# Contrastive Learning Relies More on Spatial Inductive Bias Than Supervised Learning: An Empirical Study

Yuanyi Zhong[1†] Haoran Tang[2†] Jun-Kun Chen[1†] Yu-Xiong Wang[1]
[1]University of Illinois at Urbana-Champaign [2]University of Pennsylvania [†]Equal Contribution
{yuanyiz2, junkun3, yxw}@illinois.edu thr99@seas.upenn.edu

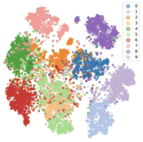
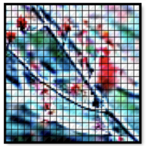
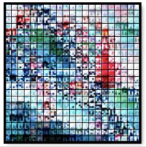
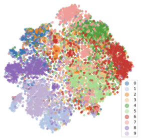
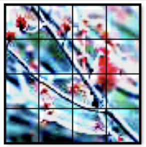
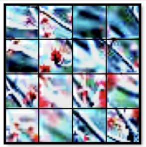
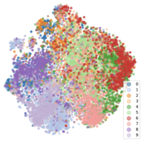
| Pre-Train Dataset | | | | Supervised Learning | Contrastive Learning | CL Feature Space | |
|---|---|---|---|---|---|---|---|
| | | | | | | Uniformity | Visualization |
| | Original Dataset | | | 89.53 | ≈ 88.73 | 2.82 |  |
|  | Destroy Local Spatial Structure ⟹ |  | Mild Corruption | 77.49<br>13% ↓ | 76.90<br>13% ↓ | 2.62<br>7% ↓ |  |
| | | | Severe Corruption | 65.94<br>26% ↓ | 63.35<br>29% ↓ | 2.42<br>14% ↓ | |
|  | Destroy Global Spatial Structure ⟹ |  | Mild Corruption | 76.06<br>15% ↓ | 67.18<br>24% ↓ | 2.41<br>15% ↓ |  |
| | | | Severe Corruption | 65.88<br>26% ↓ | 60.51<br>32% ↓ | 2.45<br>13% ↓ | |

Figure 1: We study the reliance on spatial inductive bias for both contrastive learning (CL) and supervised learning (SL), by pre-training them on dataset with destroyed global or local spatial structure to prevent such inductive bias, and compare the performance drop in evaluation. Our experiments show that compared with SL, CL has a higher performance drop without either global or local spatial inductive bias, which has a significantly lower uniformity and a more messy, entangled, and non-separable feature space (Sec. 4.3), clearly leading to the conclusion that CL relies more on spatial inductive bias than SL.

## Abstract

*Though self-supervised contrastive learning (CL) has shown its potential to achieve state-of-the-art accuracy without any supervision, its behavior still remains under-investigated. Different from most previous work that understands CL from learning objectives, we focus on an unexplored yet natural aspect: the spatial inductive bias which seems to be implicitly exploited via data augmentations in CL. We design an experiment to study the reliance of CL on such spatial inductive bias, by destroying the global or local spatial structures of an image with global or local patch shuffling, and comparing the performance drop between experiments on original and corrupted dataset to quantify the reliance on certain inductive bias. We also use the uniformity of feature space to further research how CL-pre-trained models behave with the corrupted dataset. Our results and analysis show that CL has a much higher reliance on spatial inductive bias than SL, regardless of specific CL algorithm or backbones, opening a new direction for studying the behavior of CL.*

## 1. Introduction

Self-supervised contrastive learning (CL) has demonstrated tremendous potential in learning generalizable representations from unlabeled datasets [4, 18, 16, 2, 8, 35] in recent years. Current state-of-the-art CL algorithms learn representations from ImageNet [11] that match or even exceed the accuracy of their supervised learning (SL) counterparts on ImageNet and downstream tasks. Understanding the behavior and the power of contrastive learning is therefore becoming an interesting and crucial topic in academia. Previous work on understanding CL [6, 29, 27] mostly focuses on the *loss functions*, while few work focuses on the data-centric inductive bias in CL – which information CL focuses on exploiting when lacking semantic supervision.

A typical CL method learns the semantic information of images by applying data augmentation-based contrasts, which generate variants of images with the same semantic information via specific random data augmentations, and trains the model toward distinguishing images from different variants. Data augmentation is crucial in a CL algorithm, which provides diversified variants for the model to learn the invariant semantic information. The choice of data augmentation includes random resized cropping, color jitter, Gaussian blurry, flipping, etc, where only random resized cropping (RSC) changes the geometry and the display scopes of the image, being the data augmentation that alters the image most severely. With RSC, the model is required to fetch the same semantic feature with two different scopes of an image, requiring a special focus on *spatial* information. A natural suspicion therefore arises: Does CL highly rely on *spatial* inductive bias? Furthermore, does CL relies more on this than supervised learning?

This paper focuses on the spatial inductive bias in learning algorithms including SL and CL. To investigate the inductive bias on spatial information, we use two types of data corruptions, local and global patch shuffling [15, 23, 33, 20], to respectively *destroy* the local and global information of an image preventing the learning algorithm from using them. For each corruption setting with a fixed patch size, we sample one specific corruption, and apply such deterministic corruption operation for all images in the dataset to construct a corrupted dataset. We pre-train a backbone model on the training split of such dataset with a learning algorithm, and evaluate the accuracy on the testing split. We quantify the inductive bias reliance with the performance drop between the testing accuracy of the original and the corrupted datasets.

We conduct extensive experimental results on both CIFAR-10 [22] and large-scale ImageNet [11] datasets, for SL and commonly-used CL algorithms with both CNN-based and Transformer-based backbone models. We obtain a consistent backbone-independent conclusion from our empirical study, that CL does rely more on spatial inductive bias than CL, while higher reliance on global spatial information than local. We also advise that SL does not have such clear reliance on other inductive biases studied with other types of corruptions, *e.g*. gamma distortion, by showing that no clear difference in performance drops of CL and SL occurs under different observations. To deeper analyze the impact of the corruptions, we use the uniformity metric proposed in [30] to quantify the potential of unsupervised classification ability. We observe that there is a high uniformity drop in the corrupted dataset with patch shuffling, higher than the uniformity drop in the dataset under other corruptions, showing that the CL relies most on spatial inductive bias.

