On Learning Discriminative Features from Synthesized Data for Self-Supervised Fine-Grained Visual Recognition

Zihu Wang[1], Lingqiao Liu[2], Scott Ricardo Figueroa Weston[1], Samuel Tian[3], and Peng Li[1]

[1] University of California, Santa Barbara CA 93106, USA
[2] The University of Adelaide, Adelaide South Australia 5001, Australia
[3] Carnegie Mellon University, Pittsburgh PA 15213, USA
{zihu_wang,scottricardo,lip}@ucsb.edu {lingqiao.liu}@adelaide.edu.au
{samuel.tian7}@gmail.com

Abstract. Self-Supervised Learning (SSL) has become a prominent approach for acquiring visual representations across various tasks, yet its application in fine-grained visual recognition (FGVR) is challenged by the intricate task of distinguishing subtle differences between categories. To overcome this, we introduce an novel strategy that boosts SSL's ability to extract critical discriminative features vital for FGVR. This approach creates synthesized data pairs to guide the model to focus on discriminative features critical for FGVR during SSL. We start by identifying non-discriminative features using two main criteria: features with low variance that fail to effectively separate data and those deemed less important by Grad-CAM induced from the SSL loss. We then introduce perturbations to these non-discriminative features while preserving discriminative ones. A decoder is employed to reconstruct images from both perturbed and original feature vectors to create data pairs. An encoder is trained on such generated data pairs to become invariant to variations in non-discriminative dimensions while focusing on discriminative features, thereby improving the model's performance in FGVR tasks. We demonstrate the promising FGVR performance of the proposed approach through extensive evaluation on a wide variety of datasets.

Keywords: Self-Supervised representation learning · fine-grained visual recognition · learning from generated data

1 Introduction

In computer vision, Fine-grained Visual Recognition (FGVR) focuses on identifying subcategories of visual data, such as bird species [1, 38], aircraft variants [29], and vehicle models [25]. Different from studies on large-scaled general image datasets [26, 32, 37], FGVR tasks highlight the challenge of distinguishing subtle visual patterns.

Self-Supervised Learning (SSL) methods have recently largely advanced the domain of visual representation learning, circumventing the necessity for human-provided annotations. In SSL, many contrastive learning approaches achieve state-of-the-art performance by learning the similarities between data pairs derived from augmentations of identical source images [3, 4, 6, 15, 16, 27, 40]. These methods facilitate the transferability of the learned representations across a wide range of visual recognition problems [4,13,14,16]. Despite these advancements, it has been suggested that SSL may prioritize general visual similarities, instead of critical subtle features in FGVR tasks, which leads to SSL's 'coarse-label bias' [9]. Furthermore, recent studies [24, 35, 36] have highlighted a tendency among existing SSL methods to be distracted by task-irrelevant features, consequently failing to capture FGVR-relevant patterns.

To this end, we propose an innovative self-supervised learning strategy focusing on selectively extracting highly discriminative features while disregarding less informative, noisy ones. Our approach involves the generation of new contrastive data pairs from the latent feature space of the encoder, training the encoder to prioritize critical objects within these pairs. To facilitate this, a decoder is employed to generate data based on the latent feature space, reconstructing an image's feature vector and its perturbed counterpart to form each data pair. As the data pairs are generated to guide the encoder to learn key features and to be invariant to variations in non-discriminative features, in each feature vector, only those dimensions associated with non-discriminative patterns are perturbed.

Therefore, at the core of our methods are two non-discriminative feature identification techniques. Firstly, Grad-CAM [33] induced from the SSL loss to the latent feature space highlights dimensions relevant to FGVR [35, 36]. Thus, we introduce greater perturbation to those less highlighted dimensions. Despite the conventional view that dimensional collapse in SSL—characterized by some encoder dimensions producing constant outputs—is undesirable [6, 23, 28], recent literature [10, 45] suggests that inducing such collapse in task-irrelevant feature dimensions yields beneficial outcomes. As shown in Fig. 2, our empirical studies indicate that in encoders pre-trained by SSL methods, there are always dimensions with low variance across the dataset which cannot effectively separate data from different categories. We thus treat these low-variance dimensions as task-irrelevant and introduce perturbations to them. The two aforementioned perturbation components are then combined and applied to the latent feature vector of each image. Images are then reconstructed from the perturbed and original feature vectors to form contrastive pairs for a contrastive loss [4, 30]. Such a framework encourages the encoder to learn the key features highlighted by Grad-CAM and to reduce variance and induce collapse in those non-discriminative dimensions with low variance across the dataset in the latent space.

Our proposed fine-grained feature learning method can be incorporated into various existing SSL methods. We use SimSiam [6] and MoCo v2 [16] as baseline methods and incorporate our proposed technique into these methods. Experiments across various fine-grained visual datasets show the effectiveness of our method. The proposed method provides a great improvement over baseline

methods. Our methods built on MoCo v2 outperforms existing state-of-the-art Self-Supervised fine-grained visual recognition methods in numerous downstream tasks.

2 Related Works

2.1 Self-Supervised Contrastive Learning

Self-Supervised Learning (SSL) facilitates the learning of visual representations without the need for labeled data. Among various SSL methodologies, contrastive learning has emerged as a promising technique. With the InfoNCE loss [30] and its variants [3, 4, 6, 7, 16] being introduced as the objectives for optimization, contrastive approaches treat different views of the same image as positive data pairs, while views from different images are considered negative data pairs. The goal for the encoder is to minimize the distance between positive pairs and maximize it between negative pairs within its representation space [2, 4, 16, 39]. Methods such as BYOL [15] and SimSiam [6] rely exclusively on positive pairs. Additionally, the issue of dimensional collapse, where some encoder dimensions output constant values, is discussed in [6,23,28], along with proposed solutions to mitigate this phenomenon. Nonetheless, recent studies [10, 45] have shown that the collapse of dimensions associated with task-irrelevant features can enhance the performance in downstream visual recognition tasks.

