China
[c] Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, China
Jionglong.Su@xjtlu.edu.cn
skevinzhou@ustc.edu.cn

Abstract

Self-supervised learning is well known for its remarkable performance in representation learning and various downstream computer vision tasks. Recently, Positive-pair-Only Contrastive Learning (POCL) has achieved reliable performance without the need to construct positive-negative training sets. It reduces memory requirements by lessening the dependency on the batch size. The POCL method typically uses a single objective function to extract the distortion invariant representation (DIR) which describes the proximity of positive-pair representations affected by different distortions. This objective function implicitly enables the model to filter out or ignore the distortion variant representation (DVR) affected by different distortions. However, some recent studies have shown that proper use of DVR in contrastive can optimize the performance of models in some downstream domain-specific tasks. In addition, these POCL methods have been observed to be sensitive to augmentation strategies. To address these limitations, we propose a novel POCL framework named Distortion-Disentangled Contrastive Learning (DDCL) and a Distortion-Disentangled Loss (DDL). Our approach is the first to explicitly and adaptively disentangle and exploit the DVR inside the model and feature stream to improve the representation utilization efficiency, robustness and representation ability. Experiments demonstrate our framework's superiority to Barlow Twins and Simsiam in terms of convergence, representation quality (including transferability and generalization), and robustness on several datasets.

[1]These authors contributed equally to this work.
[2]Corresponding authors

1. Introduction

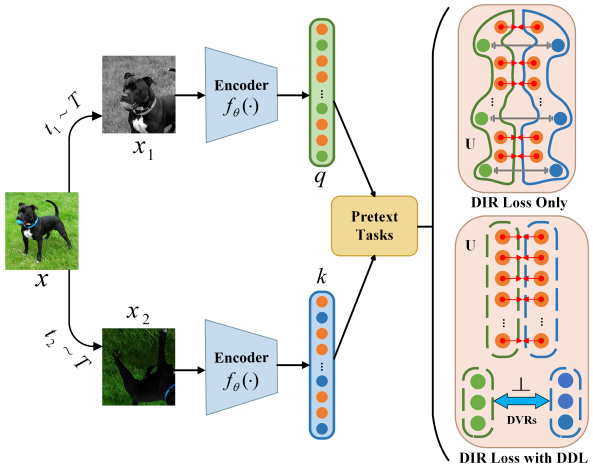

Figure 1. **U** denotes a two-dimensional feature space constructed using *query* (q) and *key* (k) based on pretext tasks. The current POCL methods attempt to minimize the distance between the DIRs of positive sample pairs (represented by orange dots) in **U**. The proposed **DDL** further ensures that the DVRs for each positive-pair (represented by green and blue dots) are orthogonal within the selected dimension of the overall representation, making them uncorrelated to each other.

High-quality representation learning has been a core topic in deep learning research, which is challenging for computer vision due to the low information density [24, 25]. In recent years, label-free un/self-supervised contrastive learning methods with instance discrimination as a pretext task undergo steady development, rapidly clos-

ing the performance gap with supervised learning methods and demonstrating reliable generalization capabilities [8, 12, 22, 25, 59].

Some previous studies have used different augmented views of the same instance as positive samples with other instances in the mini-batch as negative samples. They employ a series of *query-key* operations to classify samples in the mini-batch and train models to extract the discriminative representations. However, such methods require a large set of positive and negative samples to provide sufficient negative information, such as using large batch sizes [7–9], memory banks [20, 54], and memory queues [11, 13, 25, 47].

Recent POCL methods have achieved superior performance without the use of negative samples [12, 22, 59]. These methods aim to minimize the distance between positive sample pairs in the feature space and extract the distortion invariant representation (DIR) by explicitly supervising the Euclidean distance [22], cross-correlation [59], or vector angle [12] of the positive sample pairs. The implicit objective of the POCL methods is to filter out or ignore the distortion variant representation (DVR) using the aforementioned supervision to extract the DIR of the positive sample pairs. However, some recent studies [16, 18, 55] have shown that the appropriate use of complex distortion strategy and DVR in contrastive learning can improve the performance of models in downstream domain-specific tasks. In addition, the performance of these POCL methods is sensitive to augmentations. As such, augmentation strategies need to be carefully selected for generating positive sample pairs [22]. We do observe that these methods have unstable performance on the same test set with different augmentations. Motivated by these, we argue that filtering out or ignoring the DVR using a single loss (*i.e.*, DIR loss) that only supervises the DIR is straightforward but insufficient for representation supervision and utilization in the POCL method. Such inadequately supervised representation may result in the DIR and DVR being entangled in the learned representation, leading to reduced performance.

Therefore, we propose a novel POCL framework, named Distortion-Disentangled Contrastive Learning (DDCL). Unlike previous studies, DDCL does not use augmented information, multi-head structures, or augmentation-specific predictions. This fully adaptive training method makes DDCL applicable to more complex augmentation strategies. In DDCL, we group the last layer (*i.e.*, overall representation) of the encoder into two parts to extract the DIR and DVR. The first part is utilized to extract the DIR using the DIR loss of the corresponding original POCL method. The remaining part is utilized to extract the DVR using our novel loss. We propose a novel Distortion-Disentangled Loss (DDL) to independently supervise the DVR by making the DVRs of a positive sample pair orthogonal. As shown in Fig. 1, DDL explicitly extracts and disentangles the DVR from the overall representation. It is worth noting that the disentangled DVR is not noise or useless information but contains valuable features and distortion information. We further analyze this in Sec. 4.6. By concatenating the DIR and DVR in the following inference task, the model's convergence, representation ability, and robustness are further improved without additional parameters and computation.