**Our contributions** are three-folded. (1) We perform extensive empirical studies on different datasets, for various learning algorithms with both CNN and Transformer backbones, and show that CL has a clearly higher inductive bias on spatial information than SL. (2) We propose the method to evaluate the spatial inductive bias for learning algorithms with patch shuffling data corruptions, and use uniformity as a metric to further quantify the loss of unsupervised classification ability for SL methods trained with no spatial information. (3) We offer analyses and explanations for such observations, inspiring academia with a new direction for research on various types of inductive biases. Our work is *the first work* to observe, define, and analyze the dependency of CL on spatial inductive bias, as an *intrinsic property* of CL, via a *systematic* experiment design, showing novel insights on the mechanism of CL.

## 2. Related Work

**Self-Supervised Learning (SSL) and Contrastive Learning (CL).** Remarkable progress has been made in self-supervised representation learning from unlabeled datasets [4, 18, 16, 2, 8]. This paper focuses on a particular kind of SSL algorithm, contrastive learning, that learns augmentation invariance with a Siamese network. To prevent trivial solution, contrastive learning pushes negative examples apart (MoCo [18, 7, 9], SimCLR [4, 5]), makes use of stop-gradient operation or asymmetric predictor without using negatives (BYOL [16], SimSiam [8], DINO [3]), or leverages redundancy reduction (BarlowTwins [32]) and clustering (DeepCluster-v2 and SwAV [2]). In addition to augmentation invariance, generative pre-training [26, 1, 17] and visual-language pre-training [25] are promising ways to learn transferable representations.

**Understanding SSL and CL.** There is a growing body of literature on understanding SSL. [29] decomposes the contrastive objective into alignment (between augmentations) and uniformity (across entire feature space) terms. Uniformity can be thought of as an estimate of the feature entropy, which we use to study the feature space dynamics during training. [30] makes the connection between uniformity and the temperature parameter in contrastive loss, and finds that a good temperature can balance uniformity and tolerance of semantically similar examples. [34] discovers that SSL transferring better than SL can be due to better low- and mid-level features, and the intra-class invariance objective in SL weakens transferability by causing more pre-training and downstream task misalignment. [13] studies the downstream task accuracy of a variety of pre-trained models and finds that SSL outperforms SL on many tasks. [10] investigates the impact of pre-training data size, domain quality, and task granularity on downstream performance. [6] identifies three intriguing properties of CL: a generalized version of the loss, learning with the presence of multiple objects, and feature suppression induced by competing augmentations. Our work falls into the same line of research that attempts to understand SSL better. However, we investigate from the angle of *spatial inductive bias* exploited by SL and CL.

**Patch Shuffling.** Recent work on self-supervised learning starts analyzing input data with patch shuffling. MAE [17] drops image patches as masked parts to train an autoencoder for reconstruction. Following the same approach, [31] introduces more guidance on masking via Grad-CAM [28] and refills the masked parts with patches at the same position from images in the same batch. We can consider such an approach as batch-level patch shuffling, which improves out-of-distribution robustness by constructing counterfactual samples. Patching shuffling within a single image is also adopted as one way to augment negative samples in a recent study [15]. In an empirical study [33], pixel shuffling is used to study the generalization of deep supervised learning. In our study, we apply our designed patch shuffling to destroy spatial information of the dataset, and create an ablated version to help understand contrastive feature learning.

# 3. Method

## 3.1. Global/Local Patch Shuffling

We investigate the reliance of learning algorithms on spatial inductive bias, by evaluating the performance drop when trained and evaluated on a *corrupted* dataset with only spatial structure and information destroyed. To obtain such a dataset, we introduce random patch shuffling previously studied in [15, 23, 33, 20]. Consistent with previous work, with a preset patch size $P$ divisible by $W$ and $H$, where $W \times H$ is the image size of all images in a dataset, we uniformly cut the image into $(W/P) \times (H/P)$ patches of $P \times P$. We then define global and local shuffling as follows visualized in Fig. 1. In our experiments, we use Global/Local $N$ to indicate a global or local patch shuffling with $N \times N$ patches, where $N = W/P = H/P$.

- A **global** patch shuffling operation is defined as a permutation $\sigma$ of the $(W/P) \times (H/P)$ patches, representing the operation to *permute patches* in the permutation order of $\sigma$. It destroys global spatial structures, *e.g.*, each part's absolute location w.r.t. the whole image, making the corrupted image look like a jigsaw puzzle. The image becomes less structured with *smaller* patch size.
- A **local** patch shuffling operation is defined as a permutation $\sigma$ of $(P \times P)$ pixels in each patch, representing the operation to *permute pixels of each patch* in the *same* permutation order of $\sigma$. It destroys local spatial structures within each patch, making the corrupted image look like a blur version of the image. As the opposite, the image becomes less structured with *larger* patch size.

## 3.2. Evaluation Scheme and Metrics

We conduct experiments on various global or local shuffling operations with different patch size settings on various datasets. For each corruption setting, we *first* randomly sample the permutation $\sigma$, then apply the patch shuffling w.r.t. $\sigma$ for *each* image in the dataset, including both training and testing split. We use such a more *deterministic* way to corrupt dataset in a more stable, controllable way, while obtaining consistent observations as the fully random implementation (i.e., use different random permutations for different images) in Sec. 4.6.

We pre-train the backbone on the corrupted training set of images with the learning algorithm we study, then evaluate the pre-trained backbone on the corrupted testing set with linear evaluation, to evaluate the model within the same domain. In Sec. 4.5, we show that this setting faithfully reflects the influence of destroyed spatial structure without being affected by domain shift.

Our main metric to quantify the reliance on a certain inductive bias is the performance drop rate $\Delta_A$, defined on a pair of a certain corruption $c$ and a learning algorithm (with a backbone architecture of either CNN-based ResNet [19] or Transformer-based ViT [12]). Denote the accuracy of the original dataset and corrupted dataset are $A$ and $A_c$ respectively, we define $\Delta_A$ as $\frac{A-A_c}{A}$. We compare the $\Delta$ of SL and CL under the same corruption, to find out whether SL relies more on the inductive bias that is unable to be utilized under the same corruption.

