When Does Contrastive Visual Representation Learning Work?

Elijah Cole[1] Xuan Yang[2] Kimberly Wilber[2] Oisin Mac Aodha[3,4] Serge Belongie[5]
[1]Caltech [2]Google [3]University of Edinburgh [4]Alan Turing Institute [5]University of Copenhagen

Abstract

Recent self-supervised representation learning techniques have largely closed the gap between supervised and unsupervised learning on ImageNet classification. While the particulars of pretraining on ImageNet are now relatively well understood, the field still lacks widely accepted best practices for replicating this success on other datasets. As a first step in this direction, we study contrastive self-supervised learning on four diverse large-scale datasets. By looking through the lenses of data quantity, data domain, data quality, and task granularity, we provide new insights into the necessary conditions for successful self-supervised learning. Our key findings include observations such as: (i) the benefit of additional pretraining data beyond 500k images is modest, (ii) adding pretraining images from another domain does not lead to more general representations, (iii) corrupted pretraining images have a disparate impact on supervised and self-supervised pretraining, and (iv) contrastive learning lags far behind supervised learning on fine-grained visual classification tasks.

1. Introduction

Self-supervised learning (SSL) techniques can now produce visual representations which are competitive with representations generated by fully supervised networks for many downstream tasks [18]. This is an important milestone for computer vision, as removing the need for large amounts of labels at training time has the potential to scale up our ability to address challenges in domains where supervision is currently too difficult or costly to obtain. However, with some limited exceptions, the vast majority of current state-of-the-art approaches are developed and evaluated on standard datasets like ImageNet [40]. As a result, we do not have a good understanding of how well these methods work when they are applied to other datasets.

Under what conditions do self-supervised contrastive representation learning methods produce "good" visual representations? This is an important question for computer vision researchers because it adds to our understanding of SSL and highlights opportunities for new methods. This is

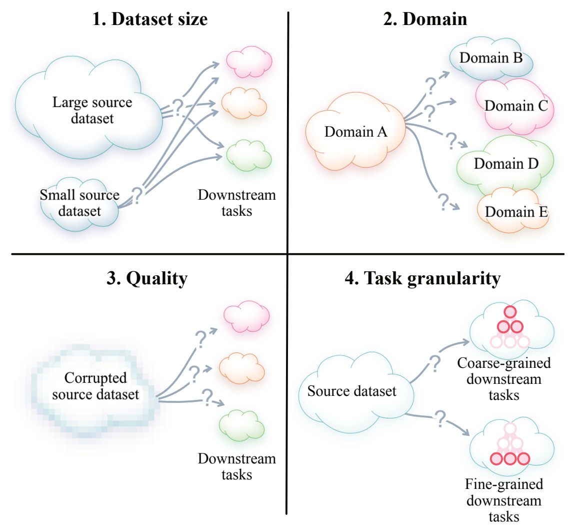


Figure 1. **What conditions are necessary for successful self-supervised pretraining on domains beyond ImageNet?** We investigate the impact of self-supervised and supervised training *dataset size*, the downstream *domain*, *image quality*, and the *granularity* of downstream classification tasks.

also an important question for domain experts with limited resources who might be interested in applying SSL to real-world problems. With these objectives in mind, we attempt to answer the following questions:

(i) What is the impact of data quantity? How many unlabeled images do we need for pretraining, and when is it worthwhile to get more? How much labeled data do we need for linear classifier training or end-to-end fine-tuning on a downstream task? In which regimes do self-supervised features rival those learned from full supervision?

(ii) What is the impact of the pretraining domain? How well do self-supervised representations trained on one domain transfer to another? Can we learn more general representations by combining datasets? Do different pretraining datasets lead to complementary representations?

(iii) What is the impact of data quality? How robust are self-supervised methods to training time image corruption such as reduced resolution, compression artifacts, or noise? Does pretraining on corrupted images lead to poor downstream performance on uncorrupted images?

(iv) What is the impact of task granularity? Does SSL result in features that are only effective for "easy" classification tasks, or are they also useful for more challenging, "fine-grained" visual concepts?

We address the above questions through extensive quantitative evaluation across four diverse large-scale visual datasets (see Figure 1). We make several interesting observations and recommendations including:

- For an ImageNet-scale dataset, decreasing the amount of unlabeled training data by half (from 1M to 500k images) only degrades downstream classification performance by 1-2% (Figure 2). In many contexts this trade-off is reasonable, allowing for faster and cheaper pretraining. This also indicates that current self-supervised methods coupled with standard architectures may be unable to take advantage of very large pretraining sets.
- Self-supervised representations that are learned from images from the same domain as the test domain are much more effective than those learned from different domains (Table 1). Self-supervised training on our current datasets may not be sufficient to learn representations that readily generalize to many contexts.
- Neither (i) combining datasets before pretraining (Table 2) nor (ii) combining self-supervised features learned from different datasets (Table 3) leads to significant performance improvements. More work may be required before self-supervised techniques can learn highly generalizable representations from large and diverse datasets.
- Pretraining on corrupted images affects supervised and self-supervised learning very differently (Figure 4). For instance, self-supervised representations are surprisingly sensitive to image resolution.
- Current self-supervised methods learn representations that can easily disambiguate coarse-grained visual concepts like those in ImageNet. However, as the granularity of the concepts becomes finer, self-supervised performance lags further behind supervised baselines (Figure 5). The contrastive loss may lead to coarse-grained features which are insufficient for fine-grained tasks.