We conduct experiments on two existing POCL methods, *i.e.*, Barlow Twins [59] and Simsiam [12]. They both use a two-branch architecture and their losses are only designed to extract DIR. Experiments demonstrate that our DDCL and DDL are adaptable to different POCL architectures and corresponding DIR losses, achieving improved performance compared to the original POCL methods. Our main contributions are as follows:

- We propose a novel POCL framework, named DDCL, which explicitly disentangles the representation into DIR and DVR. DDCL can adaptively extract the DVR without any other manual designs, which makes DDCL applicable and flexible to more complex augmentation strategies. We can utilize the DVR to enhance the model's performance and robustness.
- In DDCL, we propose a novel objective function, DDL, which explicitly supervises and extracts the DVR for its efficient use in the downstream task.
- Our proposed DDCL can be adapted to current popular POCL methods. It improves the convergence, representation ability, and robustness without additional inference parameters or computation.

2. Related Work

2.1. Contrastive Learning

As the core strategy of self-supervised learning, contrastive learning undergoes rapid development in recent years due to its simplicity and efficiency. The main idea of initial contrastive learning [8, 25, 54, 56] can be summarized as follows: constructing a set of **positive and negative samples**; using instance discrimination as a **pretext task**; and utilizing NCE Loss [23] or its variants [38, 57] as a **loss function**. Training an encoder under these settings aims to minimize the distance between positive samples in the feature space while pushing them apart from negative samples. These methods seek sufficient negative samples and appropriate data augmentation strategies for positive samples. Large batch sizes [7–9], memory banks [20, 54], memory queues [11, 13, 25, 47], and clustering structures [7, 20, 32, 47] are utilized to provide sufficient negative sample information. There have been a number of efforts [14, 40, 42] to alleviate this negative sample starvation problem in different ways. Recently, some POCL

methods that do not use negative samples have been proposed [12, 22, 59]. It is worth noting that [2] cleverly uses other instance information in the mini-batch. These methods do not rely on the constraints of negative samples, and use designed architecture [5], pretext tasks, loss functions and robustness tricks to prevent model collapse.

2.2. Distortion Variant Contrastive Learning

Recent studies [16, 55] have shown that the inductive bias of extracting the DIR of an image is not always optimal. The performance of the model on downstream tasks is jointly determined by distortion sensitivity and domain-specific tasks. In order to extract representations that are beneficial for domain-specific tasks in contrastive learning, [55] designs multiple distortion-dependent prediction heads to obtain multiple distortion-varying subspaces. [16] suggests distortion prediction for a particular operation of distortion to make the model sensitive to that distortion. [18] uses the information of the distortion operation to make the model have the specific distortion equivariance in the image space and the feature space. These methods require the manual design of objective functions or model structures for specific distortion operations. This means that these frameworks are pre-trained case-by-case. In contrast, our proposed DDCL can be directly used to adapt a variety of complex distortions without any case-specific modification. Experiments show that DDCL adapts well to affine transformations and elastic transformations.

2.3. Disentangled Representation

This topic has become a long-desired goal in the deep learning community subsequently [36, 43, 46, 48, 51]. The purpose of disentangling is to explicitly decompose the factors of variation from the target representation in a high-dimensional feature space [3, 35]. The disentangled representation has several advantages, including better interpretability [3, 30], utilization efficiency [35], robustness [3, 52, 60], and generalization capacity [6, 27, 52]. Several works analyze and explicitly use disentangled representations to improve model performance from the perspective of information theory [10, 19] and group theory [51]. Disentangling/Decoupling is also widely used in research fields such as image editing and generation [30], transfer learning [21], and fairness [34, 41]. However, research on disentangled representations in contrastive learning-based self-supervised learning is still in its infancy.

2.4. Orthogonality

Orthogonality is usually used in the kernel of deep neural networks to learn more diverse weight matrices and feature vectors [31, 33, 49, 50]. Several works apply orthogonality in disentangled representation learning [52], model initialization and training [26], and supervised learning with contrastive properties [39]. To the best of our knowledge, our proposed DDCL and DDL are the first to exploit the orthogonality of representations in a contrastive learning task to disentangle distortion information. Unlike previous works on orthogonality, our method does not require additional computations such as singular value decomposition [29, 44] and iteration [1]. The DDL extracts DVR by simply supervising the orthogonality between partial representations.

3. Proposed Methods

As mentioned in [16], the model's sensitivity to certain distortions can effectively improve the feature quality, our purpose of extracting DVR is not to exclude a certain distortion. We hope our model can sensitively capture the heterogeneous features brought by distortions. In contrastive learning with positive and negative samples, many or specific hard negative samples are essential to provide sufficient distortion information and hard samples. In POCL method, only the augmented positive sample pairs provide distortion information for each contrast operation. We find that the performances of existing POCL methods, such as BYOL [22], Barlow Twins [59], and Simsiam [12], are sensitive to augmentation strategies during training, and become unstable when making inference on unseen distorted inputs. The reason for this may be that POCL methods do not utilize the rare but valuable distortion information.

3.1. Revisiting Positive-Only Contrastive Learning

Existing POCL methods generally comprise of four main factors, *i.e.*, the model architecture, the pretext task, the loss function, and the robustness trick. These POCL methods can be categorized into either symmetric or asymmetric architecture depending on their designs. Barlow Twins (BT) [22] employs two entirely symmetric and parameter-sharing encoders. The cross-correlation matrix of the two branches' outputs is used as the pretext task. Furthermore, the distance between the cross-correlation and identity matrix is utilized as the loss function to extract the DIR and eliminate redundancy. Notably, BT uses only batch normalization for robust training. The symmetry loss of BT is as follows:

$$C = Norm(z)^T \cdot Norm(z') \tag{1}$$

$$Norm(z) = (z - \mu)/\sigma \tag{2}$$

$$L_{BT} = \sum_i (1 - C_{ii})^2 + \lambda \sum_i \sum_{j \neq i} C_{ij}^2 \tag{3}$$

where C refers to the cross-correlation matrix, z and z' refer to the projection of two branches of BT model as shown in Fig. 2. In addition, $C \in \mathbb{R}^{D \times D}$, z,$z' \in \mathbb{R}^{B \times D}$, and $\mu, \sigma \in \mathbb{R}^{1 \times D}$.

