On the Efficacy of Small Self-Supervised Contrastive Models without Distillation Signals

Haizhou Shi[2*], Youcai Zhang[1], Siliang Tang[2†], Wenjie Zhu[3*], Yaqian Li[1], Yandong Guo[1], Yueting Zhuang[2]

[1] OPPO Research Institute, [2] Zhejiang University, [3] New York University
{zhangyoucai, liyaqian, guoyandong}@oppo.com, {shihaizhou, siliang, yzhuang}@zju.edu.cn, wz2140@nyu.edu

Abstract

It is a consensus that small models perform quite poorly under the paradigm of self-supervised contrastive learning. Existing methods usually adopt a large off-the-shelf model to transfer knowledge to the small one via distillation. Despite their effectiveness, distillation-based methods may not be suitable for some resource-restricted scenarios due to the huge computational expenses of deploying a large model. In this paper, we study the issue of training self-supervised small models without distillation signals. We first evaluate the representation spaces of the small models and make two non-negligible observations: (i) the small models can complete the pretext task without overfitting despite their limited capacity and (ii) they universally suffer the problem of over clustering. Then we verify multiple assumptions that are considered to alleviate the over-clustering phenomenon. Finally, we combine the validated techniques and improve the baseline performances of five small architectures with considerable margins, which indicates that training small self-supervised contrastive models is feasible even without distillation signals. The code is available at *https://github.com/WOWNICE/ssl-small*.

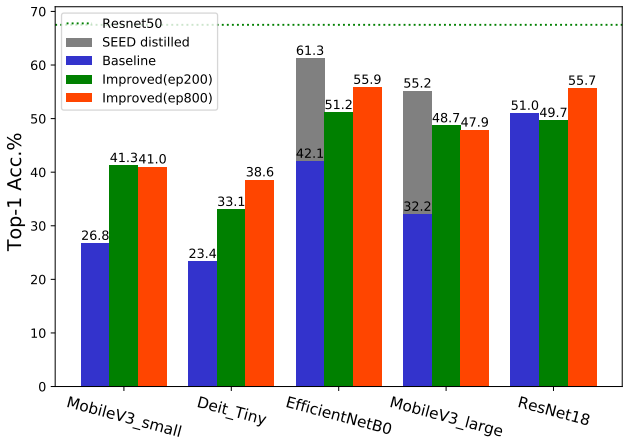

Figure 1: Baseline and improved performance of five popular small models. The gray boxes represent the SEED distillation results with ResNet50 as teacher network.

1 Introduction

Recently, the development of self-supervised contrastive learning has empowered the models to learn a good representation space without the guidance of labels. Among them, the large models, e.g., ResNet50, ResNet152, ViT, have achieved comparable results as the supervised learning methods (Chen et al. 2020a,b; He et al. 2020; Chen, Xie, and He 2021).

The small models, however, could not gain such good performance under the same training paradigm. In supervised learning, a small model is outperformed by its large counterpart by 9% in terms of accuracy (MobileNetV3_small's 67.7% v.s ResNet50's 76.1%). However, in self-supervised learning, the gap between the small and the large is dramatic (MobileNetV3_small's 26.8% v.s ResNet50's 67.5%). One simple and widely accepted assumption explaining small models' failure is proposed in many concurrent works (Fang et al. 2021; Xu et al. 2021; Gu, Liu, and Tian 2021). It argues that the pretext instance discrimination task is too challenging for the small models due to their limited capacity. Based on this assumption, the most popular framework of training self-supervised small models utilizes the technique of knowledge distillation (Hinton, Vinyals, and Dean 2015), where the problem of training small self-supervised models boils down to two phases. It first trains a large learner in a self-supervised fashion and then trains the small learner to mimic its representation distribution (Koohpayegani, Tejankar, and Pirsiavash 2020; Fang et al. 2021; Xu et al. 2021; Gu, Liu, and Tian 2021; Gao et al. 2021).

Although proven to be effective, this method is not applicable in some resource-restricted scenarios. For example, on mobile devices, due to the limited computational resource and data privacy, we cannot deploy a large model to guide the small model's learning or transfer the local data back to the server for distillation. Therefore, training small models with only contrastive learning signals is a challenge worth addressing. Besides, we argue that distilling the knowledge from an already-known-working large model to a small model weakens the significance of the self-supervision setting since the large model, although trained without la-

*This work was done when Haizhou Shi and Wenjie Zhu were interning at OPPO Research Institute, Shanghai.
†Corresponding author.

bels, is a strong supervision signal for the small one.

In this paper, we study the critical question of whether the small models can learn under the self-supervised contrastive signals without the guidance of the high-capacity teacher. To this end, we first extensively study the properties of the representation spaces generated by the small and large model. We show that, contradictory to the aforementioned assumption that "the small learners cannot complete the pretext task as well as the large learners", small models can achieve comparable (or even better) pretext-task performance as the large ones. Despite this, the small learners seem to get stuck at the "over-clustering" position where the augmented views of the same sample are tightly clustered, whereas the samples of the same semantic class are not well clustered in the representation space.

