# A Broad Study on the Transferability of Visual Representations with Contrastive Learning

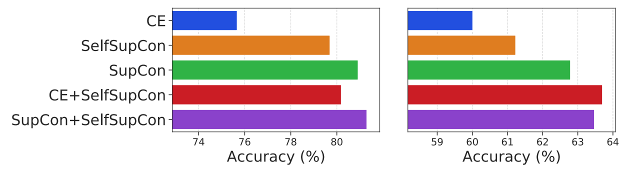
Ashraful Islam[1*], Chun-Fu Chen[2,3], Rameswar Panda[2,3], Leonid Karlinsky[3], Richard Radke[1], Rogerio Feris[2,3]
[1]Rensselaer Polytechnic Institute [2]MIT-IBM Watson AI Lab [3]IBM Research
islama6@rpi.edu, chenrich@us.ibm.com, rpanda@ibm.com
leonidka@il.ibm.com, rjradke@ecse.rpi.edu, rsferis@us.ibm.com

## Abstract

*Tremendous progress has been made in visual representation learning, notably with the recent success of self-supervised contrastive learning methods. Supervised contrastive learning has also been shown to outperform its cross-entropy counterparts by leveraging labels for choosing where to contrast. However, there has been little work to explore the transfer capability of contrastive learning to a different domain. In this paper, we conduct a comprehensive study on the transferability of learned representations of different contrastive approaches for linear evaluation, full-network transfer, and few-shot recognition on 12 downstream datasets from different domains, and object detection tasks on MSCOCO and VOC0712. The results show that the contrastive approaches learn representations that are easily transferable to a different downstream task. We further observe that the joint objective of self-supervised contrastive loss with cross-entropy/supervised-contrastive loss leads to better transferability of these models over their supervised counterparts. Our analysis reveals that the representations learned from the contrastive approaches contain more low/mid-level semantics than cross-entropy models, which enables them to quickly adapt to a new task. Our codes and models will be publicly available to facilitate future research on transferability of visual representations.* [1]

## 1. Introduction

Self-supervised learning is an important research area whose goal is to learn superior data representations without any labelled supervision. Recently, self-supervised contrastive learning has shown promising results in image classification tasks [21, 7, 4]. In the contrastive learning paradigm, a model is trained to recognize different augmentations of the same image (commonly referred as positives) while discriminating them from other random images (referred as negatives) in the dataset. The promising performance of self-supervised contrastive learning led to the idea of leveraging label information in the contrastive learning paradigm. To this end, Khosla *et al*. [30] proposed a supervised contrastive learning framework that achieves better ImageNet accuracy than the standard cross-entropy model.




(a) Linear evaluation (b) Few-shot classification

Figure 1: **Average top-1 accuracy of different models on the downstream datasets.** (a) Linear evaluation with a fixed feature extractor and (b) 5-way 5-shot few-shot classification. In both cases, we observe that contrastive pretrained models achieve superior performance compared to cross-entropy pretrained models. Adding a self-supervised contrastive loss (`SelfSupCon`) improves the performance for both supervised cross-entropy and supervised contrastive pretrained models. We argue that incorporating a self-supervised contrastive loss (`SelfSupCon`) increases the variability within the same-class features and makes the models learn both high-level semantics and low-level cues.

Representations learned from contrastive learning have been shown to perform better than supervised cross-entropy models in various downstream tasks, particularly the object detection task [21, 7, 30, 45, 54]. Despite recent progress, it is unclear why contrastive representations transfer better to other tasks, since most prior work focuses on in-domain evaluation, particularly ImageNet classification accuracy. In this paper, *our goal is to understand the underlying mechanism of the superior transferability of contrastive learning*. Towards this end, we conduct a comprehensive study regarding transfer learning of contrastive approaches on downstream image classification, few-shot evaluation, and object detection. We rigorously benchmark five methods with different training objective losses: cross-entropy, self-supervised contrastive, supervised con-

[*]This work was done while the author was an intern at IBM.
[1]https://github.com/asrafulashiq/transfer_broad

trastive, joint cross-entropy/self-supervised contrastive, and joint supervised/self-supervised contrastive.

We first compare the transfer performance of different ImageNet pretrained models on a collection of 12 downstream datasets from various domains. We find that contrastive methods perform much better than the supervised cross-entropy models, particularly in fixed feature transfer learning; however, the performance gap becomes smaller after full-network fine-tuning. We observe similar trends on other downstream tasks, including few-shot recognition and object detection and instance segmentation on the VOC0712 [15] and MS COCO [36] datasets. In particular, our results indicate that the joint objective of self-supervised contrastive loss and supervised cross-entropy/contrastive loss consistently outperforms the standard trained counterparts in different downstream tasks. Figure 1 shows the average top-1 accuracy of the different ImageNet pretrained methods we studied on the downstream datasets, for both fixed-feature linear evaluation and few-shot classification. Both the self-supervised contrastive model (denoted `SelfSupCon`) and supervised contrastive model (denoted `SupCon`) perform better than the cross-entropy model (denoted `CE`). Moreover, the combination of cross-entropy and self-supervised contrastive (denoted `CE+SelfSupCon`) performs better than cross-entropy or self-supervised contrastive alone. The same goes for the combination of self-supervised contrastive and supervised contrastive (denoted `SupCon+SelfSupCon`).

We next investigate why contrastive approaches show superior transferability by analyzing the similarity between hidden representations, intra-class separation, and robustness to image corruption. We find that contrastive approaches learn more low-level and mid-level information that can be easily adapted to a different domain than the supervised cross-entropy models, which mostly learns high-level semantics in the penultimate layers. Zhao *et al*. [54] hypothesized that one of the limiting factors of supervised cross-entropy models is the objective of minimizing intra-class variation. Our analysis also suggests that a model should have sufficient intra-class variation in the source domain to better transfer the learned representations to a different domain. Most standard supervised loss functions aim to increase inter-class distance and decrease intra-class variation, which might be harmful for transferability of features. We infer that contrastive approaches have larger within-class separation than the standard cross-entropy models, which could be one of the factors underlying their superior transferability. We also analyze the robustness and calibration of different models, and find that contrastive losses are more robust to different image corruptions and predict well-calibrated class probabilities that are more representative of true correctness likelihoods than cross-entropy models. Our key contributions in this work are as follows:

