Diverse Imagenet Models Transfer Better

Niv Nayman*
Technion - Israel Institute of Technology
nivn@campus.technion.ac.il

Avram Golbert*
Google Research
agolbert@gmail.com

Asaf Noy
Mobileye
asaf.noy@mobileye.com

Lihi Zelnik-Manor
Technion - Israel Institute of Technology
lihi@technion.ac.il

Abstract

A commonly accepted hypothesis is that models with higher accuracy on Imagenet perform better on other downstream tasks, leading to much research dedicated to optimizing Imagenet accuracy. Recently this hypothesis has been challenged by evidence showing that self-supervised models transfer better than their supervised counterparts, despite their inferior Imagenet accuracy. This calls for identifying the additional factors, on top of Imagenet accuracy, that make models transferable. In this work we show that high diversity of the filters learnt by the model promotes transferability jointly with Imagenet accuracy. Encouraged by the recent transferability results of self-supervised models, we use a simple procedure to combine self-supervised and supervised pretraining and generate models with both high diversity and high accuracy, and as a result high transferability. We experiment with several architectures and multiple downstream tasks, including both single-label and multi-label classification.

1. Introduction

The success of Deep Neural Networks (DNNS) in a variety of computer vision tasks is largely related to their ability to transfer feature representations learned on a pre-trained task to leverage others. A common practice is to pre-train a model on a large-scale supervised dataset such as ImageNet [70] and fine-tune it on the downstream (target) dataset that is typically of a smaller scale. This practice has systematically advanced the state-of-the-art in tasks such as image classification [52, 66], object detection [53, 65] and semantic segmentation [34,53]. The pursuit after better pre-trained models coincided with pushing the state-of-the-art performance on ImageNet, as it was shown that supervised pre-trained models that perform better on ImageNet tend to perform better when fine-tuned on other tasks [44].

Recent works demonstrate that self-supervised pre-training (SSL) without any label information can also learn effective representations from upstream data (e.g., ImageNet) and even surpass supervised methods when transferring to downstream tasks. This success in transfer learning, despite their relatively poor performance on ImageNet [13,15,31,33,83] calls for identifying the additional factors, on top of Imagenet accuracy, that make models transferable.

While supervised training focuses on class-level discrimination, SSL focuses on instance discrimination, and models are trained to keep variants of the same instance close together in the representation space, and sometimes also, separated from different instances. On the other hand, supervised models learn meaningful high-level semantic features that are shared between instances of the same class, while SSL might capture irrelevant low level visual features (e.g., related to instance background). Thus high ImageNet performance guarantees that the features learnt are semantically meaningful and SSL learns diverse features. Those features are extracted for an input image by applying a composition of filters on it, that are determined by the model's structure and learnt weights. This observation is supported by recent work that combines both supervised and self-supervised losses to improve transferability [37, 40], yet those require intervention in the self-supervision stage and the transferability is attributed to other less important factors than filter diversity such as the abstraction of the representations learned (measured by Centered Kernel Alignment (CKA) [43] between layers) and intra-class variations.

Our contribution is two-fold: (1) We suggest *Filter Diversity* as a calibration for the Imagenet accuracy for assessing the transferability of models.

$$CIS = \text{Imagenet Accuracy} \times \text{Filter Diversity} \quad (1)$$

* Equal contribution

As we empirically validate that *Calibrated Imagenet Score (CIS)* better correlates with trasferability, i.e. the averaged log odds [44] over many downstream tasks (see Figure 1).

(2) We use a simple scheme for a *Controlled Label Injection (CLI)*, that enables the injection of label information into any self-supervised pre-trained model in a controlled manner, for generating models of different filter diversity and Imagenet accuracy. The resulted models increase ImageNet performance while either improving or maintaining filter diversity of the self-supervised model. This both allows us to make observations about the connection between the CIS and transferability, while at the same time this leads to models with higher transferability.

We validate our approach over both CNNs (ResNets [35]) and vision transformers (ViT [25]), several self-supervised pre-training methods (e.g., MoCo-v2 [14], SimCLR [13], SwAV [10], DINO [10] and MAE [32]), two formulations of Filter Diversity, several downstream vision tasks, including multi-label classification on the MS-COCO [51] dataset and a variety of 14 single-label classification datasets.

2. Related Work

2.1. Transfer learning

Transfer learning was shown to be highly effective in transferring knowledge from upstream (source) datasets to typically much smaller datasets given that their domains are similar [59]. [36] searched for the properties that make a dataset a good choice for transfer learning. [44] showed that when it comes to supervised models, ImageNet accuracy score is highly correlated with performance over downstream tasks, confirming the common practice of selecting pre-trained models for transfer learning based on their Imagenet accuracy. We show that when self-supervised models are included, the correlation significantly drops, calling for improved measures for selection. The architecture and depth of CNNs were also shown to impact transfer performance [6]. The effects of the pre-training loss function were studied by [37,40,42], and the importance of projector heads design and data augmentation to control the trade-off between performance on the upstream task and transferability is shown by [71,79]. Improved Imagenet score might actually lead to worse transfer learning results when used as fixed feature extractor, while the choice of the loss has little effect when networks are fully fine-tuned on the new tasks as shown by [42]. A combination of contrastive and supervised learning was shown to improve transfer leaning performance [37, 40], but the factors driving the performance are still not completely understood. Centered Kernel Alignment (CKA) [43] was utilized by [42] to show that differences among loss functions are apparent only in the last few layers of the network, and [37] further showed that contrastive models contain more low level and mid-level features in those layers. [28, 37, 42] connect this to intra-class variations, concluding that representations with higher class separation obtain higher accuracy on the upstream task, but their features are less useful for downstream tasks. In this work, we identify filter diversity as a more important factor that implies on the transferability of the model, even in the more practical use-case of fine-tuning the pretrained models on the downstream tasks. [62] quantifies transferability between a source model and a target dataset by class separation measures over the embeddings of the target images, while filter diversity introduced in this paper is an intrinsic property of the model that does not depend on the data.