2.2 Fine-Grained Visual Recognition in Self-Supervised Learning

While encoders pre-trained by Self-Supervised Learning (SSL) methods demonstrate transferability and generalizability in many tasks [4, 6, 16, 21, 41], studies [9, 24, 35] reveal SSL's limitations in capturing essential features for Fine-Grained Visual Recognition (FGVR). To enhance SSL's capability in identifying critical features, several works concentrate on refining data augmentations. Approaches such as SAGA [43], CAST [34], and ContrastiveCrop [31] adopt attention-guided heatmaps to locate and better crop key objects in images. DiLo [44] introduces a novel augmentation by merging key image objects with different backgrounds to generate additional views. Contrary to methods that modify images directly, our approach involves perturbing feature vectors and generating realistic images from the latent feature space to enhance the encoder's discriminative capacity. Another line of research employs auxiliary neural networks connected to the encoder's convolutional layers for improving encoder's attention on salient regions. LEWEL [20] trains an additional head to adaptively aggregate features. Techniques such as CVSA [12] and [42] train a network to fit segmentation annotations or outputs of pre-trained saliency detectors. Similarly, LCR [35] and SAM [36] train the network to align with Grad-CAM, treating Grad-CAM as the ground truth for the encoder's attention maps. Our method proposes training the encoder on generated data pairs to learn critical features. In addition to Grad-CAM, we use dimension variance as

a criterion for identifying non-discriminative features. Low-variance dimensions where data points across the dataset are not well separated are treated less crucial. Besides, SimCore [24] pre-trains an encoder on a target dataset, then using it to select more relevant data from a large-scaled dataset to expand the training set, upon which a new encoder is retrained for downstream tasks.

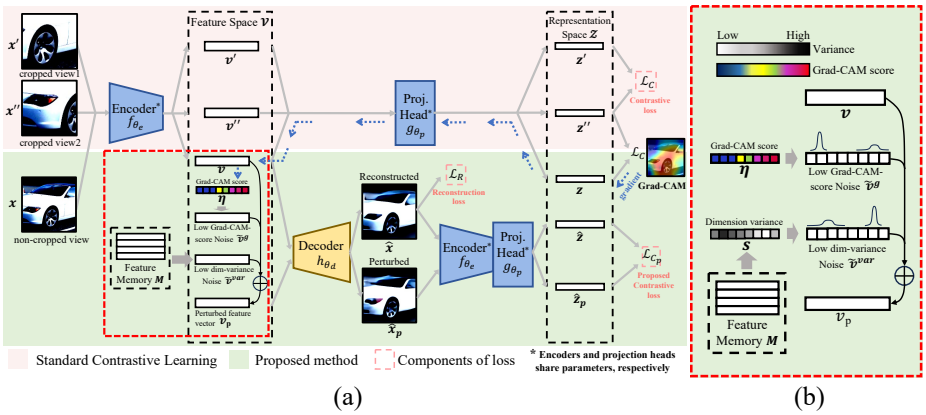

Fig. 1: The overview of the proposed method. (a) Our method can be incorporated into various existing SSL methods. A decoder is utilized to generate images from both the original feature vector and its perturbed counterpart to form data pairs. The overall loss consists three terms: a conventional contrastive loss, a reconstruction loss (ensuring the decoder evolves with the encoder), and a proposed contrastive loss on the generated pairs. (b) We propose two techniques to identify and perturb non-discriminative features in a feature vector, i.e., features with low variance that fail to effectively separate data and those deemed less important by Grad-CAM induced from the SSL loss.

3 Method

3.1 Background

Self-Supervised Contrastive Learning Without need for labels, self-supervised contrastive learning learns to represent data $\mathbf{x} \in \mathbb{R}^m$ in a lower dimensional space $\mathbb{R}^n$ by learning the similarities among data samples. Typically, in contrastive learning, the model consists two components, an encoder $f_{\theta_e} : \mathbb{R}^m \to \mathbb{R}^n$ that maps data to a latent feature space $\mathcal{V} \subseteq \mathbb{R}^n$, and a projection head $g_{\theta_p} : \mathbb{R}^n \to \mathbb{R}^k$ which projects the latent feature vectors in $\mathcal{V} \subseteq \mathbb{R}^n$ to a lower dimensional representation space $\mathcal{Z} \subseteq \mathbb{R}^k$ where the contrastive loss is applied. Formally, given a batch $\mathcal{B}$ of unlabeled data, every image $\mathbf{x}$ in it is augmented by two random augmentation $\mathcal{T}_1$ and $\mathcal{T}_2$ to acquire two views, i.e., $\mathbf{x}' = \mathcal{T}_1(\mathbf{x})$, $\mathbf{x}'' = \mathcal{T}_2(\mathbf{x})$. Views augmented from the same image are considered a positive

pair, while those acquired from different images form negative pairs. Two augmented views of all images form a new batch $\mathcal{B}_a$ which doubles the size of $\mathcal{B}$. The encoder and projection head are then used to represent the views in $\mathcal{Z}$, i.e., $\mathbf{z}' = g_{\theta_p}(f_{\theta_e}(\mathbf{x}'))$, $\mathbf{z}'' = g_{\theta_p}(f_{\theta_e}(\mathbf{x}''))$. A contrastive loss $\mathcal{L}_C$ is then defined in $\mathcal{Z}$ space:

$$\mathcal{L}_C(\mathbf{z}', \mathbf{z}'') = -\log \frac{\exp(\mathbf{z}' \cdot \mathbf{z}''/\tau)}{\sum_{\mathbf{z_i} \in \mathcal{B}_a, \mathbf{z_i} \neq \mathbf{z}', \mathbf{z_i} \neq \mathbf{z}''} \exp(\mathbf{z}' \cdot \mathbf{z_i}/\tau)} \tag{1}$$

where τ is the temperature hyperparameter.