2. Related Work

SSL for visual representations. Early self-supervised representation learning methods typically centered around solving hand-designed "pretext tasks" like patch location prediction [16], rotation prediction [20], inpainting [37], cross-channel reconstruction [59], sorting sequences of video frames [33], solving jigsaw puzzles [35], or colorization [58]. However, more recent work has explored *contrastive learning-based* approaches where the pretext task is to distinguish matching and non-matching pairs of augmented input images [28, 36, 48]. The prototypical example is SimCLR [8, 9], which is trained to identify the matching image using a cross-entropy loss. Other variations on the contrastive SSL framework include using a momentum encoder to provide large numbers of negative pairs (MoCo) [11, 25], adaptively scaling the margin in MoCo (EqCo) [62], and contrasting clustering assignments instead of augmented pairs (SwAV) [6]. Moving beyond the contrastive loss entirely, some papers recast the problem in a "learning-to-rank" framework (S2R2) [52], use simple feature prediction (SimSiam) [12], or predict the output of an exponential moving average network (BYOL) [24]. [4] investigates the role of negatives in contrastive learning, though we note that BYOL and SimSiam avoid using negatives explicitly. In this work, our focus is on self-supervised visual classification. We do not explore alternative settings such as supervised contrastive learning [31], contrastive learning in non-vision areas like language [39] or audio [41], or other methods that aim to reduce the annotation burden for representation learning such as large-scale weak supervision [34].

SSL beyond ImageNet. ImageNet classification has long been viewed as the gold standard benchmark task for SSL, and the gap between supervised and self-supervised performance on ImageNet has steadily closed over the last few years [6, 8, 24, 25]. There is now a growing expectation that SSL should reduce our dependence on manual supervision in challenging and diverse domains which may *not* resemble the traditional object classification setting represented by ImageNet. A number of papers have studied how well self-supervised representations pretrained on ImageNet perform on downstream tasks like fine-grained species classification [56], semantic segmentation [5], scene understanding [24], and instance segmentation [25].

More recently, researchers have begun to study the effectiveness of contrastive learning when *pretraining* on datasets other than ImageNet. In the case of remote sensing, the unique properties of the data have motivated the development of domain-specific contrastive learning techniques [2, 30]. In the medical domain, where images tend to be very dissimilar to ImageNet, it has been shown that contrastive pretraining on domain-specific images leads to significant gains compared to pretraining on ImageNet [9, 43]. [32] compared the representations learned from five different datasets, and showed that in most cases the best performing representations came from pretraining on similar datasets to the downstream task. In the case of fine-grained data, [51] found that contrastive pretraining on images of animals and plants did not lead to superior performance on downstream bird classification compared to pretraining on ImageNet. These apparently conflicting observations may be explained by the relationship between the pretraining and downstream data distributions, which we investigate in our experiments. [60] and [50] pretrained on several different datasets and showed that there was surprisingly little impact on downstream detection and segmentation performance, unless synthetic data was used for pretraining [60].

[47] pretrained on very large datasets (JFT-300M [44] and YFCC100M [46]), but did not observe an improvement over ImageNet pretraining in the standard regime.

We build on the above analysis by performing controlled, like-for-like, comparisons of SSL on several large datasets. This allows us to separate dataset-specific factors from general patterns in SSL performance, and deliver new insights into the necessary conditions for successful pretraining.

Analysis of SSL. A number of works have explored questions related to the conditions under which SSL is successful. [42] showed that self-supervised representations generalize better than supervised ones when the downstream concepts of interest are less semantically similar to the pretraining set. [18] showed that contrastive pretraining on ImageNet performs well on downstream tasks related to object recognition in natural images, while leaving more general study of pretraining in different domains to future work. While these works show that SSL on ImageNet can be effective, our experiments demonstrate that current SSL methods can perform much worse than supervised baselines on non-ImageNet domains, e.g. fine-grained classification.

Existing work has also investigated other aspects of SSL, e.g. [38] examined the invariances learned, [10] showed that easily learned features can inhibit the learning of more discriminative ones, [8, 50, 60] explored the impact of different image augmentations, [10, 50] compared representations from single vs. multi-object images, and [8, 23] varied the backbone model capacity. Most relevant to our work are studies that vary the amount of data in the pretraining dataset, e.g. [32, 50, 57, 60]. We extend this analysis by presenting a more detailed evaluation of the impact of the size of the unlabeled and labeled datasets, and investigate the role of data quality, data domain, and task granularity.

3. Methods

Datasets. We perform experiments on four complementary large-scale datasets: ImageNet [15], iNat21 [50], Places365 [61], and GLC20 [13]. Collectively, these datasets span many important visual properties, including: curated vs. "in-the-wild" images, fine- vs. coarse-grained categories, and object-centric images vs. scenes. Each dataset has at least one million images, which allows us to make fair comparisons against the traditional ImageNet setting. ImageNet (1.3M images, 1k classes) and Places365 (1.8M images, 365 classes) are standard computer vision datasets, so we will not describe them in detail. For ImageNet, we use the classic ILSVRC2012 subset of the full ImageNet-21k dataset. For Places365, we use the official variant "Places365-Standard (small images)" where all images have been resized to 256x256. iNat21 (2.7M images, 10k classes) contains images of plant and animal species and GLC20 (1M images, 16 classes) consists of remote sensing images. As both are recent datasets, we discuss them in the supplementary material.

Fixed-size subsets. For some experiments we control for dataset size by creating subsampled versions of each dataset with sizes: 1M, 500k, 250k, 125k, and 50k images. We carry out this selection only once, and the images are chosen uniformly at random. We refer to these datasets using the name of the parent dataset followed by the number of images in parentheses, e.g. ImageNet (500k). Note that subsets of increasing size are *nested*, so e.g. ImageNet (500k) includes all of the images in ImageNet (250k). These subsets are also *static* across experiments, e.g. ImageNet (500k) always refers to the same set of 500k images. With the exception of Figures 2 and 3, we use the full dataset for any type of supervised training (i.e. linear evaluation, fine tuning, or supervised training from scratch). We always report results on the same test set for a given dataset, regardless of the training subset used.