Next, we study some of the common assumptions that are considered to solve the problem of over clustering, including higher temperature, fewer negative samples during training, more aggressive augmentation scheme, better weight initialization, and various projector architectures. We find that although some of the tricks work, there is no universal rationale behind the setting of these hyper-parameters, which showcases that the reason for the success of self-supervised contrastive learning still needs to be further understood. Nevertheless, we combine the validated tricks and finally improve the baseline performance of 5 often-used small learners by a large margin (on average $15+$ absolute linear evaluation accuracy points, measured on ImageNet1k (Deng et al. 2009)), as shown in Fig.1. Although there is still a performance gap between the self-supervised small models and their distillation-based counterparts, we show a huge potential of the research direction where the self-supervised small models are trained without distillation signals. Our contributions are summarized as follows:

- To the best of our knowledge, our paper for the first time studies why the small models perform poorly under the current self-supervised contrastive learning framework.
- We identify that small models can complete the pretext instance discrimination task as large ones. However, the representation spaces yielded by the small models universally suffer the problem of over clustering.
- We improve the baseline performance of 5 different small models by a large margin (on average 15+ absolute points), showing the potential of training contrastive small models in the resource-restricted scenario.

2 Background and Related Work

2.1 Self-Supervised Contrastive Learning

Contrastive learning adopts the multi-view hypothesis, which regards the augmented views of the data sample as positive and requires the model to distinguish them from other (negative) samples (Arora et al. 2019). The contrastive learning optimizes the following InfoNCE (Oord, Li, and Vinyals 2018) loss:

$$\mathcal{L}_{\text{con}}(x_i) = -\log \frac{\exp(s_{i,i}/\tau)}{\exp(s_{i,i}/\tau) + \sum_{k\neq i}^{K} \exp(s_{i,k}/\tau)}, \quad (1)$$

where $s_{i,i} = f(x_i)^\top f(x_i^+)$ and (x_i, x_i^+) is the positive pair consisting of two randomly augmented views. $s_{i,j} = f(x_i)^\top f(x_j)$ is the similarity between two negative samples. The negative sample set $\{x_j\}_{j\neq i}^K$ is constructed by sampling for K times independently from the data distribution.

Based on the primary form of contrastive learning, multiple contrastive-based methods have been proposed to train the networks without supervision (Wu et al. 2018; Oord, Li, and Vinyals 2018; Tian, Krishnan, and Isola 2019; Chen et al. 2020a; He et al. 2020). Many of them achieve the SOTA performance on the downstream linear classification task with the backbone network fixed (Zhang, Isola, and Efros 2016; Oord, Li, and Vinyals 2018; Bachman, Hjelm, and Buchwalter 2019). However, little attention has been paid to training small models (Howard et al. 2017; Tan and Le 2019) solely under the contrastive learning framework, for its failure has been widely observed (Koohpayegani, Tejankar, and Pirsiavash 2020; Fang et al. 2021; Gao et al. 2021; Xu et al. 2021; Gu, Liu, and Tian 2021). In this paper, we want to fill in the void of training small models with and only with contrastive learning signals.

2.2 Self-Supervised Small Models

Currently, knowledge distillation (Hinton, Vinyals, and Dean 2015) becomes a widely acknowledged paradigm to solve the slow convergence and difficulty of optimization in self-supervised pretext task for small models (Koohpayegani, Tejankar, and Pirsiavash 2020; Fang et al. 2021; Gao et al. 2021; Xu et al. 2021; Gu, Liu, and Tian 2021). ComPress (Koohpayegani, Tejankar, and Pirsiavash 2020) and SEED (Fang et al. 2021) distill the small models based on the similarity distributions among different instances randomly sampled from a dynamically maintained queue. DisCo (Gao et al. 2021) removes the negative sample queue and straightforwardly distills the final embedding to transmit the teacher's knowledge to a lightweight model. BINGO (Xu et al. 2021) proposes a new self-supervised distillation method by aggregating bags of related instances to overcome the low generalization ability to highly related samples. SimDis (Gu, Liu, and Tian 2021) establishes the online and offline distillation schemes and builds two strong baselines for the distillation-based training paradigm.

We appreciate the efforts made by the previous researchers, and it is advisable to address the problem of training small self-supervised models in a divide-and-conquer way: [*Self-Supervised Small Model*] = [*Self-Supervised Large Model*] + [*Supervised Distillation for Small Model*]. However, since the former and the latter tasks are both technically well-studied, we must point out that, by doing so, the existing works are somewhat evasive about the crucial problem "why the small models benefit from the self-supervised method far less than the large models" and "how to improve them". Furthermore, distillation-based methods are not applicable to some resource-restricted scenarios (*e.g.*, self-supervised learning on mobile devices) due to the huge computational expenses of deploying a large model, which is another core motivation of this work.