To further analyze the reliance of inductive bias in SL, we further use the uniformity of feature space studied in [30] as a metric. Uniformity measures how uniformly distributed the features learned by a learning algorithm, which is a potential of how well a linear classifier is able to classify the images, implicitly maximized in the training of SL. We also consider the uniformity drop rate $\Delta_U$ defined as $\frac{U-U_c}{U}$, where $U$ and $U_c$ are the uniformity of the original and $c$-corrupted dataset respectively, as another clue of how SL relies on the destroyed inductive bias.

## 3.3. Experiment Setup

**Models and Algorithms.** We benchmark a variety of self-supervised contrastive learning algorithms. These methods are carefully sampled to be representative. They include contrastive learning with negatives: SimCLR-v2 [4, 5], MoCo-v2 [18, 7]; without negatives: SimSiam [8], the momentum based, BYOL [16]; with redundancy reduction: BarlowTwins [32]; and with clustering assignments: DeepCluster-v2 [2], SwAV [2]. We test both CNN (standard ResNet-18/50 [19]) and Vision Transformer (ViT) [12] backbones. For transformers, we leverage pre-trained models on ImageNet [11] from vanilla ViT [12], DINO [3], and MoCo-v3 [9].

**Datasets.** We pre-train on CIFAR-10 [22], ImageNet [11] and its variants to evaluate the performance drop of CL and SL and uniformity of SL, under certain data corruptions. We keep the CIFAR-10 images at original size 32 × 32, but up-sample ImageNet images to 256 × 256 to support more patch sizes.

For fair comparisons, we use the same data augmentations across methods and datasets when we need to train any model.

# 4. Results and Analysis

## 4.1. Main results: CL relies more on both global and local spatial inductive bias than SL

We show the results of CIFAR-10 in Tab. 1 and ImageNet in Tab. 2. The results of both datasets show that SL algorithms consistently rely more on spatial inductive bias than CL, regardless of specific SL algorithm or backbone architecture. There is a clear gap in average performance drop $\Delta_A$ of at least $4\%$ between SL and CL in CIFAR-10 experiments, and an even more significant gap of at least $15\%$ in ImageNet Experiments.

Between spatial and global spatial inductive biases, the SL relies more on global inductive biases than local ones. As clearly shown in Tab. 2, even with the same size of permutations of 16 ($4 \times 4$ or $\frac{256}{64} \times \frac{256}{64}$ in global and local shuffling), there was a huge gap of at least $35\%$ between $\Delta_A$ these two settings. In addition to experiment results, we also visualize the Grad-CAM activation maps [28] of pre-trained models. Fig. 2 shows the activation mapping of both SL and MoCo-v2 pre-trained on both clean and globally corrupted ImageNet-100. We can visually notice the larger impact of corrupting such spatial inductive bias on CL methods, that MoCo-v2 trained with corrupted images has the most diffused and messy attention maps, while SL trained with the same setting still obtains similar attention maps as the model trained with original images. This shows in the opposite way that a model that learned to exploit spatial inductive bias during pre-training can even deal well with images with destroyed spatial structure.

Beyond the above main results, we perform thorough empirical analysis from various perspectives to trace down the potential causes of our observations. Further experiments are designed and presented to evaluate the impact of different data information, backbone architectures, domain, and implementations. We empirically verify that CL relies more on spatial inductive bias attacked by our designed corruptions, with good consistency and stability.

## 4.2. Why not use other corruptions?

Previous work has proposed several other corruptions from a visual or adversarial perspective, *e.g*. gamma distortion, flipping, and JPEG compressing [20]. However, most of these corruptions modify the image globally while keeping the spatial structure unchanged. Even if the image's appearance may change a lot, most of the absolute or relative spatial relationships between parts in the images remain unchanged or change in a well-predictable pattern. We therefore cannot use these corruptions to destroy spatial information. Also as shown in the next section and last two columns of Tab. 3, these corruptions do not affect much on CL.

## 4.3. What happens in CL when the dataset is corrupted?

Previous work [30] has verified that contrastive learning objective favors and thus implicitly optimizes towards a uniformly distributed feature space to preserve information of the data maximally, and has proposed an approximate of feature uniformity as log-mean of Gaussian potentials. We can utilize this quantification as a metric to evaluate the effectiveness of data corruptions to CL, where a feature space with high uniformity is easier to be linearly separated well in downstream tasks, showing a more powerful model, and potentially leading to better accuracy. As shown in Fig. 3, with the same backbone and learning algorithm, the model trained on the original dataset with high uniformity has a clearer boundary of classes in classification, and therefore can be classified better.

Rewriting with normalized features, we compute the uniformity of a model as

$$U(f_t, \mathcal{D}) = -\log \mathbb{E}_{x_0,x_1 \sim \mathcal{D}} \left[ e^{-2\|f_t(x_0)-f_t(x_1)\|_2^2} \right], \quad (1)$$

where $f_t$ is the network at epoch $t$, $\mathcal{D}$ is the dataset, and $x_0, x_1$ are images sampled from the dataset. Therefore, we expect an effective data corruption to shrink the contrastive feature space and output a smaller uniformity.

As shown in Tab. 3, there is a huge uniformity drop of at least $7\%$ for each global or local patch shuffling, showing that the obtained model is highly degraded without the inductive bias of spatial structures, while global spatial structures are more important to local ones with a higher uniformity drop. Also as shown in Fig. 3, the models trained on datasets with destroyed space structures have a much lower uniformity and a more messy, entangled, and non-separable feature space and cannot be classified well, where global patch shuffling has a higher impact than local, while the model trained on dataset under other corruptions (defocusing blur) has a similar uniformity and a feature space with clear boundaries between classes.

