Revisiting the Transferability of Supervised Pretraining: an MLP Perspective

Yizhou Wang[1,3*†], Shixiang Tang[2†], Feng Zhu[3], Lei Bai[2‡], Rui Zhao[3,4], Donglian Qi[1], Wanli Ouyang[2]
[1]Zhejiang University, [2]The University of Sydney,[3]SenseTime Research,
[4]Qing Yuan Research Institute, Shanghai Jiao Tong University, Shanghai, China
{yizhouwang, qidl}@zju.edu.cn, stan3906@uni.sydney.edu.au, baisanshi@gmail.com,
{zhufeng, zhaorui}@sensetime.com, wanli.ouyang@sydney.edu.au

Abstract

The pretrain-finetune paradigm is a classical pipeline in visual learning. Recent progress on unsupervised pretraining methods shows superior transfer performance to their supervised counterparts. This paper revisits this phenomenon and sheds new light on understanding the transferability gap between unsupervised and supervised pretraining from a multilayer perceptron (MLP) perspective. While previous works [6, 8, 17] focus on the effectiveness of MLP on unsupervised image classification where pretraining and evaluation are conducted on the same dataset, we reveal that the MLP projector is also the key factor to better transferability of unsupervised pretraining methods than supervised pretraining methods. Based on this observation, we attempt to close the transferability gap between supervised and unsupervised pretraining by adding an MLP projector before the classifier in supervised pretraining. Our analysis indicates that the MLP projector can help retain intra-class variation of visual features, decrease the feature distribution distance between pretraining and evaluation datasets, and reduce feature redundancy. Extensive experiments on public benchmarks demonstrate that the added MLP projector significantly boosts the transferability of supervised pretraining, e.g. ***+7.2%** top-1 accuracy on the concept generalization task, **+5.8%** top-1 accuracy for linear evaluation on 12-domain classification tasks, and **+0.8%** AP on COCO object detection task, making supervised pretraining comparable or even better than unsupervised pretraining.*

1. Introduction

While Supervised Learning with the cross-entropy loss[1] (SL) were the de facto pretraining paradigm in computer vision [14, 20, 26, 39] for a long period, recent unsupervised learning methods [3–9, 15, 17, 18, 49, 52] show better transfer learning performance on various visual tasks [17, 22, 53]. This raised the question of why unsupervised pretraining surpasses supervised pretraining even though supervised pretraining uses annotations with rich semantic information.

Several works have attempted to explain the better transferability of unsupervised pretraining than supervised pretraining by the following two reasons: (1) *Learning without semantic information in annotations* [16, 37, 45, 53], which makes the backbone less overfit to semantic labels to preserve instance-specific information which may be useful in transfer tasks, and (2) *Special design of the contrastive loss* [22, 23, 53], which helps the learned features to contain more low/mid-level information for effective transfer to downstream tasks. Starting from the perspective of supervision and loss design, these works provide intuitive explanations for better transferability.

In this paper, we shed new light on understanding transferability by considering the multilayer perception (MLP) projector. While previous works [6, 8, 17] verified its effectiveness on the unsupervised image classification task: unsupervised training and evaluating the model on the same ImagNet-1K dataset, they did not explore its effectiveness on transfer tasks thoroughly and rigorously. It is not straightforward to extend the effectiveness of MLP on the unsupervised image classification task to downstream tasks if not supported by rigorous experiments or theoretical analysis, because the performance on the pretraining task is not always predictive of the performance on transfer tasks when there exists a large semantic gap [16, 35, 43]. To our best knowledge, we are the first to identify the MLP projector as the core factor for the transferability with deep empirical and theoretical analysis. With this new viewpoint, we find that a simple yet effective method, adding an MLP projector, can promote the transferability of the conventional

[1]In the paper, we specifically use the notation "SL" to indicate the conventional supervised learning with the cross-entropy loss.

∗ The work was done during an internship at SenseTime.
† Equal Contribution.
‡ Corresponding author.

supervised pretraining methods with the cross-entropy loss (SL) to be comparable or even better than representative unsupervised pretraining methods.

Specifically, we use the *concept generalization task* [37] on ImageNet-1K, where the pretraining and the evaluation datasets have a large semantic distance, as a probe to analyze the transferability of different models. Our experimental results and corresponding analysis indicate that the MLP projector in unsupervised pretraining methods is important for their better transferability. Motivated by this observation, we insert an MLP projector before the classifier in SL, forming SL-MLP. The added MLP can improve the transferability of supervised pretraining, making supervised pretraining comparable or even better than unsupervised pretraining. Experimental results on SL and SL-MLP show three interesting findings: 1) The added MLP preserves the intra-class variation on the pretraining dataset. 2) The added MLP decreases the feature distribution distance between the pretraining and the evaluation dataset; 3) The added MLP decreases the feature redundancy in the pretraining dataset. We also provide theoretical analysis on how the preserved intra-class variation and the decreased feature distribution distance improve the performance on the target dataset, by adding an MLP projector.

Extensive experimental results confirm that adding an MLP projector into the supervised pretraining method (SL) can consistently improve the transferability of the model on various downstream tasks. Specifically, on the concept generalization task [37], SL-MLP boosts the top-1 accuracy compared to SL (55.9%→63.1%) by **+7.2%**. It also achieves better performance (64.1%) than Byol (62.3%) by **+1.8%** on the 300-epochs pretraining setting. In classification tasks on 12 cross-domain datasets [22], SL-MLP improves SL by **+5.8**% accuracy on average. Moreover, SL-MLP shows better transferability than SL on COCO object detection [25] by **+0.8**% AP. These improvements brought by the MLP projector can largely bridge the transferability gap between supervised and unsupervised pretraining as detailed in Sec. 5.2.

The main contributions of our paper are three-fold. (1) We reveal that the MLP projector is the main factor for the transferability gap between existing unsupervised and supervised learning methods. (2) We empirically demonstrate that, by adding an MLP projector, supervised pretraining methods can have comparable or even better transferability than representative unsupervised pretraining methods. (3) We theoretically prove that the MLP projector can improve transferability of pretrained models by preserving intra-class feature variation.