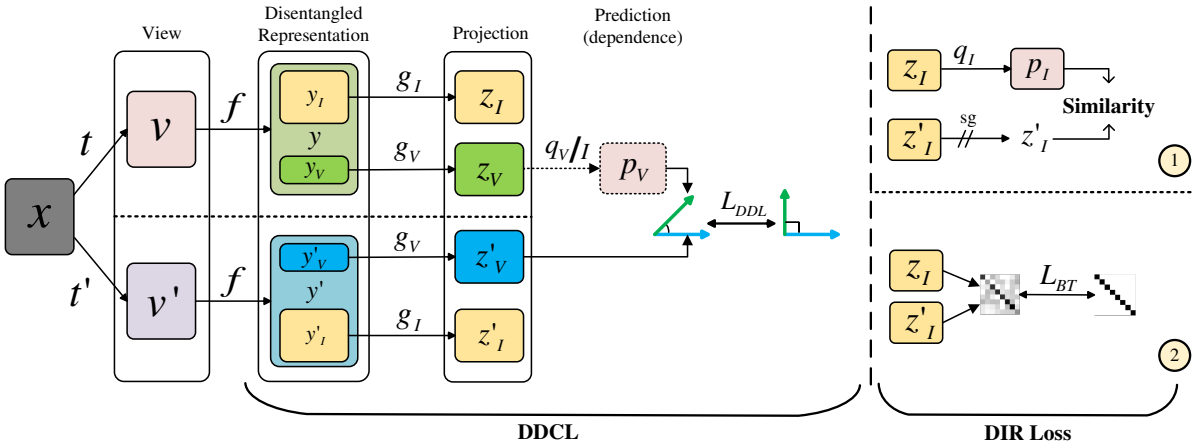

Figure 2. The DDCL framework for symmetric and asymmetric architectures. The encoder denoted as f, projector as g, and predictor as q; the subscripts I and V denote a process or a feature related to DIR and DVR, respectively; p_V and p_I only coexist in the asymmetric version. Two designs of DIR loss correspond to Simsiam [12] and Barlow Twins [59]. We explicitly group and disentangle the overall representation into DIR (y_I and y'_I) and DVR (y_V and y'_V) after the encoder, then supervise the mapped features of DIR and DVR using the DIR loss and DDL, respectively.

BYOL [22] and Simsiam [12] adopt an online-target asymmetric architecture that use regression prediction as the pretext task. BYOL uses the Euclidean distance between the two branches' outputs as the loss function and uses stop-gradient and momentum update strategies to prevent model collapse. Simsiam (Sims) uses the cosine similarity between the two branches' outputs as the loss function and only uses stop-gradient to avoid trivial solutions. From the perspective of model stability and robustness, Simsiam can be regarded as an optimized version based on BYOL. Therefore, we mainly discuss Simsiam in terms of the asymmetric design. The asymmetric loss function is as follows:

$$L_{Sims} = \frac{1}{2}S(p, sg(z')) + \frac{1}{2}S(p', sg(z)), \tag{4}$$

$$\mathcal{S}(p, z') = -\frac{p \cdot z'}{\| p \|_2 \| z' \|_2}, \tag{5}$$

where p refers to prediction, z refers to projection given in Fig. 2-1, $\| \cdot \|_2$ is an L2-norm, and $sg(\cdot)$ is the stop-grad.

Notably, the primary goal of these POCL methods is to train an encoder that robustly extracts the DIR by narrowing the proximity between different distortion views of the same instance in a high-dimensional feature space, while attempting to ignore or remove the DVR. The loss function of Barlow Twins is designed to explicitly remove redundant information and only retain the DIR, while that of Simsiam is to implicitly ignore the DVR.

We find that, in the original POCL methods that lack distortion information and hard samples, ignoring or eliminating the DVR, either implicitly or explicitly, decreases the overall representation utilization. Moreover, implementing only a single DIR loss to supervise high-dimensional projected representations may be inadequate. Moreover, the models trained using these methods are sensitive to the augmentation strategy during training, making more difficult to infer unseen distorted instances, which leads to reduced robustness. Therefore, we propose the DDCL and DDL to explicitly supervise high-dimensional representations in order to disentangle the DIR and DVR, resulting in the sufficient utilization of concatenated overall representation.

3.2. Distortion-Disentangled Contrastive Learning

This paper proposes a novel POCL framework, named Distortion-Disentangled Contrastive Learning (DDCL). When training the model, we group the output of the last layer of each encoder. In this case, the overall representation is grouped into two parts for the DIR and DVR. In addition, the following mapping processes for these two parts are synchronized and supervised by the DIR loss and DDL, respectively. Formulated instructions are as follows:

$$y_I = M_{n,I} \cdot f_{1:n-1}(v) \tag{6}$$

$$y_V = M_{n,V} \cdot f_{1:n-1}(v) \tag{7}$$

$$y = cat(y_I, y_V) \tag{8}$$

$$M_{n,I} \in \mathbb{R}^{DR \cdot d \times H \cdot W}, M_{n,V} \in \mathbb{R}^{(1-DR) \cdot d \times H \cdot W}$$

where $f_{1:n-1}(\cdot)$ refers to the function of the encoder except the last layer, and M_n refers to the matrix of last layer of the encoder. y refers to the overall representation, and $cat(\cdot)$ is concatenation. d is the output dimension of $f_{1:n-1}$, and DR is the disentangling ratio used to group the overall representation into DIR and DVR parts for separate supervision.