We also observe that as SL does not implicitly optimizes uniformity, there is no pattern between its uniformity and corruption. Corruptions other than spatial patch shuffling, including gamma distortion and 'natural corruptions' (shot noise, defocusing blur, and JEPG compression) of ImageNet-C ('IN-C') [20] do not lead to an obvious drop of the uniformity, showing that these corruptions are not

Table 1: The experiment results on CIFAR-10 with ResNet18/50 and ViT-Small/Base backbones ('B.B.') show that all CL algorithms consistently rely higher on both local and global spatial inductive bias, regardless of the backbone architecture. Supervised contrastive learning (SupCon) as the mixture of CL and SL has a halfway performance drop between CL and SL. Each number within the bracket is the $\Delta_A$ under the corruption of its column. The darkness of the color represents the severity of each performance drop.

| B.B. | Alg. | Orig. | Global 4 | | Global 8 | | Local 4 | | Local 8 | | Avg. $\Delta_A$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ResNet18 | Sup | 89.53 | 76.06 | (15.0%↓) | 65.88 | (26.4%↓) | 65.94 | (26.3%↓) | 77.49 | (13.4%↓) | 20.3% |
| | SupCon | 86.09 | 70.14 | (18.5%↓) | 60.46 | (29.8%↓) | 64.19 | (25.4%↓) | 73.36 | (14.8%↓) | 22.1% |
| | MoCo-v2 | 88.73 | 67.18 | (24.3%↓) | 60.51 | (31.8%↓) | 63.35 | (28.6%↓) | 76.90 | (13.3%↓) | 24.5% |
| | BYOL | 88.39 | 67.47 | (23.7%↓) | 60.63 | (31.4%↓) | 62.64 | (29.1%↓) | 75.15 | (15.0%↓) | 24.8% |
| | Barlow | 88.89 | 68.34 | (23.1%↓) | 61.13 | (31.2%↓) | 62.53 | (29.7%↓) | 75.28 | (15.3%↓) | 24.8% |
| | DINO | 84.75 | 64.26 | (24.2%↓) | 55.83 | (34.1%↓) | 58.57 | (30.9%↓) | 68.96 | (18.6%↓) | 27.0% |
| ResNet50 | Sup | 89.24 | 74.75 | (16.2%↓) | 62.85 | (29.6%↓) | 62.71 | (29.7%↓) | 75.59 | (15.3%↓) | 22.7% |
| | MoCo-v2 | 88.78 | 67.38 | (24.1%↓) | 59.19 | (33.3%↓) | 61.30 | (31.0%↓) | 75.56 | (14.9%↓) | 25.8% |
| Vit-Small | Sup | 76.04 | 63.18 | (16.9%↓) | 55.72 | (26.7%↓) | 58.38 | (23.2%↓) | 67.90 | (10.7%↓) | 19.4% |
| | MoCo-v3 | 73.84 | 55.16 | (25.3%↓) | 47.62 | (35.5%↓) | 55.61 | (24.7%↓) | 64.32 | (12.9%↓) | 24.6% |
| | DINO | 62.68 | 46.39 | (26.0%↓) | 41.48 | (33.8%↓) | 50.13 | (20.0%↓) | 56.88 | (9.3%↓) | 22.3% |
| Vit-Base | Sup | 74.35 | 60.00 | (19.3%↓) | 54.05 | (27.3%↓) | 57.27 | (23.0%↓) | 66.60 | (10.4%↓) | 20.0% |
| | MoCo-v3 | 73.18 | 55.49 | (24.2%↓) | 47.53 | (35.1%↓) | 55.56 | (24.1%↓) | 63.64 | (13.0%↓) | 24.1% |

Table 2: The experiment results on large-scale ImageNet with ResNet50 and ViT-Small/Base backbones show that CL relies higher on the spatial inductive bias, consistent with the conclusions from CIFAR-10 experiments. With the same size of permutations, global patch shuffling causes significantly higher performance drops than local patch shuffling.

| B.B. | Alg. | Orig. | Global 4 | | Local 64 | | Avg. $\Delta_A$ |
|---|---|---|---|---|---|---|---|
| ResNet50 | Sup | 71.79 | 62.59 | (12.8%↓) | 66.24 | (7.7%↓) | 10.3% |
| | MoCo-v2 | 64.06 | 35.02 | (45.3%↓) | 57.63 | (10.0%↓) | 27.7% |
| | BYOL | 64.19 | 35.00 | (45.5%↓) | 58.72 | (8.5%↓) | 27.0% |
| Vit-Small | Sup | 67.43 | 52.80 | (21.7%↓) | 62.66 | (7.1%↓) | 14.4% |
| | MoCo-v3 | 56.04 | 26.35 | (53.0%↓) | 51.54 | (8.0%↓) | 30.5% |
| Vit-Base | Sup | 68.73 | 53.77 | (21.8%↓) | 64.14 | (6.7%↓) | 14.2% |
| | MoCo-v3 | 57.79 | 25.04 | (56.7%↓) | 48.61 | (15.9%↓) | 36.3% |

Table 3: The results of uniformity of CL with ResNet18 backbone show that destroying spatial inductive bias cause a significant uniformity drop in the CIFAR-10 dataset, where global has a higher impact than local. On the contrary: (1) There is no obvious pattern in the uniformity of SL, and (2) other corruptions including gamma distortion ('$\gamma$') and ImageNet-C [20] natural corruptions ('IN-C') do not cause a clear drop in uniformity.

| Method | Ori. | Global 4 | | Global 8 | | Local 4 | | Local 8 | | Avg. $\Delta_U$ | $\gamma$0.2 | | Avg. IN-C | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sup | 2.11 | 2.19 | (3.8%↑) | 2.15 | (1.9%↑) | 2.28 | (8.1%↑) | 2.33 | (10.4%↑) | -6.0% | 2.07 | (1.9%↓) | 2.21 | (4.7%↑) |
| MoCo-v2 | 2.82 | 2.41 | (14.5%↓) | 2.45 | (13.1%↓) | 2.42 | (14.2%↓) | 2.62 | (7.1%↓) | 12.2% | 2.66 | (5.7%↓) | 2.68 | (5.0%↓) |

critical for CL, consistent with the discussions in the previous section.

Fig. 4 also provides a temporal glimpse of uniformity through training time. We compare the uniformity of MoCo-v2 [7], supervised contrastive learning (SupCon) [21], and supervised learning, during pre-training on CIFAR-10. We are interested in SupCon because it bridges CL and SL by leveraging a similar contrastive loss. As illustrated, the overall feature uniformity of MoCo-v2 is greater than 2.5 and approaching 3, while the overall uniformity of SupCon and supervised methods range from 1.25 to 2.2. Note that the class-wise uniformity of MoCo-v2 is also increasing. This further shows that features from CL methods are more uniformly distributed.

