Boost Supervised Pretraining for Visual Transfer Learning: Implications of Self-Supervised Contrastive Representation Learning

Jinghan Sun[1,2,*] , **Dong Wei**[2,*] , **Kai Ma**[2] , **Liansheng Wang**[1,†] , **Yefeng Zheng**[2]
[1] Xiamen University, Xiamen, China
jhsun@stu.xmu.edu.cn, lswang@xmu.edu.cn
[2] Tencent Healthcare (Shenzhen) Co., LTD, Tencent Jarvis Lab, Shenzhen, China
{donwei,kylekma,yefengzheng}@tencent.com

Abstract

Unsupervised pretraining based on contrastive learning has made significant progress recently and showed comparable or even superior transfer learning performance to traditional supervised pretraining on various tasks. In this work, we first empirically investigate when and why unsupervised pretraining surpasses supervised counterparts for image classification tasks with a series of control experiments. Besides the commonly used accuracy, we further analyze the results qualitatively with the class activation maps and assess the learned representations quantitatively with the representation entropy and uniformity. Our core finding is that it is the amount of information effectively perceived by the learning model that is crucial to transfer learning, instead of absolute size of the dataset. Based on this finding, we propose **C**lassification **A**ctivation **M**ap guided con**trast**ive (CAMtrast) learning which better utilizes the label supervision to strengthen supervised pretraining, by making the networks perceive more information from the training images. CAMtrast is evaluated with three fundamental visual learning tasks: image recognition, object detection, and semantic segmentation, on various public datasets. Experimental results show that our CAMtrast effectively improves the performance of supervised pretraining, and that its performance is superior to both unsupervised counterparts and a recent related work which similarly attempted improving supervised pretraining.

1 Introduction

Deep convolutional neural networks (DCNNs) are the current state of the art (SOTA) of many fundamental tasks in computer vision. However, to achieve satisfactory performance, DCNNs often need large amounts of data for effective training, which can be difficult to collect in practice. Transfer learning is an effective solution to the data deficiency problem (Tan et al. 2018), where the model is first pretrained on pretext tasks with sufficient data and then finetuned on the target downstream task with limited data. Pretraining on the ImageNet dataset (Deng et al. 2009) with one million images and corresponding categorical labels has been the de facto standard for visual transfer learning and

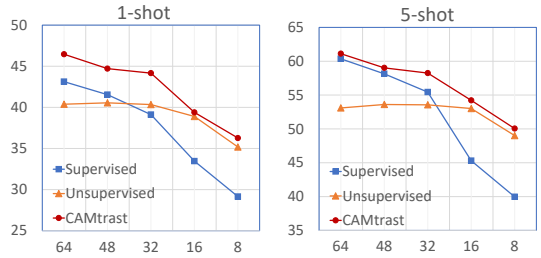

Figure 1: Effects of number of pretraining classes (x-axis) of the supervised, unsupervised (SimSiam; Chen and He 2021), and our proposed CAMtrast pretraining on miniImageNet on downstream 5-way few-shot classification accuracy (%; y-axis), with the total number of images (4,800) unchanged.

achieved SOTA performance on many tasks. Recently, a series of self-supervised representation learning frameworks (He et al. 2020; Chen et al. 2020a; Caron et al. 2020; Grill et al. 2020; Chen and He 2021) emerged that adopted contrastive learning (Hadsell, Chopra, and LeCun 2006). Without using any label, these frameworks enforced certain invariance across different augmented views of an image. Surprisingly, unsupervised contrastive pretraining beat supervised counterparts on various downstream tasks while being comparable on many others. This naturally raises some intriguing questions, including: Why does the unsupervised pretraining transfer better than supervised counterparts?[1] Are manual labels still useful to pretraining?

In their pioneering work, Zhao et al. (2021) empirically found that the intra-category invariance enforced by supervised models weakened transferability by causing information loss and increasing task misalignment, and proposed that supervised pretraining could be strengthened by an exemplar-based approach (the "Exemplar"). Concretely, Exemplar used the labels to filter true negatives in contrastive learning to recover much low- and mid-level information, without explicitly enforcing intra-category invariance. However, the authors did not explicitly answer when supervised pretraining would fall short especially when there was only minimal task misalignment (e.g., transfer within the same

[*] J. Sun and D. Wei—Contributed equally; J. Sun contributed to this work during an internship at Tencent.
[†] Corresponding author.

[1] Without causing confusion, the terms "supervised" and "unsupervised" pretraining in this work means traditional supervised classification and self-supervised contrastive learning, respectively.

dataset), nor explain why it would happen. Meanwhile, their findings were only based on qualitative analysis by visually comparing image reconstructions from the features of supervised and contrastive models, without any quantitative validation of the information loss as evidence. In addition, as a supervised variant of MoCo (He et al. 2020), Exemplar discarded much of the innate ability of traditional supervised pretraining to transfer high-level semantics, which was proved useful in many transfer learning tasks (Zhao et al. 2021; Yu et al. 2018; Zhang et al. 2018).