Since the overall representation (y) has been disentangled, the DIR can be used independently for downstream inference tasks. We find that the performance when using only DIR (*i.e.*, y_I) is similar to that of the corresponding original POCL method with the full representation. In addition, since the overall representation can be considered as a concatenation of the DIR and DVR, this overall representation achieves even better performance in the subsequent linear evaluation. Furthermore, the overall representation of DDCL is more robust to unseen distorted data. Details of performance are given in Sec. 4.

3.3. Distortion-Disentangled Loss

To supervise the DVR, we propose a novel loss function, named Distortion-Disentangled Loss (DDL). As shown in Fig. 2, the purpose of DDL is to supervise the orthogonality of projected representation vectors, which extracts the DVR from the same instance under different augmentation views. We use the DDL for both symmetric and asymmetric architectures. The formula of DDL in **symmetric** architecture is as follows:

$$\mathcal{D}(z_V, z'_V) \triangleq \mid \frac{z_V \cdot z'_V}{\parallel z_V \parallel_2 \parallel z'_V \parallel_2} - \xi \mid \tag{9}$$

$$L_{DDL}^{Sym} \triangleq \mathcal{D}(z_V, z'_V) \tag{10}$$

where the hyperparameter ξ is set to 0 in our default setting. The value z_V refers to the projection of DVR (*i.e.*, y_V) in **symmetric** architecture as given in Fig. 2. Therefore, the overall loss in **symmetric** architecture can be written as:

$$L^{Sym} = \gamma L_{DDL}^{Sym} + L_{BT}^{I} \tag{11}$$

where L_{BT}^{I} is the DIR loss of L_{BT} given in Eq. (3). The hyperparameter γ is set to 1 in our setting. L_{BT}^{I} is defined as:

$$L_{BT}^{I} \triangleq \sum_i (1 - C_{ii}^{I})^2 + \lambda \sum_i \sum_{j \neq i} {C_{ij}^{I}}^2 \tag{12}$$

where

$$C^I \triangleq Norm(z_I)^T \cdot Norm(z'_I) \tag{13}$$

Furthermore, the formulas of DDL in **asymmetric** architecture is given as follows:

$$L_{DDL}^{Asy} \triangleq \frac{1}{2}\mathcal{D}(p_V, sg(z'_V)) + \frac{1}{2}\mathcal{D}(p'_V, sg(z_V)) \tag{14}$$

$$L^{Asy} = \gamma L_{DDL}^{Asym} + L_{Sims}^{I} \tag{15}$$

Referring to the definition in Eq. (12), L_{Sims}^{I} is defined as follows:

$$L_{Sims}^{I} \triangleq \frac{1}{2}S(p_I, sg(z'_I)) + \frac{1}{2}S(p'_I, sg(z_I)) \tag{16}$$

As shown by these equations, DDL can be applied to both symmetric and asymmetric architectures. Our design simply groups the overall representation of the last layer of each encoder into two parts. The original loss of the corresponding POCL method is utilized to supervise the DIR part, and DDL is used to supervise the DVR part. Therefore, DDL can be considered to be a plug-in without adding extra dimensions to the overall representation.

4. Experiments

Based on the Simsiam [12] and Barlow Twins [59] methods, we conduct experiments on the convergence, representation quality, and robustness using different scale datasets. These two methods represent the design of asymmetric and symmetric architecture, respectively. In addition, we encounter unstable performance when reproducing BYOL [22]. Since Simsiam is more robust and designed for general purposes based on BYOL, our experiments only report the performance of Simsiam. Notably, VICReg [2] uses other instances in the mini-batch to do contrastive. We cannot ensure it is a POCL method, so we do not explore its performance in this experiment. All reported results in this paper are from our reproductions. Since the overall performance of Simsiam is better than that of Barlow Twins, we only use Simsiam as the baseline in downstream tasks, robustness experiments, and ablation studies. Some related previous work are based on the CL with positive-negative pairs [55], and some of them have not been verified by downstream tasks [16, 18]. No other studies utilize more complex distortions, such as elastic transformation. We believe that the fairest comparison is to compare with baseline methods under various identical distortion settings.

4.1. Implementation Details

Image augmentations. This paper reports three augmentation strategies: basic augmentation (BAug), complex augmentation (CAug) and CAug with elastic transformation (CAug$^+$). For the BAug strategy, we refer to the parameters and transformations in Simsiam [12]: random resized cropping, horizontal flipping, color jittering, converting to grayscale, and Gaussian blurring. Similar to [12], Gaussian blurring is only used to augment the ImageNet datasets (IN-100 and IN-1k) [17]. In CAug, we add random rotations from -90 to 90 degrees to test the impact of a more complex augmentation strategy on the DDCL and original POCL methods.

Method	CIFAR-10		CIFAR-100		STL-10	
	Top 1	Top 3	Top 1	Top 3	Top 1	Top 3
100 Epochs						
Barlow Twins	82.89	96.49	54.98	74.37	86.90	97.31
DDCL_Sym_DIR	82.78	96.72	55.06	74.58	86.78	97.21
DDCL_Sym	**83.64**	**96.81**	**56.03**	**75.31**	**87.10**	**97.39**
SimSiam	75.34	93.68	36.87	56.90	86.38	97.46
DDCL_Asy_DIR	76.25	94.36	39.77	60.12	86.32	97.44
DDCL_Asy	**76.85**	**94.58**	**40.52**	**61.01**	**86.44**	**97.49**
200 Epochs						
Barlow Twins	86.17	**97.46**	59.60	78.34	88.53	97.56
DDCL_Sym_DIR	86.33	97.17	58.43	77.71	88.63	97.66
DDCL_Sym	**86.72**	97.44	**59.61**	**78.41**	**89.02**	**97.76**
SimSiam	86.20	97.65	56.35	77.13	89.80	98.17
DDCL_Asy_DIR	87.82	97.91	56.23	77.05	89.34	98.19
DDCL_Asy	**88.18**	**98.05**	**57.04**	**78.07**	**89.84**	**98.27**

Table 1. Linear evaluation results of two POCL and the corresponding DDCL architectures pre-trained with 100 and 200 epochs. The best and the second best performance are in bold and underlined, respectively. DIR with a smaller dimension achieves comparable performance compared to the vanilla BT and Sims. Using the overall representation for linear evaluation according to Eq. (8) achieves the best convergence performance.