2.2. Self-Supervised Learning

SSL is a subset of unsupervised learning, where neural networks are explicitly trained with automatically generated labels. In earlier works, labels were generated by diverse pre-text tasks such as prediction of rotation [41], colorization [85], pathces positions [24] and others [38]. More recent methods can be roughly divided to contrastive methods [13, 14, 31, 33] and clustering methods [2, 10, 50]. Notably, MoCo-v2 [14], SimCLR [13], SwAV [10] have shown a dramatic improvement in representation quality learned from unlabeled Imagenet images, surpassing the performance of modern supervised methods over various downstream tasks [27]. They also showed that while self-supervised features seems to discard color information, their attentive focus is higher compared to their supervised counterparts. This motivated the proposals of hybrid methods. [40] proposed a new contrastive loss to leverage the label information and [37] combined it with both constrastive and cross-entropy losses. However, it is yet unclear what self-supervised features should be maintained and how in order to improve resulting models' transferability. In this work, we propose a measure that captures the diversity of the information encoded in different networks, and a simple method to inject supervised label information to a pretrained SSL model, in a way that maintains this diversity in order to improve transfer learning performance.

2.3. Filter Diversity

It has been shown that a significant portion of filters extracted by DNNs are redundant [5, 7, 11, 21, 68]. By simply training on a low-rank decomposition of the weight matrices, [20] demonstrated that a fraction of the parameters is sufficient to reconstruct the entire network. [3] estimated the number of redundant filters in each layer, by hierarchically clustering those according to their relative cosine distances in filter space. [4, 67] proposed regularizing correlated filters based on their relative cosine distances to yield a network with diverse filters, with less overfitting, and better generalization. [1, 58] use determinantal point pro-

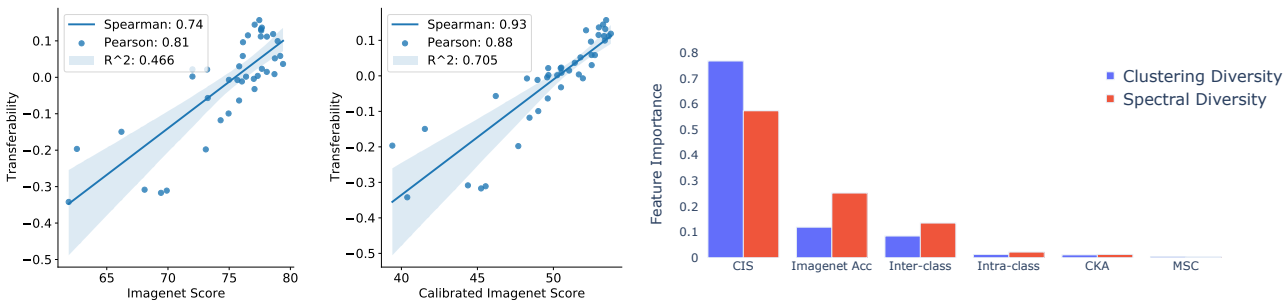

Figure 1. Transferability vs (Left) Imagent Score, and (Middle) the *Calibrated* Imagenet Score (CIS) for 40 models that were pre-trained with supervised learning, self-supervised learning or their combination. CIS correlates with transferability significantly better than imagenet score. (Right) The relative importance of different factors are quantified by the popular feature importance derived from XGBoost. Most of the importance is attributed to the Calibrated Imagenet Score (CIS).

cesses to select a subset of diverse neurons or connections and subsequently fuse the redundant ones into the selected ones for the purpose of pruning. Differently from the aforementioned, which deal mainly with reducing overfitting and pruning, in this work we focus on the importance of learning diverse filters for the purpose of transfer learning.

3. Filter Diversity Measures

Previous work connected the transferability of a model to data-dependent measures, such as the abstraction of representations learnt for the upstream data and the variations in its embedding space [37, 42], the number of non-zero elements in the activations [42] and robustness to corrupted data [37]. Considering that transfer learning deals with different, sometimes unknown in advance, downstream tasks, we instead search for a connection to an intrinsic data-independent property of the model. Intuitively, the more diverse the information learnt by the pre-trained model is, the more likely this information can be utilized in transfer learning to a larger variety of downstream tasks. With this intuition, since the information learnt by the pre-trained model is encoded in its weights, we need a way to quantify the diversity of filters learnt by the pre-trained model.

We adapt two measures, illustrated in Figure 2, both of which view the weights of the various neural layers as vectors in a metric space. The measures quantify the scatter of those vectors in the filter space. We describe in more detail the first measure, which is based on clusterability properties of the filters. The second measure, based on spectral analysis of the filters distribution, is presented in more details in Appendix E for brevity. Empirical evaluation with both measures leads to similar conclusions and validates the importance of filter diversity for transferability (Figure 1).

3.1. Clustering Filter Diversity

The first measure we adapt to evaluate filter diversity is based on assessing the organization of the filters into clusters, and is inspired by [3]. Filters that are grouped together into tight clusters imply low diversity, while filters that are sparsely spread imply high diversity. We next extend this idea to measure the overall clusterability of a deep neural network's filters across all of its layers.

Let $W = [w_i, \ldots, w_n] \in \mathbb{R}^{d \times n}$ be a weight matrix whose columns $\{w_i\}_{i=1}^{n}$ are its filters. We apply the agglomerative clustering approach of [23, 78], while adjusting it to fit our purpose. The clustering continues agglomeratively, merging two clusters C_a and C_b as long as their average mutual cosine similarity $\mathcal{S}_C(C_a, C_b)$ [49, 57] crosses some threshold τ:

$$\mathcal{S}_C(C_a, C_b) = \frac{\sum_{\{w_a, w_b\} \in C_a \otimes C_b} cos(w_a, w_b)}{|C_a| \cdot |C_b|} > \tau \quad (2)$$

The *cluster ratio* between the number of clusters and the number of filters for a given threshold τ quantifies the resulted clusterability, as illustrated in Figure 2 (Left). Due to different neural layers of the same model learning different levels of abstractions, a single threshold τ value does not fit all. Hence, differently from [3], we evaluate the clusterability of the entire model by averaging the cluster ratio of all layers across a spectrum of threshold values. For the full technical details and illustrations see Appendix D.