Training details. All experiments in this paper are based on a ResNet-50 [26] backbone, which is standard in the contrastive learning literature [6, 8, 25]. We primarily perform experiments on SimCLR [8], a simple and popular contrastive learning method that contains all the building blocks for state-of-the-art self-supervised algorithms. We follow the standard protocol of first training with self-supervision alone and then evaluating the learned features using linear classifiers or end-to-end fine-tuning. Unless otherwise specified, we use hyperparameter settings based on [8] for all methods and datasets. While this may not lead to maximal performance, it is likely to be representative of how these methods are used in practice – due to the high computational cost of contrastive pretraining, extensive hyperparameter tuning is not feasible for most users. We also consider MoCo [25] and BYOL [24] in Figure 3. Full training details are provided in the supplementary material.

4. Experiments

We now describe our experiments in which we investigate the impact of data quantity, data domain, data quality, and task granularity on the success of contrastive learning.

4.1. Data quantity

First we consider the question of how much data is required to learn a "good" representation using SSL. There are two important notions of data quantity: (i) the number of *unlabeled images* used for pretraining and (ii) the number of *labeled images* used to subsequently train a classifier. Since labels are expensive, we would like to learn representations that generalize well with as few labeled images as possible. While unlabeled images are cheap to acquire, they still incur a cost because pretraining time is proportional to the size of the pretraining set. To understand when SSL is cost-effective, we need to understand how performance depends on these two notions of data quantity.

To study this question, we pretrain SimCLR using different numbers of unlabeled images. Each pretrained representation is then evaluated using different numbers of labeled images. In Figure 2 we present these results for iNat21 (left column), ImageNet (center column), and Places365 (right column). We also include results for supervised training from scratch (in black). We show linear evaluation results in the top row and corresponding fine-tuned results in the bottom row. Each curve in a figure corresponds to a different pretrained representation. The points along a curve correspond to different amounts of supervision used to train a linear classifier or fine-tune the network.

There is little benefit beyond 500k pretraining images. The gap between the 500k (blue) and 1M (orange) pretraining image curves is typically less than 1-2% in top-1 accuracy. This means that for a dataset with one million images, we can trade a small decrease in accuracy for a 50% decrease in pretraining time. If a 2-4% top-1 accuracy drop is acceptable, then the pretraining set size can be reduced by a factor of four (from 1M to 250k). However, the difference between 50k (pink) pretraining images and 250k (green) pretraining images is substantial for each dataset, often in excess of 10% top-1 accuracy. We conclude that SimCLR seems to saturate well before we get to ImageNet-sized pretraining sets. This is consistent with observations from the supervised learning literature, though more images are required to reach saturation [34].

Self-supervised pretraining can be a good initializer when there is limited supervision available. In the bottom row of Figure 2 we see that when only 10k or 50k labeled images are available, fine-tuning a SimCLR representation is significantly better than training from scratch. When supervision is plentiful, fine-tuned SimCLR representations achieve performance similar to supervised training from scratch. It is interesting to compare this to findings from the supervised setting which suggest that networks which are initially trained on distorted (i.e. augmented) images are unable to recover when subsequently trained with undistorted ones [1].

Self-supervised representations can approach fully supervised performance for some datasets, but only by using lots of labeled images. The ultimate goal of SSL is to match supervised performance without the need for large amounts of labeled data. Suppose we consider the right-most point on the black curves in Figure 2 as a proxy for "good" supervised performance. Then in both the linear and fine-tuned cases, the gap between SimCLR (pretrained on 1M images) and "good" supervised performance is quite large unless well over 100k labeled images are used. For instance, the gap between "good" supervised performance and a classifier trained using 50k labeled images on top of SimCLR (1M) is around 11% (11%) for Places365, 23% (21%) for ImageNet, and 58% (56%) for iNat21 in the linear (and fine-tuned) case. Although SSL works well when lots of supervision is available, further innovation is needed to improve the utility of self-supervised representations in the low-to-moderate supervision regime.

iNat21 is a valuable SSL benchmark. Figure 2 shows a surprisingly large gap ($\sim$ 30%) between supervised and self-supervised performance on iNat21 in the high supervision regime. In Figure 3 we see that other SSL methods exhibit similar limitations. The newer BYOL outperforms MoCo and SimCLR, but a considerable gap ($\sim$ 25%) remains. The high supervised performance shows that the task is possible, yet the self-supervised performance remains low. It seems that iNat21 reveals challenges for SSL that are not apparent in ImageNet, and we believe it is a valuable benchmark for future SSL research.

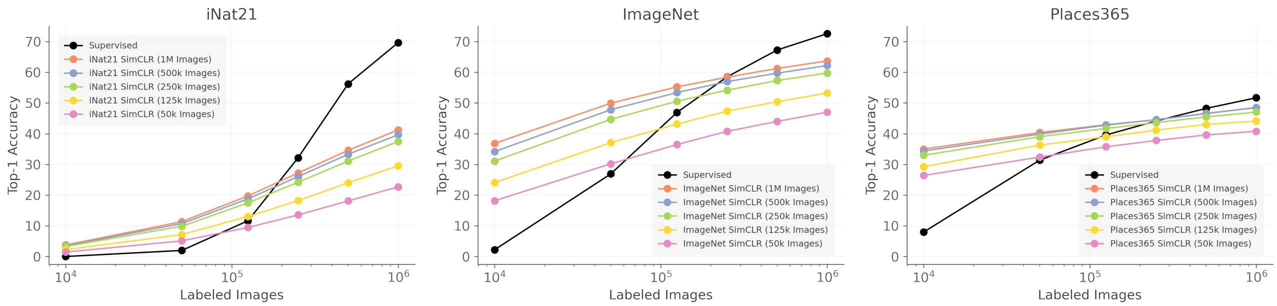
4.2. Data domain

In the previous section we observed that increasing the pretraining set size yields rapidly diminishing returns. In this section we consider a different design choice: *what kind of images* should we use for pretraining? Since most contrastive learning papers only pretrain on ImageNet, this question has not received much attention. We take an initial step towards an answer by studying the properties of SimCLR representations derived from four pretraining sets drawn from different domains.