In this work, we first empirically investigate when supervised pretraining will fall short of unsupervised counterparts on image classification tasks, and analyze why the former collapses in some circumstances. To this end, we conduct a series of control experiments to compare the performance of supervised and unsupervised pretraining. Notably, the 5-way few-shot experiments on miniImageNet show that with a fixed number of training images, the superiority of supervised pretraining to unsupervised counterpart gradually diminishes as the number of classes decreases, and eventually turns into inferiority as the supervised models collapse with fewer than 32 training classes (Fig. 1). On the other hand, this superiority is unaffected by the decreasing number of images given a fixed number of 64 classes. To uncover the underlying reasons, we employ the class activation maps (CAMs; Zhou et al. 2016), and the representation entropy (Devijver and Kittler 2012) and uniformity (Wang and Isola 2020) for qualitative and quantitative analysis, respectively. Our findings are threefold: 1) supervised learning focuses on small class-discriminative regions—the fewer classes the more so, resulting in sparse feature representations and great information loss, while unsupervised pretraining attends to broad areas, 2) representations learned by unsupervised pretraining are more informative and uniform, and 3) both informativeness and uniformity of the supervised pretraining decrease with the decreasing number of classes while those of the unsupervised are relatively consistent.

All in all, our core finding is that it is the amount of information effectively perceived by the learning model that is crucial to the transfer performance, instead of absolute size of the dataset. An important implication is that we can strengthen supervised pretraining by encouraging the model to perceive more information from the images. To this end, we propose **C**lassification **A**ctivation **M**ap guided con**tras**tive (CAMtrast) learning, a novel supervised pretraining framework integrating CAM-guided activation suppression and self-supervised contrastive learning for more effective information perception. Concretely, we use supervised CAMs to locate and suppress the most discriminative image regions, forcing the network to identify secondary discriminative regions in the suppressed images for correct classification. In addition, the pair of original and suppressed images are input to Siamese networks (Chopra, Hadsell, and LeCun 2005) for contrastive learning, for effective transfer of low- and mid-level semantics. Strengthened by the CAM-guided suppression, CAMtrast is able to retain even more high-level semantics than traditional supervised pretraining, and use it to guide where to contrast. We evaluate CAMtrast on three different types of downstream tasks involved with semantic information (image classification, object detection, and semantic segmentation) on several public datasets (miniImageNet, tieredImageNet, CIFAR-FS, PASCAL VOC, and Cityscapes) to demonstrate its efficacy in boosting supervised pretraining and preventing collapse, and superiority to supervised, unsupervised, and Exemplar pretraining. In conclusion, this work provides important new knowledge to the community, and demonstrates an effective tool motivated by the new knowledge.

2 Related Work

Supervised Pretraining: The well established transfer learning paradigm of pretraining on the ImageNet (Deng et al. 2009) classification task followed by fine-tuning on target tasks has achieved remarkable success on a wide variety of downstream tasks such as object detection (Sermanet et al. 2013; Girshick et al. 2014), semantic segmentation (Long, Shelhamer, and Darrell 2015) and others. Accordingly, ImageNet-pretrained models are available for many popular DCNN structures and routinely used nowadays. However, as the distance between the pretraining and target tasks plays a critical role in transfer learning (Zhang, Wang, and Zheng 2017), these models may become less effective when transferred to vastly different tasks (e.g., medical image analysis) due to large domain gaps. In addition, it may be difficult or costly to obtain sufficient annotations for effective pretraining on a new dataset. Therefore, self-supervised pretraining has attracted great attention recently.

Self-Supervised Pretraining: Self-supervised learning can pretrain networks on unlabeled data with pretext tasks defined by certain properties of the data (Jing and Tian 2020). Despite the progress made, performance of earlier self-supervising methods (Doersch, Gupta, and Efros 2015; Noroozi and Favaro 2016; Zhang, Isola, and Efros 2016) was still not comparable with that of supervised counterparts. Recently, we have witnessed a surge of self-supervised visual representation learning methods (Chen et al. 2020a,b; Caron et al. 2020; Grill et al. 2020; He et al. 2020; Chen and He 2021) that instantiated the notion of contrastive learning (Hadsell, Chopra, and LeCun 2006). Despite different original motivations, these methods generally defined two augmentations of one image as the input to Siamese-like networks (Chopra, Hadsell, and LeCun 2005), and maximize the similarity between the output. The contrastive learning paradigm significantly narrowed the gap between unsupervised and supervised pretraining: on many downstream tasks, unsupervised pretraining achieved comparable or even superior performance to supervised counterparts. The strikingly excellent performance of unsupervised contrastive pretraining has stirred the interest of the research community to explore the underlying reasons for insights.

Implication from Contrastive Learning: Wang and Isola (2020) verified that features of contrastive learning were more uniformly distributed on the unit hypersphere than those of supervised learning. Their work mainly conducted research from the perspective of contrastive learning. In contrast, our focus is to understand the cause for the superior performance of unsupervised to supervised pretraining, and improve supervised pretraining based on the understanding.

Based on a visual comparison of the image reconstructions from supervised and contrastive models, Zhao et al. (2021) concluded that the intra-category invariance enforced by supervised models weakened transferability by causing information loss. Accordingly, they proposed Exemplar, an approach that made use of labels in the contrastive learning framework MoCo (He et al. 2020) to filter the true negatives for loss computation, such that no intra-category invariance was explicitly enforced. Exemplar demonstrated encouraging improvements upon supervised pretraining on various downstream tasks. However, as a supervised variant of MoCo in essence, it primarily transferred low- and mid-level representations, thus the high-level semantics were lost. In contrast, our CAMtrast strives to retain even more high-level semantics than supervised pretraining. In addition, it is not obvious to apply Exemplar to contrastive learning frameworks that are not based on instance discrimination, e.g., SimSiam (Chen et al. 2020b), whereas it is straightforward to do so with CAMtrast.