3.2. Spectral Filter Diversity

Another way to evaluate the distribution of filters is suggested next, based on spectral analysis of the filter vectors. Principal component analysis (PCA) [29] is an effective approach for evaluating variance along principal directions in the filter space. Low diversity implies that the filter distribution can be captured by a small number of principal directions, while high diversity implies requiring many principal directions, as illustrated in Figure 2 (Right) while the technical details and exact calculations are provided in Appendix E for brevity. Table 1 shows that both measures of Filter Diversity improve the correlation of the Calibrated Imagenet Score with the transferability in both cases of linear probing and finetuning, while for the former the spectral version is favourable, and for the latter the clustering based

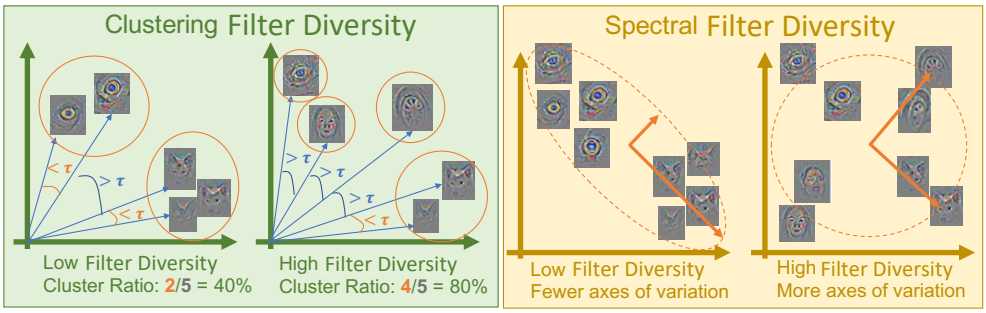

Figure 2. Illustration of the two adapted measures for Filter Diversity. (Left) Low and high diversity result in low and high clustering ratio respectively. (Right) For low and high diversity the variance of filters is explained by fewer or more directions respectively.

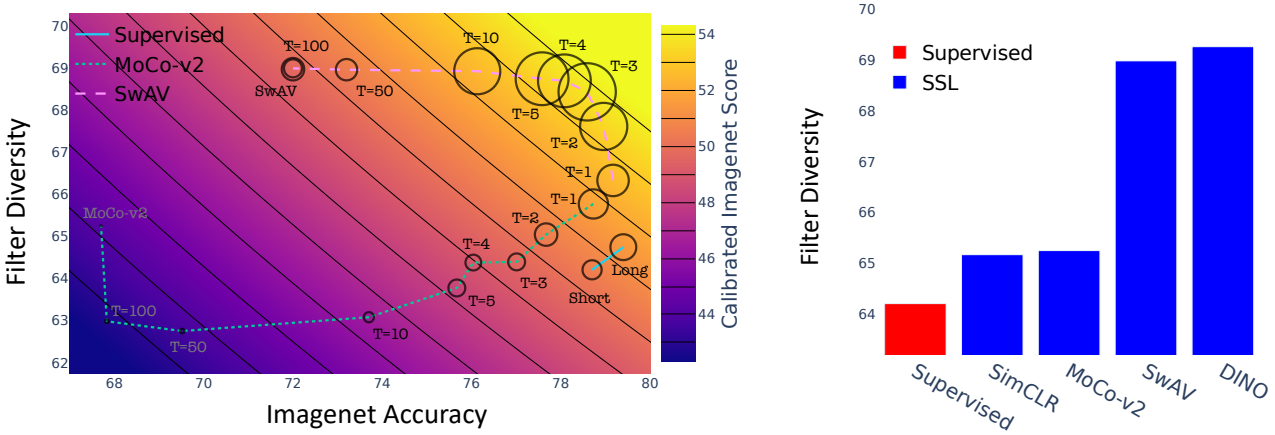

Figure 3. (Left) Circle size corresponds to the transferability averaged over 14 downstream tasks, as a function of Imagenet top-1 accuracy (x axis) and filter diversity (y axis). Results shown for 3 different training methods (see legend). The background colors and curves show the Calibrated Imagenet Score. Evidently, models with both high Imagenet accuracy and high Filter Diversity, that together result in high Calibrated Imagenet Score (in yellow), transfer better (larger circles). (Right) SSL methods learn more diverse filters.

one is. For the corresponding Figure 1 in the case of linear probing see Appendix J.2.

4. Scheme for High CIS

We present a simple scheme that produces models with both high filter diversity as well as high ImageNet accuracy. The scheme is generic in the sense that it can be applied to any type of model and training method. The proposed scheme, illustrated in Figure 4, is composed of two stages. First, we train a model using Self-Supervised Learning (SSL). Then, we gradually introduce supervision by injecting label information in a controlled manner, while fine-tuning the model.

4.1. Controlled Label Injection

The scheme has two stages. It starts with Self-supervised learning (SSL) that typically yields pre-trained models with higher diversity, because the underlying contrastive learning views each sample as a unique class. We observe that filters learnt by SSL methods are richer and more diverse compared to supervised models that effectively capture much fewer classes, as shown in Figure 3 (Right). The second stage injects label information into any self-supervised pre-trained model in a controlled manner, and through that increases ImageNet accuracy while at the same time either improving or maintaining filter diversity of the self-supervised model. We show empirically that this scheme leads to models with higher transferability.

Denote by f_{W_B} and $g_{W_{FF}}$ the backbone model and the classification model on top of it, with weights W_B and W_{FF} respectively, such that $\hat{y} = g_{W_{FF}}(f_{W_B}(x))$ holds for an input sample x and its predicted label $\hat{y}$. The backbone weights W_B were trained by any self-supervised method and W_{FF} is randomly initialized.

We fine-tune the weights by training with controlled supervision, according to Algorithm 1. Considering that the classifier for pre-training is to be replaced eventually, the backbone weights W_B are updated once every $\mathcal{T}$ updates of the classifier W_{FF}, thus the classifier is encouraged to undertake most of the classification burden, and only some of it is passed through to the backbone, whose weights change more slowly. While other valid implementation alternatives for CLI might result this desired effect, e.g., assigning a higher learning rate to the classifier, the chosen implemen-

	Linear Probing				Finetune			
	ρ	r	R^2	τ	ρ	r	R^2	τ
ImageNet Score	0.55	0.73	0.13	0.38	0.74	0.81	0.47	0.53
Spectral CIS	**0.91**	**0.89**	**0.74**	**0.75**	0.89	0.83	0.56	0.72
Cluster CIS	0.89	0.88	0.71	0.72	**0.93**	**0.88**	**0.70**	**0.77**

Table 1. The Spearman (ρ), Pearson (r), R^2 and Kendall-tau (τ) correlation coefficients between transferability and standard or Calibrated Imagenet Score (CIS) computed with the proposed filter diversity measures. Both diversity based CIS measures show significantly higher correlation than the plain Imagenet accuracy, with an advantage to Spectral Filter Diversity for linear probing and to Clustering Filter Diversity for finetune.

tation is analysed empirically (Section 4.2) and shows to be effective (Section 5.2). Effectively, $\mathcal{T} = 1$ is a standard fine-tuning and $\mathcal{T} \to \infty$ is linear probing. Hence, the diversity is maintained for a large enough control cycle $\mathcal{T}$, when starting from models of high filter diversity.