3 Representation Space Analysis

To better understand why the small models perform poorly in self-supervised contrastive learning, we introduce several evaluation metrics that help us diagnose the problems. There are generally two types of metrics: the first being pretext-task-related metrics that do not require any human annotation and reflect how well the model performs on the instance discrimination, including (i) **alignment**, (ii) **uniformity**, and (iii) **instance discrimination accuracy**; the second category being downstream-task-related metrics that require semantic labels, including (i) **intra-class alignment**, (ii) **best-NN**, and (iii) **linear evaluation protocol**.

Alignment. The alignment of the model is defined to measure the average squared distance of the samples' representations within a positive pair. It is one of the two training objectives of the contrastive loss under certain conditions (Wang and Isola 2020), and is the core of the multi-view hypothesis (Tian, Krishnan, and Isola 2019):

$$\mathcal{L}_{\text{align}} \triangleq \mathbb{E}\left[\|f(x_1) - f(x_2)\|_2^2\right], \tag{2}$$

where $(x_1, x_2) \sim \mathcal{P}_{\text{pos}}(x)$ denotes two augmented positive samples under the given data augmentation scheme.

Uniformity. The uniformity is defined to measure how uniform the representation distribution is in the hyper-spherical space (Wang and Isola 2020). It is the other objective the contrastive learning framework actively optimizes:

$$\mathcal{L}_{\text{uniform}} \triangleq \log \mathbb{E}\left[\exp\left(-t\|f(x) - f(y)\|_2^2\right)\right], \tag{3}$$

where $(x, y) \sim \mathcal{X}^2$ denotes two data points sampled i.i.d from the data distribution $\mathcal{X}$. The alignment and the uniformity of a representation space are essential to the model training; failing either one of them would cause the model to learn a non-generalizable representation space.

Instance discrimination accuracy. The contrastive learning framework follows the instance discrimination pretext task (Wu et al. 2018; Ye et al. 2019). It regards the augmented views of the same sample as of the same class and other samples as from different classes. Suppose there are N samples in total, then the pretext instance discrimination task can be viewed as an N-way classification problem. Some methods measure this metric within a batch of samples during training (He et al. 2020), which could be problematic since the batch size hugely influences the value of the accuracy. This paper measures the above three metrics on a static pre-generated dataset from both training and validation set. We use the standard data augmentation scheme used in MoCoV2 (He et al. 2020) to create positive pairs. For a fair comparison, we sample 50 images per class for both the training set and the validation sets, making it a $50,000$-way classification task for the pre-trained models.

Intra-class alignment. By alignment, we measure whether augmented views of the same sample are mapped into a small and tight cluster. Similarly, we want to measure whether the samples of the same semantic classes are mapped to a close neighborhood in the representation space.

	res-50	res-18	mob-l	mob-s	eff-b0	deit-t
#params (M)	25	11	5.5	2.5	5.3	5
training time (h)	42.4	40.9	40.0	38.9	40.0	39.8

Table 1: Parameter number and the training time of various models. The training times are evaluated on a single 8-card V100 GPU server for 200 epochs of training.

Following the form of the alignment term, we define the Intra-class alignment:

$$\mathcal{L}_{\text{intra-align}} \triangleq \mathbb{E}_{c,(x^{(c)},y^{(c)})}\left[\|f(x^{(c)}) - f(y^{(c)})\|_2^2\right], \tag{4}$$

where $(x^{(c)}, y^{(c)}) \sim \mathcal{P}_c$ denotes two samples that belong to the same semantic class c sampled independently. Note our work does not originally propose this metric. The metric of *tolerance* has been proposed in Wang and Liu (2021), where $\textit{tolerance} = 1 - \mathcal{L}_{\text{intra-align}}/2$ under the constraint of the hyper-spherical representation space. However, the definition of intra-class alignment complies more with our intuition since it has a similar tendency as alignment. What's more, these two metrics have a more in-depth causal relationship: we often regard the optimization of intra-class alignment as the consequence of the optimization of alignment.

Linear evaluation protocol. The linear evaluation protocol (also known as linear probing accuracy) is the standard metric that evaluates the linear separability of the representation space (Zhang, Isola, and Efros 2016; Chen et al. 2020a; He et al. 2020). It follows a basic assumption that the samples of different semantic labels should be easily separated in good representation spaces. In practice, we freeze the backbone network's parameter and train a simple linear classifier on top of it to measure the linear separability.

Best-NN. In this paper, we extend the k-NN metric to the best-NN metric. Many works adopt k-NN as the indicator for the downstream task performance (Fang et al. 2021; Wu et al. 2018) for it runs faster than the standard linear evaluation protocol and is deterministic once the hyper-parameter k is determined. However, different methods may generate different types of representation space favoring different k and cause unfair comparison. To solve this problem, we propose to use the best-NN metric, which picks the best k-NN accuracy out of the range $\{1, 3, \cdots, K\}$. In our setting, $K = 101$:

$$\text{best-NN} = \max_{k \in \{1,3,\cdots,K\}} \{k\text{-NN}\} \tag{5}$$

3.1 Differences Between Large and Small Models

We base our research on the MoCoV2 algorithm (Chen et al. 2020c) since it's computationally efficient and stable. To better utilize the computational resource, we set the batch size as 1024, and the learning rate as 0.06. The rest of the hyper-parameters are kept the same as the original MoCoV2 paper. We run the baseline experiments on five small models, including ResNet18 (He et al. 2016)