- We benchmark five methods including cross-entropy, self-supervised contrastive, supervised contrastive, and their combinations on downstream image classification, object detection, and few-shot recognition. All results show a similar trend that contrastive learning extracts better features for transfer learning.
- We show that combining supervised loss with self-supervised contrastive loss improves transfer learning performance. Specifically, learned representation from the joint objective of self-supervised contrastive and supervised contrastive loss significantly outperforms the model trained with cross-entropy by 5.63% under linear evaluation protocol and 3.46% in few-shot recognition (5-shot) on the 12 downstream datasets, 1.37% AP50 under object detection on VOC0712, and $\sim$0.8% on MS COCO. The improvement of the joint objective over supervised contrastive model is small but consistent across all downstream tasks. The joint objective of cross-entropy and self-supervised contrastive loss also consistently performs better than the models trained with the individual objectives.
- We apply Centered Kernel Alignment (CKA) [33] and show that contrastive models contain more low-level and mid-level information in the penultimate layers than standard cross-entropy models. Furthermore, our analysis suggests that the contrastive models have higher intra-class variation than the standard cross-entropy models, even if the network is not explicitly trained to increase intra-class distance.

## 2. Related Work

**Transfer Learning.** Early results on transfer learning showed that convolutional neural networks (CNNs) trained on large-scale datasets could be used to extract features to train SVMs and logistic regression models that outperformed hand-crafted feature-based approaches [6, 14, 44]. Transfer learning can be a powerful tool to train a significantly smaller dataset than the base dataset without overfitting. However, the factors driving the performance are still not completely understood. Huh *et al*. [27] investigated the effect of the source dataset on transfer learning. Simon *et al*. [34] found that pretrained models with higher ImageNet accuracy also tend to perform well in the downstream task. Azizpour *et al*. [2] investigated the effect of network depth on the transfer performance. In this work, we show that contrastive training can improve transfer learning performance, and investigate the underlying principle behind it.

**Self-Supervised and Supervised Contrastive Learning.** Earlier work on self-supervised learning generated pseudo labels by patch position [13], image colorization [53], image inpainting [43], rotation [17], predictive coding [23, 23] and other pretext tasks. Recently, contrastive learning has led to significant performance enhancement in self-supervised image representation learning. In particular, MoCo [21], SimCLR [7], SwAV [4], and others have shown dramatic improvement in representation quality learned from unlabeled ImageNet images. Khosla *et al.* [30] proposed a new contrastive loss to leverage the label information. Moreover, representations learned from contrastive learning have been shown to perform better than supervised cross-entropy models in various downstream tasks [52, 21, 7, 30, 29, 31, 28]. However, most of the studies perform limited comparison, particularly with regard to fixed-feature transfer, few-shot learning and robustness, and the underlying principle of why contrastive learning transfers better still remains unclear.

## 3. Analysis Setup

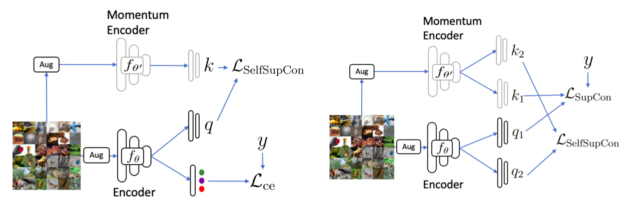
Given a source domain $\mathcal{D}_s = \{(\mathbf{x}_{S_1}, y_{S_1}), (\mathbf{x}_{S_2}, y_{S_2}), \ldots, (\mathbf{x}_{S_N}, y_{S_N})\}$ with a marginal distribution $\mathcal{P}_S$ and a target domain $\mathcal{D}_T = \{(\mathbf{x}_{T_1}, y_{T_1}), (\mathbf{x}_{T_2}, y_{T_2}), \ldots, (\mathbf{x}_{T_N}, y_{T_N})\}$ with a marginal distribution $\mathcal{P}_T$, where $(\mathbf{x}_i, y_i)$ is the image-label pair, and, in general, $\mathcal{P}_S \neq \mathcal{P}_T$, the objective of transfer learning is to learn a target prediction function $f_T(\cdot)$ using the knowledge of $\mathcal{D}_S$. We study various target prediction tasks, namely, linear evaluation over fixed network for image classification, full-network fine-tuning for image classification, object detection, and few-shot image classification tasks.

### 3.1. Loss Functions

**Supervised Cross-Entropy Loss.** Supervised cross-entropy loss [3] is the standard loss function for multi-class classification. Given an input image $\mathbf{x}$ and one-hot encoded target label $\mathbf{y}$, denote the output representation of the encoder network as $\mathbf{v} = f_\theta(\mathbf{x})$. The class logits are calculated as $\mathbf{l} = \mathbf{W}\mathbf{v} + \mathbf{b}$, where $\mathbf{W} \in \mathbb{R}^{K \times D}$ contains the weights and $\mathbf{b} \in \mathbb{R}^K$ is the bias of the final linear layer. The supervised cross-entropy loss is defined as:

$$\mathcal{L}_{\text{CE}}(\mathbf{l}, \mathbf{y}) = -\sum_{i=1}^{K} y_i \log \left( \frac{\exp(l_i)}{\sum_{j=1}^{K} \exp(l_j)} \right) \quad (1)$$

**Self-Supervised and Supervised Contrastive Loss.** In the contrastive learning paradigm, the network is trained by distinguishing between similar and dissimilar instances. We use Momentum Contrast (MoCo) [21], particularly MoCov2 [8], for studying the efficacy of contrastive representations for transfer learning. The encoder $f_\theta(\cdot)$ of MoCo is a convolutional neural network (CNN), followed by a multi-layer perceptron (MLP) head to embed the encoded features in a contrastive subspace. MoCo has two base networks; one is actively trained to extract query features, and the other is the moving average of the query encoder to extract positive and negative features (commonly known as keys). Denote the query as $\mathbf{q}$ and the set of keys in the queue as $\{\mathbf{k}_1, \mathbf{k}_2, \ldots, \mathbf{k}_M\}$. Assuming the key denoted by $\mathbf{k}_+$ matches with the query $\mathbf{q}$, the objective of MoCo is:



(a) `CE+SelfSupCon` (b) `SupCon+SelfSupCon`

Figure 2: Overview of the `CE+SelfSupCon` and `SupCon+SelfSupCon` models, both of which contain additional headers to optimize the corresponding loss terms.

$$\mathcal{L}_{\texttt{SelfSupCon}}(\mathbf{q}) = -\log \frac{\exp(\mathbf{q} \cdot \mathbf{k}_+ / \tau)}{\sum_{j=1}^{M} \exp(\mathbf{q} \cdot \mathbf{k}_j / \tau)}, \quad (2)$$

where $M$ is the queue size. In self-supervised contrastive loss, the positive key $\mathbf{k}_+$ is obtained from the augmented view of the same input image.