2. Related Works

MLP in unsupervised learning methods. Adding a multi-layer perceptron (MLP) projector after the encoder was first introduced in SimCLR [6] and followed by recent unsupervised learning frameworks [3, 7–9, 17, 49]. SimCLR claims that the MLP can reduce the loss of information caused by the contrastive loss, and various works [6, 8] have verified that the MLP projector can enhance the discriminative ability of unsupervised models on the unsupervised image classification task, where unsupervised training and evaluation are conducted on the same dataset. However, the relation between the MLP and the transferability of unsupervised learning methods is under-explored. In this paper, we reveal that the MLP projector is also important for the desirable transferability of unsupervised learning.

MLP in supervised learning methods. The typical supervised learning method (SL) only uses the cross-entropy loss and shows inferior performance on various transfer tasks than recent unsupervised learning methods. Inspired by [15, 48], recent works [22, 23] introduced the contrastive loss equipped with an MLP projector into SL to improve its transferability. Nonetheless, those works ignored the ablation on the MLP and attributed the better transfer performance to the contrastive mechanism in the loss. In this paper, we propose that the MLP is important for the improved transferability of recent supervised learning methods [22, 23] with empirical and theoretical analysis.

Transferability gap between supervised and unsupervised learning. Previous works attributed the superior transferability of unsupervised learning to *lack of annotation* [16, 37, 45, 53] or *special design of the contrastive loss* [22, 23, 40, 53]. Different from both reasons, we explain the transferability gap by considering the architectural difference between the supervised and the unsupervised learning frameworks. From this perspective, we analyze the role of the MLP projector in both supervised and unsupervised learning methods, and are the first to identify its key importance to model transferability to the best of our knowledge.

3. Transferability Analysis of the Unsupervised and Supervised Pretraining Methods

3.1. The Concept Generalization Task

We use the concept generalization task [37] to analyze the transferability gap between the unsupervised and supervised pretraining methods.

Data preparation. Sariyildiz *et al*. [37] evaluated the transferability of methods when the pretraining and evaluation dataset have semantic distance. Their experimental results show that larger semantic distance will lead to more accuracy differences among different pretraining methods. Therefore, we enlarge the semantic gap between the pretraining and the evaluation dataset to help us compare different pretraining methods. Sariyildiz*et al*. [37] use the hierarchy in WordNet [29] and divide ImageNet-21K [13] into six class-exclusive datasets with different semantic distance

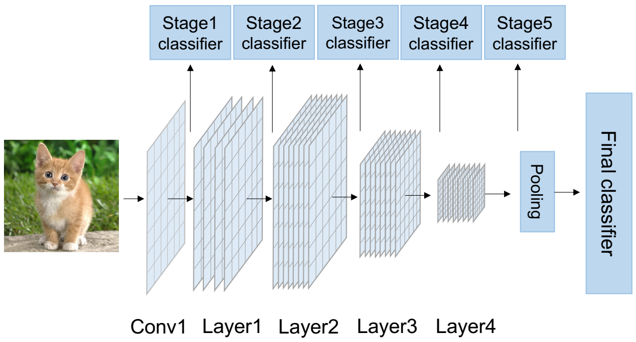


Figure 1. Schematic illustration of stage-wise evaluation. We flatten intermediate feature maps from different stages and then use them to train stage-wise classifiers. Top-1 accuracy is reported by evaluating images in eval-D with the stage-wise classifiers.

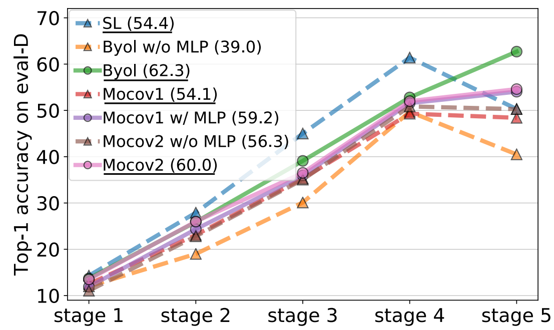


Figure 2. Top-1 accuracy of stage-wise evaluation. All methods use ResNet50 as their backbones and are trained by 300 epochs with the setting in original papers. The results of linear evaluation of layer4-pooled-features (see Fig. 1) are reported in the legend.

– one for pretraining, and others for evaluation. Without loss of generality, we construct a smaller pretraining dataset (pre-D) and evaluation dataset (eval-D) based on ImageNet-1K [36] to reduce the experimental burden. Pre-D contains 652 classes mostly of organisms, and eval-D contains the other 348 classes of instrumentality.

Transferability evaluation. Following [37], to assess the transferability, we freeze all parameters in the pretrained backbone [2], and finetune the classifier with the ImageNet-1K training samples in eval-D for reporting top-1 accuracy on ImageNet-1K validation samples in eval-D.

3.2. Stage-wise Evaluation on Existing Methods

Motivated by works [22, 50, 53] we make a thorough stage-wise investigation of the conventional supervised pretraining method (SL) and the existing representative unsupervised pretraining methods (Mocov1 [18], Mocov2 [8], Byol [17]) by evaluating the transferability of intermediate feature maps (Fig. 1). After pretraining the model on pre-D, we freeze all model parameters and use the extracted intermediate feature maps of images in eval-D to finetune a stage-wise classifier for a stage-wise linear evaluation.

The evaluation results of these existing methods are depicted in Fig. 2 (underlined on the legend). Our stage-wise evaluation shows two new findings that have not been reported by existing works. First, on stage-wise evaluation from stage 1 to stage 4, SL is consistently higher than Byol, Mocov1, and Mocov2, which suggests that the semantic information in annotations can benefit the transferability of low/middle-level feature maps. Second, on stage-wise evaluation from stage 4 to stage 5, the performance of Byol and Mocov2 still increase while SL and Mocov1 have a transferability drop. By carefully inspecting these methods, we notice an architectural difference between SL, Mocov1, Mocov2, and Byol after stage 5: An MLP projector is inserted after stage 5 in Byol and Mocov2, which does not exist in SL and Mocov1. Such difference, together with the experimental results in Fig. 2, leads to a new hypothesis that the MLP projector might be the core factor of the desirable transferability of unsupervised pretraining.