Although there are subtle differences between different contrastive methods, the contrastive loss are defined similarly. MoCo [5,16] introduces a large memory of negative representations. SimSiam [6] and BYOL [15] discard negative pairs and learn solely from positive pairs.

3.2 Overview

The overview of our method is illustrated in Fig. 1. The essence of the proposed method is learning discriminative fine-grained visual features from synthesized data pairs which are reconstructed from latent feature vectors by a decoder h_{θ_d}. During training, to ensure that the decoder follows the evolution of the encoder, we use Mean Square Error (MSE) as the loss function to optimize the decoder. Our empirical studies show that the decoder produces images of better quality when it is trained on non-augmented images. The following reconstruction loss $\mathcal{L}_R$ is calculated for each image.

$$\mathcal{L}_R = \frac{1}{m} \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 \tag{2}$$

where $\mathbf{x} \in \mathbb{R}^m$ is non-augmented image, and $\hat{\mathbf{x}} = h_{\theta_d}(f_{\theta_e}(\mathbf{x}))$ is the reconstruction of it.

In addition to producing $\hat{\mathbf{x}}$, we generate $\hat{\mathbf{x}}_p$ from $\mathbf{v}_p \in \mathcal{V}$, which is a perturbed version of $\mathbf{v}$ where non-discriminative features are perturbed. A positive data pair is formed between $\hat{\mathbf{x}}$ and $\hat{\mathbf{x}}_p$ from which the encoder learns discriminative features while disregarding task-irrelevant ones. In Sec. 3.3 and Sec. 3.4, we introduce two methods of identifying crucial discriminative dimensions in the latent feature space $\mathcal{V}$.

3.3 Identifying key dimensions via Grad-CAM

Grad-CAM [33], a widely used saliency detection technique, uses the gradient of the target loss with respect to intermediate features of the network to produce an attention map highlighting regions in the features that contribute to minimizing the loss. As in our proposed self-supervised method, labels are not available during training, we thus choose the contrastive loss as the target. To identify important features within an image's feature vector $\mathbf{v} = f_{\theta_e}(\mathbf{x}) \in \mathbb{R}^n$, we form positive pair between $\mathbf{x}$ and an augmented view $\mathbf{x}''$ to calculate a contrastive loss

$\mathcal{L}_C(\mathbf{z}, \mathbf{z}'')$, where $\mathbf{z} = g_{\theta_p}(f_{\theta_e}(\mathbf{x}))$, $\mathbf{z}'' = g_{\theta_p}(f_{\theta_e}(\mathbf{x}''))$. As it is shown in Fig. 1, the Grad-CAM score vector $\boldsymbol{\eta} = \{\eta_i\}_{i=1}^n \in \mathbb{R}^n$ is calculated by gradient of the contrastive loss with respect to the feature vector $\mathbf{v}$.

$$\eta_i = \text{ReLU}(\frac{\partial \mathcal{L}_C(g_{\theta_p}(\mathbf{v}), g_{\theta_p}(\mathbf{v}''))}{\partial v_i} \cdot v_i) \tag{3}$$

Here, $\mathbf{v} = f_{\theta_e}(\mathbf{x})$, $\mathbf{v}'' = f_{\theta_e}(\mathbf{x}'')$. v_i denotes the i_{th} element of $\mathbf{v}$. The application of ReLU$(\cdot)$ makes $\eta_i > 0$ for all $i \in \{1, 2, \ldots, n\}$. Note that the original Grad-CAM calculates gradient with respect to feature maps of the last convolutional layer. To measure the saliency of feature dimensions, we calculate gradient with respect to feature vectors. Higher Grad-CAM scores in $\boldsymbol{\eta}$ represent corresponding dimension's higher contribution to data discrimination in contrastive learning.

With the Grad-CAM scores, random Gaussian noise is then introduced as perturbation to $\mathbf{v}$. We first scale all elements in $\boldsymbol{\eta}$ to $[0, 1]$ by min-max normalization.

$$\bar{\eta}_i = \frac{\eta_i - \min\{\eta_j, j = 1, 2, \ldots, n\}}{\max\{\eta_j, j = 1, 2, \ldots, n\} - \min\{\eta_j, j = 1, 2, \ldots, n\}} \tag{4}$$

where $\bar{\eta}_i$ is the i_{th} element of the normalized Grad-CAM score vector $\bar{\boldsymbol{\eta}}$. After the normalization, a random Gaussian noise perturbation vector $\tilde{\mathbf{v}}^g \in \mathbb{R}^n$ is calculated as follows.

$$\tilde{\mathbf{v}}^g = \{\tilde{v}_i^g : \tilde{v}_i^g \sim \mathcal{N}(0, \epsilon_g \cdot (1 - \bar{\eta}_i))\}_{i=1}^n \tag{5}$$

where ϵ_g is a hyperparameter that controls the standard deviation of Gaussian noise. In such a perturbation, all elements are sampled from i.i.d. zero-mean Gaussian distributions. Importantly, dimensions with lower normalized Grad-CAM scores $\bar{\eta}_i$ receive noise from a Gaussian distribution with a greater standard deviation, implying a higher likelihood of more significant noise affecting dimensions with lower Grad-CAM scores. And on average, key dimensions with higher Grad-CAM scores are affected less which helps preserve crucial features in the original images.

3.4 Determining feature's task-relevance via dimension variance

In addition to the feature perturbation technique elaborated in Sec. 3.3, we propose another technique to determine and perturb task-irrelevant dimensions.

In SSL, dimensional collapse is a phenomenon where some encoder dimensions output constant values [6,19,23,28]. As data points are not separated along such dimensions, these dimensions can not be used to perform downstream visual recognition. However, recent works [10, 45] suggest that collapse of dimensions which are related to downstream task-irrelevant features can be beneficial. In spite of the potential benefit, how to induce beneficial dimensional collapse is not illustrated by existing SSL studies.