3 Exploratory Experiments

To understand when unsupervised contrastive pretraining outperforms supervised counterparts and the underlying reasons, we conduct a series of control experiments. We also employ two quantitative metrics to analyze the learned feature representations from the perspectives of informativeness and distribution uniformity, in addition to the analysis of transfer performance.

Basic Experimental Protocol: We conduct few-shot recognition on the miniImageNet (Vinyals et al. 2016) and tieredImageNet (Ren et al. 2018) datasets. MiniImageNet contains 100 classes sampled from ImageNet, which are split into 64 base, 16 validation and 20 novel classes. TieredImageNet is another subset of ImageNet. It contains 34 categories, each including 10–30 fine-grained classes. The 34 categories are divided into 20 base categories (351 classes), 6 validation categories (97 classes), and 8 novel categories (160 classes). The hierarchy of categories and classes allows convenient control of the degree of inter-class variation/similarity (and thus classification difficulty), as we do later.

For implementation, we follow Chen et al. (2018b) for supervised pretraining with a standard cross-entropy loss, and MoCo_v2 (Chen et al. 2020b) and SimSiam (Chen and He 2021) for unsupervised pretraining, respectively, using the base classes. The ResNet-50 (He et al. 2016) is used as backbone. After pretraining, a linear evaluation protocol is adopted following the vast literature (Tian et al. 2020; Chen et al. 2020b; Chen and He 2021). Specifically, the networks are frozen as a feature extractor; a logistic regression classifier is then fit on the features of the few support samples and tested on the query samples. The few-shot evaluation tasks are randomly sampled from the novel classes in 5-way 1- and 5-shot settings. A total of 600 tasks are sampled and the mean classification accuracy is reported. Note that the novel classes are never seen during pretraining, hence the experimental setting represents a typical transfer learning scenario.

Effects of Training Data Size: It is generally believed that the success of the ImageNet pretraining is attributed to the size of the dataset or the great number of classes (Mahajan et al. 2018). Here, we begin by investigating the effects of the size of the pretraining dataset on miniImageNet. We fix the number of training classes ($C = 64$) and reduce the number of images per class from 600 (all available) to 75, resulting in 38,400 to 4,800 total images. The supervised pretraining is compared to two representative self-supervised contrastive learning methods: MoCo_v2 (Chen et al. 2020b) and SimSiam (Chen and He 2021). The results are shown in Table 1. We observe that despite the expected performance drops with fewer images for training, the supervised pretraining consistently outperforms the two unsupervised counterparts in all cases. This suggests that the absolute data size does not cause inferiority of supervised pretraining to unsupervised counterparts.

No. images		Supervised		MoCo_v2		SimSiam	
Per class	Total	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
600	38,400	**55.31**	**73.73**	52.04	73.56	48.81	69.27
300	19,200	**51.02**	**69.61**	47.99	67.23	48.51	67.29
150	9,600	**48.92**	**66.58**	43.39	60.38	43.02	59.65
75	4,800	**43.12**	**60.36**	36.60	52.41	40.38	53.10

Table 1: 5-way few-shot recognition accuracies (%) on miniImageNet, as the number of images decreases and the number of classes remains unchanged ($C = 64$) for pretraining.

Effects of Training Class Number: We then study the effects of training class number by doing the opposite: we fix the total number of pretraining images while decreasing the number of classes on miniImageNet. As we are constrained by the number of images per class (600) of miniImageNet, we use a relatively small number of total images (4,800). Accordingly, SimSiam is used for the following experiments, considering its superior performance to MoCo_v2 with this number of images (last row in Table 1). The results are shown in Fig. 1. As we can see, the transfer performance of the unsupervised pretraining is relatively insensitive to the exact number of classes (fluctuates within the range of 6%, whereas that of the supervised pretraining severely degrades about 14–20% as the number of classes decreases from 64 to 8. When the number of classes is lower than 32, the unsupervised pretraining overtakes the supervised counterpart. Zhao et al. (2021) suggested that supervised representations mainly modeled the discriminative object parts of each class, which were central to classification tasks, but at the cost of information loss in other regions. We thus hypothesize that, when trained with few classes, the supervised representations only focus on limited object parts exclusive to these specific classes, and become too sparse for effective transfer to novel classes. The loss of information in turn causes collapse of the supervised pretraining. In contrast, the contrastive representations are learned to discriminate instances based on broader regions and independent of the class labels, thus insensitive to the number of classes.

To verify the above hypothesis, we visualize the CAMs (Zhou et al. 2016) for both the supervised and unsupervised models pretrained with different number of classes in Fig. 2. It can be seen that: 1) the activated regions of the supervised models are concentrated whereas those of the unsupervised are dispersed; and 2) for the supervised models, the acti-

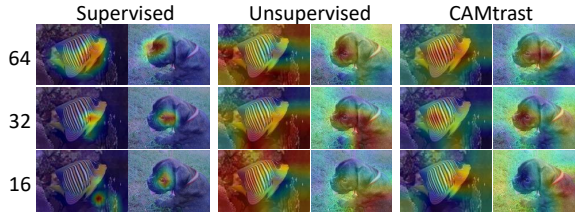

Figure 2: Class activation maps of supervised, unsupervised (SimSiam), and our CAMtrast models pretrained with different number of classes (64, 32, and 16).

vated regions obviously shrink when the number of classes decreases, while for the unsupervised models the changes are less appreciable. These observations qualitatively prove our hypothesis.

Combining the experimental findings so far, we thus conclude that it is the amount of information perceived by the learning model that is crucial to the transfer learning performance, instead of absolute size of the dataset. It is worth noting that Zhao et al. (2021) concluded similarly that the information loss in supervised pretraining resulted in the degradation of transfer performance, although without any rigorous quantitative validation. Next, we quantify the amount of information in the learned representations with two metrics.