	Retrieval Acc.		$E[d]$		$E[d_{ij}]$		$log E[e^{-t\|x-y\|_2^2}]$	
	inst-disc		align		intra-class align		uniformity	
	train	val	train	val	train	val	train	val
res-50	83.4	85.4	0.25	0.25	0.91	0.90	-2.89	-2.85
res-18	81.2	82.8	0.24	0.24	0.95	0.93	-2.77	-2.74
mob-l	77.2	78.8	0.25	0.25	1.24	1.23	-3.29	-3.28
mob-s	86.2	88.5	0.25	0.25	1.32	1.30	-3.40	-3.32
eff-b0	81.0	82.1	0.26	0.26	1.29	1.27	-3.48	-3.51
deit-t	64.7	68.1	0.46	0.46	1.34	1.34	-3.06	-3.09

Table 2: Self-supervised baseline models trained by MoCoV2. Different colors demonstrate different semantic classes. The downstream-task performance of the baseline models are shown in Tab.7.

dubbed as res-18, MobileNetV3_large dubbed as mob-l, MobileNetV3_small (Howard et al. 2019) dubbed as mob-s, EfficientNet_B0 (Tan and Le 2019) dubbed as eff-b0, and one small vision transformer model DeiT_Tiny Touvron et al. (2021) dubbed as deit-t. For DeiT_Tiny, we set the head number of the transformer layer to 6 instead of 3 in the original work (Dosovitskiy et al. 2020; Touvron et al. 2021). During training, we fix the linear mapping layer as suggested in Chen, Xie, and He (2021). We train all the models on a single 8-card V100 GPU server, which well satisfies the computational requirement in all cases. We summarize the model parameter number (#params) and their training time as in Tab.1. Then we compare all the small models' behavior with the ResNet50 large model (dubbed as res-50). All the metrics are evaluated on the penultimate output of the networks (refer to Tab.2) and the ImageNet1k dataset (Deng et al. 2009). To yield a more intuitive understanding, we further visualize the representation distributions using the t-SNE (Van der Maaten and Hinton 2008) algorithm (refer to Fig.2). We list our observations and corresponding conclusions as follows.

OBSERVATION. The large and small models perform similarly on the pretext instance discrimination task.

The rationale behind self-supervised learning is that completing a human-designed challenging pretext task forces the model to learn the ability to extract generalizable features from the data. In the instance discrimination setting, models are required to do a N-way classification task where $N \gg C$, N is the number of the samples, and C is the number of the semantic classes. Therefore, when given the fact that the small models perform way worse than the large models, it's natural to conjecture that the pretext contrastive learning task might be too difficult for them. However, we are surprised to find that under the same training setting, the small learners, despite the architecture differences, can complete the instance discrimination task as well as the large model, in some cases (Tab.2, **mob-s**' 86.2% v.s **r-50**'s 83.4%) even better than

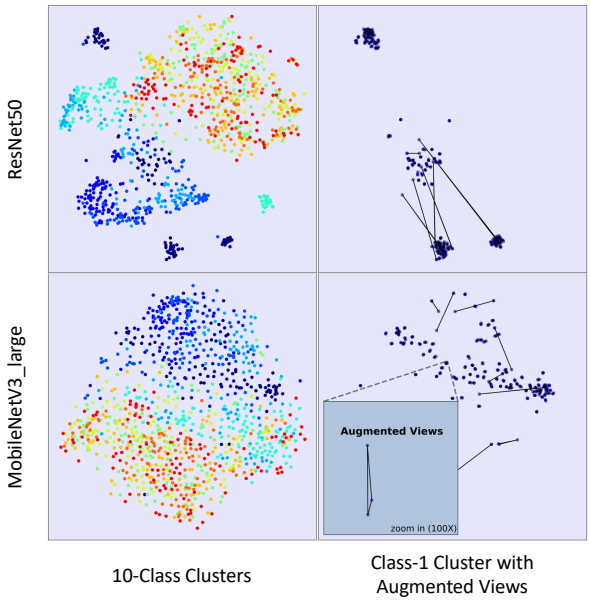

Figure 2: t-SNE representation space visualization of the large and small model under the same setting, evaluated on 10 classes of ImageNet. **Top**: ResNet50; **Bottom**: MobileNetV3_large. We connect the original data point and its two augmented views as a triangle as shown in the bottom-right zoom-in window, which indicates the small model's representations are over-clustered.

it. This phenomenon concludes that the pretext contrastive learning task is an easy task, (except for DeiT_Tiny). It also leads to a weird conclusion that accomplishing the pretext contrastive learning task doesn't have an absolute positive correlation with the model's downstream generalization ability. If this conclusion holds, then more critical questions about contrastive learning should be posed: (i) what else is optimized during self-supervised contrastive learning, and (ii) what is critical to the generalization other than the instance-discriminative ability?

OBSERVATION. For the small models, there is no overfitting problem when trained on the pretext task.