To leverage label information in contrastive learning, we follow the loss function from [30], where the positives are sampled from the contrastive features of the same classes as the query. The supervised contrastive loss is:

$$\mathcal{L}_{\texttt{SupCon}}(\mathbf{q}) = \sum_{\substack{1 \leq j \leq M \\ y(\mathbf{q}) = y(\mathbf{k_j})}} -\log \frac{\exp(\mathbf{q} \cdot \mathbf{k_j} / \tau)}{\sum_{j'=1}^{M} \exp(\mathbf{q} \cdot \mathbf{k_{j'}} / \tau)}. \quad (3)$$

### 3.2. Methods

We examine five different methods to analyze the effectiveness of contrastive learning in tranferability, namely, `CE`, `SelfSupCon`, `SupCon`, `CE+SelfSupCon`, and `SupCon+SelfSupCon`. For all the models, we use the same backbone (ResNet-50 for linear evaluation and fine-tuning; ResNet-18 for few-shot recognition) when evaluating to the downstream datasets. `CE`, `SelfSupCon`, and `SupCon` are trained from the loss functions defined in Eq. 1, 2, and 3, respectively. `SelfSupCon` contains a ResNet backbone and an MLP projection head with two linear layers (2048-d hidden layer with ReLU and 128-d output layer). Note that the MLP projection head is only used during the pretraining stage. See [8] for more details. `CE+SelfSupCon` (Figure 2a) contains two branches on top of the shared backbone; one branch minimizes the supervised cross-entropy loss ($\mathcal{L}_{\texttt{CE}}$), and the other minimizes the self-supervised contrastive loss ($\mathcal{L}_{\texttt{SelfSupCon}}$). The supervised branch contains a linear layer to produce the class

| Category | Dataset | Train Size | Test Size | Classes |
|---|---|---|---|---|
| Natural | CropDisease [38] | 43456 | 10849 | 38 |
| | Flowers [40] | 1020 | 6149 | 102 |
| | DeepWeeds [41] | 12252 | 5257 | 9 |
| Satellite | EuroSAT [22] | 18900 | 8100 | 10 |
| | Resisc45 [9] | 22005 | 9495 | 45 |
| Symbolic | Omniglot [35] | 9226 | 3954 | 1623 |
| | SVHN [39] | 73257 | 26032 | 10 |
| Medical | ISIC [11] | 7007 | 3008 | 7 |
| | ChestX [50] | 18090 | 7758 | 7 |
| Illustrative | Kaokore [46] | 6568 | 821 | 8 |
| | Sketch [49] | 35000 | 15889 | 1000 |
| Texture | DTD [10] | 3760 | 1880 | 47 |

Table 1: Datasets used for downstream image classification.

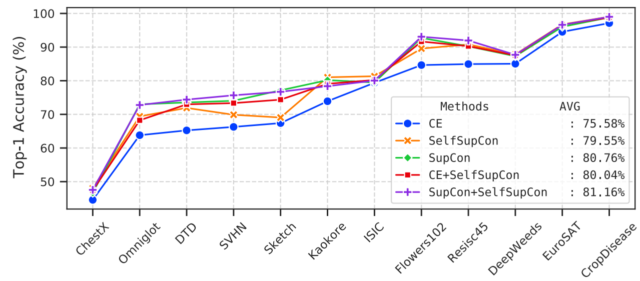



Figure 3: Top-1 accuracy of different models on the downstream datasets for fixed-feature extractor transfer learning (averaged over 5 runs). The models are pretrained on the ImageNet1K dataset and we only train the final linear layer on top of the pretrained backbones. Best viewed in color.

logits. The self-supervised branch consists of an MLP projection head containing two linear layers similar to MoCo. `SupCon+SelfSupCon` (Figure 2b) contains two headers with the shared backbone, one of which optimizes the self-supervised loss ($\mathcal{L}_{\texttt{SelfSupCon}}$), and the other minimizing the supervised contrastive loss ($\mathcal{L}_{\texttt{SupCon}}$). Both branches consist of an MLP projection header similar to MoCo. We also experimented with a single header instead of using two separate branches; however, it diverges during training.

## 4. Experimental Results

In the following sections, we describe the datasets, experimental setup, and results of our analysis. More details about datasets, implementation, and hyperparameter tuning are provided in Appendix A.

### 4.1. Datasets

For source dataset training, we use ImageNet1K training set [12] with 1.28M images for downstream linear evaluation and full-network fine-tuning, and Mini-ImageNet [48] training set with 38k images for few-shot classification. For the downstream task of image classification, we use 12 datasets from different domains to evaluate the transferability of different models. The datasets are categorized as natural, satellite, symbolic, illustrative, medical, and texture. Table 1 describes statistics of the benchmark datasets.

### 4.2. Experimental Setup

We set the temperature parameter of MoCo to $\tau = 0.07$, and queue size to 65596 for ImageNet pretraining and 16384 for Mini-ImageNet pretraining for all contrastive models. For both the cross-entropy and contrastive models, we use the standard data augmentations used in the literature for the respective models. For `CE+SelfSupCon`, we use the same MoCo augmentation for both the cross-entropy and contrastive branches. We also provide ablations on `CE` with MoCo augmentation in Sec. 6. Unless otherwise mentioned, we use top-1 accuracy as the evaluation metric. When performing transfer learning to the downstream datasets, we created a separate validation set from the training set, and swept the hyperparameters (learning rate, batch size, and weight decay) for each dataset. Then we use the optimal hyperparameters to train the full training set (train+val), and evaluate on the test set. We also perform multiple runs with different random seeds, and report mean score among different runs. See Appendix B for detailed results with confidence intervals over multiple runs.

### 4.3. Linear Evaluation over Fixed Network

We first use the linear evaluation over fixed network to test the learned visual representations. For all methods, we freeze the backbone, only train a linear layer on top of the backbone, and optimize the cross-entropy loss with the SGD optimizer. The learning rate, weight decay, and batch size are selected by hyperparameter tuning on the validation set. Figure 3 shows the performance of different models on the 12 downstream datasets in terms of top-1 accuracy (averaged over 5 runs). Note that our reproduced cross-entropy trained model performs slightly differently than the pretrained ResNet-50 model from PyTorch [42].