As it is illustrated in Fig. 2, our empirical studies show that, in the latent feature space of encoders pre-trained by SSL methods, typically, data points are

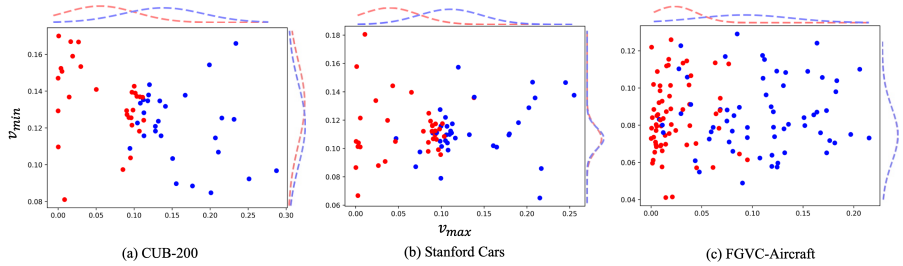

Fig. 2: An illustration of data distribution in the feature space of encoders pre-trained by MoCo v2 [5]. Blue and red dots represent feature vectors of two categories' data from 3 fine-grained datasets, CUB-200 [38], Stanford Cars [25], and FGVC-Aircraft [29]. v_{min} and v_{max} are the dimensions in the feature space where data has the minimal and maximal variance across the dataset. Probability density curve fitting of each category along each dimension is attached to the corresponding axis. Different classes are separated much better along v_{max} than v_{min}.

not well separated along dimensions with low variance. Variance along these dimensions thus introduces noise to downstream classification. Therefore, we treat such dimensions as task-irrelevant and propose a technique to induce collapse in these dimensions to guide the encoder's to be invariant to variations of such features. This technique start with estimating the dataset's variance along each feature dimension in a feature vector memory bank $\mathbf{M} \in \mathbb{R}^{D \times n}$ of size D. During training, whenever the encoder is provided with a batch of data, its feature vectors will be stored in the memory to replace the oldest batch of feature vectors in it. Variance of each dimension across the dataset can be approximated in $\mathbf{M}$ to get a variance vector $\mathbf{s} = \{s_i : \sigma^2(\bar{\mathbf{w}}_i), i = 1, 2, \ldots, n\} \in \mathbb{R}^n$. Here $\bar{\mathbf{w}}_i \in \mathbb{R}^D$ is the ℓ_2-normalized i_{th} column vector of $\mathbf{M}$. A feature represented by dimension i is considered less discriminative if its corresponding variance $s_i < \kappa$ where κ is a threshold hyperparameter. To introduce random noise to those less discriminative dimensions, similar to Eq. (4), we first apply min-max normalization to $\mathbf{s}$ to acquire $\bar{\mathbf{s}} = \{\bar{s}_i\}_{i=1}^n$. We then calculate the random noise vector $\tilde{\mathbf{v}}^{var} = \{\tilde{v}_i^{var}\}_{i=1}^n$ to be applied.

$$\tilde{v}_i^{var} = \begin{cases} u_i \sim \mathcal{N}(0, \epsilon_{var} \cdot (1 - \bar{s}_i)) & \text{if } s_i < \kappa \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Here, ϵ_{var} defines the standard deviations of the i.i.d. Gaussian distributions. Similar to $\tilde{\mathbf{v}}^g$, $\tilde{\mathbf{v}}^{var}$ introduces greater noise to lower-variance dimensions.

3.5 Learning from reconstructed data pairs

With the two feature perturbation techniques proposed in Sec. 3.3 and Sec. 3.4, we can finally perturb the feature vector $\mathbf{v}$ by adding the two random Gaussian noise to it to obtain its perturbed version $\mathbf{v}_p$, i.e., $\mathbf{v}_p = \mathbf{v} + \tilde{\mathbf{v}}^g + \tilde{\mathbf{v}}^{var}$. As it is shown in Fig. 1, $\mathbf{v}$ and $\mathbf{v}_p$ are reconstructed by the decoder to produce

$\hat{\mathbf{x}}$ and $\hat{\mathbf{x}}_p$, respectively. In $\hat{\mathbf{x}}_p$, key patterns from $\hat{\mathbf{x}}$ are preserved, while those contribute less to Grad-CAM or deemed less discriminative across the dataset by the low-variance criterion are perturbed.

We then form a positive pair between $\hat{\mathbf{x}}$ and $\hat{\mathbf{x}}_p$, addressing extracting key features and disregarding non-discriminative ones. To this end, we propose to pass the representations $\hat{\mathbf{z}} = g_{\theta_p}(f_{\theta_e}(\hat{\mathbf{x}}))$ and $\hat{\mathbf{z}}_p = g_{\theta_p}(f_{\theta_e}(\hat{\mathbf{x}}_p))$ to a contrastive loss $\mathcal{L}_{C_p}$.

$$\mathcal{L}_{C_p}(\hat{\mathbf{z}}, \hat{\mathbf{z}}_p) = -\log \frac{\exp(\hat{\mathbf{z}} \cdot \hat{\mathbf{z}}_p / \tau)}{\sum_{\hat{\mathbf{z}}_\mathbf{i} \in \mathcal{B}_r, \hat{\mathbf{z}}_\mathbf{i} \neq \hat{\mathbf{z}}, \hat{\mathbf{z}}_\mathbf{i} \neq \hat{\mathbf{z}}_p} \exp(\hat{\mathbf{z}} \cdot \hat{\mathbf{z}}_\mathbf{i} / \tau)} \tag{7}$$

where $\mathcal{B}_r$ is the set of all reconstructed and perturbed images from the current batch $\mathcal{B}$. By forming a positive pair between $\hat{\mathbf{x}}$ and $\hat{\mathbf{x}}_p$, Eq. (7) requires the encoder to be invariant to those perturbed less discriminative features.