3.3. MLP Improves the Transferability of Unsupervised Pretraining Methods

To confirm our hypothesis of the effectiveness on unsupervised pretraining methods, we ablate the MLP projectors on existing unsupervised methods, [3] using stage-wise evaluation. Specifically, we remove the MLP projector in Byol and Mocov2 as Byol w/o MLP and Mocov2 w/o MLP, and add an MLP projector in Mocov1 as Mocov1 w/ MLP. The stage-wise evaluation results of these ablations are summarized in Fig. 2. We use solid lines for methods that have an MLP projector and dash lines for those that do not.

These ablation results offer us two observations. First, when evaluating the layer4-pooled-features (depicted in the legend), unsupervised learning methods with an MLP projector achieve better transferability than their variants without the MLP projector, *e.g.*, Byol, Mocov1 w/ MLP, Mocov2 achieve higher accuracy than Byol w/o MLP, Mocov1, and Mocov2 w/o MLP by $+23.3\%$, $+5.1\%$ and $+3.7\%$, respectively. Second, on stage-wise evaluation from stage 4 to stage 5, the MLP projector can help unsupervised learning methods without the MLP projector to avoid the transferability drop. These consistent improvements by adding an MLP projector empirically show that the MLP projector is important for the transferability of unsupervised pretraining. While there might exist some other non-linear structures that can boost the transferability, we only explore from an MLP perspective in this paper because of its simplicity and demonstrated effectiveness.

[2] All experiments in Sec. 3 and Sec. 4 are conducted with ResNet50.

[3] We do not directly compare Mocov1 with Mocov2 because Mocov2 has more augmentations and the different learning rate schedule.

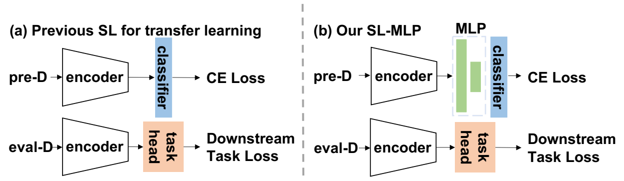


Figure 3. The difference between SL and SL-MLP. Our SL-MLP adds an MLP before the classifier compared to SL. Only the encoders in both methods are utilized for downstream tasks.

4. MLP Can Enhance Supervised Pretraining

4.1. SL-MLP: Adding an MLP Projector to SL

Motivated by the empirical results in Sec. 3, an interesting question is whether the MLP projector can also promote the transferability of supervised pretraining? We attempt to insert an MLP projector before the classifier on SL for better transferability. We denote this supervised pretraining method as SL-MLP (see Fig. 3 for their comparison). Specifically, SL-MLP includes a feature extractor $f(\cdot)$, an MLP projector $g(\cdot)$, and a classifier $\mathbf{W}$. Given an input image $\mathbf{x}$, the feature extractor outputs a feature $\mathbf{f} = f(\mathbf{x})$. For example, $f(\mathbf{x})$ transforms an image $\mathbf{x}$ to a 2048 dimensional feature $\mathbf{f}$ when using the ResNet-50 backbone. The MLP projector maps $\mathbf{f}$ into a projection vector $\mathbf{g} = g(\mathbf{f})$. Following Byol, the MLP projector consists of two fully connected layers, a batch normalization layer, and a ReLU layer, which can be mathematically formulated as $g(\mathbf{f}) = fc_2(ReLU(BN(fc_1(\mathbf{f})))) \in \mathbb{R}^{D_g}$, where fc_1 and fc_2 are fully connected layers, the hidden feature dimension in the MLP projector is set to 4096, and D_g is set to 256. Given the label denoted by y for image $\mathbf{x}$, the objective function for SL-MLP can be formulated as

$$\mathcal{L}(\mathbf{x}) = \text{CE}(\mathbf{W} \cdot g(f(\mathbf{x})), y), \tag{1}$$

where CE$(\cdot)$ is the cross-entropy loss. Same as SL, only the learned feature extractor $f(\cdot)$ is utilized in downstream transfer tasks after supervised pretraining.

4.2. Empirical Findings of MLP in SL-MLP

MLP projector avoids transferability drop at stage 5 in supervised pretraining. We conduct stage-wise evaluation as Sec. 3.2 again to see whether the transferability drop from stage 4 to stage 5 exists in SL-MLP. In Fig. 6(a), the transferability of SL-MLP continuously increases from stage 1 to 5, avoiding a decrease at stage 5 as SL. Besides, we observe that the transferability of SL-MLP is higher than that of Byol from stage 1 to 4, indicating that annotations benefit the transferability of intermediate feature maps.

MLP projector enlarges the intra-class variation of features. According to [22, 53], features with large intra-class variation can preserve more instance discriminative information, which is beneficial for transfer learning. We reveal that adding an MLP projector also can enlarge the intra-class variation. We compare two supervised pretraining methods, *i.e.,* SL, SupCon [23], and one unsupervised pretraining method, *i.e.,* Byol, with their variants with/without MLP. Qualitatively, we visualize their features learned on pre-D by t-SNE in Fig. 4. The intra-class variation of features from SL-MLP, SupCon, and Byol are larger than that from SL, SupCon w/o MLP, and Byol w/o MLP, respectively. Quantitatively, following LDA [1], we utilize a discriminative ratio $\phi(I^{pre})$ to measure intra-class variation on pre-D, where I^{pre} denotes pre-D (mathematically defined in Sec. 4.3). Smaller discriminative ratio ϕ usually means larger intra-class variation[4]. Comparing Fig. 6(c) with Fig. 6(b), we can see Byol and SL-MLP have smaller $\phi(I^{pre})$ but higher accuracy on eval-D than SL, which shows larger intra-class variation can benefit transferability. Furthermore, when inspecting SL only, we can see a process where the accuracy on eval-D first rises and then descends (after 210 epochs) along with $\phi(I^{pre})$ increasing. This phenomenon can be theoretically explained in Sec. 4.3. We additionally provide the visualization of intra-class variation on different pretraining epochs in Supplementary B.1.