Quantification of Learned Representations: *Representation entropy* (Devijver and Kittler 2012) is a measure of informativeness of a set of features. Let $X : N \times D$ be a feature set, where N is the number of features, and D is the feature dimension, and λ_j be the the j^{th} eigenvalue of X's covariance matrix $\Sigma : D \times D$. Then the representation entropy is defined as:

$$H_R = -\sum_{j=1}^{D} \tilde{\lambda}_j \log \tilde{\lambda}_j, \tag{1}$$

where $\tilde{\lambda}_j = \lambda_j / \sum_{j=1}^{D} \lambda_j$. The smaller H_R is, the less information is contained in the feature set. Besides H_R, we also quantify the amount of information from the perspective of feature distribution. With theoretical motivations, Wang and Isola (2020) empirically verified that *uniformity*, i.e., the uniform distribution on the unit hypersphere, is a desirable property for representations, which prefers a feature distribution that preserves maximal information. The uniformity is defined as (Wang and Isola 2020):

$$U(f;t) = -\log \mathop{\mathbb{E}}_{x,y \overset{\text{i.i.d.}}{\sim} p_{\text{data}}} \left[e^{-t \left\| f(x) - f(y) \right\|_2^2} \right], \tag{2}$$

where x and y are two samples, p_{data} is the data distribution over $\mathbb{R}^n$, $f(\cdot)$ is a feature extractor, and t is set to 2 following Wang and Isola (2020). A higher U indicates a more uniform distribution on the unit hypersphere.

Based on both criteria, a good representation should be informative and uniform enough to effectively model the information from the training data.[2] We compute H_R and U on the full training datasets, to measure the informativeness and uniformity of the representations learned by supervised and unsupervised pretraining. The values are shown in Fig. 3 (corresponding to the results in Fig. 1). It is apparent that the feature representations learned by the unsupervised models are more informative (higher H_R) and more uniform (higher U) than the supervised models. In addition, H_R and U of unsupervised models are stable as the number of classes decreases, whereas for supervised models both metrics drop dramatically. These results suggest that the information extracted by the supervised models is relatively limited and the corresponding features are unevenly distributed, and the situation worsens with fewer number of classes. The quantitative analysis validates our conclusion in the previous section.

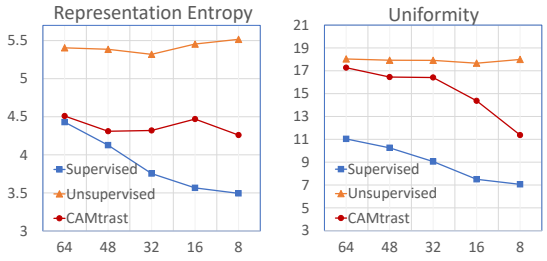

Figure 3: Effects of number of pretraining classes (x-axis) on the representation entropy and uniformity (y-axis) of the supervised, unsupervised (SimSiam), and our proposed CAMtrast pretraining on miniImageNet, while keeping the total number of images (4,800) unchanged.

Effects of Inter-Class Variation: So far we are only concerned about the apparent factors such as numbers of images and classes. A more insidious factor is the inter-class variation, whose impact on supervised and unsupervised pretraining still remains unclear.[3] Therefore, we make use of tieredImageNet to conduct two more control experiments. In the first experiment, the pretraining is done with the same number of fine-grained classes (C_f) belonging to different numbers of general categories (C_g), e.g., $C_f(C_g) = 50(20)$ means the training set includes 50 fine-grained classes from 20 general categories. Intuitively, fewer general categories indicate smaller inter-class variations and more challenging tasks. The fine-grained labels are used for supervised pretraining. As shown in Table 2 (top), with fewer general categories, the performance is better by 2.84–8.28% and the representation entropy and uniformity are generally higher for supervised pretraining, whereas unsupervised pretraining is largely unaffected.

The second experiment is aimed to answer this question: although the absolute amount of information contained in a set of images should remain constant, is it possible to make the supervised model perceive different amounts of information from it and yield different transfer performance? Accordingly, we train supervised models using the same set of images but supervise them with the general category labels and fine-grained class labels, respectively. Intuitively, the inter-class variation is lower for the fine-grained classification tasks, meaning higher inter-class similarity and diffi-

[2] It is worth noting that due to potential noise and redundancy, higher H_R or U does not always correspond to better performance.

[3] In this work, we do not concern ourselves with intra-class variation which was already discussed by Zhao et al. (2021).

	Supervised				Unsupervised			
C_f (C_g)	1-shot	5-shot	H_R	U	1-shot	5-shot	H_R	U
50 (20)	41.28	55.26	3.95	10.90	45.48	63.03	5.10	21.88
50 (5)	44.12	61.77	3.95	13.18	45.93	63.31	5.51	21.96
30 (20)	37.84	51.15	4.26	10.61	45.98	63.43	5.07	19.75
30 (3)	45.07	59.43	4.71	12.89	45.13	63.12	4.18	20.95
200 (**20**)	41.38	58.79	4.38	12.95	43.11	59.25	4.48	18.33
200 (20)	43.94	59.10	4.66	13.24				
100 (**10**)	38.65	50.82	4.15	11.75	39.27	55.54	4.98	17.11
100 (10)	42.25	58.61	4.69	12.73				

Table 2: Effects of inter-class variation on tieredImageNet 5-way few-shot classification accuracy (%). Top: pretraining with same number of fine-grained classes (C_f) from different numbers of general categories (C_g) (fixed total number of images: 35,100). Bottom: pretraining with coarse- versus fine-grained labels (as bolded) on the same sets of images (fixed number of images per class: 100).

culty. The results are shown in Table 2 (bottom). As we can see, using fine-grained class labels consistently outperforms using category labels for supervised pretraining, with absolute margins about 0.31–7.79%. Meanwhile, representations of fine-grained pretraining are also more informative and uniform (higher H_R and U) than those of coarse-grained. Compared to unsupervised counterparts, the performances of coarse-grained pretraining are apparently worse, whereas those of fine-grained pretraining are generally better. Therefore, more information is effectively perceived from the same set of images by the fine-grained pretraining, leading to superior transfer performance.