Algorithm 1 Controlled Label Injection (CLI)

input Self-supervided pretrained weights W_B,
Upstream train set $\mathcal{D}_{train}$, Control cycle $\mathcal{T}$
Fine-tuning steps T, Learning rate η

$W_{FF} \leftarrow$RandomInit()
for $t = 1, \ldots, T$ **do**
  Sample an i.i.d train batch $(x_t, y_t) \sim \mathcal{D}_{train}$
  $W_{FF} \leftarrow W_{FF} - \eta \nabla_{W_{FF}} \mathcal{L}_{CE}\left(g_{W_{FF}}(f_{W_B}(x_t)), y_t\right)$
  if mod(t, $\mathcal{T}$) == 0 **then**
    $W_B \leftarrow W_B - \eta \nabla_{W_B} \mathcal{L}_{CE}\left(g_{W_{FF}}(f_{W_B}(x_t)), y_t\right)$
  end if
end for

output W_B

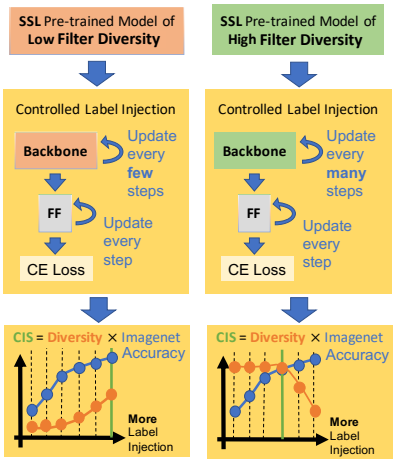

Figure 4. The CLI Scheme: control the injected label information to the backbone by updating it more/less frequently for less/more diverse pretrained SSL models respectively.

4.2. Empirical Analysis of the CLI Procedure

Figure 5 shows the impact of the control cycle. Starting from a SSL pretrained model of high filter diversity (SwAV), we apply Algorithm 1 with different control cycle values. At the left side we show the similarity between the learnt representations of the final model and: (i) the original SSL model, and (ii) a fully supervised model. The similarity is measured by the average Centered Kernel Alignment (CKA) [43] between all pairs of stages of two Resnet50 models. When the label injection is high (low control cycle $\mathcal{T}$) the similarity to the initial SSL model is low and the similarity to a fully supervised model is high. Our experiments show that the best transferability is obtained when those similarities are similar. At the right side of the figure, we empirically validate that the proposed label injection scheme improves Imagenet accuracy while maintaining most of the filter diversity of the input SSL model (SwAV). This is true up to a certain tipping point ($\mathcal{T} = 3$ in this case), where the tranferability is the highest and right after the aggressive label injection ruins the initial filter diversity and thus the Calibrated Imagenet Score drops together with the transferability.

Figure 3 (Left) shows how the control label injection can start off from different SSL pre-trained models of both low (MoCo-v2) and high (SwAV) filter diversity and generate models of different levels of Imagenet accuracy and filter diversity for different control cycle values. Those generated models allow us to make observations about the connection between Imagenet Score and Filter Diversity to the transferability through the Calibrated Imagenet Score, as shown in Figure 1. The trajectory for every origin SSL model traverses the Calibrated Imagenet Score contour lines towards more transferable regions, as expressed by the circle's size.

This is further shown for more SSL methods applied on CNN in Figure 6, where the connection between the control cycle, Filter Diversity, CIS and transferability is shown. Noteably, SSL models of low filter diversity benefit from the maximal label injection, while the filter diversity of highly diverse models is to be maintained by strenghtning label injection for increasing the Imagenet accuracy right up to the point where the diversity drops. Those control cycle values are aligned with the points of highest CIS and ultimately highest transferability. Besides, SSL method undergoing CLI, two supervised models are presented, the first is trained shortly for 200 epochs and the second is trained longer for 600 epochs. Longer supervised training increases both Imagenet accuracy and Filter Diversity.

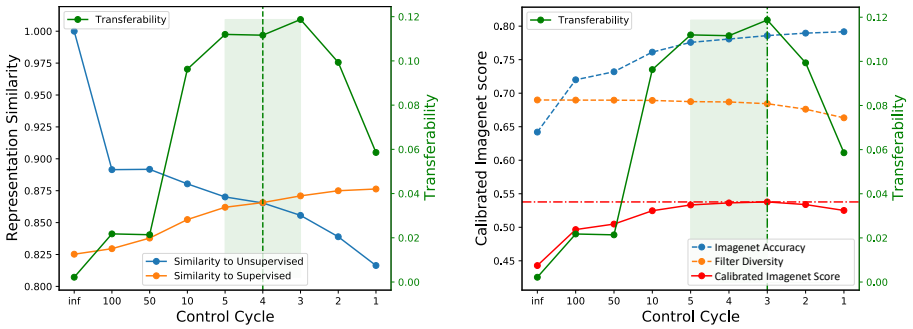

Figure 5. The impact of controlling label injection. (Left) Representation similarity measured by CKA between models with label injected of different control cycle values and fully supervised and SSL (SwAV) models. The best transferability is obtained when the similarity to supervised and unsupervised models is balanced in shaded area. (Right) Filter Diversity, Imagenet accuracy and the resulted Calibrated Imagenet Score are plotted for the same models. Label injection improves Imagenet accuracy, but when too aggressive it can ruin Filter Diversity. The point of the best transferability is obtained when the Calibrated Imagenet Score is the highest in the same shaded area.