Architecture. In small and medium-scale datasets such as CIFAR-10, CIFAR-100 [28], and STL-10 [15], we use the Lightly version [45] of ResNet-18 as the backbone. The output dimension of the backbone is 512 in CIFAR experiments and 4608 in STL-10 experiments. In experiments on IN-100 and IN-1k [17], we use ResNet-50 as the backbone, and the output dimension of the backbone is 2048. All output dimensions of the following mapping processes, including projection and prediction, are 2048. For fair comparison, the original POCL methods and their corresponding DDCL use the same hyperparameters.

Optimization. Referring to Simsiam [12], when pre-training the model, we use SGD with base lr = 0.03 on CIFAR and STL-10 with batch size (bs) = 512, base lr = 0.05 on IN-100 (bs = 512) and IN-1k (bs = 256), weight decay = 0.0001, momentum = 0.9, and a cosine decay schedule. For linear evaluation on CIFAR and STL-10, we use an SGD optimizer with 100 epochs, lr = 30.0, weight decay = 0, momentum = 0.9, and batch size = 256. For linear evaluation on IN-100, we use an SGD optimizer with 200 epochs, base lr = 30.0, and batch size = 256. For linear evaluation on IN-1k, we employ a LARS optimizer [58] with 90 epochs, base lr = 0.1, and batch size = 4096 (similar to Simsiam [12]).

4.2. Convergence Study

We evaluate the performance of the DDCL on CIFAR and STL datasets with small-epoch pre-training (100, 200 epochs). As shown in Tab. 1, the linear evaluation using only DIR (y_I in Eq. (6)) achieves approximately on-par convergence performance in smaller dimensions compared to vanilla Barlow Twins and Simsiam. The linear evaluation after concatenating DIR and DVR (y_V in Eq. (7)) achieves the best convergence performance (*i.e.*, DDCL_Sym and DDCL_Asy in Tab. 1). This suggests that disentanglement and efficient use of the DVR improve the convergence performance of the model.

Method		CIFAR-10	CIFAR-100	STL-10
Barlow Twins	Acc.	87.82	59.66	90.68
	KNN	84.78	51.63	**83.61**
DDCL_Sym_DIR	Acc.	88.56	59.83	90.65
	KNN	/	/	/
DDCL_Sym	Acc.	**88.70**	**60.95**	**90.83**
	KNN	**84.91**	**51.97**	82.77
Simsiam	Acc.	91.56	66.29	91.02
	KNN	87.46	52.38	83.82
DDCL_Asy_DIR	Acc.	92.01	65.66	91.28
	KNN	/	/	/
DDCL_Asy	Acc.	**92.19**	**66.49**	**91.39**
	KNN	**87.90**	**52.41**	**84.05**

Table 2. Linear evaluation and KNN results of two POCL and the corresponding DDCL architectures pre-trained with 800 epochs. The best and the second best performance are in bold and underlined, respectively. Since the dimensionality of the DIR part is lower than the representation of vanilla methods, we do not report and compare the KNN performance of DDCL_Sym/Asy_DIR.

Dataset	Train/Epoch	Method	CUB-200	Flowers-102	Food-101
IN-100	BAug/500	Simsiam	30.53	76.73	62.78
		DIR_only	30.29	76.13	61.91
		DDCL_Asy	30.53	77.51	63.02
	CAug/500	Simsiam	34.79	78.57	65.37
		DIR_only	34.88	78.18	64.79
		DDCL_Asy	35.11	78.92	65.67
	CAug$^+$/500	Simsiam	34.05	77.79	64.63
		DIR_only	34.93	78.00	64.35
		DDCL_Asy	**35.50**	**79.88**	**65.74**
IN-1k	CAug/100	Simsiam	39.35	79.57	71.41
		DIR_only	40.39	80.40	70.65
		DDCL_Asy	40.73	81.98	71.79
	CAug/200	Simsiam	40.44	80.35	71.93
		DIR_only	40.30	80.22	71.87
		DDCL_Asy	**41.34**	**82.05**	**72.91**

Table 3. Downstream task performance on domain-specific datasets. Complex distortions and the adaptive DVR extracting method in DDCL significantly improve the transferability and generalization capability.

4.3. Representation Evaluation

To assess the representation quality extracted by well-trained DDCL models, we evaluate the **linear evaluation accuracy and KNN performance** of DDCL (DDCL_Sym and DDCL_Asy) and the corresponding vanilla POCL methods (Barlow Twins and Simsiam) after 800 epochs of pre-training in CIFAR and STL datasets. Tab. 2 shows that the classification performance of DIR parts extracted by DDCL (*i.e.*, DDCL_Sym_DIR and DDCL_Asy_DIR) is still on-par with the vanilla methods after sufficient training. The overall representation extracted by DDCL achieves al-