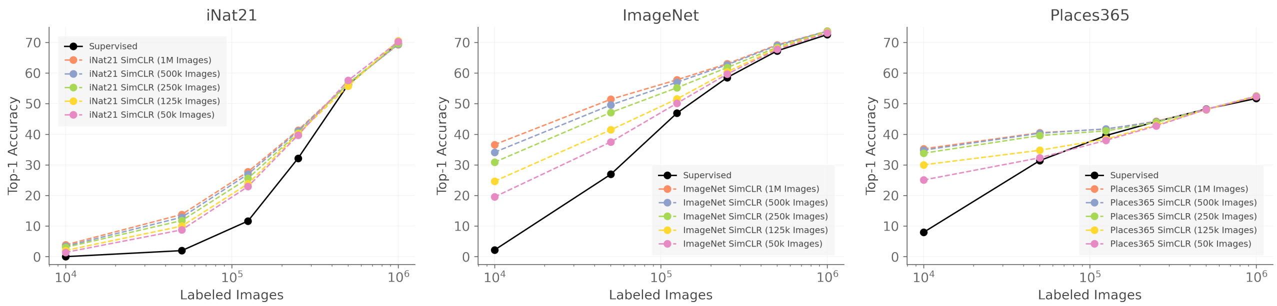
We train SimCLR on iNat21 (1M), ImageNet (1M), Places365 (1M), and GLC20 (1M). By holding the pretraining set size constant, we aim to isolate the impact of the different visual domains. We present in-domain and cross-domain linear evaluation results for each representation in Table 1. In Table 2 we consider the effect of pretraining on *pooled datasets*, i.e. new image collections built by shuffling together existing datasets. Finally, in Table 3 we study different *fused representations*, which are formed by concatenating the outputs of different feature extractors.

Pretraining domain matters. In Table 1 we see that in-domain pretraining (diagonal entries) consistently beats cross-domain pretraining (off-diagonal entries). The gap can be surprisingly large, e.g. in-domain pretraining provides a 12% boost on iNat21 compared to the best cross-domain pretraining (ImageNet). One might have expected that a visually diverse dataset like ImageNet would lead to a better self-supervised representation than a more homogeneous dataset like GLC20 (even when evaluating on GLC20) but this is not what we observe.

The off-diagonal entries of Table 1 show that training SimCLR on ImageNet leads to the best cross-domain performance, while GLC20 leads to the worst cross-domain performance. Since the pretraining protocols and dataset sizes are held constant, we suggest that the characteristics of the image sets themselves are responsible for the differences we observe. The strong cross-domain performance of



(a) Linear Evaluation



(b) Fine-Tuning

Figure 2. **How much data does SimCLR need?** Linear evaluation results (top row) and fine-tuning results (bottom row) as a function of the number of *unlabeled images* used for pretraining and the number of *labeled images* used for downstream supervised training. The "Supervised" curve (black) corresponds to training from scratch on different numbers of labeled images. It is the same for the top and bottom plots in each column. Most SSL papers focus on the "high data" regime, using $\sim 10^6$ images (e.g. all of ImageNet) for both pretraining and classifier supervision, but there are significant opportunities for improvement in the "low-data" regime. Even with 10^6 labeled images for linear classifier training, SimCLR performs far worse than supervised learning on iNat21, suggesting that iNat21 could be a more useful SSL benchmark than ImageNet in future.

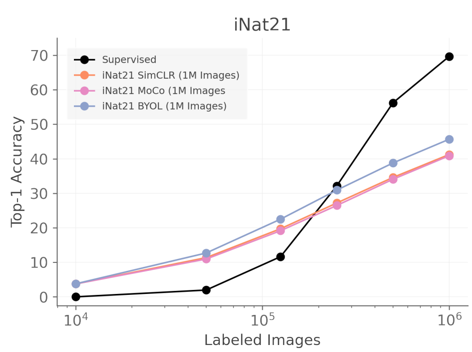


Figure 3. **How does SimCLR compare to other self-supervised methods?** Linear evaluation results on iNat21 for SimCLR, MoCo, and BYOL. All methods are pretrained on 1M images for 1000 epochs and follow the same linear evaluation protocol. The more recent BYOL performs better than the others, but a large gap remains to supervised performance.

SimCLR pretrained on ImageNet may be due to *semantic similarity* – perhaps it is better to pretrain on a dataset that is semantically similar to the downstream task, even in a self-supervised context. This makes sense because there are classes in ImageNet that are similar to classes in iNat21 (animals) and Places365 (scenes). This also explains the weak performance of GLC20, since remote sensing imagery is

Pretraining	iNat21	ImageNet	Places365	GLC20
iNat21 (1M) SimCLR	**0.493**	<u>0.519</u>	0.416	0.707
ImageNet (1M) SimCLR	<u>0.373</u>	**0.644**	<u>0.486</u>	<u>0.716</u>
Places365 (1M) SimCLR	0.292	0.491	**0.501**	0.693
GLC20 (1M) SimCLR	0.187	0.372	0.329	**0.769**
Supervised (All Images)	0.791	0.741	0.539	0.826

Table 1. **Does pretraining domain matter?** Linear evaluation results for representations derived from different million-image datasets. We train the linear classifiers using the full training sets. The results in the "Supervised" row correspond to supervised training from scratch on the full training set. We report MAP for GLC20 and top-1 accuracy for other datasets. In all cases, in-domain pretraining outperforms cross-domain pretraining. In each column we highlight the **best** and <u>second-best</u> results.

not similar to the other datasets.