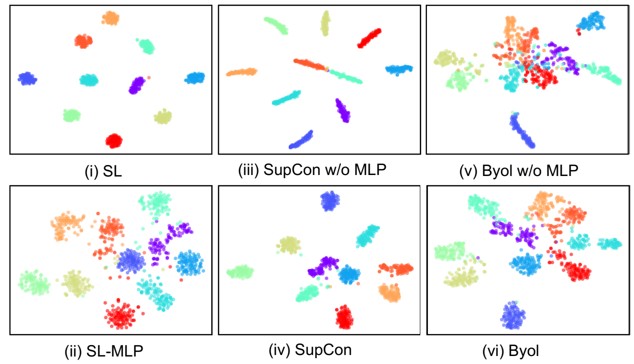


Figure 4. Visualization of different methods with 10 randomly selected classes on pre-D. Different colors denote different classes. Features extracted by pretrained models without an MLP projector (top row) have less intra-class variation than those extracted by pretrained models with an MLP projector (bottom row).

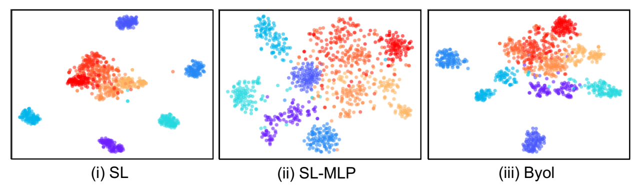
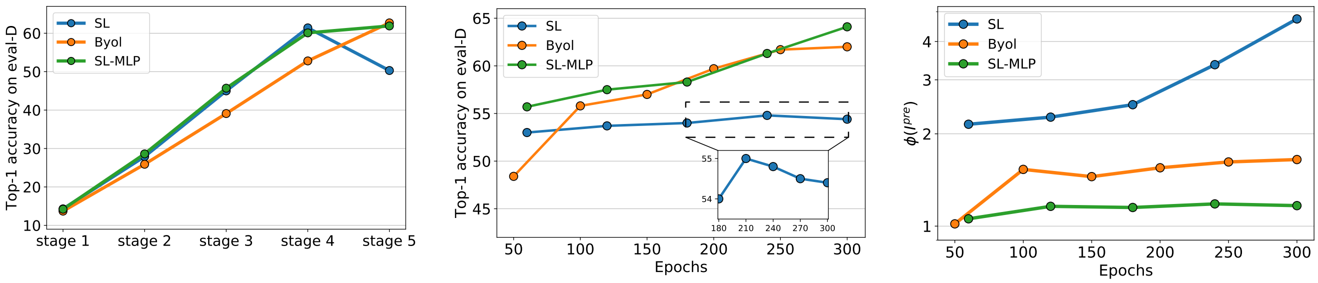


Figure 5. Visualization of Feature Mixtureness between pre-D and eval-D. Cold colors denote features from 5 classes that are randomly selected from pre-D, and warm colors denote features from 5 classes that are randomly selected from eval-D.

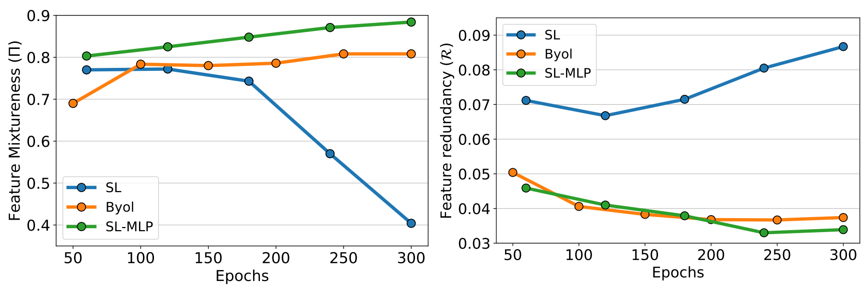
MLP projector reduces feature distribution distance be-

[4]Strictly speaking, larger intra-class variation is relative to inter-class distance, which is theoretically defined as discriminative ratio. We use "intra-class variation" to be consistent with previous work [22, 53].



(a) (b) (c)

Figure 6. **(a)** Stage-wise evaluation on eval-D. **(b)** Linear evaluation accuracy on eval-D. **(c)** Discriminative ratio of features on pre-D. Following [17, 19], we pretrain SL, SL-MLP, and Byol for 300 epochs.



(a) (b)

Figure 7. **(a)** Feature Mixtureness between pre-D and eval-D. **(b)** Redundancy $\mathcal{R}$ of pretrained features during different epochs. Following [17, 19], we pretrain SL, SL-MLP, and Byol for 300 epochs.

Method	EP	Top-1(↑)	$\mathcal{R}$(↓)
SL	100	55.9	0.078
SL-MLP	100	**63.1**	**0.035**
SL	300	54.4	0.087
SL-MLP	300	**64.1**	**0.034**
Byol w/o MLP	300	39.0	0.247
Byol	300	**62.3**	**0.037**
Mocov1	300	54.1	0.069
Mocov1 w/ MLP	300	**59.2**	**0.058**

Table 1. Redundancy $\mathcal{R}$ of pretrained features. Methods with an MLP obtain lower channel redundancy and transfer better.

tween pre-D and eval-D. According to [2, 27], decreasing the feature distribution distance between pre-D and eval-D in the representation space can benefit transfer learning. Intuitively, the distribution distance between two sets of features is small when features are well mixed (visualization provided in Supplementary A). Therefore, we compare the mixtureness of features in pre-D and eval-D to indicate the feature distribution distance between SL and SL-MLP. Graphically, we visualize features from pre-D and eval-D by t-SNE in Fig. 5. We observe that features from pre-D and eval-D are more mixed comparing SL and SL-MLP, indicating that MLP projector can mitigate the distribution distance between pre-D and eval-D. Quantitatively, we define *Feature Mixtureness* Π in the feature space as

$$\Pi = 1 - \frac{1}{C}\sum_{i=1}^{C}\left|\frac{top_k^{eval}(i)}{k} - \frac{C^{eval}}{C}\right|, \tag{2}$$

where $C = 1000$ is total number of classes in ImageNet-1K, C^{eval} represents the number of classes in eval-D, and $top_k^{eval}(i)$ represents the number of classes in eval-D found by top k neighbors search of any class $i \in C$. Since the percentage of finding a sample from eval-D in k nearest neighbors is C^{eval}/C when pre-D and eval-D are uniformly mixed, Feature Mixtureness measures the similarity of the current and the uniformly mixed distribution between pre-D and eval-D in the feature space. We examine Feature Mixtureness of SL, SL-MLP, and Byol during different pretraining epochs in Fig. 7(a). Feature Mixtureness of SL gradually decreases, which indicates that SL will enlarge the distribution difference between pre-D and eval-D. In contrast, SL-MLP and Byol show consistently high Feature Mixtureness, indicating that the MLP projector can reduce the distribution distance between pre-D and eval-D. We visualize the evolution of Feature Mixtureness in Supplementary B.2.