Dataset	Train/Epoch	Method	CJ	CJ+Flip	CJ+90°	CJ+90°+ET
IN-100	BAug/500	Simsiam	81.31	81.40	50.18	27.34
		DIR_only	81.24	81.33	49.52	26.55
		DDCL_Asy	**81.60**	**81.64**	50.02	26.76
	CAug/500	Simsiam	78.99	78.95	76.95	51.88
		DIR_only	79.11	78.97	77.29	49.00
		DDCL_Asy	**79.33**	**79.40**	**77.32**	48.49
	CAug$^+$/500	Simsiam	77.67	77.69	75.11	74.06
		DIR_only	77.64	77.65	**75.44**	74.09
		DDCL_Asy	**78.19**	**78.20**	75.37	**74.27**
IN-1k	CAug/100	Simsiam	65.59	65.51	62.42	28.57
		DIR_only	65.30	65.34	62.36	28.74
		DDCL_Asy	**66.07**	**66.11**	**62.48**	29.47
	CAug/200	Simsiam	67.44	67.57	64.17	31.16
		DIR_only	66.88	66.86	**64.35**	30.34
		DDCL_Asy	**67.69**	**67.79**	64.34	30.83

Table 4. Robustness evaluation of models pre-trained by different data augmentation strategies on ImageNet datasets. CJ is random color jetter, 90° denotes randomly applying -90° to 90° rotation, and ET is random elastic transformations ($\alpha = 100$, $\sigma = 5$).

most optimal performance in both classification and KNN.

Transferability and generalization capability are essential properties for evaluating the representation quality of models. We pre-train Simsiam and DDCL_Asy on IN-100 (500 epochs) and IN-1k (100/200 epochs) and then evaluate the transferability and generalization capability in downstream tasks (linear probe) on domain-specific datasets whose distributions are far from ImageNet, such as CUB-200 [53], Flowers-102 [37], and Food-101 [4]. As shown in Tab. 3, DDCL has demonstrated significant improvements on these downstream datasets by utilizing different training distortion strategies and dataset complexities, compared to the original Simsiam with BAug.

4.4. Robustness

Since unseen distortions during pre-training decrease the linear evaluation performance, a common approach to improve model robustness is to apply complex augmentations. We perform a series of experiments to evaluate the robustness of the proposed DDCL. The pre-trained models of Simsiam and DDCL utilized in Tab. 4 are exactly the same ones used in Tab. 3.

As shown in Tab. 4, when models are pre-trained using only the commonly used augmentation strategy (BAug), the models perform poorly in dealing with rotation distortions (*i.e.*, unseen distortions). After applying a random -90° to 90° rotation to BAug (*i.e.*, CAug), the performance of original POCL (Simsiam) on rotation distortion improves, whereas that on basic distortions decreases by large margins. This performance reduction is justified by the fact that the training process is more difficult (from BAug to CAug) while the training epochs remain unchanged. In contrast, the proposed DDCL has better compatibility with various augmentation strategies. As given in Tab. 4, when CAug and CAug$^+$ strategies are used for pre-training, DDCL improves the robustness evaluation performance on rotation and elastic distortions while alleviating performance reduction in basic distortions.

DR	CIFAR-10		CIFAR-100		STL-10	
	DDCL_Asy	DIR_only	DDCL_Asy	DIR_only	DDCL_Asy	DIR_only
0.2	91.96	91.15	66.26	62.87	91.08	90.17
0.4	92.11	91.58	66.01	63.56	91.23	90.63
0.6	91.72	91.43	**66.60**	65.45	**91.47**	91.22
0.8	**92.19**	**92.01**	66.49	**65.66**	91.39	**91.28**

Table 5. Linear evaluation performance with different disentangling ratios (DR).

Method\Batch Size	32	64	128	256	512
Simsiam	91.66	91.44	91.25	90.85	91.56
DIR_only	91.60	91.51	90.96	91.41	92.01
DDCL_Asy	**91.75**	**91.66**	**91.26**	**91.64**	**92.19**

Table 6. Linear evaluation on CIFAR-10 with different batch sizes.

The attention map in Fig. 3 also visually supports this argument and our design goals. The DIR focuses on the region correlated to the target object. It also achieves similar attention performance using a smaller feature dimension (2048$\times$DR) than Simsiam (2048). The DVR focuses on the region that complements the DIR and can further contribute to linear evaluation. In addition, Fig. 3 demonstrates the robustness of our DDCL, as the attention maps of DIR and DVR disentangled by DDCL are highly consistent with different distortions.

4.5. Ablation Study

Disentangling Ratio. The disentangling ratio (DR) mentioned in Eq. (6) and Eq. (7) describes the ratio of DIR and DVR in the overall representation. As shown in Tab. 5, the performance of DIR_only improves as the ratio of DIR increases. Since DR = 0.8 achieves the general best performance for DDCL, we set it as the default value.

Batch Size. As shown in Tab. 6, DDCL achieves stability and optimal performance at multiple batch sizes.

4.6. Brick Study

We design this novel brick study to further explore the impact of DVR on the overall representation. As the DIR and DVR parts have been grouped and disentangled from the overall representation, they can be utilized independently or concatenated to others (as bricks). Therefore, in Tab. 7, we concatenate the DIR (row) with the DVR (column) disentangled from various instances and distortions, and evaluate the linear classification performance of this new overall representation. This part of the research is carried out on the STL-10 dataset.