Finally, we write the training loss $\mathcal{L}$ of our method as follows.

$$\mathcal{L} = \mathcal{L}_C + \alpha \cdot \mathcal{L}_R + \nu \cdot \mathcal{L}_{C_p} \tag{8}$$

α and ν are hyperparameters that control the weight of $\mathcal{L}_R$ and $\mathcal{L}_{C_p}$ in training. The pseudocode of our method is provided in Appendices.

Table 1: Performance comparison on three datasets. Our method is compared with MoCo v2 [5] and ResNet50 supervised pre-trained on ImageNet-1k [11]. Top-1 classification accuracy (in%) is reported when model is evaluated on 100%, 50%, and 20% of all labels. Rank-1, rank-5, and mAP (in%) in image retrieval are reported.

		Classification			Image Retrieval		
Dataset	Methods	100%	50%	20%	rank-1	rank-5	mAP
CUB-200	ResNet50	63.06	55.71	42.24	40.39	68.94	15.88
	MoCo V2	63.98	56.35	42.63	39.72	67.14	15.91
	Ours	**66.17**	**60.84**	**49.69**	**42.06**	**69.59**	**19.70**
Stanford Cars	ResNet50	61.41	49.85	33.57	29.28	54.66	7.01
	MoCo V2	62.02	51.08	35.44	30.51	56.15	7.13
	Ours	**65.60**	**54.36**	**40.24**	**35.81**	**61.94**	**10.02**
FGVC-Aircraft	ResNet50	49.79	41.02	33.61	29.48	51.39	10.17
	MoCo V2	51.13	44.34	36.42	30.02	52.87	11.24
	Ours	**55.28**	**49.37**	**41.10**	**33.27**	**56.80**	**12.69**

4 Experiments

4.1 Experiment Settings

Datasets. Experiments are conducted across five fine-grained visual datasets. We adopt on three widely used fine-grained datasets. `Caltech-UCSD Birds 200 -2011 (CUB-200)` dataset [38] contains 5994 training data and 5794 testing data

of 200 categories of birds. `Stanford Cars (Cars)` [25] has 196 classes of car models where 8144 data and 8041 data are in its training and testing split respectively. `FGVC-Aircraft (Aircraft)` [29] has 100 classes where 6667 images are for training and 3333 images are for testing. Additionally, we consider `German Traffic Sign Recognition Benchmark (GTSRB)` [18] which contains 43 classes of traffic signs. This dataset is usually used in autonomous driving and smart cities development. We take 4800 images for training and 3750 images for testing from `GTSRB`. We also evaluate the effectiveness of our method on `ISIC2017` [8], a medical image dataset with 3 categories of skin lesion analysis where 2000 images are for training and 600 images are for testing.

Training Settings. All methods adopt ResNet-50 [17] as the encoder backbone where weights are initialized by loading ImageNet-1k [32] pre-trained model. Using two state-of-the-art SSL method, MoCo v2 [5] and SimSiam [6], as the baseline methods, we incorporate our proposed method in their framework. For the sake of fair comparison, encoders of all methods are pre-trained for 100 epochs. And the training batch size is set to 128 for all methods. More detailed encoder pre-training settings are provided in Appendices.

Additionally, in our method, we use a feature vector memory bank of size $D = 5632$. For the training loss in Eq. (8), we choose $\alpha = 1$ and $\nu = 0.5$. When introducing noise described in Eq. (5) and Eq. (6), we choose $\epsilon_g = 0.1$, $\epsilon_{var} = 0.05$, and $\kappa = 0.02$. To ensure the quality of reconstructed images, before encoder training, we freeze the encoder parameters and pre-train the decoder on target datasets by a reconstruction loss. We provide decoder pre-training details in Appendices.

Performance Evaluation Protocols. Linear evaluation is a widely adopted protocol for assessing the performance of learned representations in visual recognition. This approach freezes the parameters of the pre-trained encoder and attaches a linear classifier to it. The classifier is then trained to perform classification. Our linear evaluation setup follows [16], detailed further in the Appendices.

The task of image retrieval [22, 35, 41] serves as another pivotal method for evaluating the performance of representation learning. Without adjusting any model parameters, it searches the nearest neighbors of a query image in the latent feature space for images that share the same label with the query image. The effectiveness of the evaluated model is quantified by recording the proportion of retrieved images that fall into the same categories as the query image. We present three commonly utilized metrics of retrieval performance: rank-1, rank-5, and mean Average Precision (mAP).

Furthermore, to illustrate the effect of the proposed method, attention maps generated by Grad-CAM on encoders trained by different methods are compared. Images from the training datasets, along with their corresponding reconstructed and perturbed versions, are also provided.

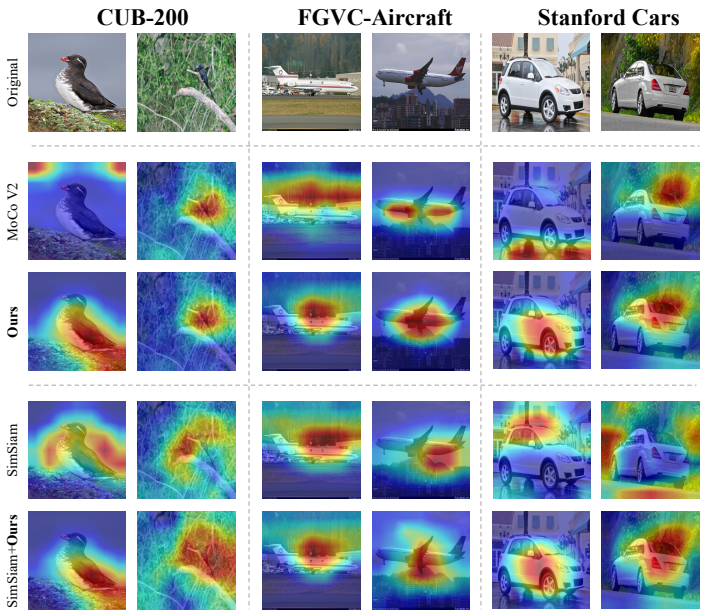

Fig. 3: Grad-CAM attention visualized on images. Our proposed method is incorporated into MoCo v2 and SimSiam and compared with them.