MLP projector reduces feature redundancy. Inspired by [52], high channel redundancy limits the capability of feature expression in deep learning. Mathematically, we compute Pearson correlation coefficient among feature channels to evaluate *feature redundancy* $\mathcal{R}$, *i.e,*

$$\mathcal{R} = \frac{1}{d^2}\sum_{i=1}^{d}\sum_{j=1}^{d}|\rho(i,j)|, \quad \rho(i,j) = \frac{\sum_{n=1}^{N}\mathbf{f}_{n,i}\cdot\mathbf{f}_{n,j}}{\sqrt{\sum_{n=1}^{N}||\mathbf{f}_{n,i}||^2}\sqrt{\sum_{n=1}^{N}||\mathbf{f}_{n,j}||^2}} \tag{3}$$

where $d = 2048$ is the feature dimension, $\rho(i,j)$ is Pearson correlation coefficient of feature channel i and j. As shown in Fig. 7(b), SL-MLP has lower feature redundancy than SL, which indicates that the MLP projector can reduce feature redundancy. In Tab. 1, we further confirm that the MLP projector can reduce the feature redundancy and thus increase the accuracy on eval-D by ablating the MLP projector on various pretraining methods.

4.3. Theoretical Analysis for Empirical Findings

In this section, we provide a theoretical analysis to reveal that: 1) maximizing the discriminative ratio $\phi(I^{pre})$ of

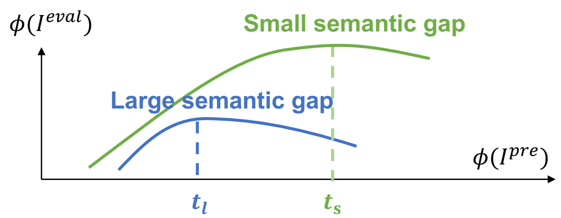


Figure 9. Insights for transferability. $\phi(I^{pre})$ and $\phi(I^{eval})$ are the discriminative ratios (Eq. 4) on the pretraining and evalution datasets. Higher $\phi(I^{eval})$ indicates better model transferability. Green and Blue line show the performance curve on the evaluation dataset with small and large semantic gap, respectively.

a model on the pretraining dataset above a certain threshold will lead to a transferability descrease (shown by blue/green lines in Fig. 9); 2) the threshold is smaller when the semantic gap between the pretraining and evalutaion dataset is larger ($t_l < t_s$ in Fig. 9).

Mathematically, the discriminative ratio $\phi(I)$ on the dataset I can be defined by LDA [1] as

$$\phi(I) = D_{inter}(I)/D_{intra}(I), \tag{4}$$

where $D_{inter}(I) = \frac{1}{C(C-1)} \sum_{j=1}^{C} \sum_{k=1,k\neq j}^{C} ||\mu(I_j) - \mu(I_k)||^2$ is the inter-class distance, $D_{intra}(I) = \frac{1}{C} \sum_{j=1}^{C} (\frac{1}{|I_j|} \sum_{(x_i,y_i)\in I_j} ||\mathbf{f}_i - \mu(I_j)||^2)$ is the intra-class distance, and C is the number of classes. $\mu(I_j) = \frac{1}{I_j} \sum_{(x_i,y_i)\in I_j} \mathbf{f}_i$ is the center of features in class I_j, and $\mathbf{f}$ is the feature in Sec. 4.1. Higher discriminative ratio ϕ indicates higher classification accuracy. Inspired by [26], we analyze the relation between $\phi(I^{pre})$ and $\phi(I^{eval})$ in Theorem 1 (Supplementary C).

Theorem 1 *Given $\phi_1(I^{pre}) < \phi_2(I^{pre})$, $\phi_1(I^{eval}) > \phi_2(I^{eval})$ when $\phi_1(I^{pre}) > t$, where t is a threshold that is negatively related to the feature distribution distance.*

Insights for the intra-class variation. Theorem 1 reveals that continuously minimizing the intra-class variation (maximizing the discriminative ratio) on the pretraining dataset will decrease the transferability of the model when the discriminative ratio $\phi(I^{pre})$ is larger than t. It explains the observation in Fig. 6(b) and Fig. 6(c) that training with more than 210 epochs leads to better performance on pre-D, but a worse transferability on eval-D. This insight suggests that we should not make the intra-class variation on the pretraining dataset too small when designing the objective function or network architecture (*e.g.,* adding an MLP projector).

Insights for the relation between the feature distribution distance and threshold t. When the feature distribution distance between the pretraining and evaluation dataset is large, the threshold t is small, in which case it is easier to have the undesirable effect of increasing the discriminative ratio $\phi(I^{pre})$ on pre-D leading to decreasing the discriminative ratio $\phi(I^{eval})$ on eval-D (and thus the accuracy on the evaluation data). This insight suggests that we should maintain more intra-class variation on the pretraining dataset when transferring the model to a target dataset which has a larger semantic distance from the pretraining dataset.

5. Experiment

5.1. Experimental Setup

Datasets. For backbone analysis, we keep using the concept generalization setting described in Sec. 3.1. For generalization to other classification tasks, we follow the setup in [22], which pretrains the models on the whole ImageNet-1K dataset and then evaluates the transferability on 12 classification datasets [10–12, 21, 24, 31–34, 41, 42, 44] from different domains. Furthermore, the COCO [25] dataset is used to evaluate the performance of SL-MLP pretrained by ImageNet-1K [36] on object detection task.