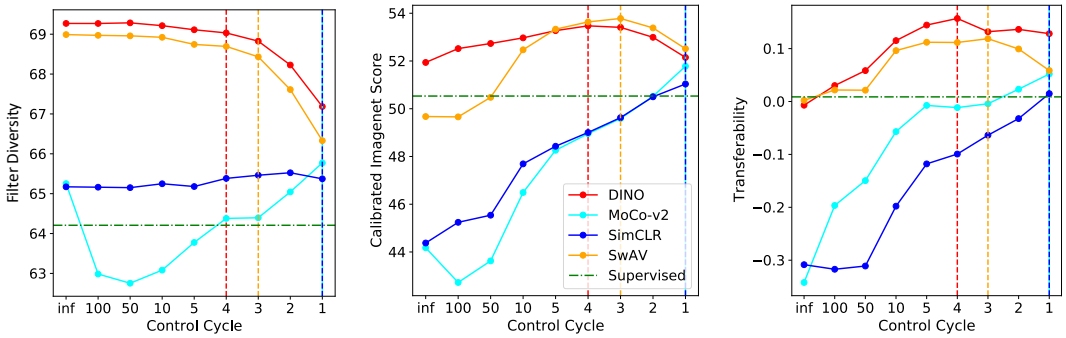

Figure 6. Demonstrating the effect of the Controlled Label Injection (CLI) on the Filter Diversity (Left) Calibrated Imagenet Score (Middle) and Transferability (Right) on different SSL pre-trained models. SSL models of low filter diversity benefit from the maximal label injection, while the filter diversity of highly diverse models is to be maintained. The points of highest transferability are aligned with those of the highest CIS (vertical lines).

5. Experimental Settings

5.1. Downstream Datasets

We evaluated models for multi-label image classification on the popular MS-COCO [51] dataset and another 14 single-label image classification datasets ranging in training set size from 1,020 to 80,800 images (20 to 5,000 images per class; Table 2). These datasets covered a wide range of image classification tasks, including superordinate-level object classification (CIFAR-10 [46], CIFAR-100 [46], Caltech-256 [30]); fine-grained object classification of different kinds (Food-101 [8], NABirds [75], Stanford Cars [45], FGVC Aircraft [56], OxfordIIIT Pets [63], Oxford Flowers-102 [61], Stanford Dogs [39], CUB-200 [77]); texture classification (DTD [16]); and scene classification (MIT indoor 67 [64], SUN397 [81]).

5.2. Comparison to Other Supervised and SSL Combining Methods

5.2.1 Improved Transferability for CNN

Tables 3 and 4 compare the transfer learning performance of Resnet50 pretrained models across 15 downstream tasks, specified in Table 2, and averaged following [44] (see Appendix A) for linear probing and finetuning respectively. The compared models include pure supervised and unsupervised (MoCO-v2, SwAV, SimCLR, DINO) learning, supervised constrastive learning (SupCon [40]), a pretraining combining supervised and self-supervised losses (CE+SelfSupCon [37]) and label injected models following Algorithm 1. The behaviour for linear probing and finetuning is similar. Specifically, certain label injected models transfer best. As SwAV and DINO benefit from high filter diversity (see Figure 3), once label injected to the point of high Calibrated Imagenet score (see Figure 6) it transfers better than all the rest both in terms of overall transferability score and for the most downstream datasets individually. Since MoCo-v2 and SimCLR have relatively low filter diversity (see Figure 3), those benefit from more label injection that increases both their Imagenet accuracy and filter diversity together, as shown in Figure 6. Indeed, those attain the best transferability at $\mathcal{T} = 1$. Those observations call for further future research about the underlying mechanisms that make different SSL methods resulting in different levels of filter diversity, as discussed in section 7.

Category	Name	Symbol	Classes	Train Size	Test Size
Upstream	Imagenet [47]	ImNet	1000	1,281,167	100,000
Superordinate-level object classification	CIFAR-10 [46]	CIFAR10	10	50,000	10,000
	CIFAR-100 [46]	CIFAR100	100	50,000	10,000
	Caltech-256 [30]	Caltech	256	24,581	6,026
Fine-grained object classification	Food-101 [8]	Food	101	80,800	20,200
	NABirds [75]	Birds	555	24,615	23,912
	Stanford Cars [45]	Cars	196	8,041	8,144
	FGVC Aircraft [56]	Aircraft	100	3,334	3,333
	OxfordIIIT Pets [63]	Pets	37	3,680	3,669
	Oxford Flowers-102 [61]	Flowers	102	1,020	6,149
	Standord Dogs [39]	Dogs	120	12,000	8,580
	CUB-200 [77]	CUB	200	5,994	5,794
Texture	DTD [16]	DTD	47	1,880	1,880
Scene classification	MIT indoor 67 [64]	Indoor	67	5360	1,340
	SUN397 [81]	SUN	397	19,850	19,850
Multi-label	MS-COCO [51]	COCO	80	82,081	40,137