This conclusion is supported by the fact that each model's metrics have no significant difference on both the training and validation sets. According to recent studies, one of the most significant differences between large networks and small networks is their generalization ability. The phenomenon of "double descent" has been observed by the deep learning researchers (Nakkiran et al. 2019). The large networks can reduce the generalization error when the number of the model parameters increases to a certain extent. However, the concept of the generalization error, which measures the difference between the training and the testing error, cannot explain the small self-supervised models' poor performance on the downstream tasks. Based on this observation, we will only measure the models on the validation set in the following sections.

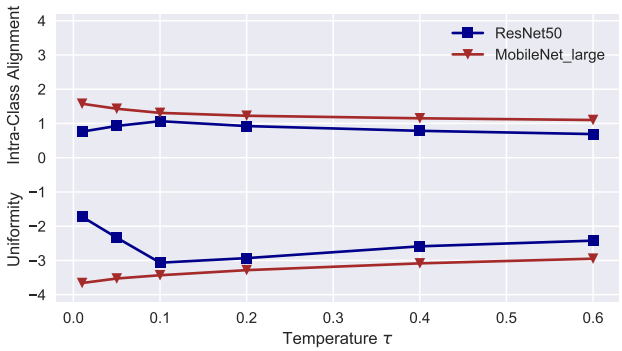

Figure 3: Intra-class alignment and uniformity yielded by different temperature τ.

		Temperature τ					
		0.01	0.05	0.1	0.2	0.4	0.6
mob-l	linear	24.3	33.5	**36.7**	31.3	21.4	15.9
	NN	25.5	32.8	**32.9**	28.0	21.2	17.3
	inst-disc	91.6	**93.3**	90.8	78.8	63.2	50.1
r-50	linear	43.9	60.6	63.8	**67.5**	64.3	61.8
	NN	31.3	41.0	47.8	**50.5**	47.4	43.9
	inst-disc	90.0	89.2	**91.3**	84.0	80.1	76.1

Table 3: The performance of the large and small models trained by different temperatures. "**linear**": top-1 accuracy of the linear evaluation protocol; "**NN**": top-1 accuracy of the Best-NN; "**inst-disc**": top-1 accuracy of the pretext instance discrimination task.

OBSERVATION. The small models' representation spaces are universally over-clustered compared to the large model.

According to Tab.2, we can see that the alignment of the small models is almost the same as the large model. However, the intra-class alignment is significantly higher than the large model. It showcases that the small models address the pretext-instance discrimination problem by trivially mapping augmented views of the same sample into a tight cluster without effectively clustering the samples of the same semantic label. This conclusion is supported by Fig.2 as well. By comparing the subfigures in the left column, we can see that ResNet50 achieves a representation space where same-class samples are well clustered. While MobileNetV3_large's representations are loosely scattered, the decision boundaries of which are not clear. As defined in Wang et al. (2021), this could be caused by either under-clustering (positive samples are not properly aligned) or over-clustering (positive samples are well aligned, but not leading to proper class-level clustering). To identify the cause, we further visualize the representation with their augmented views in the right column. The triangles are formed by connecting the samples and their positive pair. The zoom-in window (100X) shows that MobileNetV3_large can perfectly align the positive pairs of data, which proves the existence of over-clustering problem.

4 Solve Over Clustering with Simple Assumptions

Based on the above observations, we approach the over-clustering problem by applying simple tricks that seem to have a direct influence in this section. The assumptions are presented in the order of their importance: the temperature has the greatest influence and therefore we study it and fix it for the remaining assumptions; the negative samples and data augmentation correspond to the fundamental setting of negative/positive construction for contrastive learning; finally we deal with the network structure and weight initialization. At the end of the statement of each assumption, we use *[✗]* & *[✓]* to pre-indicate the correctness based on its following experiments.

4.1 Temperature

ASSUMPTION. The temperature trades off the intra-class alignment and the uniformity. Higher temperatures can alleviate the problem of over-clustering. [✗]

Wang et al. (2021) partially demonstrates how the temperature influences the contrastive model's behavior. When the temperature is infinitely large or infinitely close to zero, the original contrastive loss will degrade to two simple forms:

$$\mathcal{L}_{\text{simple}}(x_i) = -s_{i,i} + \lambda \sum_{k \neq i} s_{i,k}, \quad (\tau \to +\infty) \tag{6}$$

$$\mathcal{L}_{\text{triplet}}(x_i) = \max\left[s_{\max} - s_{i,i}, 0\right], \quad (\tau \to 0^{+}) \tag{7}$$

where in the first case, the model pays equal attention to the negative samples. In contrast, the model ignores all the negative samples but the one with the maximum similarity in the second case. Based on the analysis, lower temperature causes the model to pay more attention to the negative samples that are close to the current data point; higher temperature causes the model to pay more uniform attention to all the negative samples. Their experiments show that a higher temperature can improve the intra-class alignment by sacrificing the overall uniformity of the representation space, which means alleviating the problem of over-clustering. To validate this idea, we run experiments on different temperature setups on MobileNetV3_large and ResNet50. The results are shown in Tab.3.