We note that the cross-entropy model performs the worst as a fixed feature extractor, both in terms of the accuracy for an individual dataset and the final average across all datasets. All of the contrastive approaches perform better than cross-entropy. We also find that training the cross-entropy model with strong augmentation as MoCo does not help much either. The best performing model `SupCon+SelfSupCon` performs, on average, **5.63**% better than `CE`. For datasets that are different than ImageNet, e.g., SVHN, Sketch, Omniglot, and DTD, `CE` performs much worse than contrastive models. We infer that features learned from the cross-entropy model are not directly helpful for datasets that are much different than ImageNet, while features learned from contrastive approaches are applicable to the datasets from different domains. Moreover, `CE+SelfSupCon` performs better than

| | CropDisease | DeepWeeds | Flowers102 | EuroSAT | Resisc45 | ISIC | ChestX | Omniglot | SVHN | Kaokore | Sketch | DTD | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Full Dataset | | | | | | | | | | | | | |
| `CE` | 99.92 | 97.37 | 93.47 | 98.78 | 96.03 | 88.06 | 55.67 | 90.34 | 97.03 | 88.13 | 79.49 | 73.26 | 88.13 |
| `SelfSupCon` | 99.91 | 97.39 | 93.36 | 98.85 | 96.28 | 88.13 | 56.41 | 91.10 | 97.25 | 88.92 | 77.16 | 75.49 | 88.35 |
| `SupCon` | 99.91 | 96.89 | 95.37 | 98.67 | 96.01 | 87.92 | 55.08 | 90.56 | 97.11 | 87.88 | 80.14 | 74.19 | 88.31 |
| `CE+SelfSupCon` | 99.88 | 97.28 | 95.30 | 98.91 | 96.18 | 88.29 | 54.77 | 90.20 | 97.14 | 88.25 | 80.85 | 74.12 | **88.43** |
| `SupCon+SelfSupCon` | 99.91 | 97.38 | 96.17 | 98.75 | 96.21 | 88.55 | 55.12 | 90.86 | 97.04 | 87.73 | 80.17 | 74.86 | **88.56** |
| 1000 training samples | | | | | | | | | | | | | |
| `CE` | 93.89 | 87.46 | 88.31 | 94.68 | 79.33 | 78.26 | 40.42 | 44.95 | 77.65 | 77.28 | 14.87 | 60.43 | 69.79 |
| `SelfSupCon` | 93.95 | 88.10 | 88.92 | 95.47 | 81.14 | 78.27 | 43.09 | 45.33 | 82.37 | 79.41 | 10.57 | 62.95 | 70.80 |
| `SupCon` | 93.93 | 86.89 | 91.53 | 94.73 | 81.95 | 78.51 | 41.97 | 44.48 | 79.25 | 77.41 | 16.31 | 64.71 | **70.97** |
| `CE+SelfSupCon` | 93.60 | 86.83 | 91.01 | 94.51 | 80.43 | 78.68 | 41.65 | 44.08 | 79.85 | 79.14 | 15.72 | 65.07 | 70.88 |
| `SupCon+SelfSupCon` | 93.82 | 87.42 | 91.93 | 95.12 | 81.28 | 78.93 | 41.77 | 44.06 | 80.51 | 78.84 | 15.57 | 66.02 | **71.27** |

Table 2: Performance of different models on the downstream datasets in terms of top-1 accuracy (%) (averaged over 5 runs) for full-network fine-tuning. Contrastive pretrained methods are slightly more effective in a limited data regime than cross-entropy based models.

| Datasets | VOC0712 | | | MS COCO (Trained with 1× schedule) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Detectors | FasterRCNN-R50-C4 | | | MaskRCNN-R50-C4 | | | | | | MaskRCNN-R50-FPN | | | | | |
| Methods | $AP^{bb}_{50}$ | $AP^{bb}$ | $AP^{bb}_{75}$ | $AP^{bb}$ | $AP^{bb}_{50}$ | $AP^{bb}_{75}$ | $AP^{mk}$ | $AP^{mk}_{50}$ | $AP^{mk}_{75}$ | $AP^{bb}$ | $AP^{bb}_{50}$ | $AP^{bb}_{75}$ | $AP^{mk}$ | $AP^{mk}_{50}$ | $AP^{mk}_{75}$ |
| `CE` | 81.58 | 54.63 | 60.17 | 38.91 | 59.05 | 42.14 | 33.98 | 55.39 | 36.10 | 39.84 | 60.54 | 43.60 | 36.54 | 57.65 | 39.25 |
| `SelfSupCon` | 82.48 | 57.33 | 64.03 | 39.14 | 58.74 | 42.40 | 34.29 | 55.53 | 36.62 | 39.19 | 59.54 | 42.74 | 36.03 | 56.79 | 38.75 |
| `SupCon` | 82.64 | 56.12 | 62.45 | 39.63 | 59.76 | 42.77 | 34.52 | 56.26 | 36.54 | 40.40 | 61.20 | 44.35 | 36.97 | 58.31 | 39.80 |
| `CE+SelfSupCon` | **83.06** | 57.07 | 63.97 | 39.68 | 59.84 | 42.80 | 34.62 | 56.31 | 36.74 | **40.65** | 61.38 | 44.65 | **37.21** | 58.57 | 40.06 |
| `SupCon+SelfSupCon` | 82.95 | 57.26 | 64.12 | **39.95** | 60.02 | 43.12 | **34.93** | 56.59 | 37.27 | 40.33 | 60.96 | 44.20 | 36.95 | 58.19 | 39.73 |

The results of VOC0712 is the average of 5 runs. $AP^{bb}$: AP of objection detection; $AP^{mk}$: AP of instance segmentation.

Table 3: Object detection and instance segmentation results on VOC0712 and MS COCO (averaged over 5 runs).

`CE` or `SelfSupCon` and `SupCon+SelfSupCon` performs better than the individual `SupCon` or `SelfSupCon` model, suggesting that *self-supervised contrastive learning improves the transferability of supervised cross-entropy and supervised contrastive learning*.

## 4.4. Full-Network Fine-Tuning

We further fine-tune the full network to study the transferability of all the methods. Here we include the results of image classification and object detection.