Adding cross-domain pretraining data does not necessarily lead to more general representations. We have seen that pretraining on different domains leads to representations with significantly differing capabilities. This leads to a natural question: *what happens if we combine our datasets and then learn a representation?*

Table 2 gives linear evaluation results for SimCLR pretrained on different "pooled" datasets. In each row, n images from dataset A and m images from dataset B are shuffled together to produce a pretraining set of size $n + m$. For instance, the pretraining dataset in the first row of Table 2

Pretraining				Evaluation		
iNat21	*ImageNet*	*Places365*	*GLC20*	*iNat21*	*ImageNet*	*Places365*
250k	250k	-	-	0.444	0.597	0.467
-	250k	250k	-	0.334	0.596	0.490
250k	-	250k	-	0.428	0.531	0.483
250k	250k	250k	250k	0.410	0.574	0.482
In-Domain (250k)				0.451	0.608	0.485
In-Domain (500k)				0.477	0.629	0.499
In-Domain (1M)				0.493	0.644	0.501

Table 2. **The effect of dataset pooling.** Linear evaluation results for self-supervised representations derived from *pooled datasets*, where two or more datasets are shuffled together. We train the linear classifiers using the full training sets. The "In-Domain" results correspond to pretraining on subsets of the dataset named at the top of the column. Pooling datasets increases pretraining set size and diversity, but we find that performance *decreases* relative to comparable in-domain pretraining. The "In-Domain (1M)" row corresponds to the diagonal entries of Table 1.

consists of 250k iNat21 images and 250k ImageNet images shuffled together.

If we compare the "In-Domain (500k)" row against the (equally sized) pooled datasets in the first three rows of Table 2, we see that the in-domain pretraining on 500k images is always better. Similarly, the "In-Domain (1M)" row beats the 1M-image pooled dataset (consisting of 250k images from the four datasets). The more diverse pooled pretraining sets always lead to worse performance compared to the more homogeneous pretraining sets of the same size.

Table 2 also allows us to say whether it is worthwhile to *add* pretraining data from a different domain (as opposed to swapping out some in-domain data for some data from a different domain, as we have been discussing so far). The "In-Domain (250k)" row is better than the 1M-image pooled dataset and almost all of the 500k-image pooled datasets. It seems that adding pretraining data from a different domain typically *hurts* performance. In contrast, Figure 2 shows that increasing the amount of *in-domain* pretraining data consistently improves performance.

We hypothesize that the reason for this lackluster performance is that diverse images are easier to tell apart, which makes the contrastive pretext task easier. If the contrastive task is too easy, the quality of the representation suffers [4, 10]. While more investigation is needed, the fact that increasing pretraining data diversity can hurt performance suggests a "diversity-difficulty trade-off" that should be considered when creating pretraining sets for SSL.

Self-supervised representations can be largely redundant. From Table 1 it is clear that pretraining on different datasets leads to representations that differ significantly. For instance, iNat21 SimCLR beats ImageNet SimCLR on iNat21 (+12.4%) and ImageNet SimCLR beats iNat21 SimCLR on ImageNet (+12.7%). Do these representations learn complementary information, or do they just capture the same information to different degrees?

ImageNet	iNat21	Dim.	ImageNet	iNat21
SimCLR	-	2048	0.647	0.380
-	SimCLR	2048	0.520	0.506
Sup.	-	2048	0.711	0.434
-	Sup.	2048	0.490	<u>0.769</u>
Sup.	Sup.	4096	0.712	**0.772**
SimCLR	SimCLR	4096	0.641	0.520
SimCLR & Sup.	-	4096	**0.720**	0.472
-	SimCLR & Sup.	4096	0.527	**0.772**
SimCLR	Sup.	4096	0.605	<u>0.769</u>
Sup.	SimCLR	4096	<u>0.717</u>	0.553

Table 3. **The effect of representation fusion.** Linear evaluation results for different combinations of supervised and self-supervised representations on ImageNet and iNat21. We train the linear classifiers using the full training sets. For comparability, the in-domain supervised results in this table (ImageNet Sup. evaluated on ImageNet and iNat21 Sup. evaluated on iNat21) are for linear classifiers trained on representations learned from full supervision. "Dim." is the representation dimensionality. In each column we highlight the **best** and <u>second-best</u> results.

To probe this question we concatenate features from different pretrained networks and carry out linear evaluation on these "fused" representations. In Table 3 we present linear evaluation results for fused representations on ImageNet and iNat21. Combining ImageNet SimCLR and iNat21 SimCLR is worse than ImageNet SimCLR alone on ImageNet (-0.6%), but better than iNat21 SimCLR alone on iNat21 (+1.4%). These effects are small relative to the $> 12\%$ difference between ImageNet SimCLR and iNat21 SimCLR. This suggests that the two self-supervised representations are largely redundant.

There is a larger effect when combining supervised and self-supervised representations. For iNat21, adding ImageNet Sup. (i.e. supervised ImageNet features) on top of iNat21 SimCLR improves performance significantly (+4.7%). However, adding iNat21 Sup. on top of ImageNet SimCLR actually decreases performance (-4.2%). These results are consistent with the hypothesis that dataset semantics are important even for SSL. Since ImageNet is semantically broader than iNat21 (ImageNet has animal classes, but also many other things), features learned from ImageNet (supervised or self-supervised) should be more helpful for iNat21 than vice-versa.

4.3. Data quality

We have seen that the characteristics of the pretraining data can have a significant impact on the quality of self-supervised representations. In this section we dig deeper into this question by studying the impact of pretraining on artificially degraded images. This serves two purposes. First, this is a practical question since there are many settings where image quality issues are pervasive e.g. medical imaging [45] or camera trap data [3]. Second, it can help us understand the robustness properties of SSL.