There are two main takeaways. First, the temperature interacts with the contrastive learning methods in a more complicated way than we expected. Increasing the temperature does lead to a less uniform representation space with better intra-class alignment for the small model. However, for the large model, this principle does not fully apply: when the temperature is in the range of $(10^{-2}, 10^{-1})$, increasing the temperature causes a more uniform representation space, which is against the previously accepted understanding (Wang and Liu 2021). Secondly, for the small learner, although it suffers from the problem of over-clustering, further trading-off the intra-class alignment for better uniformity leads to better downstream-task performance. It indicates that generally, we want to keep the temperature lower than the large model for the small model. The reason for this

phenomenon might be due to the slow convergence of the small model. As shown in (Wang and Liu 2021), the gradient w.r.t the positive pair and the negative pairs are proportional to the reciprocal of the temperature $1/\tau$. We conjecture that lowering the temperature might have a similar effect of increasing the learning rate and thus help the model to converge faster. We leave the in-depth study on this subject for future. We set the temperature $\tau = 0.05$ based on its best-NN and inst-disc performance for the rest of this section.

4.2 Negative Sample Size

ASSUMPTION. The over-clustering problem is caused by the instance discrimination setting. False-negative samples make it hard for the small models to cluster data of the same semantic class. [✗]

A lot of work have focused on analyzing the influence of the current negative sampling strategies (Wu et al. 2020; Wang et al. 2021; Wang and Liu 2021; Huynh et al. 2020). According to (Wang et al. 2021), reducing the negative sample size would alleviate the problem of over clustering since it reduces the probability of colliding into false-negative samples in the mini-batch of data. Although it seems not to be a problem for the large model (larger the negative sample size, better the performance (Chen et al. 2020a)), we cannot exclude the possibility that the small models might be more prone to the problem of false-negative samples. We try different negative sample sizes in this section and find that, contrary to the large model's benefiting from a larger number of the negative sample (He et al. 2020), the small model's performance fluctuates with the number of negative samples, which counters the proposed assumption.

	65536	16384	4096	1024
linear	33.5	**34.2**	33.0	33.5
NN	32.8	**33.3**	32.6	33.1

Table 4: Results of different negative sample sizes.

4.3 Data Augmentation

ASSUMPTION. The current data augmentation scheme is not challenging enough. It cannot drive the small model to learn from the self-supervised signal continually. [✓]

Following Tian et al. (2020), we mainly consider the influence of the color distortion. As shown in Tab.5, we can see that increasing the strength of the data augmentation can help the small model to achieve better linear separability. However, there is also a sweet spot in selecting the augmentation scheme. If the augmentation is too strong, it may violate the basic "multi-view" assumption, and the model will learn the hand-crafted noises. If the augmentation is too weak, the model will easily align the positive samples without learning high-level visual patterns.

For the small model, we conjecture that due to its limited capacity, although it can perform comparably on the pretext instance discrimination task as the large model, it might achieve this goal by taking a different learning path.

CJ				GS	GB	linear	NN
brightness	contrast	saturation	hue				
0.4	0.4	0.4	0.1	0.2	0.5	33.5	32.8
0.4	0.4	0.4	0.2	0.2	0.5	34.8	34.8
0.8	0.8	0.8	0.4	0.5	0.5	**36.6**	**35.7**
0.8	0.8	0.8	0.4	0.5	1	35.2	35.0

Table 5: Results of different data augmentation. "**CJ**": ColorJitter; "**GS**": Grayscale; "**GB**": GaussianBlur.

It has been validated that if we do not apply the color distortion when creating the positive pairs, the model can fit the training objective quickly by taking the shortcut of color abbreviation (Chen et al. 2020a; Tian et al. 2020). Therefore we argue that the small models might be more prone to the shortcut existing in contrastive learning. Thus adding more randomness into the data augmentation is conducive to alleviating this problem.

4.4 Weight Initialization

ASSUMPTION. The small models are more prone to getting stuck at the local minimum during self-supervised contrastive training. [✓]

Usually, a good initialization point helps the model converge to a better minimum point. To validate this idea, we first train the small model under the supervision of a pre-trained large model using the SEED loss (Fang et al. 2021) for certain epochs. Then we load its weights and train the small model purely under the contrastive learning framework (same hyperparameter setting, no distillation signal). We benchmarked three sets of weights, which are trained under SEED for 2, 10, and 100 epochs, respectively.

	0	2	10	100
NN-init	-	30.2	43.1	47.5
NN	32.8	35.0	37.8	**39.3**
linear	33.5	35.4	39.4	**42.0**

Table 6: Results of initialization using weights trained by SEED. "**NN-init**": the best-NN performance after initialized with the SEED weights, before SSL training.