4.2 Main Results

Comparison between the proposed method and baseline methods As among all configurations of our method, the one based on MoCo v2 achieves the best overall performance, this configuration is thus called **'ours'** in the experiments. We first compare our method with MoCo v2 in Tab. 1. 'ResNet50' in Tab. 1 represents the encoder supervised pre-trained on ImageNet-1k [11].

In linear evaluation, we evaluate pre-trained encoders using different proportions of labels. Across all three datasets, our proposed method achieves an average top-1 accuracy improvement of 3.31%, 4.27%, and 5.51% over MoCo v2, with 100%, 50%, and 20% of all labels, respectively. This advancement highlights the efficacy of our data pairs generation technique, enabling great performance in classification tasks. Notably, the remarkable improvement in label-insufficient scenarios highlights our method's ability to learn more generalizable features from unlabeled data.

We further evaluate the effectiveness of our method by image retrieval tasks. Performance in these tasks serves as a measure of semantic consistency in the learned latent feature space. Our method outperforms MoCo v2 on all datasets in rank-1, rank-5, and mAP. Invariant to irrelevant patterns, our approach ensures that images from the same category, which exhibit discriminative patterns, are more closely clustered in the feature space. This distribution enhances performance in image retrieval.

Table 2: Performance comparison on ISIC2017 [8] and GTSRB [18]. Top-1 classification accuracy (in %) and rank-1 (in %) of image retrieval is reported.

	ISIC2017		GTSRB	
Methods	Classification	Retrieval	Classification	Retrieval
MoCo V2 [5]	64.33	66.34	88.25	97.46
Ours	**66.92**	**67.90**	**91.67**	**98.59**
SimSiam [6]	58.83	64.46	81.10	92.81
SimSiam + Ours	**63.78**	**65.55**	**86.94**	**94.42**

Additionally, we visualize Grad-CAM attention in MoCo v2 and our method in Fig. 3. Unlike MoCo v2, which may concentrate on background regions irrelevant to visual recognition, our method exhibits enhanced precision in identifying and focusing on pivotal objects within the images, effectively minimizing the influence of background distractions.

Furthermore, we also integrate the proposed technique into a state-of-the-art negative pair-free method SimSiam [6], denoted as 'SimSiam+ours' in our experiments. Our framework significantly outperforms the original SimSiam in both linear evaluation and image retrieval tasks, as demonstrated in Tab. 3. Attention comparison is also visualized by Grad-CAM and compared with SimSiam in Fig. 3.

To comprehensively assess the effectiveness of our proposed method, we utilize two more fine-grained datasets: the traffic sign visual dataset `GTSRB` and the medical image dataset `ISIC2017`. The results, as shown in Tab. 2, indicate that our approach enhances the performance of Self-Supervised Learning (SSL), demonstrating its potential in real-world applications for FGVR tasks.

Table 3: Comparison with state-of-the-art self-supervised FGVC methods. Supervised training is also included. Top-1 classification accuracy (in %) and rank-1 image retrieval (in %) are reported.

	Classification			Image Retrieval		
Method	CUB-200	Cars	Aircraft	CUB-200	Cars	Aircraft
supervised	77.46	88.60	85.93	-	-	-
Dino [3]	16.74	14.33	12.07	-	-	-
SimSiam [6]	46.75	45.72	38.52	16.24	12.45	18.49
MoCo V2 [5]	63.98	62.02	51.13	39.72	30.51	30.02
DiLo [44]	62.97	-	-	-	-	-
CVSA [12]	63.02	-	-	-	-	-
LEWEL [20]	64.59	62.91	51.90	39.91	32.36	31.09
ContrastiveCrop [31]	64.23	63.29	52.04	39.84	32.71	30.37
SAM-SSL-Bilinear [36]	64.94	62.85	52.83	40.08	33.19	30.52
LCR [35]	65.24	63.96	53.22	41.26	34.74	31.55
SimSiam+Ours	57.80	53.63	47.50	24.67	19.72	24.56
Ours	**66.17**	**65.60**	**55.28**	**42.06**	**35.81**	**33.27**

Comparison with state-of-the-art self-supervised FGVR methods In this section, we compare our proposed method against existing self-supervised learning (SSL) techniques renowned for their enhanced fine-grained visual recognition capabilities [12, 20, 31, 35, 36, 44], as detailed in Tab. 3. Methods like Dino [3], SimSiam [6], and MoCo v2 [16], which are not optimized for fine-grained feature extraction are also listed. Supervised training performance is included to provide a comprehensive comparison. Top-1 accuracy in linear evaluation and rank-1 in image retrieval tasks are reported.

In Tab. 3, our proposed method achieves the best overall performance in both image retrieval and linear evaluation tasks. DiLo [44], CVSA [12], and ContrastiveCrop [31] innovate with novel data augmentation techniques which directly modify the original images. In contrary, our method generates more realistic images from the learned feature space, highlighting the learning of FGVR-related features. And unlike SAM [36] and LCR [35] which train an auxiliary network to directly fit the encoder's attention to Grad-CAM, our method learns from generated data to highlight dimensions with high Grad-CAM scores and introduce dimensional collapse to non-discriminative features.