To create a corrupted dataset we apply a particular image

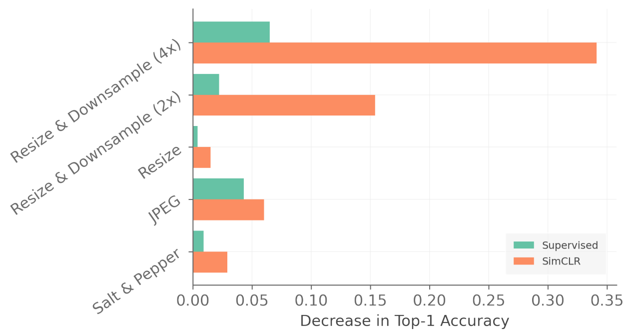


Figure 4. **What is the effect of pretraining image corruption?** Decrease in linear evaluation accuracy on ImageNet due to pretraining on corrupted versions of the ImageNet training set. The zero point corresponds to pretraining (supervised or SimCLR) on uncorrupted images followed by linear evaluation. "Supervised" and "SimCLR" have different zero points. All linear classifiers are trained using the full uncorrupted ImageNet training set.

corruption to each image in the dataset. This is a one-time offline preprocessing step, so corruptions that have a random component are realized only once per image. Given a corrupted dataset we then pretrain as normal. During linear evaluation, we use the original clean images for training and testing, i.e. the corrupted images are only used for pretraining.

In Figure 4 we present linear evaluation results on ImageNet for a simple but diverse set of corruptions. The zero point corresponds to pretraining on uncorrupted images, and we measure how much performance drops when pretraining on corrupted images. The "Salt and Pepper" corruption is salt and pepper noise applied independently to each pixel, in each channel, with probability 0.01. The "JPEG" corruption is JPEG compression with a very low quality level of 10. For "Resize", we resize each image so that the short side is 256 pixels while preserving the aspect ratio. This reduces the resolution of the crops used for training. For our downsampling corruptions, we follow the resize operation with downsampling by 2x or 4x and then upsampling by the same factor. This holds constant the image size and the fraction of the image occupied by each object, but reduces resolution. Implementation details and examples can be found in the supplementary.

Image resolution is critical for SSL. "Downsample (2x)" and "Downsample (4x)" are by far the most damaging corruptions for SimCLR, reducing accuracy by around 15% and 34%, respectively. Since SimCLR already involves extreme cropping, we might expect more robustness to changes in image resolution. This finding could be partially explained by the difficulty of generalizing to higher-resolution images during linear classifier training [49]. However, supervised pretraining faces the same challenge but the effect of downsampling is much less dramatic. This suggests that the performance drop is due to deficiencies in the features learned by SimCLR.

SSL is relatively robust to high-frequency noise. "JPEG" and "Salt & Pepper" both add high-frequency noise to the image. For SimCLR, these corruptions have a much milder impact than the downsampling corruptions. One possible explanation is that downsampling destroys texture information, which is known to be a particularly important signal for convolutional neural networks [19, 29]. For supervised pretraining the ranking of corruptions is very different, with "JPEG" landing between 2x and 4x downsampling.

4.4. Task granularity

We have seen that the properties of pretraining datasets are important for determining the utility of self-supervised representations. But are there downstream tasks for which self-supervised representations are particularly well or poorly suited? We consider *fine-grained classification* and show that classification performance depends on *task granularity*, i.e. how fine or coarse the labels are. While there are formal methods for measuring dataset granularity [14], we claim by intuition that iNat21 is more fine-grained than ImageNet, which is more fine-grained than Places365.

In Figure 5 we use label hierarchies (which are available for ImageNet, iNat21, and Places365) to explicitly study how performance depends on label granularity. We treat "distance from the root of the hierarchy" as a proxy for granularity, so labels further from the root are considered to be more fine-grained. We perform (i) linear classifier training (for SimCLR) and (ii) end-to-end training from scratch (for "Supervised") using the labels at the finest level of the taxonomy and re-compute accuracy values as we progressively coarsen the predictions and labels. We do not re-train at each level of granularity. A complete description of this process can be found in the supplementary materials.

The performance gap between SSL and supervised learning grows as task granularity becomes finer. We start with the iNat21 results in Figure 5. The supervised and SimCLR pretrained models perform similarly at the coarsest levels of the label hierarchy ("Kingdom"). Both models perform worse as task granularity increases, but the SimCLR model degrades much more rapidly ("Species"). This suggests that SimCLR may fail to capture fine-grained semantic information as effectively as supervised pretraining. We also observe a growing supervised/self-supervised gap for ImageNet and Places365. The magnitude of this gap seems to track dataset granularity, since iNat21 (most fine-grained) has the largest gap and Places365 (least fine-grained) has the smallest gap. The fact that supervised learning achieves high performance on iNat21 while SSL lags behind suggests that iNat21 could be a valuable benchmark dataset for the next phase of SSL research.