The results of both experiments suggest that reducing inter-class variations can force the supervised models to perceive more information from the images for the more challenging classification tasks, leading to more generalizable representation learning towards better transfer.

Summary: We present the following findings in this section: 1) The absolute data size is unlikely to be the key to the difference in transfer performance between supervised and unsupervised pretraining. 2) It is the amount of information effectively perceived by the learning model that is crucial to the transfer learning performance, which is confirmed by our CAM-based qualitative analysis, and representation entropy and uniformity based quantitative analysis. 3) For supervised pretraining, the effectively perceived information can be affected by the inter-class variation. The last finding suggests that it is possible to strengthen supervised pretraining by making it perceive more information. In the next section, we propose a novel framework that achieves this goal via effective training strategy.

4 Better Label Supervision

Towards Informative and Uniform Representation: From Section 3 we know that, effective representation for transfer learning requires broad attention to various characteristics of the images, yet supervised pretraining only focuses on the small, task-relevant regions while ignoring others. An intuitive countermeasure is to force the network to discover characteristic regions other than the most discriminative ones (e.g., dog head for an image labeled as dog) for correct classification. To this end, we propose to suppress

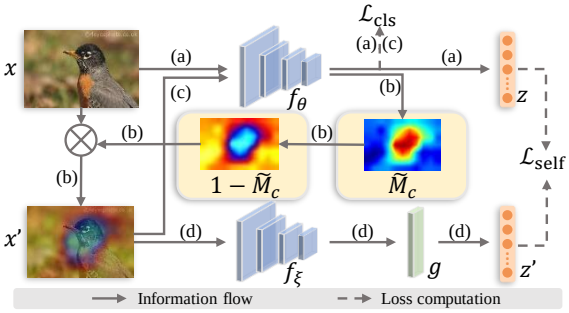

Figure 4: Pipeline of the proposed CAMtrast framework. Processing pathways: (a) process the original image x with the classification network f_θ; (b) class activation map (CAM) guided image suppression; (c) and (d): process the suppressed image x' with f_θ and $g(f_\xi)$, respectively.

the most discriminative regions in an image as perceived by the network, and force the network to correctly classify the image based on other regions.

Concretely, as illustrated in Fig. 4, we first warm up the classification network f_θ for t epochs by minimizing a standard cross-entropy loss $\mathcal{L}_{\text{cls}}\big(f_\theta(x)\big)$ (Pathway (a)), where x is the input image. Then, from the $t+1^{\text{th}}$ epoch onwards we obtain the CAM of x, denoted by M_c (Pathway (b)). M_c localizes the most discriminative regions in x as perceived by the network, where higher values indicate higher activation for classification. Despite the existence of much potentially useful information in the image, the highly activated regions in M_c is often very concentrated (Fig. 2). Next, M_c is upsampled to the size of the input and normalized to the range [0, 1], denoted by $\tilde{M}_c$, which is subsequently used to suppress the high-response regions by:

$$x' = (1 - \tilde{M}_c) \otimes x, \tag{3}$$

where $\otimes$ is element-wise multiplication. Finally, the suppressed image x' is fed into f_θ to force discovery of new information for correct recognition with the loss $\mathcal{L}_{\text{cls}}\big(f_\theta(x')\big)$ (Pathway (c)).

Integrate Supervised and Unsupervised Pretraining: Although the suppressed image can force the network to dig other regions, the network still relies on class-relevant features for pretraining. As shown in Section 3, the representations learned by self-supervised contrastive learning are more representative and uniform, and yield superior performance to supervised counterparts when the training data lack diversity. Hence, we propose to integrate self-supervised contrastive learning (He et al. 2020; Chen and He 2021) with the CAM-guided supervised learning to further improve the informativeness and uniformity of the learned representations. We thus call our framework the CAM-guided contrastive (CAMtrast) learning. Without losing generality, two different views of the same image are input to the Siamese networks (Chopra, Hadsell, and LeCun 2005) (Fig. 4). It is worth noting that we directly use the original image x and its suppressed version x' as the paired input, and get rid of the extensive augmentations that are commonly adopted in self-supervised contrastive learning. x and

x' are processed by the Siamese networks f_θ and f_ξ, respectively, one of which is followed by a prediction function g, to produce two projections $z = f_\theta(x)$ and $z' = g(f_\xi(x'))$ (Pathways (a) and (d)). Then, a consistency loss $\mathcal{L}_{\text{self}}$ is enforced on z and z' for contrastive learning.