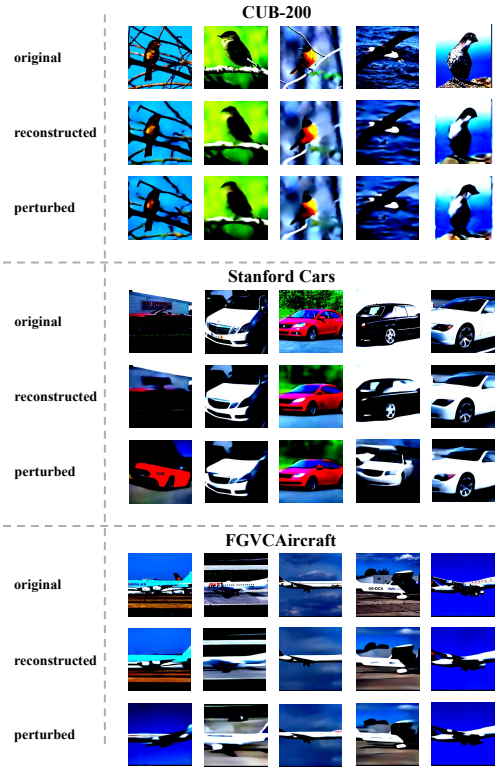

Fig. 4: Generated data pairs on CUB-200, Stanford Cars, and FGVC-Aircraft. The original images are also included.

Generated data pairs of the proposed method To better understand why the proposed technique enhances the fine-grained visual feature extraction capability, we show the generated data pairs from different datasets in Fig. 4. The perturbed images, as illustrated in Fig. 4, show that they largely retain the original data's key objects, with modifications primarily appearing as subtle changes in background or less important regions, e.g., changes of the tree's branch behind a bird, alterations in vehicle light's textures and adjustments in an airplane's exterior finish. These modifications do not affect the defining features of the subjects. Remarkably, some perturbations lead to entirely new objects that maintain the original's identity. For instance, transforming a side view of a car into a front view, or depicting an aircraft in flight from a grounded position. These images, obtained by modifying latent semantics of original images, are difficult to obtain through traditional data augmentation techniques defined in the original data space. They efficiently guide the encoder in identifying which features to prioritize and which to ignore, enhancing its learning performance in FGVR.

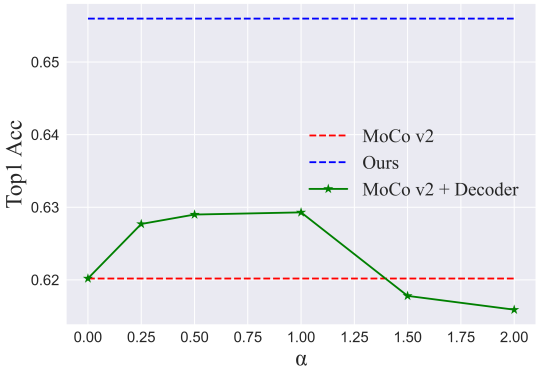

Fig. 5: Performance comparison of encoders trained by $\mathcal{L}_C + \alpha \cdot \mathcal{L}_R$ with respect to different α value (green solid line). Top-1 classification accuracy on Stanford Cars is reported. MoCo v2 (red dashed line) and our method (blue dashed line) are included for comparison.

4.3 Effectiveness of the reconstruction loss in contrastive learning

As described in Eq. (8), our model's overall training loss, $\mathcal{L}$, includes a reconstruction loss term, $\mathcal{L}_R$. To assess $\mathcal{L}_R$'s effect on self-supervised contrastive learning, we incorporate a decoder into MoCo v2 and train the encoder by a loss $\mathcal{L}_C + \alpha \cdot \mathcal{L}_R$ on the Stanford Cars dataset, varying α in the loss function. The results, shown in Fig. 5, indicate that α values of 0.5 and 1.0 enhance top-1

classification accuracy the most over MoCo v2. Generally, $\mathcal{L}_R$ provides a modest improvement (less than 1%) to Self-Supervised FGVC.

4.4 Effectiveness of the two feature perturbation techniques

As detailed in Sec. 3.3 and Sec. 3.4, two non-discriminative feature perturbation techniques are proposed to introduce noise $\tilde{\mathbf{v}}^g$ and $\tilde{\mathbf{v}}^{var}$, respectively. We conduct further experiments to assess the effectiveness of each technique in identifying and perturbing task-irrelevant features. In Tab. 4, we evaluate three different configurations of our method: (1) Ours: $\mathbf{v}_p$ is obtained by adding both noise components to $\mathbf{v}$, i.e., $\mathbf{v}_p = \mathbf{v} + \tilde{\mathbf{v}}^g + \tilde{\mathbf{v}}^{var}$; (2) Ours - Grad-CAM: $\mathbf{v}_p$ is obtained by adding only the noise generated by the low Grad-CAM scores criterion, i.e., $\mathbf{v}_p = \mathbf{v} + \tilde{\mathbf{v}}^g$; (3) Ours - low-var: $\mathbf{v}_p = \mathbf{v} + \tilde{\mathbf{v}}^{var}$.

As shown in Tab. 4, all configurations achieve competitive results comparing with existing state-of-the-art self-supervised FGVR methods. And when combining two noise components $\tilde{\mathbf{v}}^g$ and $\tilde{\mathbf{v}}^{var}$, our methods achieves the best overall performance.

Table 4: Comparison with state-of-the-art self-supervised FGVC methods. Supervised training is also included. Top-1 classification accuracy (in %) and rank-1 image retrieval (in %) are reported.

	Classification			Image Retrieval		
Method	CUB-200	Cars	Aircraft	CUB-200	Cars	Aircraft
MoCo V2 [5]	63.98	62.02	51.13	39.72	30.51	30.02
Ours - Grad-CAM	66.04	64.18	54.19	41.02	35.08	32.14
Ours - low-var	65.91	64.34	53.11	41.69	34.55	31.23
Ours	**66.17**	**65.60**	**55.28**	**42.06**	**35.81**	**33.27**

5 Conclusion

To enhance the performance of Self-Supervised Learning (SSL) in Fine-grained Visual Recognition (FGVR) tasks, this paper introduces a novel approach where an encoder learns discriminative features from generated images. By introducing noise to features deemed non-discriminative by two proposed criteria, we generate synthetic data from both the original and perturbed feature vectors by a decoder, thus forming data pairs that emphasize learning key features for FGVR. Our approach outperforms existing methods across various datasets in many downstream tasks. While the proposed approach also offers a modest boost to SSL performance in non-fine-grained visual recognition tasks—as detailed in the Appendices—the gains are notably more substantial in FGVR contexts. The refinement of our methodology for application to large-scaled general datasets remains an avenue for future research works.