Are the augmentations destructive? State-of-the-art con-

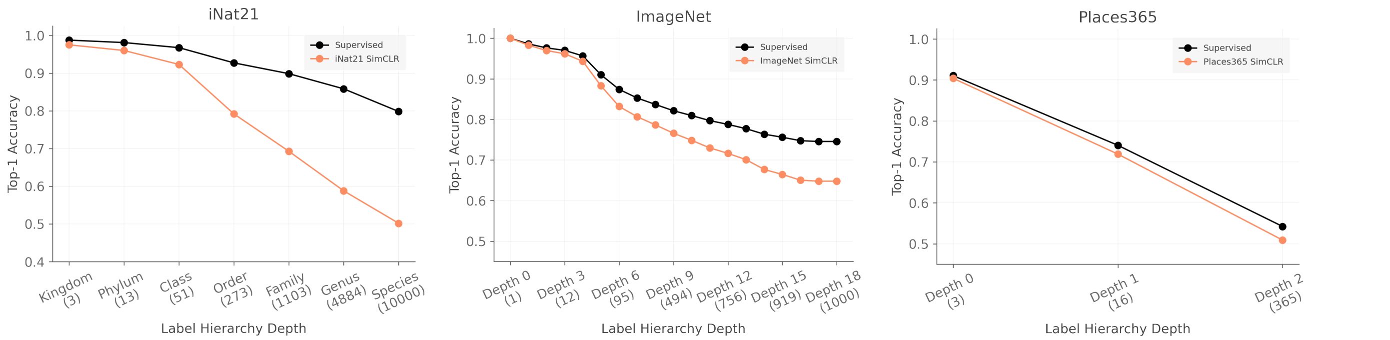


Figure 5. **How does performance depend on label granularity?** Linear evaluation at different levels of label granularity for iNat21, ImageNet, and Places365. Each plot compares supervised learning from scratch against a linear classifier trained on top of in-domain SimCLR. Both are trained using the full training sets. We plot top-1 accuracy against label granularity, which is more fine-grained as we move from left to right. The numbers on the x-axis are the class counts at a given level of the label hierarchy. We do not re-train at coarser granularity levels, we just change the evaluation label set. The definitions of the hierarchy levels are given in the supplementary material.

trastive learning techniques are designed for ImageNet, so the default augmentation policy may be poorly tuned for other datasets [56]. For instance, if color is a key fine-grained feature for species classification then the "color jitter" augmentation used by SimCLR may destroy important information for iNat21 classification. Could this explain the rapid drop in performance exhibited by iNat21 SimCLR for fine-grained classes? Notice that there is a similar, though less extreme, fine-grained performance drop for ImageNet SimCLR in Figure 5. Since the ImageNet-tuned augmentations are presumably not destructive for ImageNet, it does not seem likely that this fully explain our observations.

Does contrastive learning have a coarse-grained bias? We hypothesize that the contrastive loss tends to cluster images based on overall visual similarity. The intuition is that fine-grained features are often subtle, and subtle features are unlikely to be very useful for distinguishing between pairs of images in the contrastive pretext task. If our hypothesis is correct then the boundaries between different clusters would not be well-aligned with the boundaries between fine-grained classes. This effect could be overlooked when evaluating on coarse-grained classes, but would become apparent on a more fine-grained task. Additional analysis is required to fully understand this "granularity gap" in SSL, which we leave to future work.

5. Conclusion

We have presented a comprehensive set of experiments to address several aspects of the question: *when does contrastive visual representation learning work?* In Section 4.1 we found that we need fewer than 500k pretraining images before encountering severe diminishing returns. However, even the best self-supervised representations are still much worse than peak supervised performance without hundreds of thousands of labeled images for classifier training. In Section 4.2 we found that self-supervised pretraining on 1M images from different domains results in representations with very different capabilities, and that simple methods for combining different datasets do not lead to large gains. In Section 4.3 we showed that image resolution is critical for contrastive learning and, more broadly, that some image corruptions can degrade a self-supervised representation to the point of unusability while others have almost no impact. Finally, in Section 4.4 we found that supervised pretraining retains a substantial edge when it comes to fine-grained classification. These experiments highlight several areas where further research is needed to improve current SSL algorithms, most of which were not evident from traditional evaluation protocols, i.e. top-1 accuracy on ImageNet.

Limitations. We mainly perform experiments using one self-supervised method. We focus on SimCLR because it reflects the essence of state-of-the-art contrastive learning methods without introducing additional architectural complexities. While our MoCo and BYOL experiments are not much different from SimCLR, it is important to validate our results on other self-supervised methods. It would also be interesting to explore alternative backbone architectures [7, 17], though after controlling for training settings, ResNet-50 remains competitive with newer architectures [54, 55]. We study only classification tasks, so additional work is also required to understand how these results translate to segmentation [53] or detection [27, 63]. Finally, we only consider datasets up to roughly ImageNet scale. We believe this is the most practical setting for most use cases, but it is possible that some patterns may be different for significantly larger datasets and models [21, 22].

Acknowledgements. We thank Mason McGill for detailed feedback, and Grant Van Horn, Christine Kaeser-Chen, Yin Cui, Sergey Ioffe, Pietro Perona, and the rest of the Perona Lab for insightful discussions. This work was supported by the Caltech Resnick Sustainability Institute, an NSF Graduate Research Fellowship (grant number DGE1745301), and the Pioneer Centre for AI (DNRF grant number P1).

References

[1] Alessandro Achille, Matteo Rovere, and Stefano Soatto. Critical learning periods in deep neural networks. In *ICLR*, 2019. 4

[2] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning. In *ICCV*, 2021. 2

[3] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *ECCV*, 2018. 6

[4] Tiffany Tianhui Cai, Jonathan Frankle, David J Schwab, and Ari S Morcos. Are all negatives created equal in contrastive instance discrimination? *arXiv:2010.06682*, 2020. 2, 6

[5] Yue Cao, Zhenda Xie, Bin Liu, Yutong Lin, Zheng Zhang, and Han Hu. Parametric instance classification for unsupervised visual feature learning. In *NeurIPS*, 2020. 2

[6] 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. 2, 3

[7] 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. 8

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

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

[10] Ting Chen, Calvin Luo, and Lala Li. Intriguing properties of contrastive losses. In *NeurIPS*, 2021. 3, 6

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

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

[13] Elijah Cole, Benjamin Deneu, Titouan Lorieul, Maximilien Servajean, Christophe Botella, Dan Morris, Nebojsa Jojic, Pierre Bonnet, and Alexis Joly. The geolifeclef 2020 dataset. *arXiv:2004.04192*, 2020. 3