**Image Classification.** We fine-tune the pretrained models along with the final linear header on downstream datasets. We perform different hyperparameter sweeping on the validation set and report the scores on the test set for the optimal hyperparameters. Table 2 shows the top-1 accuracy of all datasets for full-network fine-tuning for 5 runs. While `CE` performs much worse than other contrastive models in the linear evaluation experiments, we did not observe similar behavior for full-network fine-tuning. `SupCon+SelfSupCon` achieves 88.56% top-1 accuracy, just 0.43% better than `CE`, which achieves 88.13% top-1 accuracy. We also report the performance of fine-tuning with only 1000 training samples, where `CE+SelfSupCon` performs 1.09% better than `CE`, and `SupCon+SelfSupCon` performs 1.48% better than `CE`. We infer that contrastive pretrained methods are slightly more effective in a limited data regime than cross-entropy based models. However, when we have a sufficient amount of data, all models achieve similar performance.

**Object Detection and Instance Segmentation.** We conduct experiments on object detection and instance segmentation on VOC0712 [15] and MS COCO [36] to validate the learned representations from different models based on Detectron2 [51]. We follow the settings in [21] to fine-tune the whole network while only training a few epochs (1× schedule in the Detectron2 setting). The results are shown in Table 3. `SupCon` provides slightly better results than `CE` and `SelfSupCon`. Furthermore, `CE+SelfSupCon` and `SupCon+SelfSupCon` achieved consistent improvement on AP over their counterparts by ∼0.8% and ∼0.5% on MS COCO, respectively. Again, the results echo our previous observations in linear evaluation experiments.

## 4.5. Few-shot Classification

For few-shot learning, we use the ResNet-18 backbone, and pretrain all models on the Mini-ImageNet dataset. As suggested by [47], a model providing good embedding is essential for few-shot learning; thus, we simply trained a logistic regression classifier on top of the fixed network for the few-shot classification. Table 4 shows the average top-1 accuracy of 600 episodes for 5-way 5-shot and 20-shot experiments. We observe a similar trend that contrastive approaches consistently perform better than cross-entropy across all downstream datasets. `CE+SelfSupCon` and `SupCon+SelfSupCon` achieve the best scores, which suggests that self-supervised contrastive learning improves upon both supervised cross-entropy and supervised contrastive learning. On the other hand, when performing in-

| | Mini-IN* | CropDisease | DeepWeeds | Flowers102 | EuroSAT | Resisc45 | ISIC | ChestX | Omniglot | SVHN | Kaokore | Sketch | DTD | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5-shot | | | | | | | | | | | | | | |
| `CE` | 72.47 | 86.58 | 48.33 | 81.31 | 78.51 | 72.86 | 44.28 | 26.16 | 94.26 | 27.98 | 37.63 | 64.64 | 57.64 | 60.01 |
| `SelfSupCon` | 67.71 | 83.29 | 49.90 | 84.29 | 81.65 | 72.36 | 45.20 | 26.91 | 93.61 | 27.10 | 42.75 | 66.32 | 61.35 | 61.23 |
| `SupCon` | 75.20 | 83.44 | 47.74 | 82.93 | 80.94 | 74.48 | 42.97 | 26.23 | 96.78 | 33.22 | 45.10 | 74.24 | 65.42 | 62.79 |
| `CE+SelfSupCon` | 76.13 | 84.68 | 50.02 | 86.88 | 82.63 | 75.11 | 44.66 | 27.93 | 96.19 | 31.36 | 45.32 | 72.38 | 67.21 | **63.70** |
| `SupCon+SelfSupCon` | 72.81 | 84.26 | 50.35 | 86.72 | 82.57 | 74.94 | 45.82 | 28.19 | 96.72 | 30.67 | 45.26 | 71.09 | 65.08 | **63.47** |
| 20-shot | | | | | | | | | | | | | | |
| `CE` | 80.81 | 92.51 | 58.43 | 89.00 | 84.77 | 82.08 | 52.88 | 29.86 | 97.76 | 35.69 | 46.87 | 76.43 | 67.85 | 67.84 |
| `SelfSupCon` | 76.95 | 90.42 | 58.99 | 90.77 | 87.72 | 82.20 | 53.57 | 32.01 | 97.70 | 34.70 | 52.71 | 77.94 | 71.00 | 69.14 |
| `SupCon` | 83.32 | 90.73 | 55.90 | 90.45 | 87.71 | 83.40 | 51.11 | 31.06 | 98.87 | 44.98 | 54.94 | 85.23 | 74.11 | 70.71 |
| `CE+SelfSupCon` | 84.24 | 91.27 | 58.63 | 92.70 | 89.13 | 84.54 | 52.80 | 33.42 | 98.63 | 41.60 | 54.72 | 83.32 | 75.78 | **71.38** |
| `SupCon+SelfSupCon` | 82.07 | 91.54 | 59.24 | 92.51 | 89.04 | 84.18 | 54.08 | 33.92 | 98.71 | 40.84 | 55.25 | 82.84 | 74.58 | **71.39** |

*: test on the novel classes.

Table 4: Few-shot classification accuracies (average score over 600 episodes) for 5-shot and 20-shot on the Mini-ImageNet and 12 downstream datasets. Contrastive approaches consistently perform better than cross-entropy across all downstream datasets.

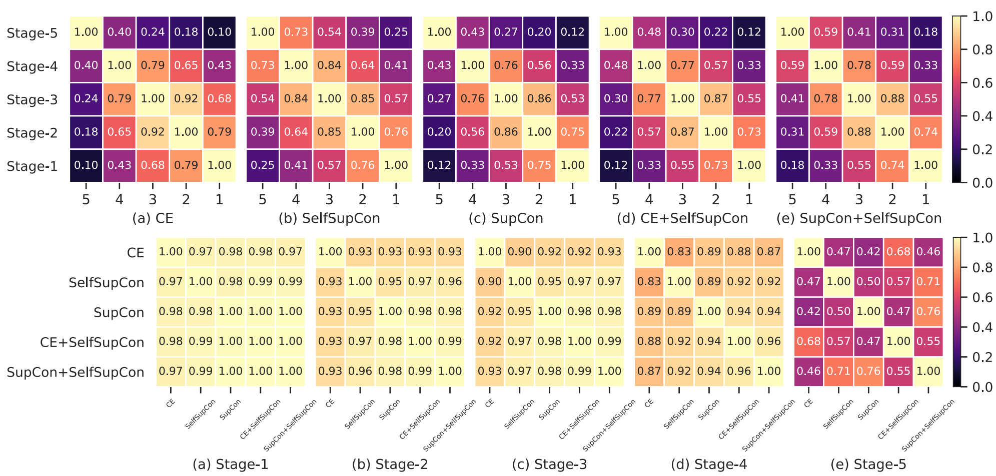



Figure 4: Centered Kernel Alignment (CKA) scores. Top: between different stages of the same model, which shows that final layers of a contrastive model are more similar to initial layers than in the cross-entropy model. Bottom: the same stage between different models, showing that different models learn similar representation in the initial layers but diverge drastically in the final layers.