Our CAMtrast framework is generic and can incorporate different self-supervised contrastive learning methods. In this work, we experiment with the MoCo_v2 (Chen et al. 2020b) and SimSiam (Chen and He 2021). For MoCo_v2, g is an identity mapping, and f_θ and f_ξ share network structure but not parameters. The network parameters of f_θ are updated by stochastic gradient descent, whereas those of f_ξ are updated by exponential moving average: $\xi \leftarrow m\xi + (1-m)\theta$, where $m \in [0, 1)$. A form of contrastive loss (Hadsell, Chopra, and LeCun 2006) called InfoNCE (Oord, Li, and Vinyals 2018) is employed:

$$\mathcal{L}_{\text{self}}(z, z') = -\log\left[e^{z\cdot z'/\tau}/\left(e^{z\cdot z'/\tau} + \sum\nolimits_{j=1}^{K} e^{z\cdot z'_j/\tau}\right)\right], \quad (4)$$

where τ is a temperature hyper-parameter and K is the number of negative samples. For SimSiam, g is implemented as a multilayer perceptron, and f_θ and f_ξ share parameters: $f_\theta = f_\xi$. A cosine similarity loss is employed to minimize the distance between z and z':

$$\mathcal{L}_{\text{self}}(z, z') = -(z \cdot z')/(\|z\|_2 \|z'\|_2). \quad (5)$$

The overall optimization objective of CAMtrast is:

$$\mathcal{L}(x) = \mathcal{L}_{cls}\big(f_\theta(x)\big) + \mathcal{L}_{cls}\big(f_\theta(x')\big) + \mathcal{L}_{\text{self}}\big(f_\theta(x), g(f_\xi(x'))\big). \quad (6)$$

5 Experiments

In this section, we evaluate our CAMtrast on three types of downstream tasks: few-shot recognition, object detection, and semantic segmentation. We also conduct ablation studies to evaluate effectiveness of key components. The source code for this paper is available at: https://github.com/jinghanSunn/CAMtrast.

Few-Shot Recognition: We evaluate our approach with both standard (pretraining on base classes and transferring to novel classes) as well as cross-domain (pretraining on a dataset and transferring to another) few-shot recognition tasks. For the former, the miniImageNet and tieredImageNet datasets (already described in Section 3) are used. For the latter, the tieredImageNet and CIFAR-FS (Bertinetto et al. 2018) are used. CIFAR-FS includes 100 classes that are divided into 64 base, 16 validation, and 20 novel classes. A total of 600 tasks are randomly sampled from the novel classes for performance evaluation. All results are based on a standard ResNet-50 (He et al. 2016) backbone. Four NVIDIA Tesla V100 GPUs are used for training. For our proposed CAMtrast, we train 30 epochs ($t = 30$) for warmup on miniImageNet and CIFAR-FS, and 50 epochs ($t = 50$) on tieredImageNet. The protocol of linear classification with the logistic regression classifier on frozen features is adopted (Tian et al. 2020). More implementation details are provided in the supplement.

Standard few-shot recognition: We first compare the performance of our framework to those of supervised, three SOTA unsupervised (MoCo_v2 (Chen et al. 2020b), BYOL (Grill et al. 2020), and SimSiam (Chen and He 2021)), and Exemplar (Zhao et al. 2021) pretraining. All methods are pretrained using all base-class samples and evaluated on novel classes of miniImageNet and tieredImageNet, respectively. The results are shown in Table 3. We observe that supervised pretraining is better than unsupervised ones here, likely due to the increased information enforced by more labels. However, by integrating supervised and unsupervised pretraining, our framework achieves the highest accuracies in both 1- and 5-shot settings on both datasets, beating all competitors using labels or not. These results validate the efficacy of our framework.

	miniImageNet		tieredImageNet	
Methods	1-shot	5-shot	1-shot	5-shot
Supervised	55.31±0.78	73.73±0.63	61.06±0.84	78.90±0.66
Exemplar	53.43±0.73	74.64±0.58	58.32±0.80	75.61±0.62
BYOL	45.01±0.81	61.42±0.64	47.31±0.87	63.78±0.73
MoCo_v2	52.04±0.73	73.56±0.57	56.87±0.83	74.50±0.70
SimSiam	48.81±0.71	69.27±0.58	55.42±0.86	75.39±0.75
Ours (MoCo_v2)	54.14±0.72	74.02±0.59	**62.90**±0.84	**80.04**±0.66
Ours (SimSiam)	**58.42**±0.78	**74.68**±0.63	62.88±0.84	77.84±0.67

Table 3: Standard 5-way few-shot classification accuracies (%; mean ± 95% confidence interval (CI)) and comparison to other methods.

Few-shot recognition with fewer classes: In Section 3 the supervised pretraining falls short of unsupervised counterparts with fewer numbers of training classes. Here we repeat that experiment with CAMtrast. As shown in Fig. 1, CAMtrast outperforms both the supervised and unsupervised pretraining by margins approximately from 0.50% to 10.09%. Especially, the supervised pretraining collapses and becomes inferior to unsupervised pretraining when the number of classes is below 32. By better utilizing the labels, CAMtrast can recover the generalization ability of supervised learning and outperforms the unsupervised pretraining even with extremely low numbers of classes. We conjecture that the superiority of CAMtrast to unsupervised pretraining is because the labels can help the contrastive learning to focus on regions that are relatively more discriminative, instead of making easy contrast based on "shortcuts" (Robinson et al. 2021). In addition, CAMtrast is more stable with respect to the reduction of classes than supervised pretraining. To gain insights if CAMtrast really improves the learned representations, we also quantify the representation entropy and uniformity for CAMtrast. As shown in Fig. 3, CAMtrast indeed consistently increases both of the metrics upon supervised pretraining across different numbers of classes.