### 4.4. Are Transformer's results reasonable and consistent with CNN's?

Vision transformer [12] is a highly different backbone from the CNN's, which divides images into patches to sequentialize the image and apply a transformer encoder, where the patchify operation is the same as what we do for patch shuffling. A natural question arises: Does the patch size of the transformer model affects the results obtained by patch shuffling? Or more specifically, is the results on the

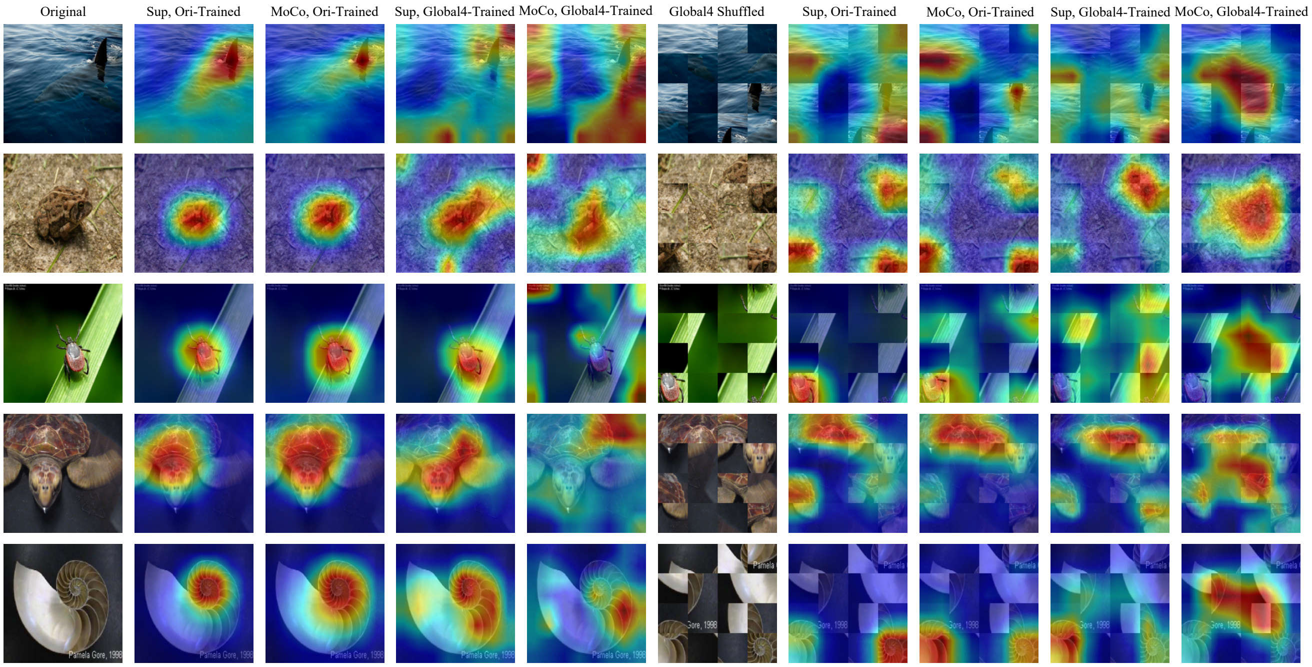


Figure 2: For several ImageNet-100 images, we visualize the Grad-CAM activation maps of 4 models: SL ('Sup') and MoCo-v2 ('MoCo') with ResNet backbone trained on original images, along with SL and MoCo-v2 trained on images corrupted by global patch shuffling. The activation maps of MoCo-v2 trained with corrupted images are mostly different from the reference (SL or MoCo-v2 trained with original ImageNet), and are the most diffused and messy ones.

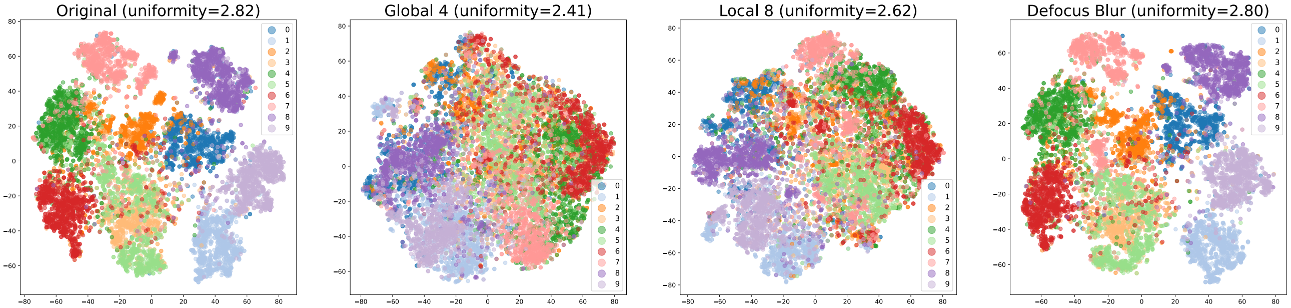


Figure 3: t-SNE visualizations of features from MoCo-v2 Original (1st), Global 4 (2nd), Local 8 (3rd) and Defocus Blur (4th) pre-trained on CIFAR-10. We observe that corrupting spatial inductive bias can effectively undermine a well-clustered feature space.

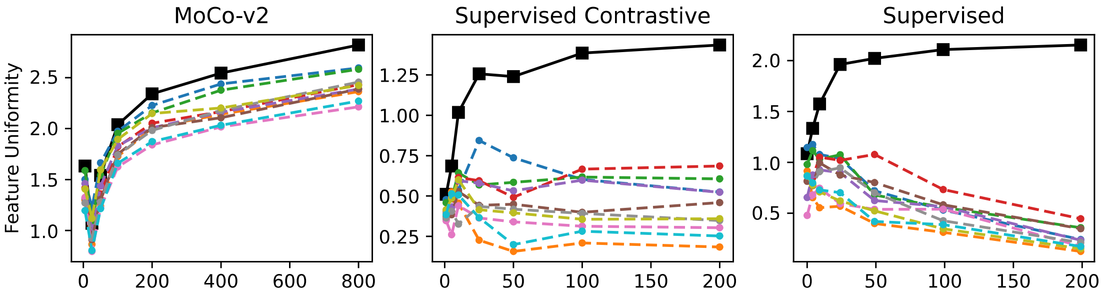


Figure 4: Uniformity of MoCo-v2, supervised contrastive learning, and SL with training epochs on CIFAR-10. Solid black line – uniformity of the overall feature space. Dashed lines – class-wise feature uniformities of the 10 classes. While the overall uniformity of all methods grows, the uniformity of each class of Sup or SupCon is shrinking as training progresses. In the end, the overall uniformity of MoCo is the largest.

transformer consistent with the results on CNN?