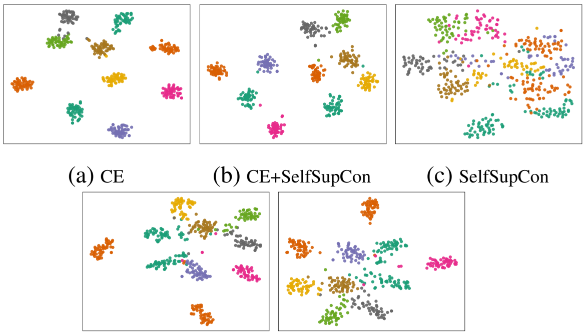
domain few-shot on the Mini-ImageNet test set, we observe that `SupCon+SelfSupCon` performs worse than `SupCon`, suggesting that *the performance of same-domain few-shot classification might not be a good proxy for cross-domain few-shot performance*.

## 5. Discussion and Analysis

**Contrastive approaches learn more low/mid-level features.** Figure 4 (top row) shows the similarity between different stages of the same ResNet-50 model in terms of the centered kernel alignment (CKA) [33]. The initial stages mostly learn low-level features, while the final stages learn more semantic information. The `CE` model has the least similarity between the representations of the final ResNet stage and the initial stage with a CKA score of 0.10. We infer that the final ResNet stages of `CE` contain mostly domain-specific high-level semantics. While `SupCon` is also trained in a supervised fashion to increase inter-class distance, it has a higher CKA score of 0.12 between the final and initial layers, which suggests that the final layers of `SupCon` contain more low-level and mid-level information than `CE`. As expected, the final stage of `SelfSupCon` contains the most low/mid-level information, and `CE+SelfSupCon` and `SupCon+SelfSupCon` show higher CKA scores between the final layers and initial layers than their supervised counterparts.

Similarly, Figure 4 (bottom row) shows the similarity between representations of the same ResNet stages between different models. We observe that the representations from ResNet stage-1 to stage-3 are highly similar between different methods, with CKA scores more than 0.9. However, the representations become slightly differentiated in ResNet stage-4 and highly dissimilar in the final ResNet stage. For example, the CKA between `CE` and `SelfSupCon` in the final layer is only 0.47. Surprisingly, CKA between `CE` and `SupCon` is 0.42 despite the fact that both models were trained with ImageNet1k with full supervision, which suggests that `SupCon` learns a much different feature representation than `CE`. Moreover, all contrastively trained models are more similar to each other than to cross-entropy model.

Supervised learning models learn feature representations using objectives that also increase the inter-class separation. However, we argue that increasing the intra-class variation, though possibly harmful for in-domain performance, is beneficial for learning rich feature representations in transfer learning. t-SNE visualizations in Figure 5 show that the clusters in contrastive methods are more spread out than in



(a) CE (b) CE+SelfSupCon (c) SelfSupCon

(d) SupCon (e) SupCon+SelfSupCon

Figure 5: t-SNE visualization of penultimate layer activations of different models on the ImageNet dataset. The projections of `CE` are much tighter, whereas the projections of contrastive approaches are spread into broad clusters. Best viewed in color.

vanilla cross-entropy, which also supports our claim.

**Contrastive models improve model calibration.** There are several metrics for measuring the models' calibration. Here we adopt the Expected Calibration Error (ECE) and Negative Log Likelihood (NLL) [19]. See Appendix C.1 for the experimental setup.

Table 5 reports the performance for ImageNet pretrained models on the ImageNet validation dataset in terms of Top-1 accuracy (higher is better), Negative Log Likelihood (lower is better), and Expected Calibration Error (lower is better). We see that well-calibrated models do not necessarily have higher accuracy. In particular, `SelfSupCon` shows the best calibration performance, but worst top-1 accuracy score. Moreover, `CE+SelfSupCon` has better ECE and NLL scores than either `CE` or `SelfSupCon`, and `SupCon+SelfSupCon` shows better calibration scores than `SupCon` or `SelfSupCon`. We also evaluate the ImageNet pretrained models on the Stylized ImageNet validation set [16] to see whether contrastive approaches learn both texture-based and shape-based representations. In Table 5, we show that contrastively trained models perform better on the Stylized ImageNet validation set than the cross-entropy model. While `SupCon+SelfSupCon` has slightly lower accuracy than `SupCon`, it improves the calibration by 2.93%. Note that `SelfSupCon` performs the worst in terms of top-1 accuracy, which is expected since the backbone has not been trained with label information, and both of the datasets in Table 5 contain ImageNet classes. However, `SelfSupCon` performs better than `CE` in terms of ECE score. Overall, our experiments suggest that contrastive approaches produce more calibrated predictions than the cross-entropy model on both in-domain evaluation and transfer learning. Calibration performance on the 12 downstream datasets is provided in Appendix C.2.

| Method | ImageNet Top-1(%) | ImageNet NLL | ImageNet ECE(%) | Stylized ImageNet Top-1(%) | Stylized ImageNet ECE(%) |
|---|---|---|---|---|---|
| `CE` | 76.60 | 0.9575 | 6.63 | 7.50 | 33.19 |
| `SelfSupCon` | 66.07 | 1.3943 | **1.60** | 5.34 | 21.59 |
| `SupCon` | **77.32** | 0.9466 | 4.49 | **10.12** | **16.35** |
| `CE+SelfSupCon` | **77.47** | **0.9106** | 5.14 | **9.77** | 24.36 |
| `SupCon+SelfSupCon` | 77.11 | **0.9306** | **4.31** | 9.70 | **13.42** |

Table 5: Performance on ImageNet validation in terms of Top-1 accuracy (higher is better), Negative Log Likelihood (lower is better), and Expected Calibration Error (lower is better). All the models were trained on the ImageNet1K training dataset.