Table 2. Datasets examined in transfer learning

Pretrain	ImNet	Aircraft	Birds	CIFAR10	CIFAR100	CUB	Caltech	Cars	DTD	Dogs	Flowers	Food	Indoor	Pets	SUN	Transfer
Supervised	78.7	46.4	60.9	93.0	77.1	71.5	89.1	67.4	69.5	90.4	86.5	70.0	78.7	93.0	63.1	7.3
SupCon [40]	77.3	50.9	56.6	94.9	79.2	69.0	88.5	72.6	70.2	90.5	89.1	68.6	79.3	92.5	63.6	12.6
CE + SelfSupCon [37]	77.3	40.2	52.8	93.3	76.5	63.0	87.3	58.1	67.6	**94.0**	85.6	67.4	76.6	92.6	61.3	-3.9
MoCo-v2	61.9	43.9	38.9	93.4	76.4	53.8	83.5	59.3	69.7	68.0	85.3	68.5	76.0	84.6	60.6	-31.1
MoCo-v2 ($\mathcal{T}=1$)	78.7	35.8	55.9	92.4	74.6	65.6	87.3	52.7	67.8	91.6	80.6	65.4	76.3	93.1	60.5	-12.4
MoCo-v2 ($\mathcal{T}=4$)	76.0	41.3	57.3	92.2	74.4	66.2	87.0	57.0	66.9	87.5	83.5	67.2	76.9	91.9	60.8	-11.8
SwAV	72.0	52.0	53.3	93.2	77.8	66.7	86.5	71.0	71.4	76.4	90.6	73.2	81.6	88.9	65.1	0.3
SwAV ($\mathcal{T}=1$)	**79.2**	44.0	62.8	93.3	77.4	71.2	89.2	64.2	70.7	91.1	86.6	70.2	80.1	93.3	63.5	8.9
SwAV ($\mathcal{T}=4$)	78.1	53.4	65.7	93.1	78.8	73.2	**89.8**	69.8	72.2	87.0	90.4	73.2	81.6	93.1	65.8	18.1
DINO	75.0	**54.8**	54.8	93.7	78.6	68.9	87.1	74.5	72.7	75.9	**92.5**	74.7	81.7	89.3	**66.1**	8.1
DINO ($\mathcal{T}=1$)	77.6	46.0	63.4	93.5	78.3	73.2	89.2	66.2	71.0	90.9	87.0	71.2	80.9	93.8	64.2	13.1
DINO ($\mathcal{T}=4$)	77.5	53.8	**66.0**	**93.8**	**79.5**	**74.3**	89.7	70.5	**73.0**	86.9	91.8	**74.8**	**82.2**	93.3	66.0	**22.6**
SimCLR	68.1	43.4	35.3	89.1	69.0	50.5	82.4	56.2	65.4	65.4	85.2	62.2	72.4	83.8	58.2	-48.0
SimCLR ($\mathcal{T}=1$)	78.1	47.8	61.1	**93.8**	77.7	71.3	88.7	65.8	70.9	88.9	87.5	70.5	80.2	92.9	64.0	8.9
SimCLR ($\mathcal{T}=4$)	75.0	50.4	55.3	93.7	77.6	67.5	87.5	66.8	69.8	82.4	88.0	69.2	77.0	91.3	63.0	-1.5

Table 3. Linear probing performance of different **CNN** models, including different levels of label injected models) fit on the downstream datasets in terms of top-1 accuracy (%) and the overall transferability score. The models are grouped by the underlying base SSL method. The best performance of each column appears in **bold** and the best in each group is underlined. Label injected models transfer best.

5.2.2 Improved Transferability for ViT

Similar results are shown for vision transformers (ViT) in Table 5 and Appendix L. Specifically, label injected ViT models obtain better transfer learning performance than their pure SSL counterparts. Notably, for all the SSL methods examined for ViT, the maximal label injection strength results in the best transferability. Interstingly, this is also true for DINO applied to ViT, when this is not true when applied to CNNs. This observation invites future research on the reasons why ViT tend to learn less diverse filters than CNNs when pre-trained with the same SSL method, see section 7 for a further discussion.

6. Feature Importance for Transfer Learning

In this section we empirically analyze the importance of different factors to transferability. We consider the *Calibrated Imagenet Score* (CIS) and the previously suggested [37,42] CKA, intra-class variance and class separation. The relative importance of those factors is quantified in Figure 1 (Right) by the ability to predict the transfer learning performance after finetuning, by the popular XGBoost [12], that allows feature importance analysis [86], for more technical details see Appendix H. It is evident that CIS is highly predictive of the transferability, and specifically significantly more important than the other factors inspected. Notably, both measures of Filter Diversity capture the importance of CIS compared to other factors. This shows that the very notion of filter diversity and the way it calibrates Imagenet accuracy are valid. Moreover, calibrating Imagenet accuracy by any other of the inspected factors is not predictive of the transferability (see Appendix H). Similar results are shown in Appendix J.2 for the case of linear probing.

Pretrain	ImNet	Caltech	CIFAR10	CIFAR100	CUB	DTD	Aircraft	Food	Indoor	Birds	Flowers	Pets	Cars	Dogs	SUN	Transfer	COCO
Supervised	78.7	86.9	97.6	86.0	85.9	69.3	82.0	85.3	81.3	74.9	99.1	92.4	94.4	82.9	65.3	0.012	81.9
SupCon [40]	77.3	86.3	97.6	85.7	85.0	69.8	**84.1**	86.1	81.4	73.1	99.0	91.9	94.7	83	65.2	0.007	81.2
CE+SelfSupCon	76.4	86.3	97.6	85.8	85.9	69.5	83.3	86.3	80.9	74.1	98.7	92.4	**94.8**	**85.3**	64.4	0.003	82.0
MoCo-v2	61.9	78.1	96.7	82.0	79.4	66.3	80.1	85.4	75.6	61.1	98.5	86.8	93.5	73.4	56.3	-0.352	78.7
MoCo-v2 ($\mathcal{T}=1$)	78.7	87.3	97.6	86.3	85.9	70	83.0	86.1	82.2	73.9	99.2	**92.8**	94.4	84.4	65.1	0.054	81.9
MoCo-v2 ($\mathcal{T}=4$)	76.0	86.4	97.6	86.1	85.7	69.7	82.8	86.1	82.3	73.3	98.9	92.0	94.3	81.5	64.8	-0.010	81.7
SwAV	64.2	87.0	97.8	86.6	84.3	72.1	82.3	87.4	83.1	75.1	98.9	90.3	93.6	80.6	**67.8**	0.005	82.8
SwAV ($\mathcal{T}=1$)	**79.2**	87.5	97.7	86.5	86.5	71.3	83.4	86.8	82.3	75.8	99.0	92.5	94.6	83.6	66.4	0.062	82.3
SwAV ($\mathcal{T}=4$)	78.1	**88.4**	97.8	87.1	**86.6**	**72.9**	82.8	87.6	**84.4**	76.1	99.3	91.9	94.6	82.1	67.8	0.118	**83.2**
DINO	75.0	87.2	97.8	86.9	83.7	72.1	80.6	87.5	83.2	74.5	98.7	89.6	93.8	80.1	67.6	-0.024	82.7
DINO ($\mathcal{T}=1$)	77.6	87.4	97.7	86.7	86.3	71.1	83	87.1	82.7	76.2	99.4	92.5	94.7	82.9	66.2	0.095	82.3
DINO ($\mathcal{T}=4$)	77.5	88.2	97.8	**87.5**	86.5	72.1	82.8	**87.6**	84.1	**76.5**	**99.4**	92.1	94.7	82.1	67.6	**0.126**	83.1
SimCLR	68.1	85.4	**97.9**	86.3	78.3	65.9	77.2	84.0	75.4	62.6	97.6	86.6	91.6	73.8	62.9	-0.318	80.5
SimCLR ($\mathcal{T}=1$)	78.1	86.3	97.6	86.3	85.6	70	82.8	85.9	80.7	72.8	99.1	91.4	94.5	80.8	65.5	-0.013	81.2
SimCLR ($\mathcal{T}=4$)	75.0	86.6	97.7	86.7	82.7	69.5	81.1	84.9	79.3	68.0	98.7	89.5	93.5	77.9	64.9	-0.120	81.5