As shown in Tab.6, the model initialized by a better SEED checkpoint yields better downstream performance. When the weights are trained by SEED for 10 and 100 epochs, the final downstream accuracy doesn't surpass the accuracy before SSL training. It also demonstrates that the convergence point of the distillation-based methods and the SSL methods are different. One may argue that it's pointless to reload the small model from the distilled checkpoint and train it on the pure SSL signal. However, in reality, the small models are often deployed on the edge devices whose data distributions are much different from the central server. We can first train the small models using distillation loss and then deploy them onto edge devices if the distilled weights can help the small model to converge faster.

	MobileNetV3_large		MobileNetV3_small		EfficientNet_B0		ResNet18		DeiT_Tiny	
	linear	NN	linear	NN	linear	NN	linear	NN	linear	NN
baseline (200)	32.2	28.0	26.8	23.5	42.1	31.5	51.1	39.3	23.4	17.4
improved (200)	**48.7** (+16.5)	39.1 (+11.5)	**41.3** (+14.5)	32.8 (+9.3)	51.2 (+9.1)	41.3 (+9.8)	49.7(-1.3)	39.6 (+0.3)	33.1 (+9.7)	30.5 (+13.1)
improved (**800**)	47.9 (+15.7)	**41.6** (+13.5)	41.0 (+14.3)	**35.6** (+12.0)	**55.9** (+13.8)	**44.2** (+12.7)	**55.7** (+4.7)	**44.3** (+5.0)	**38.6** (+15.2)	**33.4** (+16.0)

Table 7: Linear evaluation results of improved baseline performance of five different small self-supervised contrastive models, pretrained and evaluated on ImageNet. The number listed in the parenthesis indicates the the epochs of pretraining.

	MobileNetV3_large			MobileNetV3_small			EfficientNet_B0			ResNet18			DeiT_Tiny		
	C10	C100	Cal101	C10	C100	Cal101	C10	C100	Cal101	C10	C100	Cal101	C10	C100	Cal101
baseline (200)	71.8	42.4	72.9	70.0	40.4	70.0	72.0	43.2	77.2	81.5	54.0	81.2	67.2	40.3	62.1
improved (200)	77.9	51.4	84.0	75.5	48.5	81.0	78.3	51.6	86.0	79.9	52.2	85.1	76.8	50.0	78.5
improved (**800**)	**80.3**	**53.8**	**85.3**	**78.6**	**52.2**	**82.7**	**81.5**	**55.6**	**86.9**	**82.0**	**54.4**	**86.7**	**79.7**	**54.7**	**81.0**

Table 8: Transfer learning results of the improved baselines across different architectures. "**C10**": evaluated on CIFAR10 dataset; "**C100**": evaluated on CIFAR100 dataset; "**Cal101**": evaluated on Caltech101 dataset. All the models are pretrained and evaluated on ImageNet. The number listed in the parenthesis indicates the the epochs of pretraining..

While in this work, to make sure the comparison between the loaded model and the baseline model is fair, we only use the weights that have worse downstream-task performance than the baseline model as the initialization weights in both the main experiments and the ablation study.

4.5 Projector Architecture

ASSUMPTION. Deeper (Fang et al. 2021) [✗] / Wider (Gao et al. 2021) [✓] / Dropout (Gao, Yao, and Chen 2021)[✗] MLP projector brings improvement to the (small) models.

The rationale behind the deeper/wider projector is simple: adding more parameters will make small models somewhat closer to the large model (without considering the architecture difference). Apart from this, one theory (Fang et al. 2021) states that the top layer of the network is more focused on addressing the pretext-task and loses some of the generalization ability. A deeper MLP projector would make the backbone network's representation farther away from the top layer and yield better downstream performance. A wider MLP obeys the Information Bottleneck (IB) principle (Gao et al. 2021) and thus make the small learner more capable of preserving the information, which might lead to better performance. As for the dropout layer in the projector, one can regard it as equivalent to a stronger data augmentation, creating a more challenging pretext task for the backbone network. We validate these assumptions one by one:

	MLP projector structure	linear	NN
baseline	1280, 128	33.5	32.8
deeper	1280,1280,128	29.0	28.4
wider	2560, 128	**33.9**	**33.2**
	1280,256	32.8	31.7
dropout	1280, dropout(p=0.5), 128	28.9	26.7

Table 9: Results of different MLP projector architectures.

In Tab.9, only the projector with the wider intermediate layer can improve the model with a small margin. One thing confusing about the dropout MLP is that it indeed creates a more challenging pretext task for the model and yields better alignment (0.32→0.28), but worse intra-class alignment (1.43→1.57). Usually, we assume the optimization of alignment causes the improvement of intra-class alignment since the former is the direct training objective in contrastive learning. However, the phenomenon produced by the dropout layer might challenge the common understanding and force us to rethink the efficacy of contrastive learning.

4.6 Summary

In summary, the assumptions that help the small model to achieve better downstream-task performance are listed as follows: (i) lower temperature, (ii) MLP projector with wider architecture, (iii) stronger augmentation scheme, and (iv) better weight initialization slightly trained by distillation-based methods. We will then in the next section combine the validated assumptions and apply them to 5 small models.