For a more intuitive perception, we visualize CAMs of CAMtrast in Fig. 2. As expected, activations of CAMtrast are more dispersed than those of the supervised pretraining, yet more focused than those of the unsupervised. These observations indicate that in the proposed CAMtrast, contrastive learning encourages attention to broad areas while CAM-guided supervised learning helps focus on semantically meaningful regions.

Cross-domain few-shot recognition: To further evaluate the transfer performance in a more realistic scenario, we also conduct few-shot experiments with domain shifts between tieredImageNet and CIFAR-FS. The models are pretrained on either tieredImageNet's or CIFAR-FS's base classes and

	Tiered → CIFAR		CIFAR → Tiered	
Methods	1-shot	5-shot	1-shot	5-shot
Supervised	44.56±0.78	61.11±0.76	43.43±0.70	58.48±0.71
MoCo_v2	34.94±0.72	48.13±0.70	42.07±0.72	58.51±0.69
SimSiam	35.75±0.70	50.37±0.79	43.97±0.72	59.31±0.74
Exemplar	36.63±0.74	49.99±0.75	44.13±0.74	59.07±0.70
Ours (MoCo_v2)	**47.26**±0.85	**64.48**±0.75	45.23±0.74	62.03±0.62
Ours (SimSiam)	45.23±0.79	62.32±0.77	**47.22**±0.77	**63.77**±0.64

Table 4: Cross-domain 5-way few-shot recognition accuracies (%; mean±95% CI). Tiered → CIFAR means pretrained on tieredImageNet and tested on CIFAR-FS, and vice versa.

Model	Supervised	CAM	Contrast	1-shot	5-shot	H_R	U
Baseline	✓			33.46	45.30	3.56	7.49
Ablation-1	✓	✓		34.39	47.85	3.67	8.09
Ablation-2 (MoCo_v2)	✓		✓	35.38	48.68	4.13	13.17
Ablation-2 (SimSiam)	✓		✓	36.24	49.99	4.33	13.60
CAMtrast (SimSiam)	✓	✓	✓	**39.39**	**54.22**	**4.47**	**14.37**

Table 5: Ablation study using 5-way few-shot recognition accuracy (%) on miniImageNet; corresponding losses: Supervised: $\mathcal{L}_{cls}\big(f_\theta(x)\big)$, CAM(-guided): $\mathcal{L}_{cls}\big(f_\theta(x')\big)$, and Contrast: $\mathcal{L}_{\text{self}}\big(f_\theta(x), g(f_\xi(x'))\big)$.

evaluated on the other's novel classes. As shown in Table 4, our proposed CAMtrast substantially outperforms supervised and unsupervised pretraining, as well as Exemplar about 3–16%. In fact, with the domain gap present, this setting is more challenging than standard few-shot recognition. Yet our framework demonstrates even larger improvements upon the competing methods, further validating its efficacy.

Ablation Study: We conduct ablation studies to validate the effects of the two core improvements we made to traditional supervised learning: CAM-guided activation suppression and integration with contrastive learning. We base the ablation studies on the data point of $C = 16$ in Fig. 1 (16 base classes with a total of 4,800 images on miniImageNet for training) considering the collapse of supervised pretraining from that point. As shown in Table 5, compared to the baseline of supervised pretraining, either adding CAM-guided suppression (Ablation-1) or incorporating contrastive learning (Ablation-2, implemented with either MoCo_v2 or SimSiam) improves the transfer performance, by about 1–2% and 2–4%, respectively. For the complete CAMtrast model, SimSiam is incorporated for its superior performance to MoCo_v2 in Ablation-2. Then, the combination of the CAM-guided suppression and SimSiam brings further performance improvements of approximately 3–4%, suggesting that they mutually benefit each other. Similar trends of improvements can be observed for representation entropy (H_R) and uniformity (U), too. On one hand, contrastive learning helps supervised pretraining by making the network attend to broader regions in the images. On the other hand, the CAM-enhanced supervised learning helps contrastive learning focus on more meaningful contents in the images rather than the shortcuts (Robinson et al. 2021). Besides, the CAM generated in an epoch can be different from that generated in another for the same image, serving as effective online augmentation.

Generalize for More Downstream Tasks: So far we have focused on the downstream task of image classification. To investigate the effectiveness of the proposed CAMtrast on other downstream applications, we also consider two other fundamental tasks in computer vision: PASCAL VOC (Everingham et al. 2010) object detection and Cityscapes (Cordts et al. 2016) semantic segmentation. We pretrain the models on the base classes of tieredImageNet. For PASCAL VOC object detection, we use Faster-RCNN (Ren et al. 2015) with a backbone of ResNet-50 (He et al. 2016) as the detector. For Cityscapes semantic segmentation, we follow Chen et al. (2018a) to employ the DeepLab v3+ with a ResNet-50 backbone. All layers are fine-tuned end-to-end. More implementation details are provided in the supplement. The results are presented in Table 6. We notice that for these downstream tasks, unsupervised pretraining is better than supervised counterparts, which is the opposite of classification tasks (Table 3 and Table 4). This may be explained by the larger task misalignment between pretraining and target tasks (Zhao et al. 2021). Despite that, our framework (with MoCo_v2) outperforms all competing methods, including the supervised, MoCo_v2, SimSiam, and Exemplar pretraining, with apparent margins on both object detection (about 1–2% advantages) and semantic segmentation (about 1–6% advantages) tasks. This suggests that CAMtrast is able to extract more effective information that is more transferable to a wide variety of potential downstream tasks, than both supervised and unsupervised pretraining as well as Exemplar.