Table 4. Performance of different **CNN** models fine-tuned on the downstream datasets in terms of top-1 accuracy (%) (averaged over 3 runs) and the overall transferability score. The models are grouped by the underlying base SSL method. The best performance of each column appears in **bold** and the best in each group is underlined. Label injected models transfer best.

Pretrain	ImNet	Caltech	CIFAR10	CIFAR100	CUB	DTD	Aircraft	Food	Indoor	Birds	Flowers	Pets	Cars	Dogs	SUN	Transfer
Supervised	81.0	90.9	98.6	89.6	82.3	70.8	60.8	87.9	83.2	79.6	91.3	93.8	85.4	**91.5**	68.1	-0.101
MAE	68.0	89.2	98.1	86.9	75.4	68.5	53.8	88.8	82.2	81.2	70.6	91.2	79.6	83.2	67.6	-0.425
MAE ($\mathcal{T}=1$)	**83.4**	93.0	98.8	90.6	84.3	73.7	73.1	**90.6**	85.4	**86.2**	94.4	**94.8**	89.8	89.5	71.1	0.131
MAE ($\mathcal{T}=4$)	81.3	92.2	98.8	90.1	84.5	73.2	72.4	90.2	85.0	85.6	94.6	94.6	89.2	88.2	71.0	0.095
DINO	78.2	87.2	97.8	86.9	83.7	72.1	**80.6**	87.5	83.2	74.5	**98.7**	89.6	**93.8**	80.1	67.6	-0.187
DINO ($\mathcal{T}=1$)	83.2	**93.1**	**99.0**	**91.2**	84.7	74.6	72.1	89.9	**86.3**	84.9	94.7	94.3	89.4	90.5	**71.5**	**0.148**
DINO ($\mathcal{T}=4$)	82.3	92.8	98.8	91.0	**84.7**	**75.2**	71.2	90.0	85.8	85.2	95.4	94.7	89.1	88.3	71.5	0.131

Table 5. Performance of different **ViT** models fine-tuned on the downstream datasets in terms of top-1 accuracy (%) and the overall transferability score. The models are grouped by the underlying base SSL method. The best performance of each column appears in **bold** and the best in each group is underlined. Label injected models transfer best.

7. Discussion and Future Work

While we showed that models with high diversity transfer better, a natural extension would be to better understand how and why some training methods produce higher diversity than others. Indeed [4] uses an explicit diversity regularization. We don't expect a diversity regularization during training to work well since diversity encourages high complexity models. In comparison, standard regularization restrict model complexity as a balance to the models overparamtrization. Indeed, there are many ways the network can increase the diversity metric with no real change to the model behavior. One such trivial way to increase the Spectral Filter Diversity is to scale each filter $\hat{W}_i = \frac{W_i}{\|W_i\|}$ and then insert $\|W_i\|$ into subsequent BatchNorm. Similarly, Cluster Diversity is based on cosine similarity, which is scale agnostic, thus filters that are redundant, or close to zero can be set to orthogonal vectors with epsilon scale. We hope this paper motivates future work in ways to increase real filter diversity, and transferability.

8. Conclusions

In this paper, we analyse the importance of different properties of pre-trained models to their transferability. We identify the notion of filter diversity as one of the key factors for transferability, together with the performance on the upstream task. A simple fine-tuning procedure is used for improving the transferability of given self-supervised pre-trained models, by injecting controlled supervision to those, while maintaining their filter diversity and improving their performance on the upstream task. Our study holds for different popular architectures of CNNs and ViTs and self-supervised methods, two different formulations for capturing filter diversity and many downstream tasks of multi-label and single-label classification over more than 15 different datasets.

References

[1] Rupam Acharyya, Boyu Zhang, Ankani Chattoraj, Shouman Das, and Daniel Stefankovic. Diversity based edge pruning of neural networks using determinantal point processes. In *Neural Compression: From Information Theory to Applications–Workshop@ ICLR 2021*, 2021. 2

[2] Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. *arXiv preprint arXiv:1911.05371*, 2019. 2

[3] Babajide O Ayinde, Tamer Inanc, and Jacek M Zurada. On correlation of features extracted by deep neural networks. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019. 2, 3, 14

[4] Babajide O Ayinde, Tamer Inanc, and Jacek M Zurada. Regularizing deep neural networks by enhancing diversity in feature extraction. *IEEE transactions on neural networks and learning systems*, 30(9):2650–2661, 2019. 2, 8

[5] Babajide O Ayinde and Jacek M Zurada. Nonredundant sparse feature extraction using autoencoders with receptive fields clustering. *Neural Networks*, 93:99–109, 2017. 2

[6] 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

[7] Yoshua Bengio and James Bergstra. Slow, decorrelated features for pretraining complex cell-like networks. *Advances in neural information processing systems*, 22:99–107, 2009. 2

[8] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer, 2014. 6, 7

[9] Léon Bottou. Online algorithms and stochastic approxima-p tions. *Online learning and neural networks*, 1998. 13

[10] 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. 2

[11] Soravit Changpinyo, Mark Sandler, and Andrey Zhmoginov. The power of sparsity in convolutional neural networks. *arXiv preprint arXiv:1702.06257*, 2017. 2

[12] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016. 7, 18

[13] 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

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

[15] 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

[16] 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, 7

[17] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Algorithms for learning kernels based on centered alignment. *The Journal of Machine Learning Research*, 13:795–828, 2012. 16

[18] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 13

[19] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020. 13

[20] Misha Denil, Babak Shakibi, Laurent Dinh, Marc'Aurelio Ranzato, and Nando de Freitas. Predicting parameters in deep learning. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 2148–2156, Red Hook, NY, USA, 2013. Curran Associates Inc. 2