To investigate this issue, we conduct extra experiments on the vision transformer, under global and local patch shuffling with the same patch size (16) as the transformer. In Tab. 4, we show that the conclusion is consistent with the observation from our main results. The transformer-based backbones did not achieve irregular results dealing with patch shuffling at the same patch size, since the data augmentations of both SL and CL are applied *after* the data corruption. The random resized cropping operation replaces the image with a smaller scope, making the patches of corruption unaligned with the patches used by ViT.

### 4.5. Does the performance drop result from domain shift?

We further investigate whether CL's higher dependency on spatial inductive bias comes from the domain shift by data corruptions. To draw more significant and clear conclusions, we perform the proposed corruptions on downstream tasks, where domain shifts are more explicit. Adopting CL and SL models well pre-trained on ImageNet, we conduct linear fine-tuning on corrupted downstream datasets. Tab. 5 demonstrates that the effect of domain shift is minimal: during downstream tasks, destroying spatial inductive bias is less effective to CL than SL, which means the pre-trained contrastive model has sufficiently exploited the spatial inductive bias to learn strong representations against our corruptions. Specifically, CL methods have the worst 14.1% with ResNet-18 and 32.8% with ViT-Small, while SL has 16.3% and 39.3% respectively. Our observations thus indicate the higher importance of spatial inductive bias to CL, and the domain shift does not affect our conclusion.

Also, as shown in Fig. 2, both SL and CL trained with original images still obtain consistent and reasonable attention maps even when evaluated on the corrupted images where a domain shift occurs, showing that both of the methods have already learned to exploit spatial inductive bias through the pre-training with original images. On the contrary, with the corrupted images for pre-training, SL still achieves similar attention maps as trained with original images, while CL outputs messy and inconsistent attention maps even for the original images, showing that CL pre-trained without spatial inductive bias is unable to understand the content of the image.

### 4.6. Why use the same corruption for all images?

The patch shuffling we implemented randomly samples one fixed permutation as the shuffling order of the whole dataset, so that each image in the dataset is corrupted in the same way. One may argue that such corruption introduces a very large bias, and the corruptions should follow the augmentation pipeline to apply independently randomly to each image, so that different images can be corrupted in a different way. This violates the common belief that considers data source as a black box, and data augmentations can be considered as the pre-processing step of the proposed method. We still investigate whether a totally random permutation for each individual data brings different observations, and record the results on CIFAR-10 in Tab. 6. We perform random permutations on Global 4 and Local 8, and compare the results with those in Tab. 1. While random permutation leads to different performance degradation, CL is still more vulnerable to corrupting spatial inductive bias, which holds for both CNN and Transformer-based backbones.

### 4.7. Interactions between corruption and data augmentation in SL and CL

One may also argue that the data augmentations can be regarded as an extrapolation of the dataset, so corruptions should be made on top of this extrapolated dataset. We hold the same reason as the previous section for the proposed corruption-augmentation order, that we consider the corrupted data as the new black-box source, followed by the learning algorithm starting with data augmentation. Also, augment then corrupt ('aug-corrupt') makes the images input to the model contain a fixed pattern of the shuffled patches, that the patches are always aligned for different images. This gives the learning algorithm the potential to directly learn the pattern of the corruptions and overfit to it, and even makes transformer-based backbones invariant to such corruptions with a patch size aligned with Transformer's, deviating from our goal only to corrupt global or local spatial inductive bias. We thus choose our design in this way.

We also conduct experiments on more data augmentation settings, including this 'aug-corrupt' setting as shown in Tab. 7, for further understanding the effect of data augmentation. By comparing the results of 'corrupt-aug' and 'aug-corrupt' in both CL and SL, we observe that the results in 'aug-corrupt' setting are consistently and significantly better than 'corrupt-aug', showing that both SL and CL learn to model and even inverse the fixed, aligned corruption, which deviates from our goal to study only the spatial inductive bias. We also observe that SL in 'no-aug' setting achieves lower accuracy in original dataset but higher accuracy in corrupted dataset, which shows that *data augmentations (especially random resized cropping) promote the model to utilize spatial inductive bias in pre-training*, which helps the accuracy in original dataset, but harms in corrupted dataset. As data augmentation is necessary and crucial in CL to generate positive examples for contrastive training, such spatial inductive bias is implicitly highly utilized and relied on.

Table 4: The experiment results of patch shuffling with the same patch size as ViT (16) shows that the patch size is still consistent with previous conclusions, that CL relies much more on spatial inductive bias than SL.

| Alg. | Orig. | Global 4 | Global 16 | Local 16 | Local 64 | Avg. $\Delta_A$ |
|---|---|---|---|---|---|---|
| Sup | 67.43 | 52.80 (21.7%↓) | 40.00 (40.7%↓) | 48.77 (27.7%↓) | 62.66 (7.1%↓) | 24.3% |
| MoCo-v3 | 56.04 | 26.35 (53.0%↓) | 13.20 (76.4%↓) | 36.69 (34.5%↓) | 51.54 (8.0%↓) | 43.0% |

Table 5: We adopt pre-trained models on ImageNet and sufficient fine-tune to CIFAR-10 with proposed shuffling. Despite the worse performance on ViT-Small backbone, well-trained CL methods are less vulnerable to the corruptions of spatial inductive bias during downstream tasks.

| B.B. | Alg. | Orig. | Global 4 | Global 8 | Local 4 | Local 8 | Avg. $\Delta_A$ |
|---|---|---|---|---|---|---|---|
| ResNet18 | Sup | 96.7 | 88.2 (8.8%↓) | 77.5 (19.8%↓) | 72.0 (25.5%↓) | 86.0 (11.0%↓) | 16.3% |
| | MoCo-v2 | 96.8 | 89.6 (7.5%↓) | 81.5 (15.9%↓) | 77.4 (20.1%↓) | 89.4 (7.7%↓) | 12.8% |
| | BYOL | 96.5 | 88.8 (7.9%↓) | 80.2 (16.9%↓) | 75.2 (22.0%↓) | 87.4 (9.4%↓) | 14.1% |
| | Barlow | 96.8 | 96.7 (0.1%↓) | 94.4 (2.5%↓) | 87.9 (9.2%↓) | 76.4 (21.0%↓) | 8.2% |
| Vit-Small | Sup | 94.2 | 64.1 (32.0%↓) | 52.6 (44.2%↓) | 52.5 (44.2%↓) | 59.6 (36.7%↓) | 39.3% |
| | DeiT (Sup) | 95.4 | 73.2 (23.2%↓) | 59.5 (37.6%↓) | 53.1 (44.3%↓) | 59.7 (37.5%↓) | 35.7% |
| | DINO | 96.7 | 78.0 (19.3%↓) | 64.6 (33.2%↓) | 60.8 (37.1%↓) | 68.0 (29.6%↓) | 29.8% |
| | MoCo-v3 | 96.2 | 75.3 (21.7%↓) | 61.1 (36.4%↓) | 57.6 (40.1%↓) | 64.4 (33.0%↓) | 32.8% |
| | MAE | 77.1 | 61.3 (20.5%↓) | 55.1 (28.5%↓) | 53.3 (30.8%↓) | 57.0 (26.0%↓) | 26.5% |