Based on the performance difference observed between two augmentation strategies when facing rotation (columns

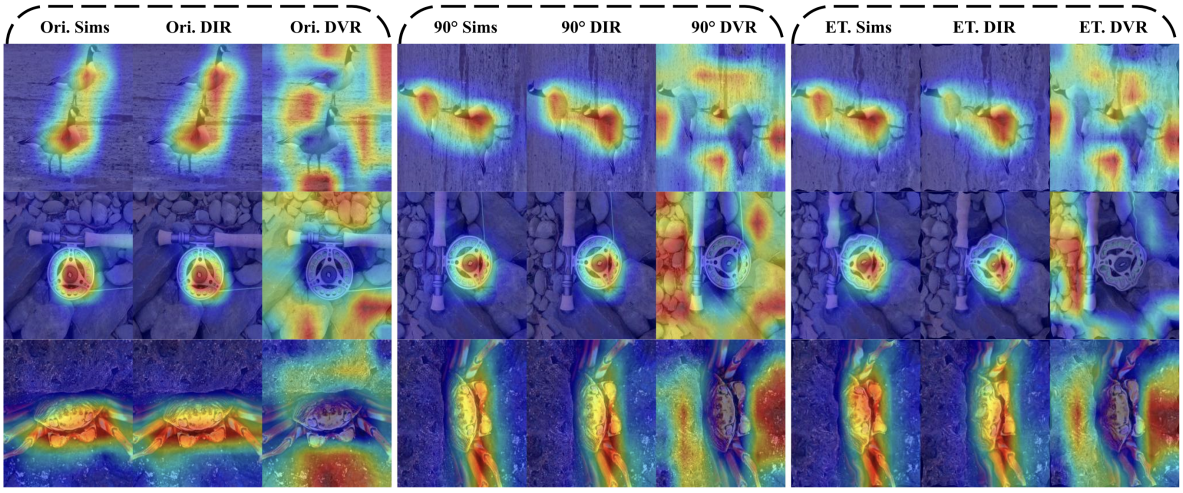

Figure 3. Attention maps of Simsiam and DIR and DVR of DDCL with different distortions on IN-100 (pre-trained using CAug^{+}).

DIR\DVR	Orig.	Flip	Flip+90°	Dif.Inst+Flip+90°	Zero DVR
Trained by BAug					
Orig.	**91.39**	91.31	90.76	90.34	89.78
Flip	91.44	**91.32**	90.76	90.32	89.95
Flip+90°	62.58	62.69	**49.85**	46.44	48.46
Trained by CAug					
Orig.	**89.98**	89.97	88.89	88.89	88.41
Flip	89.69	**89.69**	88.90	88.56	88.42
Flip+90°	87.20	87.32	**84.92**	84.66	84.58

Table 7. 'Dif.Inst' means that the DVR and DIR come from different instances, and 'Zero DVR' means to set the DVR part to zero values. The bold numbers represent the DIR and DVR originating from the same pre-training model (*i.e.*, without any altered bricks).

2 and 3 in Tab. 7), we argue that DVR can extract content features when encountering the unseen distortion that the model cannot handle (third row). When the model is trained with the rotation distortion (*i.e.*, CAug), DVR can mainly extract the distortion-related features as designed. However, the current DVR cannot clearly decompose the content and distortion-related features.

In the columns 'Flip+90°' and 'Dif.Inst+ Flip+90°' (Tab. 7), the linear evaluation performance of concatenated DIR and DVR is generally comparable when the DVR is generated by the same kind of distortion, regardless of whether the DVR and DIR come from the same instance. This is due to the fact that the DVR contains certain information representing the distortion itself.

As shown in the third row of Tab. 7, since the model is trained with the BAug strategy, rotation is an unseen distortion that cannot be handled. The DVR may extract content features, so the zero DVR at this time does not significantly impact the performance (third row, fifth column), which is consistent with our purpose. When using the DVR of other instances to classify this instance (third row, fourth column), the classification performance further deteriorates. This performance reduction may be caused by the influence of content features from other instances.

5. Conclusion and Discussion

This paper proposes a novel POCL framework, DDCL, and a novel objective function, DDL, to adaptively extract the DVR part from the overall representation. We apply the DDCL to both symmetric and asymmetric POCL architectures to improve model convergence, representation quality, and robustness by explicitly supervising and adaptively disentangling the DVR inside the model. Meanwhile, we analyze the composition of the DVR through a novel brick study. For DDCL, we plan to extend this design to positive and negative sample contrastive learning frameworks to explore the potential of adaptive DVR extracting. Furthermore, we believe that the information in the DVR is worthy of further analysis and even subsequent disentangling. In addition to this, the role of DVRs in dense prediction tasks such as segmentation and target detection is also very promising. We will further explore these in our future work.

Acknowledgment

This work was supported by the Key Program Special Fund in XJTLU (KSF-A-22). Zhou is supported by National Natural Science Foundation of China (NSFC) under grant No. 62271465.

References

[1] Nitin Bansal, Xiaohan Chen, and Zhangyang Wang. *Can we gain more from orthogonality regularizations in training deep networks?*, volume 31, pages 4261–4271. 2018. 3

[2] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *Le Centre pour la Communication Scientifique Directe - HAL - Diderot*, Apr 2022. 3, 5

[3] 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. 3

[4] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014. 7

[5] 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, 1993. 3

[6] 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*, volume 2019, page 2060. NIH Public Access, 2019. 3

[7] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020. 2

[8] 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. 2

[9] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020. 2

[10] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. *Infogan: Interpretable representation learning by information maximizing generative adversarial nets*. 2016. 3

[11] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2

[12] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021. 2, 3, 4, 5, 6

[13] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021. 2

[14] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *Advances in neural information processing systems*, 33:8765–8775, 2020. 2

[15] 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. 6

[16] Rumen Dangovski, Li Jing, Charlotte Loh, Seungwook Han, Akash Srivastava, Brian Cheung, Pulkit Agrawal, and Marin Soljačić. Equivariant contrastive learning. *arXiv preprint arXiv:2111.00899*, 2021. 2, 3, 5

[17] 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*, Jun 2009. 5, 6

[18] Alexandre Devillers and Mathieu Lefort. Equimod: An equivariance module to improve self-supervised learning. Nov 2022. 2, 3, 5

[19] Kien Do and Truyen Tran. *Theory and evaluation metrics for learning disentangled representations*. 2020. 3