| Method | ImageNet-R Top-1(%) | ImageNet-R ECE(%) | ImageNet-A Top-1(%) | ImageNet-A ECE(%) | ImageNet-C mCE(%) |
|---|---|---|---|---|---|
| `CE` | 35.83 | 19.45 | 3.35 | 55.03 | 60.80 |
| `SelfSupCon` | 26.51 | 25.14 | 2.85 | 48.47 | 68.87 |
| `SupCon` | **42.01** | **13.40** | **7.36** | **46.78** | **54.23** |
| `CE+SelfSupCon` | **41.01** | 17.87 | **7.71** | 51.19 | **54.27** |
| `SupCon+SelfSupCon` | 38.57 | **13.60** | 6.36 | **46.16** | 54.40 |

Table 6: Robustness tests on the ImageNet-R, ImageNet-A, and ImageNet-C datasets. ECE is the expected calibration error (lower is better) and mCE (lower is better) is the mean of the (unnormalized) corruption errors of the Noise, Blur, Weather, and Digital corruptions. Models are trained only on clean ImageNet images.

**Contrastive learning is robust to image corruption.** Many deep learning models lack robustness to natural corruptions. In Table 6, we report the robustness performance of different models on the ImageNet-R [24], ImageNet-A [26], and ImageNet-C [25] datasets. Contrastive approaches, with the exception of `SelfSupCon`, show superior performance for both the ImageNet-A and ImageNet-R datasets, in both top-1 accuracy and expected calibration error (ECE) [19]. The improvement is particularly noticeable between `CE` and `CE+SelfSupCon`. `CE+SelfSupCon` improves the accuracy over `CE` by 5.18% for ImageNet-A and 4.38% for ImageNet-R, and lowers the calibration error by 1.58% for ImageNet-A and 3.72% for ImageNet-R.

The rightmost column of Table 6 reports performance of different models on ImageNet-C in terms of (unnormalized) mean corruption error (mCE) of the Noise, Blur, Weather, and Digital corruptions. Lower mCE denotes that the model is more robust to different corruption types. All the models are trained on clean ImageNet1K dataset. We observe that `CE+SelfSupCon`, `SupCon`, and `SupCon+SelfSupCon` perform the best across different models, and also provide better representations that are transferable to different domains, as shown in the linear evaluation and few-shot experiments. We also note that there is no single contrastive model that works best for all metrics in terms of robustness; however, contrastive loss, in general, improves the neural network robustness.

## 6. Ablation Studies

**Ablations on weights of `SelfSupCon` loss for model with joint objective.** As described in Section 3.1,

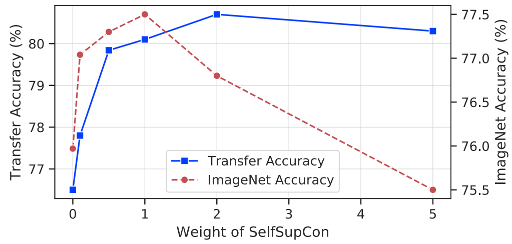



Figure 6: Effect of different weights on the $\mathcal{L}_{\texttt{SelfSupCon}}$ term in the `CE+SelfSupCon` model. Best viewed in color.

| Method | Linear-Evaluation | Finetune | Few-shot |
|---|---|---|---|
| `CE` | 75.67 | 88.13 | 60.01 |
| `CE(strong)` | 75.91 | 88.27 | 61.31 |

Table 7: Performance comparison between `CE` and `CE(strong)` for linear evaluation, full-network fine-tuning, and 5-shot few-shot classification in terms of average top-1 accuracy over the 12 downstream datasets.

`CE+SelfSupCon` is trained with the objective of $\mathcal{L}_{\texttt{CE}} + \alpha\mathcal{L}_{\texttt{SelfSupCon}}$, where $\alpha$ is the weight on self-supervised contrastive loss. Figure 6 reports the effect of $\alpha$ on average transfer accuracy of 12 downstream datasets and ImageNet1K validation accuracy. We obtain the highest ImageNet accuracy for $\alpha = 1$; however, the highest transfer accuracy is reported at $\alpha = 2$, which indicates that higher ImageNet accuracy does not always imply higher transfer accuracy. Higher values of $\alpha$ impose more intra-class variation, and there is an optimal value where transferability of the model is maximized. Imposing more intra-class distance might hurt transfer performance, as shown in figure that when $\alpha > 2$ the transfer accuracy gradually decreases.

**Effect of augmentations on `CE` model.** For the `CE` model, we adopt standard data augmentation as used in ResNet-50 ImageNet training. We also train a cross-entropy model with additional augmentations, such as color-jitter, random gray-scale, and Gaussian blur, so that all the models are pre-trained with similar augmentations. We denote this model as `CE(strong)`. Table 7 shows mean accuracy of 12 downstream datasets for `CE(strong)` in linear evaluation, full-network fine-tune, and few-shot classification (averaged over 5 runs). `CE(strong)` is just slightly better than `CE` for transfer learning; however, contrastive approaches are still significantly better than both `CE` and `CE(strong)`, particularly for fixed-feature transfer.

**Does transferability improve with longer training?** We study the effect of pretrained checkpoints from different ImageNet-training epochs on the transferability of visual representations. Figure 7 shows average transfer accuracy for fixed feature linear evaluation on the 12 downstream datasets for models from different pretraining epochs. The transfer performance of the `CE` pretrained model improves

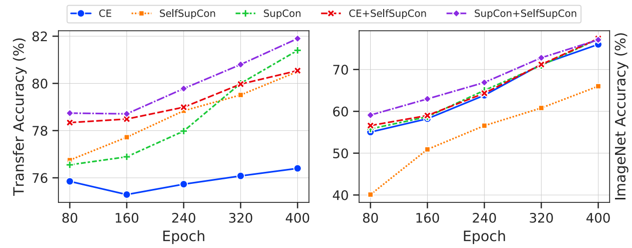



Figure 7: Left: Average linear evaluation accuracy (%) of all the downstream datasets, and Right: ImageNet validation accuracy, for intermediate checkpoints from different ImageNet-pretraining epochs. Transferable representations from the cross-entropy model do not improve much with more pretraining epochs on ImageNet. However, for contrastive approaches, longer pretraining improves transferability. Best viewed in color.

very little after 80 pretraining epochs, suggesting that `CE` learns transferable representation mostly during the initial phase of the pretraining, and it learns more source-domain specific representation at later pretraining epochs. On the other hand, transferability of contrastive methods improve gradually with longer source dataset pretraining. Moreover, during the initial pretraining stages, we find that both `CE+SelfSupCon` and `SupCon+SelfSupCon` still perform better in transfer learning than other models, suggesting that the joint objectives could be beneficial even in resource constrained environments.