[21] Misha Denil, Babak Shakibi, Laurent Dinh, Marc' Aurelio Ranzato, and Nando de Freitas. Predicting parameters in deep learning. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. 2

[22] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 13

[23] Chris Ding and Xiaofeng He. Cluster merging and splitting in hierarchical clustering algorithms. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, pages 139–146. IEEE, 2002. 3

[24] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015. 2

[25] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 2, 13

[26] Jane Elith, John R Leathwick, and Trevor Hastie. A working guide to boosted regression trees. *Journal of animal ecology*, 77(4):802–813, 2008. 18

[27] 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. 2

[28] Yutong Feng, Jianwen Jiang, Mingqian Tang, Rong Jin, and Yue Gao. Rethinking supervised pre-training for better downstream transferring. In *International Conference on Learning Representations*. 2

[29] Karl Pearson F.R.S. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. 3

[30] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007. 6, 7

[31] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc., 2020. 1, 2

[32] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv:2111.06377*, 2021. 2, 13

[33] 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, 2

[34] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1

[35] 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. 2, 13

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

[37] 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, 6, 7, 13, 16, 18, 19, 22, 23

[38] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4037–4058, 2020. 2

[39] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, volume 2. Citeseer, 2011. 6, 7

[40] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised 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 18661–18673. Curran Associates, Inc., 2020. 1, 2, 6, 7, 8, 13, 19, 22, 23

[41] Nikos Komodakis and Spyros Gidaris. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)*, 2018. 2

[42] Simon Kornblith, Ting Chen, Honglak Lee, and Mohammad Norouzi. Why do better loss functions lead to less transferable features? *Advances in Neural Information Processing Systems*, 34, 2021. 2, 3, 7, 16, 17, 18

[43] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529. PMLR, 2019. 1, 2, 5, 16

[44] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2661–2671, 2019. 1, 2, 6, 13, 20

[45] Jonathan Krause, Jia Deng, Michael Stark, and Li Fei-Fei. Collecting a large-scale dataset of fine-grained cars. 2013. 6, 7

[46] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. 6, 7

[47] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. 7, 13

[48] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989. 13

[49] Bastian Leibe, Ales Leonardis, and Bernt Schiele. Combined object categorization and segmentation with an implicit shape model. In *Workshop on statistical learning in computer vision, ECCV*, volume 2, page 7, 2004. 3

[50] Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020. 2

[51] 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, 7

[52] Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. Query2label: A simple transformer way to multi-label classification. *arXiv preprint arXiv:2107.10834*, 2021. 1

[53] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. *arXiv preprint arXiv:2111.09883*, 2021. 1

[54] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 13

[55] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 13

[56] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013. 6, 7

[57] Swami Manickam, Scott D Roth, and Thomas Bushman. Intelligent and optimal normalized correlation for high-speed pattern matching. *Datacube Technical Paper*, 2000. 3

[58] Zelda Mariet and Suvrit Sra. Diversity networks: Neural network compression using determinantal point processes. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. 2

[59] Thomas Mensink, Jasper Uijlings, Alina Kuznetsova, Michael Gygli, and Vittorio Ferrari. Factors of influence for transfer learning across diverse appearance domains and task types. *arXiv preprint arXiv:2103.13318*, 2021. 2

[60] Thao Nguyen, Maithra Raghu, and Simon Kornblith. Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. In *International Conference on Learning Representations*, 2021. 16

[61] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008. 6, 7

[62] Michal Pándy, Andrea Agostinelli, Jasper Uijlings, Vittorio Ferrari, and Thomas Mensink. Transferability estimation using bhattacharyya class separability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9172–9182, 2022. 2

[63] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 6, 7

[64] Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *2009 IEEE conference on computer vision and pattern recognition*, pages 413–420. IEEE, 2009. 6, 7

[65] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015. 1

[66] Tal Ridnik, Gilad Sharir, Avi Ben-Cohen, Emanuel Ben-Baruch, and Asaf Noy. Ml-decoder: Scalable and versatile classification head. *arXiv preprint arXiv:2111.12933*, 2021. 1

[67] Pau Rodríguez, Jordi Gonzàlez, Guillem Cucurull, Josep M. Gonfaus, and F. Xavier Roca. Regularizing cnns with locally constrained decorrelations. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. 2

[68] Pau Rodríguez, Jordi Gonzalez, Guillem Cucurull, Josep M Gonfaus, and Xavier Roca. Regularizing cnns with locally constrained decorrelations. *arXiv preprint arXiv:1611.01967*, 2016. 2

[69] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987. 18

[70] 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. 1

[71] Mert Bulent Sariyildiz, Yannis Kalantidis, Karteek Alahari, and Diane Larlus. No reason for no supervision: Improved generalization in supervised models. In *International Conference on Learning Representations*, 2023. 2

[72] N Shawe-Taylor, A Kandola, et al. On kernel target alignment. *Advances in neural information processing systems*, 14:367, 2002. 16

[73] G. W. Stewart. On the early history of the singular value decomposition. *SIAM Review*, 35(4):551–566, 1993. 14

[74] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 13

[75] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 595–604, 2015. 6, 7

[76] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 13, 14

[77] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 6, 7

[78] Bruce Walter, Kavita Bala, Milind Kulkarni, and Keshav Pingali. Fast agglomerative clustering for rendering. In *2008 IEEE Symposium on Interactive Ray Tracing*, pages 81–86. IEEE, 2008. 3

[79] Yizhou Wang, Shixiang Tang, Feng Zhu, Lei Bai, Rui Zhao, Donglian Qi, and Wanli Ouyang. Revisiting the transferability of supervised pretraining: an mlp perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9183–9193, 2022. 2

[80] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019. 13

[81] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 6, 7

[82] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *International Conference on Computer Vision (ICCV)*, 2019. 13

[83] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12310–12320. PMLR, 18–24 Jul 2021. 1

[84] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 13

[85] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016. 2

[86] Huiting Zheng, Jiabin Yuan, and Long Chen. Short-term load forecasting using emd-lstm neural networks with a xgboost algorithm for feature importance evaluation. *Energies*, 10(8):1168, 2017. 7