Details. For SL and SL-MLP pretraining, the cross-entropy is deployed as the loss function. The MLP projector deployed in SL-MLP is described in Sec. 4.1. Following [20], we use the SGD optimizer with a cosine decay learning rate of 0.4 to optimize SL and SL-MLP, and set the batch size to 1024. Byol is used as a representative method for comparisons in backbone analysis and object detection. Following [17], we use LARS optimizer [51] with a cosine decay learning rate schedule and a warm-up of 10 epochs to pretrain the network. The initial learning rate is set to 4.8. We set the batch size to 4096 and the initial exponential moving average parameter τ to 0.99. Except for the backbone analysis, we use ResNet50 as the default backbone. More detailed pretraining setups of different backbones and different methods are provided in Supplementary H.1.

5.2. Experimental Results

Generalize to unseen concepts with diverse backbones. We verify the effectiveness of the added MLP projector on SL using *concept generalization task* with different backbones. Following evaluation method mentioned in Sec. 3.1, we train a linear classifier with the frozen backbone for 100 epochs, and report the top-1 accuracy on eval-D in Tab. 2. Firstly, SL-MLP obtains better performance than SL among different backbones. Specifically, with ResNet50, SL-MLP improves SL to 63.1 (+7.2%) when we pretrain the model by only 100 epochs, which bridges the performance gap between SL and Byol. In 300 epochs setting, SL has a transferability drop compared to 100 epochs setting (55.9%→54.4%), but the transferability of SL-MLP continue to increase (63.1%→64.1%). Secondly, SL-MLP (64.1%) performs better than Byol (62.3%) when both are trained by 300 epochs. Experimental results in Tab. 2 also confirm that SL-MLP can consistently improve the trans-

Table 2. Concept generalization task. We report Top-1 accuracy on eval-D of SL-MLP, Byol, and SL on various backbones. SL-MLP and Byol share the same MLP projector.

Method	Architecture	Labels	MLP	Epochs	Top-1(↑)
SL	ResNet50	✓		100	55.9
SL-MLP	ResNet50	✓	✓	100	63.1
Byol	ResNet50		✓	300	62.3
SL	ResNet50	✓		300	54.4
SL-MLP	ResNet50	✓	✓	300	**64.1**
SL	ResNet34	✓		100	50.1
SL-MLP	ResNet34	✓	✓	100	55.0
Byol	ResNet34		✓	300	54.8
SL	ResNet34	✓		300	50.2
SL-MLP	ResNet34	✓	✓	300	**55.8**
SL	ResNet101	✓		100	56.0
SL-MLP	ResNet101	✓	✓	100	63.6
SL	ResNet101	✓		300	53.9
SL-MLP	ResNet101	✓	✓	300	**64.7**
SL	Swin-tiny	✓		100	58.9
SL-MLP	Swin-tiny	✓	✓	100	**60.6**
SL	EfficientNetb2	✓		100	57.6
SL-MLP	EfficientNetb2	✓	✓	100	**64.2**

Table 3. Object detection results. All methods are pretrained on ImageNet-1K, then finetuned on COCO using Mask-RCNN (R50-FPN) based on Detectron2 [46]. Sup. and Unsup. are short for supervised learning and unsupervised learning, respectively. Results of methods† are from [48].

Method	Sup.	Unsup.	Epoch	AP	AP50	AP75
SL	✓		100	38.9	59.6	42.7
SL-MLP	✓		100	39.7	60.4	43.1
InsDis† [47]		✓	200	37.4	57.6	40.6
PIRL† [30]		✓	200	37.5	57.6	41.0
SwAV† [3]		✓	200	38.5	60.4	41.4
Mocov2† [8]		✓	200	38.9	59.4	42.4
Byol [17]		✓	300	39.4	60.4	43.2
SL	✓		300	40.1	61.1	43.8
SL-MLP	✓		300	**40.7**	**61.8**	**44.2**

ferability of SL on various backbones, *e.g*. ResNet101 [20], Swin-tiny [28], and EfficientNetb2 [38]. Swin-tiny achieves relatively smaller gain (**+1.7%**) due to its good Feature Mixtureness (0.86), which is close to SL-MLP in Fig. 7(a).

Generalize to other classification tasks. To evaluate if the added MLP can help SL to transfer better on cross-domain tasks, following [22], we pretrain the model on ImageNet-1K, and evaluate the transferability on 12 classification datasets from different domains. As illustrated in Tab. 4, supervised pretraining methods with the MLP projector, *i.e.,* SL-MLP and SupCon [23], outperform their no MLP counterparts, *i.e.,* SL and SupCon w/o MLP on linear evaluation, by 5.79%, 13.71% on the averaged Top-1 accuracy, respectively. Consistent results can be observed on finetuning and few-shot learning settings. More results are provided in Supplementary I. Besides, by comparing SupCon, SL-MLP and SupCon w/o MLP, SL, we conclude that the MLP projector instead of the contrastive loss plays the key role in increasing transferability. Our conclusion contrasts with previous works [22, 53] because they ignore the MLP projector before the contrastive loss.

Generalize to object detection. We assess the transferability improvement by the MLP projector on COCO object detection task. We follow the settings in [18] to finetune the whole network with $1\times$ schedule. In Tab. 3, we report results using Mask-RCNN (R50-FPN), as detailed in Supplementary H. When both are pretrained over 100 epochs, SL-MLP performs better than SL (without MLP) on object detection by **+0.8 AP**. If MLP is used by both supervised and unsupervised pretraining, SL-MLP pretrained by 100 epochs achieves better performance than unsupervised pretraining (*e.g*. SwAV and Mocov2) which are pretrained with 200 epochs. When both pretrained over 300 epochs, SL-MLP shows *better* performance than Byol with **+1.3 AP**.

5.3. Ablation Study

Effectiveness of different components in MLP. In this part, we investigate the influence of different components in the MLP projector. We set the hidden units and output dimension of MLP to be 2048 to retain the dimension of output features the same as SL. Variants are constructed by adding the components incrementally: (a) no MLP projector; (b) only Input FC; (c) Input FC+BN+output FC; (d) Input FC+ReLU+output FC; (e) BN+ReLU. All experiments are pretrained on pre-D over 100 epochs. As shown in Tab. 5, SL-MLP achieves the best accuracy among all variants. We analyze the influence of different components on discriminative ratio ϕ on pre-D, Feature Mixtureness Π, feature redundancy $\mathcal{R}$ qualitatively and quantitatively in Supplementary D.3. Besides, we also observe an interesting phenomenon on Tab. 5(e) that only inserting a lightweight BN-ReLU also achieves good transfer performance. As this is not our main focus, we will investigate it in future works.