# 7. Conclusion

In this paper, we conduct extensive analysis on the transferability of contrastive learning on the downstream image classification, few-shot recognition, and object detection tasks. Our study suggests that contrastive models consistently perform better in transfer learning than standard cross-entropy models, and combining self-supervised contrastive loss with cross-entropy or supervised contrastive loss improves transfer learning performance. We find several factors that make representations from contrastive learning more transferable than supervised cross-entropy model. The penultimate layer representations of contrastive learning are much different than the cross-entropy model; in particular, contrastive models contain more low-level and mid-level information in final layers, and the contrastively trained model shows larger intra-class separation, and contrastive models are more robust to image corruptions.

# 8. Acknowledgement

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. FA8750-19-C-1001. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA).

## References


[1] Machine learning in python. https://scikit-learn.org/stable/. 13

[2] Hossein Azizpour, Ali Sharif Razavian, Josephine Sullivan, Atsuto Maki, and Stefan Carlsson. Factors of transferability for a generic convnet representation. *IEEE transactions on pattern analysis and machine intelligence*, 38(9):1790–1802, 2015. 2

[3] John S Bridle. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In *Neurocomputing*, pages 227–236. Springer, 1990. 3

[4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020. 1, 3

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020. 14

[6] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 2014. 2

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 1, 3

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

[9] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. 4, 12

[10] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2014. 4, 12

[11] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019. 4, 12

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009. 4, 12

[13] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised Visual Representation Learning by Context Prediction. In *ICCV*, 2015. 2

[14] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International Conference on Machine Learning*, pages 647–655, 2014. 2

[15] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 2, 5

[16] 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 *International Conference on Learning Representations*, 2019. 7, 16

[17] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised Representation Learning by Predicting Image Rotations. In *ICLR*. arXiv, 3 2018. 2

[18] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020. 14

[19] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. *arXiv preprint arXiv:1706.04599*, 2017. 7

[20] Yunhui Guo, Noel C Codella, Leonid Karlinsky, James V Codella, John R Smith, Kate Saenko, Tajana Rosing, and Rogerio Feris. A broader study of cross-domain few-shot learning. In *European Conference on Computer Vision*, pages 124–141. Springer, 2020. 13

[21] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 1, 3, 5, 12, 13

[22] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 4, 12

[23] 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 preprint arXiv:1905.09272*, 2019. 2

[24] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv preprint arXiv:2006.16241*, 2020. 7

[25] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 7

[26] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *arXiv preprint arXiv:1907.07174*, 2019. 7

[27] Minyoung Huh, Pulkit Agrawal, and Alexei A Efros. What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*, 2016. 2

[28] Ashraful Islam, Chun-Fu Chen, Rameswar Panda, Leonid Karlinsky, Rogerio Feris, and Richard J Radke. Dynamic distillation network for cross-domain few-shot recognition with unlabeled data. *arXiv preprint arXiv:2106.07807*, 2021. 3

[29] Ziyu Jiang, Tianlong Chen, Ting Chen, and Zhangyang Wang. Robust pre-training by adversarial contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 16199–16210. Curran Associates, Inc., 2020. 3

[30] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020. 1, 3

[31] Youngsung Kim, Jinwoo Shin, Eunho Yang, and Sung Ju Hwang. Few-shot visual reasoning with meta-analogical contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 16846–16856. Curran Associates, Inc., 2020. 3

[32] Simon Kornblith, Honglak Lee, Ting Chen, and Mohammad Norouzi. What's in a loss function for image classification? *arXiv preprint arXiv:2010.16402*, 2020. 16

[33] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. *arXiv preprint arXiv:1905.00414*, 2019. 2, 6

[34] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2661–2671, 2019. 2

[35] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015. 4, 12

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 2, 5

[37] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 14

[38] Sharada P Mohanty, David P Hughes, and Marcel Salathé. Using deep learning for image-based plant disease detection. *Frontiers in plant science*, 7:1419, 2016. 4, 12

[39] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Ng. Reading digits in natural images with unsupervised feature learning. *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, pages 722–729, 2011. 4, 12

[40] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008. 4, 12

[41] Alex Olsen, Dmitry A. Konovalov, Bronson Philippa, Peter Ridd, Jake C. Wood, Jamie Johns, Wesley Banks, Benjamin Girgenti, Owen Kenny, James Whinney, Brendan Calvert, Mostafa Rahimi Azghadi, and Ronald D. White. DeepWeeds: A Multiclass Weed Species Image Dataset for Deep Learning. *Scientific Reports*, 9(2058), 2 2019. 4, 12

[42] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 4, 17

[43] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context Encoders: Feature Learning by Inpainting. In *CVPR*, volume 2016-December, pages 2536–2544. IEEE Computer Society, 4 2016. 2

[44] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops*, pages 806–813, 2014. 2

[45] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. *arXiv preprint arXiv:2005.10243*, 2020. 1

[46] Yingtao Tian, Chikahiko Suzuki, Tarin Clanuwat, Mikel Bober-Irizar, Alex Lamb, and Asanobu Kitamoto. Kaokore: A pre-modern japanese art facial expression dataset. *arXiv preprint arXiv:2002.08595*, 2020. 4, 12

[47] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? *arXiv preprint arXiv:2003.11539*, 2020. 5, 13

[48] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638, 2016. 4

[49] Haohan Wang, Songwei Ge, Eric P. Xing, and Zachary C. Lipton. Learning robust global representations by penalizing local predictive power. *arXiv preprint arXiv:1905.13549*, 2019. 4, 12

[50] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2097–2106, 2017. 4, 12

[51] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 5

[52] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In

*Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1476–1485, 2019. 3

[53] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful Image Colorization. In *ECCV*, volume 9907 LNCS, pages 649–666. Springer Verlag, 3 2016. 2

[54] Nanxuan Zhao, Zhirong Wu, R. Lau, and Stephen Lin. What makes instance discrimination good for transfer learning? *arXiv preprint arXiv:2006.06606*, 2020. 1, 2