Acknowledgment. This material is based upon work supported by the National Science Foundation under Grant No. 1956313.

References

1. Berg, T., Liu, J., Woo Lee, S., Alexander, M.L., Jacobs, D.W., Belhumeur, P.N.: Birdsnap: Large-scale fine-grained visual categorization of birds. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2011–2018 (2014)
2. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Advances in neural information processing systems **33**, 9912–9924 (2020)
3. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
5. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
6. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15750–15758 (2021)
7. Chuang, C.Y., Robinson, J., Lin, Y.C., Torralba, A., Jegelka, S.: Debiased contrastive learning. Advances in neural information processing systems **33**, 8765–8775 (2020)
8. Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). pp. 168–172. IEEE (2018)
9. Cole, E., Yang, X., Wilber, K., Mac Aodha, O., Belongie, S.: When does contrastive visual representation learning work? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14755–14764 (2022)
10. Cosentino, R., Sengupta, A., Avestimehr, S., Soltanolkotabi, M., Ortega, A., Willke, T., Tepper, M.: Toward a geometrical understanding of self-supervised contrastive learning. arXiv preprint arXiv:2205.06926 (2022)
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR09 (2009)
12. Di Wu, S.L., Zang, Z., Wang, K., Shang, L., Sun, B., Li, H., Li, S.Z.: Align yourself: Self-supervised pre-training for fine-grained recognition via saliency alignment. arXiv preprint arXiv:2106.15788 **2**(7), 8 (2021)
13. Ericsson, L., Gouk, H., Hospedales, T.M.: How well do self-supervised models transfer? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5414–5423 (2021)
14. Gao, Y., Zhuang, J.X., Lin, S., Cheng, H., Sun, X., Li, K., Shen, C.: Disco: Remedying self-supervised learning on lightweight models with distilled contrastive learning. In: European Conference on Computer Vision. pp. 237–253. Springer (2022)

15. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33**, 21271–21284 (2020)
16. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
18. Houben, S., Stallkamp, J., Salmen, J., Schlipsing, M., Igel, C.: Detection of traffic signs in real-world images: The German Traffic Sign Detection Benchmark. In: International Joint Conference on Neural Networks. No. 1288 (2013)
19. Hua, T., Wang, W., Xue, Z., Ren, S., Wang, Y., Zhao, H.: On feature decorrelation in self-supervised learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9598–9608 (2021)
20. Huang, L., You, S., Zheng, M., Wang, F., Qian, C., Yamasaki, T.: Learning where to learn in cross-view self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14451–14460 (2022)
21. Islam, A., Chen, C.F.R., Panda, R., Karlinsky, L., Radke, R., Feris, R.: A broad study on the transferability of visual representations with contrastive learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8845–8855 (2021)
22. Jang, Y.K., Cho, N.I.: Self-supervised product quantization for deep unsupervised image retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12085–12094 (2021)
23. Jing, L., Vincent, P., LeCun, Y., Tian, Y.: Understanding dimensional collapse in contrastive self-supervised learning. arXiv preprint arXiv:2110.09348 (2021)
24. Kim, S., Bae, S., Yun, S.Y.: Coreset sampling from open-set for fine-grained self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7537–7547 (2023)
25. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 554–561 (2013)
26. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
27. Lee, H., Lee, K., Lee, K., Lee, H., Shin, J.: Improving transferability of representations via augmentation-aware self-supervision. Advances in Neural Information Processing Systems **34**, 17710–17722 (2021)
28. Li, A.C., Efros, A.A., Pathak, D.: Understanding collapse in non-contrastive siamese representation learning. In: European Conference on Computer Vision. pp. 490–505. Springer (2022)
29. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013)
30. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
31. Peng, X., Wang, K., Zhu, Z., Wang, M., You, Y.: Crafting better contrastive views for siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16031–16040 (2022)

32. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
33. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
34. Selvaraju, R.R., Desai, K., Johnson, J., Naik, N.: Casting your model: Learning to localize improves self-supervised representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11058–11067 (2021)
35. Shu, Y., van den Hengel, A., Liu, L.: Learning common rationale to improve self-supervised representation for fine-grained visual recognition problems. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11392–11401 (2023)
36. Shu, Y., Yu, B., Xu, H., Liu, L.: Improving fine-grained visual recognition in low data regimes via self-boosting attention mechanism. In: European Conference on Computer Vision. pp. 449–465. Springer (2022)
37. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: Yfcc100m: The new data in multimedia research. Communications of the ACM **59**(2), 64–73 (2016)
38. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)
39. Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: International Conference on Machine Learning. pp. 9929–9939. PMLR (2020)
40. Wang, Z., Wang, Y., Hu, H., Li, P.: Contrastive learning with consistent representations. arXiv preprint arXiv:2302.01541 (2023)
41. Xiao, T., Wang, X., Efros, A.A., Darrell, T.: What should not be contrastive in contrastive learning. arXiv preprint arXiv:2008.05659 (2020)
42. Yao, Y., Ye, C., He, J., Elsayed, G.F.: Teacher-generated spatial-attention labels boost robustness and accuracy of contrastive models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23282–23291 (2023)
43. Yeh, C.H., Hong, C.Y., Hsu, Y.C., Liu, T.L.: Saga: Self-augmentation with guided attention for representation learning. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3463–3467. IEEE (2022)
44. Zhao, N., Wu, Z., Lau, R.W., Lin, S.: Distilling localization for self-supervised representation learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 10990–10998 (2021)
45. Ziyin, L., Lubana, E.S., Ueda, M., Tanaka, H.: What shapes the loss landscape of self-supervised learning? arXiv preprint arXiv:2210.00638 (2022)