Dataset	Metric	Supervised	MoCo_v2	SimSiam	Exemplar	Ours
VOC	AP (%)	70.21	71.05	70.76	71.42	**72.16**
Cityscapes	mIoU (%)	68.26	72.35	70.32	72.93	**74.01**

Table 6: Performance on PASCAL VOC07 object detection test set and Cityscapes semantic segmentation validation set, with pretraining on base classes of tieredImageNet.

6 Conclusion

In this work, we empirically investigated when and why supervised pretraining would fall short of unsupervised contrastive pretraining for transfer learning of image classification tasks. Our core finding was that it was the amount of information effectively perceived by the learning model that was crucial to visual transfer learning, instead of absolute size of the dataset. Motivated by this finding, we proposed CAMtrast, a novel supervised framework which integrated CAM-guided activation suppression and self-supervised contrastive learning for more effective pretraining by encouraging the model to perceive more information from the images. We evaluated CAMtrast on three different downstream tasks: image classification, object detection, and semantic segmentation on several public datasets. Results showed that CAMtrast not only substantially improved transfer performance of supervised pretraining, but also outperformed unsupervised pretraining and a recent related work based on exemplars (Zhao et al. 2021). In conclusion, we took a step further towards understanding fundamental mechanisms of transfer learning in computer vision. While we plan to generalize our findings on larger and more datasets and more tasks soon, we also expect more efforts from the community to be devoted to studies along the line of fundamental understanding.

Acknowledgements

This work was supported by the National Key R&D Program of China (2020AAA0109500/2020AAA0109501) and the Fundamental Research Funds for the Central Universities (Grant No. 20720190012, 20720210121).

References

Bertinetto, L.; Henriques, J. F.; Torr, P. H.; and Vedaldi, A. 2018. Meta-learning with differentiable closed-form solvers. *arXiv preprint arXiv:1805.08136*.

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018a. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision*, 801–818.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020a. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 1597–1607. PMLR.

Chen, W.-Y.; Liu, Y.-C.; Kira, Z.; Wang, Y.-C. F.; and Huang, J.-B. 2018b. A closer look at few-shot classification. In *International Conference on Learning Representations*.

Chen, X.; Fan, H.; Girshick, R.; and He, K. 2020b. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.

Chen, X.; and He, K. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15750–15758.

Chopra, S.; Hadsell, R.; and LeCun, Y. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, 539–546. IEEE.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The Cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3213–3223.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Devijver, P. A.; and Kittler, J. 2012. *Pattern recognition theory and applications*, volume 30. Springer Science & Business Media.

Doersch, C.; Gupta, A.; and Efros, A. A. 2015. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, 1422–1430.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The PASCAL Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2): 303–338.

Girshick, R.; Donahue, J.; Darrell, T.; and Malik, J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 580–587.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P. H.; Buchatskaya, E.; Doersch, C.; Pires, B. A.; Guo, Z. D.; Azar, M. G.; et al. 2020. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*.

Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, 1735–1742. IEEE.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Jing, L.; and Tian, Y. 2020. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (01): 1–1.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3431–3440.

Mahajan, D.; Girshick, R.; Ramanathan, V.; He, K.; Paluri, M.; Li, Y.; Bharambe, A.; and Van Der Maaten, L. 2018. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European Conference on Computer Vision*, 181–196.

Noroozi, M.; and Favaro, P. 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, 69–84. Springer.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Ren, M.; Triantafillou, E.; Ravi, S.; Snell, J.; Swersky, K.; Tenenbaum, J. B.; Larochelle, H.; and Zemel, R. S. 2018. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28: 91–99.

Robinson, J.; Sun, L.; Yu, K.; Batmanghelich, K.; Jegelka, S.; and Sra, S. 2021. Can contrastive learning avoid shortcut solutions? *arXiv preprint arXiv:2106.11230*.

Sermanet, P.; Eigen, D.; Zhang, X.; Mathieu, M.; Fergus, R.; and LeCun, Y. 2013. OverFeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*.

Tan, C.; Sun, F.; Kong, T.; Zhang, W.; Yang, C.; and Liu, C. 2018. A survey on deep transfer learning. In *International Conference on Artificial Neural Networks*, 270–279. Springer.

Tian, Y.; Wang, Y.; Krishnan, D.; Tenenbaum, J. B.; and Isola, P. 2020. Rethinking few-shot image classification: a good embedding is all you need? In *Proceedings of the European Conference on Computer Vision*, 266–282. Springer.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D.; et al. 2016. Matching networks for one shot learning. *Advances in Neural Information Processing Systems*, 29: 3630–3638.

Wang, T.; and Isola, P. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, 9929–9939. PMLR.

Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; and Sang, N. 2018. Learning a discriminative feature network for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1857–1866.

Zhang, P.; Wang, F.; and Zheng, Y. 2017. Self supervised deep representation learning for fine-grained body part recognition. In *IEEE International Symposium on Biomedical Imaging*, 578–582. IEEE.

Zhang, R.; Isola, P.; and Efros, A. A. 2016. Colorful image colorization. In *European Conference on Computer Vision*, 649–666. Springer.

Zhang, Z.; Zhang, X.; Peng, C.; Xue, X.; and Sun, J. 2018. Exfuse: Enhancing feature fusion for semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, 269–284.

Zhao, N.; Wu, Z.; Lau, R. W. H.; and Lin, S. 2021. What makes instance discrimination good for transfer learning? In *International Conference on Learning Representations*.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2921–2929.