[14] Yin Cui, Zeqi Gu, Dhruv Mahajan, Laurens Van Der Maaten, Serge Belongie, and Ser-Nam Lim. Measuring dataset granularity. *arXiv:1912.10154*, 2019. 7

[15] 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. 3

[16] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, 2015. 2

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 8

[18] Linus Ericsson, Henry Gouk, and Timothy M Hospedales. How well do self-supervised models transfer? In *CVPR*, 2021. 1, 3

[19] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*, 2019. 7

[20] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018. 2

[21] Priya Goyal, Mathilde Caron, Benjamin Lefaudeux, Min Xu, Pengchao Wang, Vivek Pai, Mannat Singh, Vitaliy Liptchinsky, Ishan Misra, Armand Joulin, et al. Self-supervised pretraining of visual features in the wild. *arXiv preprint arXiv:2103.01988*, 2021. 8

[22] Priya Goyal, Quentin Duval, Isaac Seessel, Mathilde Caron, Mannat Singh, Ishan Misra, Levent Sagun, Armand Joulin, and Piotr Bojanowski. Vision models are more robust and fair when pretrained on uncurated images without supervision. *arXiv preprint arXiv:2202.08360*, 2022. 8

[23] Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. In *ICCV*, 2019. 3

[24] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, 2020. 2, 3

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

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

[27] Olivier J Hénaff, Skanda Koppula, Jean-Baptiste Alayrac, Aaron van den Oord, Oriol Vinyals, and João Carreira. Efficient visual pretraining with contrastive detection. In *ICCV*, 2021. 8

[28] Olivier J Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, SM Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv:1905.09272*, 2019. 2

[29] Katherine L Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks. In *NeurIPS*, 2020. 7

[30] Jian Kang, Ruben Fernandez-Beltran, Puhong Duan, Sicong Liu, and Antonio J Plaza. Deep unsupervised embedding for remotely sensed images based on spatially augmented momentum contrast. *Transactions on Geoscience and Remote Sensing*, 2020. 2

[31] 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. 2

[32] Klemen Kotar, Gabriel Ilharco, Ludwig Schmidt, Kiana Ehsani, and Roozbeh Mottaghi. Contrasting contrastive self-supervised representation learning pipelines. In *ICCV*, 2021. 2, 3
[33] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *ICCV*, 2017. 2
[34] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *ECCV*, 2018. 2, 4
[35] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016. 2
[36] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv:1807.03748*, 2019. 2
[37] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016. 2
[38] Senthil Purushwalkam and Abhinav Gupta. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. In *NeurIPS*, 2020. 3
[39] Nils Rethmeier and Isabelle Augenstein. Long-tail zero and few-shot learning via contrastive pretraining on and for small data. *arXiv:2010.01061*, 2020. 2
[40] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 1
[41] Aaqib Saeed, David Grangier, and Neil Zeghidour. Contrastive learning of general-purpose audio representations. In *ICASSP*, 2021. 2
[42] Mert Bulent Sariyildiz, Yannis Kalantidis, Diane Larlus, and Karteek Alahari. Concept generalization in visual representation learning. In *ICCV*, 2021. 3
[43] Hari Sowrirajan, Jingbo Yang, Andrew Y Ng, and Pranav Rajpurkar. Moco pretraining improves representation and transferability of chest x-ray models. In *Medical Imaging with Deep Learning*, 2021. 2
[44] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *ICCV*, 2017. 3
[45] Siyi Tang, Amirata Ghorbani, Rikiya Yamashita, Sameer Rehman, Jared A Dunnmon, James Zou, and Daniel L Rubin. Data valuation for medical imaging using shapley value and application to a large-scale chest x-ray dataset. *Scientific reports*, 2021. 6
[46] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 2016. 3
[47] Yonglong Tian, Olivier J Henaff, and Aaron van den Oord. Divide and contrast: Self-supervised learning from uncurated data. In *ICCV*, 2021. 3
[48] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV*, 2020. 2
[49] Hugo Touvron, Andrea Vedaldi, Matthijs Douze, and Hervé Jégou. Fixing the train-test resolution discrepancy. In *NeurIPS*, 2019. 7
[50] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Revisiting contrastive methods for unsupervised learning of visual representations. In *NeurIPS*, 2021. 2, 3
[51] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. In *CVPR*, 2021. 2
[52] Ali Varamesh, Ali Diba, Tinne Tuytelaars, and Luc Van Gool. Self-supervised ranking for representation learning. *arXiv:2010.07258*, 2020. 2
[53] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. In *ICCV*, 2021. 8
[54] Ross Wightman, Hugo Touvron, and Hervé Jégou. Resnet strikes back: An improved training procedure in timm. *arXiv:2110.00476*, 2021. 8
[55] Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross Girshick. Early convolutions help transformers see better. In *NeurIPS*, 2021. 8
[56] Tete Xiao, Xiaolong Wang, Alexei A Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. In *ICLR*, 2020. 2, 8
[57] Xingyi Yang, Xuehai He, Yuxiao Liang, Yue Yang, Shanghang Zhang, and Pengtao Xie. Transfer learning or self-supervised learning? a tale of two pretraining paradigms. *arXiv:2007.04234*, 2020. 3
[58] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016. 2
[59] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *CVPR*, 2017. 2
[60] Nanxuan Zhao, Zhirong Wu, Rynson WH Lau, and Stephen Lin. What makes instance discrimination good for transfer learning? In *ICLR*, 2021. 2, 3
[61] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 3
[62] Benjin Zhu, Junqiang Huang, Zeming Li, Xiangyu Zhang, and Jian Sun. Eqco: Equivalent rules for self-supervised contrastive learning. *arXiv:2010.01929*, 2020. 2
[63] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin D Cubuk, and Quoc V Le. Rethinking pre-training and self-training. In *NeurIPS*, 2020. 8