[20] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9588–9597, 2021. 2

[21] Yunhao Ge, Sami Abu-El-Haija, Gan Xin, and Laurent Itti. Zero-shot synthesis with group-supervised learning. *arXiv preprint arXiv:2009.06586*, 2020. 3

[22] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 2, 3, 4, 5

[23] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010. 2

[24] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 1

[25] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 1, 2

[26] Wei Huang, Weitao Du, and Richard Yi Da Xu. On the neural tangent kernel of deep networks with orthogonal initialization. *arXiv preprint arXiv:2004.05867*, 2020. 3

[27] HyeongJoo Hwang, Geon-Hyeong Kim, Seunghoon Hong, and Kee-Eung Kim. Variational interaction information maximization for cross-domain disentanglement. *Advances in Neural Information Processing Systems*, 33:22479–22491, 2020. 3

[28] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6

[29] Jose Lezama, Qiang Qiu, Pablo Muse, and Guillermo Sapiro. *OLE: Orthogonal Low-rank Embedding, A Plug and Play Geometric Loss for Deep Learning*, volume 2, page 6. IEEE, 6 2018. 3

[30] Chao Li, Kelu Yao, Jin Wang, Boyu Diao, Yongjun Xu, and Quanshi Zhang. Interpretable generative adversarial networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1280–1288, 2022. 3

[31] Shuai Li, Kui Jia, Yuxin Wen, Tongliang Liu, and Dacheng Tao. Orthogonal deep neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 43(4):1352–1368, 2019. 3

[32] Yunfan Li, Mouxing Yang, Dezhong Peng, Taihao Li, Jiantao Huang, and Xi Peng. Twin contrastive learning for online clustering. *International Journal of Computer Vision*, 130(9):2205–2221, 2022. 2

[33] Weiyang Liu, Yan-Ming Zhang, Xingguo Li, Zhiding Yu, Bo Dai, Tuo Zhao, and Le Song. *Deep hyperspherical learning*, pages 3953–3963. Number 2. 2017. 3

[34] Francesco Locatello, Gabriele Abbati, Thomas Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. On the fairness of disentangled representations. *Advances in neural information processing systems*, 32, 2019. 3

[35] 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*, pages 4114–4124. PMLR, 2019. 3

[36] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. A commentary on the unsupervised learning of disentangled representations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(09):13681–13684, 4 2020. 3

[37] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729, 2008. 7

[38] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[39] Kanchana Ranasinghe, Muzammal Naseer, Munawar Hayat, Salman Khan, and Fahad Shahbaz Khan. Orthogonal projection loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12333–12343, 2021. 3

[40] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020. 2

[41] Mhd Hasan Sarhan, Nassir Navab, Abouzar Eslami, and Shadi Albarqouni. Fairness by learning orthogonal disentangled representations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pages 746–761. Springer, 2020. 3

[42] Anshul Shah, Suvrit Sra, Rama Chellappa, and Anoop Cherian. Max-margin contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8220–8230, 2022. 2

[43] Francesco Sjoerd Van Steenkiste, Jürgen Locatello, Olivier Schmidhuber, and Bachem. *Are disentangled representations helpful for abstract visual reasoning?* 2019. 3

[44] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. *SVDNet for Pedestrian Retrieval*, pages 3800–3808. Number 2. IEEE, 10 2017. 3

[45] Igor Susmelj, Matthias Heller, Philipp Wirth, Jeremy Prescott, and Malte Ebner et al. Lightly. *GitHub. Note: https://github.com/lightly-ai/lightly*, 2020. 6

[46] Raphael Suter, Djordje Miladinovic, Bernhard Schölkopf, and Stefan Bauer. *Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness*. 2019. 3

[47] Shixiang Tang, Feng Zhu, Lei Bai, Rui Zhao, and Wanli Ouyang. Relative contrastive loss for unsupervised representation learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII*, pages 1–18. Springer, 2022. 2

[48] Luan Tran, Xi Yin, and Xiaoming Liu. *Disentangled Representation Learning GAN for Pose-Invariant Face Recognition*. IEEE, 7 2017. 3

[49] Jiayun Wang, Yubei Chen, Rudrasis Chakraborty, and Stella Yu. *Orthogonal Convolutional Neural Networks*, volume 2020. IEEE, 6 2020. 3

[50] Jiayun Wang, Yubei Chen, Rudrasis Chakraborty, and Stella X Yu. Orthogonal convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11505–11515, 2020. 3

[51] Tan Wang, Zhongqi Yue, Jianqiang Huang, Qianru Sun, and Hanwang Zhang. Self-supervised learning disentangled group representation as feature. *Advances in Neural Information Processing Systems*, 34:18225–18240, 2021. 3

[52] Yitong Wang, Dihong Gong, Zheng Zhou, Xing Ji, Hao Wang, Zhifeng Li, Wei Liu, and Tong Zhang. Orthogonal deep features decomposition for age-invariant face recognition. In *Proceedings of the European conference on computer vision (ECCV)*, pages 738–753, 2018. 3

[53] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010. 7

[54] 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. 2

[55] Tete Xiao, Xiaolong Wang, AlexeiA. Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. *International Conference on Learning Representations*, May 2021. 2, 3, 5

[56] 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. 2

[57] Chun-Hsiao Yeh, Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu, Yubei Chen, and Yann LeCun. Decoupled contrastive learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, pages 668–684. Springer, 2022. 2

[58] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017. 6

[59] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021. 2, 3, 4, 5

[60] Chaoning Zhang, Kang Zhang, Chenshuang Zhang, Axi Niu, Jiu Feng, Chang D Yoo, and In So Kweon. Decoupled adversarial contrastive learning for self-supervised adversarial robustness. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX*, pages 725–742. Springer, 2022. 3