Table 6: Using different random patch shuffling operations for different images ('Rand. Glo./Loc.') has similar behaviors and conclusions as our current setting to use the same patch shuffling operation.

| B.B. | Alg. | Orig. | Global 4 | Local 8 | Rand. Glo. 4 | Rand. Loc. 8 |
|---|---|---|---|---|---|---|
| ResNet18 | Sup | 89.53 | 76.06 (15.0%↓) | 77.49 (13.4%↓) | 70.98 (20.7%↓) | 76.76 (14.3%↓) |
| | MoCo-v2 | 88.73 | 67.18 (24.3%↓) | 76.90 (13.3%↓) | 66.31 (25.3%↓) | 73.68 (17.0%↓) |
| Vit-Small | Sup | 76.04 | 63.18 (16.9%↓) | 67.90 (10.7%↓) | 52.36 (31.1%↓) | 68.13 (10.4%↓) |
| | MoCo-v3 | 73.84 | 55.16 (25.3%↓) | 64.32 (12.9%↓) | 46.96 (36.4%↓) | 63.5 (14.0%↓) |

Table 7: Experiments on data augmentation settings show that (1) both SL and CL learn to model and even inverse the corruption in the 'aug-corrupt' setting, thus obtaining better results than the 'corrupt-aug' setting we used; and (2) in the 'no-aug' setting, SL achieves lower accuracy in original dataset but higher accuracy in corrupted dataset, showing that data augmentation promotes the model to utilizes spatial inductive bias. The dataset is CIFAR-10.

| B.B | Alg. | Setting | Orig. | Global 4 | Global 8 | Local 4 | Local 8 | Avg $\Delta_A$ |
|---|---|---|---|---|---|---|---|---|
| ResNet18 | Sup | corrupt-aug (ours) | 89.53 | 76.06 (15.0%↓) | 65.88 (26.4%↓) | 65.94 (26.3%↓) | 77.49 (13.4%↓) | 20.3% |
| | | no-aug | 87.66 | 77.37 (11.7%↓) | 71.86 (18.0%↓) | 73.30 (16.4%↓) | 82.34 (6.0%↓) | 13.0% |
| | | aug-corrupt | 92.23 | 85.92 (6.8%↓) | 80.58 (12.6%↓) | 83.61 (9.4%↓) | 89.96 (2.5%↓) | 7.8% |
| | MoCo-v2 | corrupt-aug (ours) | 82.55 | 65.43 (17.1%↓) | 59.49 (27.9%↓) | 59.62 (27.8%↓) | 70.14 (15.0%↓) | 22.0% |
| | | aug-corrupt | 82.55 | 77.63 (6.0%↓) | 73.48 (11.0%↓) | 78.12 (5.4%↓) | 81.25 (1.6%↓) | 6.0% |
| ViT-Small | Sup | corrupt-aug (ours) | 76.04 | 63.18 (16.9%↓) | 55.72 (26.7%↓) | 58.38 (23.2%↓) | 67.90 (10.7%↓) | 19.4% |
| | | no-aug | 56.47 | 54.90 (2.8%↓) | 54.16 (4.1%↓) | 57.27 (-1.4%↓) | 57.23 (-1.3%↓) | 1.0% |
| | | aug-corrupt | 76.04 | 74.18 (2.4%↓) | 68.41 (10.0%↓) | 75.43 (0.8%↓) | 75.49 (0.7%↓) | 3.5% |
| | MoCo-v3 | corrupt-aug (ours) | 73.84 | 55.16 (25.3%↓) | 47.62 (35.5%↓) | 55.61 (24.7%↓) | 64.32 (12.9%↓) | 24.6% |
| | | aug-corrupt | 73.84 | 68.29 (7.5%↓) | 64.20 (13.1%↓) | 72.60 (1.7%↓) | 73.39 (0.6%↓) | 5.7% |

Table 8: There is a very low difference between the experimental results with and without specific hyperparameter-tuning or further training, showing that our experiments are not sensitive to hyperparameters.

| Model | $\gamma$0.2 | Global 4 | Global 8 | Local 4 | Local 8 | Avg. $\Delta_A$ |
|---|---|---|---|---|---|---|
| Sup (89.53) | 87.36 (2.4%↓) | 76.06 (15.0%↓) | 65.88 (26.4%↓) | 65.94 (26.3%↓) | 77.49 (13.4%↓) | 16.7% |
| Sup-tuned | 87.46 (2.3%↓) | 76.90 (14.1%↓) | 66.24 (26.0%↓) | 66.43 (25.8%↓) | 77.93 (13.0%↓) | 16.2% |
| MoCo-v2 (88.73) | 85.84 (3.3%↓) | 67.18 (24.3%↓) | 60.51 (31.8%↓) | 63.35 (28.6%↓) | 76.90 (13.3%↓) | 20.3% |
| MoCo-v2-tuned | 86.17 (2.9%↓) | 67.92 (23.5%↓) | 62.70 (29.3%↓) | 63.81 (28.1%↓) | 77.07 (13.1%↓) | 19.4% |

Table 9: Fine-tuning segmentation task with pre-trained backbones. We experiment at two scales and observe consistent conclusions with previous findings.