Epochs and layers. Fig. 10(a) shows that adding one MLP projector achieves the optimal transferability. In addition, larger pretraining epochs benefit the transferability of SL-MLP when one MLP projector is added, but it has little influence when more MLP projectors are used.

SL-MLP is less sensitive to batch size. Most unsupervised methods depend on big mini-batches to train a representation with strong transferability. To investigate the sensitivity of SL-MLP to batch size, we do experiments with batch size from 256 to 4096 for Byol and to 1024 for SL-MLP over 300 epochs. As shown in Fig. 10(b), the transferability of Byol drops when the batch size decreases. In contrast, the transferability of SL-MLP retains when batch size changes.

SL-MLP is less sensitive to augmentation. Unsupervised methods benefit from a broader space of colors and more

Table 4. Linear evaluation on fixed backbone, full network finetuning, and few-shot learning performance on 12 classification datasets in terms of top-1 accuracy. All models are pretrained for 300 epochs with the same code base except for SelfSupCon† (Mocov2) which pretrained for 400 epochs using the results illustrated in [22]. Average results style: **best**, <u>second best</u>.

Method	ChestX	CropDisease	DeepWeeds	DTD	EuroSAT	Flowers102	Kaokore	Omniglot	Resisc45	Sketch	SVHN	ISIC	Average
linear evaluation													
SL	45.45	96.80	84.02	66.22	95.07	83.69	75.40	64.14	85.36	67.82	67.13	79.58	75.89
SL-MLP	49.89	99.02	87.86	72.61	96.63	93.46	81.12	76.73	91.66	74.51	75.16	81.53	**81.68**
SupCon w/o MLP	41.38	91.52	73.16	62.93	89.84	73.23	66.38	44.54	76.55	55.21	61.45	68.54	67.06
SupCon	47.71	98.79	85.66	74.20	95.83	92.24	79.42	73.42	91.14	76.80	74.26	79.78	<u>80.77</u>
SelfSupCon†	48.08	99.06	87.88	72.71	96.97	89.62	81.67	69.66	90.88	69.12	69.95	81.51	79.70
finetuned with 1000 training samples													
SL	40.86	94.31	86.95	62.12	94.05	88.94	78.22	46.16	80.32	14.17	82.16	78.28	70.54
SL-MLP	42.34	94.48	89.64	63.90	95.30	90.20	77.98	46.66	83.13	17.32	80.19	78.82	**71.66**
SupCon w/o MLP	41.72	93.52	84.95	58.09	95.15	88.23	78.95	45.68	80.63	14.39	82.25	77.96	70.12
SupCon	41.84	93.46	88.70	61.81	94.54	91.28	78.35	46.02	81.62	15.84	81.85	78.51	<u>71.15</u>
SelfSupCon†	43.09	93.95	88.10	62.95	95.47	88.92	79.41	45.33	81.14	10.57	82.37	78.27	70.88
5-ways 5-shots few-shot classification													
SL	25.64	89.07	54.32	78.58	82.96	93.14	46.14	92.82	84.17	87.06	38.03	41.22	67.76
SL-MLP	26.89	93.45	59.08	83.04	87.16	96.88	50.77	95.73	89.00	89.84	41.96	46.76	**71.71**
SupCon w/o MLP	23.62	75.64	49.34	73.04	73.90	82.16	38.10	67.87	75.18	81.01	34.92	35.16	59.16
SupCon	26.18	94.09	59.36	85.02	87.97	96.55	51.02	94.49	89.01	89.75	41.67	43.48	<u>71.55</u>

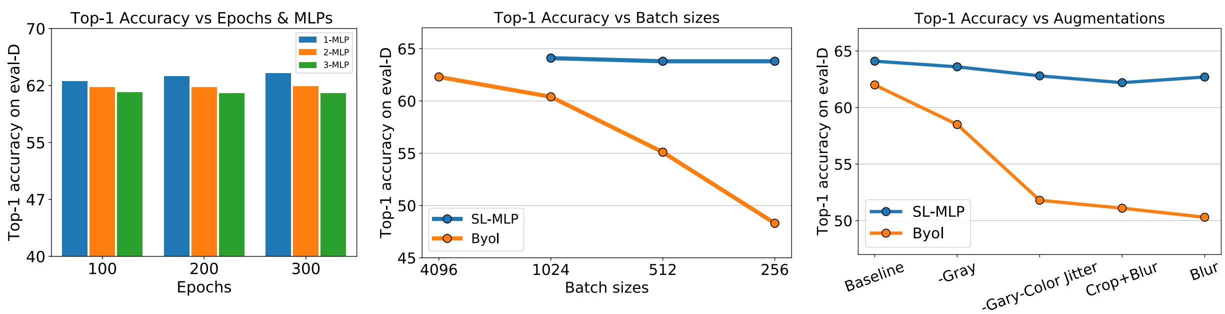


Figure 10. (Left to right) **(a)** Top-1 accuracy with different pretraining epochs and number of MLP projectors. **(b)** Top-1 accuracy with different batch sizes. **(c)** Top-1 accuracy with different pretraining augmentations.

Table 5. Empirical analysis of architectural design of the MLP projector. We pretrain models over 100 epochs and set the output dimension to 2048. Top-1 accuracy on eval-D is reported.

Exp	Components				+Params	Top-1
	Input FC	BN	ReLU	Output FC		
(a)					/	55.9
(b)	✓				4.196M	56.6
(c)	✓	✓		✓	8.395M	61.0
(d)	✓		✓	✓	8.391M	60.1
(e)		✓	✓		0.004M	60.5
SL-MLP	✓	✓	✓	✓	8.395M	62.5

intensive augmentations during pretraining, which always lead to undesirable degradation when some augmentations are missing. Supervised models trained only with horizontal flipping may perform well [53]. We set Byol's augmentations as a baseline setting for both SL-MLP and Byol. We then compare the robustness on augmentation between SL-MLP and Byol by removing augmentation step by step. Experiments of SL-MLP and Byol are all constructed on their default condition with only augmentations changed. The results are illustrated on Fig. 10(c). We find that SL-MLP inherits the robustness of SL and shows a little accuracy drop with simple augmentations.