| B.B. | Dataset | Alg. | Orig. | Global 4 | Global 8 | Local 4 | Local 8 | Avg. $\Delta_A$ |
|---|---|---|---|---|---|---|---|---|
| ResNet18 | Oxford-IIIT Pet | Sup | 81.90 | 80.50 (1.7%↓) | 78.00 (4.8%↓) | 79.47 (3.0%↓) | 78.61 (4.0%↓) | 3.4% |
|  |  | MoCo-v2 | 84.72 | 80.10 (5.5%↓) | 79.57 (6.1%↓) | 79.43 (6.2%↓) | 82.47 (2.7%↓) | 5.1% |

| B.B. | Dataset | Alg. | Orig. | Global 4 | Local 64 | Gam 0.2 | Avg. $\Delta_A$ |
|---|---|---|---|---|---|---|---|
| ResNet50 | PASCAL VOC | Sup | 70.6 | 70.2 (0.6%↓) | 68.9 (2.4%↓) | 69.5 (1.6%↓) | 1.5% |
|  |  | MoCo-v2 | 71.5 | 68.0 (4.9%↓) | 66.3 (7.3%↓) | 66.8 (6.6%↓) | 6.3% |
|  |  | BYOL | 74.0 | 65.4 (11.6%↓) | 70.7 (4.5%↓) | 72.5 (2.0%↓) | 6.0% |

### 4.8. Minimal effect of tuning hyperparameters for each corruption setting

The performance degradation can solidly reflect the impact of corrupting inductive bias. We empirically verify that while we keep the same hyperparameter settings for all corruption setting within the same algorithm and backbone, each corruption setting is sufficiently trained. We conduct a thorough hyperparameter search for each corruption setting of SL and MoCo-v2, and record the improved outcomes in Tab. 8. By comparing the improved average $\Delta_A$, we observe that SL has gained $0.5\%$ improvement while $0.9\%$ for MoCo, yet the notable gap between CL and SL still remains $(3.2\%)$ and the conclusions still hold.

### 4.9. Segmentation Task

We choose semantic segmentation as our downstream task to verify our observations. We fine-tune our pre-trained backbones on binary dataset Oxford-IIIT Pet [24] with ResNet-18, and multiclass dataset PASCAL VOC [14] with ResNet-50. At each scale, we search for a fixed hyperparameter setting that sufficiently fine-tunes on the training set, with IOU as the evaluation metric. We record experiment results in Tab. 9, and we observe consistent conclusions: CL relies more on spatial inductive bias, where a corrupted pre-training leads to inferior downstream performance with larger $\Delta$ than SL. This validates our findings beyond classification tasks.

## 5. Conclusion

This paper is an empirical study of contrastive learning's reliance on spatial inductive bias. From the results of experiments designed with random patch shuffling-corrupted datasets, we discover a consistent conclusion that CL relies much more on global or local spatial inductive bias than SL, where global spatial inductive bias is more crucial than local, regardless of the specific learning algorithm or backbones, while no high reliance on other inductive biases is shown. Our further analysis of feature space uniformity shows that spatial inductive bias is crucial for CL to learn a more powerful model, while our visualization of attention maps shows in the opposite way that a model that learned to exploit spatial inductive bias can still well understand images with destroyed spatial structure. Our results reveal an uninvestigated aspect of CL, the spatial inductive bias, and inspire more research into understanding the behavior and the learning mechanism of CL. We hope this paper could inspire future research on understanding CL and new CL methods with less dependency on spatial inductive bias.

**Acknowledgement.** This work was supported in part by NSF Grant 2106825, NIFA Award 2020-67021-32799, the Jump ARCHES endowment, the NCSA Fellows program, the Illinois-Insper Partnership, and the Amazon Research Award. This work used NVIDIA GPUs at NCSA Delta through allocations CIS220014 and CIS230012 from the ACCESS program.

## References

[1] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. In *ICLR*, 2022. 2

[2] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020. 1, 2, 3

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 2, 3

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 1, 2, 3

[5] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. In *NeurIPS*, 2020. 2, 3

[6] Ting Chen, Calvin Luo, and Lala Li. Intriguing properties of contrastive losses. In *NeurIPS*, 2021. 1, 2

[7] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2, 3, 5

[8] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021. 1, 2, 3

[9] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, 2021. 2, 3

[10] Elijah Cole, Xuan Yang, Kimberly Wilber, Oisin Mac Aodha, and Serge Belongie. When does contrastive visual representation learning work? In *CVPR*, 2022. 2

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1, 2, 3

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3, 5

[13] Linus Ericsson, Henry Gouk, and Timothy M Hospedales. How well do self-supervised models transfer? In *CVPR*, 2021. 2

[14] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, Jan. 2015. 9

[15] Songwei Ge, Shlok Mishra, Chun-Liang Li, Haohan Wang, and David Jacobs. Robust contrastive learning using negative samples with diminished semantics. In *NeurIPS*, 2021. 2, 3

[16] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Pires, Zhaohan Guo, Mohammad Azar, Bilal Piot, Koray Kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, 2020. 1, 2, 3

[17] Kaiming He, Xinlei Chen, Saining xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 2, 3

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 1, 2, 3

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[20] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2019. 2, 3, 4, 5

[21] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *NeurIPS*, 2020. 5

[22] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 2, 3

[23] Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? In *NeurIPS*, 2020. 2, 3

[24] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 9

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2

[26] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021. 2

[27] Nikunj Saunshi, Jordan Ash, Surbhi Goel, Dipendra Misra, Cyril Zhang, Sanjeev Arora, Sham Kakade, and Akshay Krishnamurthy. Understanding contrastive learning requires incorporating inductive biases. In *International Conference on Machine Learning*, pages 19250–19286. PMLR, 2022. 1

[28] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 3, 4

[29] Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *CVPR*, 2021. 1, 2

[30] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, 2020. 2, 3, 4

[31] Yao Xiao, Ziyi Tang, Pengxu Wei, Cong Liu, and Liang Lin. Masked images are counterfactual samples for robust fine-tuning. 2023. 3

[32] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *ICML*, 2021. 2, 3

[33] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *ICLR*, 2017. 2, 3

[34] Nanxuan Zhao, Zhirong Wu, Rynson WH Lau, and Stephen Lin. What makes instance discrimination good for transfer learning? In *ICLR*, 2021. 2

[35] Yuanyi Zhong, Bodi Yuan, Hong Wu, Zhiqiang Yuan, Jian Peng, and Yu-Xiong Wang. Pixel contrastive-consistent semi-supervised semantic segmentation. In *ICCV*, 2021. 1