6. Limitations and Conclusions

In this paper, we study the transferability gap between supervised and unsupervised pretraining. Based on empirical results, we identify that the MLP projector is a key factor for the good transferability of unsupervised pretraining methods. By adding an MLP projector into supervised pretraining methods, we close the gap between supervised and unsupervised pretraining and even make supervised pretraining better. Our finding is confirmed with extensive experiments on diverse backbone networks and various downstream tasks, including the concept generalization tasks, cross-domain image classifications, and objection detection. While the MLP is a simple design for better transferability, there might exist some straightforward designs on the objective function, which we leave for future work.

7. Acknowledgment

This work is supported by the Key R&D Project of Zhejiang Province (No.2022C01056), Natural Science Foundation of Zhejiang Province (No.LQ21F030017), the National Natural Science Foundation of China (No.62127803) and Hetao Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone (HZQB-KCZYZ-2021045).

References

[1] Suresh Balakrishnama and Aravind Ganapathiraju. Linear discriminant analysis-a brief tutorial. *Institute for Signal and information Processing*, 18(1998):1–8, 1998. 4, 6

[2] John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman. Learning bounds for domain adaptation. *Advances in Neural Information Processing Systems*, 20:129–136, 2007. 5

[3] 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, 2, 7

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021. 1

[5] Kai Chen, Lanqing Hong, Hang Xu, Zhenguo Li, and Dit-Yan Yeung. Multisiam: Self-supervised multi-instance siamese representation learning for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7546–7554, 2021. 1

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1, 2

[7] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020. 1

[8] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 1, 2, 3, 7

[9] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 1, 2

[10] 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. 6

[11] 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. 6

[12] 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. 6

[13] 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. 2

[14] 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. *arXiv preprint arXiv:2010.11929*, 2020. 1

[15] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. *arXiv preprint arXiv:2104.14548*, 2021. 1, 2

[16] Linus Ericsson, Henry Gouk, and Timothy M Hospedales. How well do self-supervised models transfer? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5414–5423, 2021. 1, 2

[17] 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. 1, 2, 3, 5, 6, 7

[18] 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, 7

[19] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4918–4927, 2019. 5

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 6, 7

[21] 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. 6

[22] Ashraful Islam, Chun-Fu Chen, Rameswar Panda, Leonid Karlinsky, Richard Radke, and Rogerio Feris. A broad study on the transferability of visual representations with contrastive learning. *arXiv preprint arXiv:2103.13517*, 2021. 1, 2, 3, 4, 6, 7, 8

[23] 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, 2, 4, 7

[24] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015. 6

[25] 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, 6

[26] Bin Liu, Yue Cao, Yutong Lin, Qi Li, Zheng Zhang, Mingsheng Long, and Han Hu. Negative margin matters: Understanding margin in few-shot classification. In *European*

Conference on Computer Vision, pages 438–455. Springer, 2020. 1, 6

[27] Hong Liu, Mingsheng Long, Jianmin Wang, and Michael I Jordan. Towards understanding the transferability of deep representations. *arXiv preprint arXiv:1909.12031*, 2019. 5

[28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 7

[29] George A Miller. *WordNet: An electronic lexical database*. MIT press, 1998. 2

[30] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020. 7

[31] 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. 6

[32] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, pages 722–729, 2011. 6

[33] 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. 6

[34] Alex Olsen, Dmitry A Konovalov, Bronson Philippa, Peter Ridd, Jake C Wood, Jamie Johns, Wesley Banks, Benjamin Girgenti, Owen Kenny, James Whinney, et al. Deepweeds: A multiclass weed species image dataset for deep learning. *Scientific reports*, 9(1):1–12, 2019. 6

[35] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. *Advances in Neural Information Processing Systems*, 32:3347–3357, 2019. 1

[36] 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. *International journal of computer vision*, 115(3):211–252, 2015. 3, 6

[37] Mert Bulent Sariyildiz, Yannis Kalantidis, Diane Larlus, and Karteek Alahari. Concept generalization in visual representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9629–9639, 2021. 1, 2, 3

[38] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. 7

[39] Shixiang Tang, Dapeng Chen, Lei Bai, Kaijian Liu, Yixiao Ge, and Wanli Ouyang. Mutual crf-gnn for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2329–2339, 2021. 1

[40] Shixiang Tang, Peng Su, Dapeng Chen, and Wanli Ouyang. Gradient regularized contrastive learning for continual domain adaptation. In *Proc. 35th AAAI Conf. Artif. Intell.,(AAAI)*, pages 2–13, 2021. 2

[41] 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. 6

[42] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32:10506–10518, 2019. 6

[43] Hongyu Wang, Henry Gouk, Eibe Frank, Bernhard Pfahringer, and Michael Mayo. A comparison of machine learning methods for cross-domain few-shot learning. In *Australasian Joint Conference on Artificial Intelligence*, pages 445–457. Springer, 2020. 1

[44] 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. 6

[45] Longhui Wei, Lingxi Xie, Jianzhong He, Jianlong Chang, Xiaopeng Zhang, Wengang Zhou, Houqiang Li, and Qi Tian. Can semantic labels assist self-supervised visual representation learning? *arXiv preprint arXiv:2011.08621*, 2020. 1, 2

[46] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019. 7

[47] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 7

[48] Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. *arXiv preprint arXiv:2102.04803*, 2021. 2, 7

[49] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16684–16693, 2021. 1, 2

[50] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems*, 27:3320–3328, 2014. 3

[51] Yang You, Igor Gitman, and Boris Ginsburg. Scaling sgd batch size to 32k for imagenet training. *arXiv preprint arXiv:1708.03888*, 6:12, 2017. 6

[52] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*, 2021. 1, 5

[53] Nanxuan Zhao, Zhirong Wu, Rynson WH Lau, and Stephen Lin. What makes instance discrimination good for transfer

learning? *arXiv preprint arXiv:2006.06606*, 2020. 1, 2, 3, 4, 7, 8