5 Improved Baselines for Small Self-Supervised Contrastive Models

In-domain downstream classification. We first validate the summarized measures for five different small models in the downstream classification task. To showcase their effectiveness, we adopt the same training hyper-parameter set for all the small architectures without individually tuning them. We set temperature $\tau = 0.1$, batch size $B = 512$. learning rate $\eta = 0.06$, and negative sample size $K = 65536$. For wider MLP projector, we adaptively set the intermediate layer of the MLP as twice wide as its input layer for all the architectures. For augmentation scheme, we adopt the best color distortion scheme as the strengthened "aug+". We use the checkpoints that are trained by SEED for two epochs as the weight initialization for all the models ex-

τ	mocov2. pre-train w-MLP	aug+	init	epochs	ImageNet Acc. linear	NN	Transfer Acc. cifar100
0.2				200	32.2	28.0	42.4
0.05				200	33.5	32.8	44.6
	✓			200	33.9	33.2	45.7
		✓		200	36.6	35.7	47.0
			✓	200	35.4	35.0	48.3
	✓	✓	✓	200	**38.0**	**37.6**	**49.3**
0.1				200	36.7	32.9	46.4
	✓	✓		200	42.6	37.5	50.6
	✓	✓	✓	200	46.0	40.3	51.8
	✓	✓	✓	**800**	**47.9**	**41.6**	**53.9**

Table 10: Ablation study of MobileNetV3_large trained by MoCoV2. "**w-MLP**": with a wider MLP head; "**aug+**": with augmentation having stronger color distortion as in section (Tian et al. 2020); "**init**": with better initialization trained by SEED for 2 epochs.

cept DeiT_Tiny since we observe that the small vision transformer has much slower convergence rate than other models even being trained by strong distillation signal; we train the DeiT_Tiny by SEED for 20 epochs and then use it as the weight initialization for SSL training (Dosovitskiy et al. 2020; Touvron et al. 2021). We train all the models for 800 epochs with cosine decay, and evaluate them at epoch 200 and epoch 800.

In Tab.7, all five models are improved by a large margin compared to the original baselines, which showcases the effectiveness and universality of our work. One thing to note here, training longer epochs (800) will surely improve the best-NN accuracy, while it's not the case for the linear evaluation protocol: the MobileNetV3 architectures will lose some of the linear separability to longer training.

Transfer learning for classification. We benchmark the transferability of the backbone networks on CIFAR10, CIFAR100 (Krizhevsky, Hinton et al. 2009), and Caltech101 (Fei-Fei, Fergus, and Perona 2004) image classification datasets. We evaluate the accuracy of the transfer learning in the same way as the linear evaluation protocol. To make the same set of hyper-parameters usable for different models, we follow the MoCoV2 (Chen et al. 2020c) and apply normalization to the representations before they are put into the linear classifier. We first tune the hyper-parameters, including base learning rate and the learning schedule for the pretrained ResNet50 model to make sure it is comparable to the value reported in the SimCLR (Chen et al. 2020a). Then we fix the hyper-parameters and apply them to all the small models.

We apply bicubic interpolation when resizing the image to 224x224, random cropping of scale $[0.6, 1]$, and random horizontal flipping during training. During testing, we first resize the image to 256x256 and center-crop it to 224x224. For CIFAR10/CIFAR100, we set the batch size $B = 256$, training epochs $E = 60$, base learning rate $\eta = 1.5$, and decrease the learning to one-tenth at epoch 40. For Caltech101, we set the base learning rate as 2 without scheduling. The results of transfer learning are reported in Tab.8, which reflects that the models trained by our improved setting have a more generalizable representation space.

Ablation study. We present the ablation study to verify the effectiveness of all the working measures we find, as shown in Tab.10. We can see that our simple tricks can improve the baseline by a large margin on both downstream classification tasks and transfer learning in both temperature settings. Lower temperature, better initialization, and more aggressive data augmentation are the factors that have the strongest influence on improving the small models' representation quality.

6 Conclusion

This work does not propose new technical contributions. Instead we provide an in-depth analysis on the behavior of the small self-supervised models with rigorous empirical verification: why do they fail and how can we improve them? Supported by empirical evidence, we point out a non-negligible fact that the small models can address the pretext instance discrimination task, and they do not overfit on its training data. However, they universally suffer the problem of over-clustering and therefore yield poor-quality representations. We then experiment with several assumptions that are supposed to solve this problem. Finally, we summarize and combine the practical measures, considerably improving the current baselines for five distinct small models. Our work shortens the gap between the small models and their large counterparts in self-supervised learning, highlighting that training small models without a high-capacity teacher model is a promising direction of research.

7 Acknowledgements

This work has been supported in part by the National Key Research and Development Program of China (2018AAA0101900), Zhejiang NSF (LR21F020004), Chinese Knowledge Center of Engineering Science and Technology (CKCEST).

References

Arora, S.; Khandeparkar, H.; Khodak, M.; Plevrakis, O.; and Saunshi, N. 2019. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*.

Bachman, P.; Hjelm, R. D.; and Buchwalter, W. 2019. Learning representations by maximizing mutual information across views. *arXiv preprint arXiv:1906.00910*.

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, T.; Kornblith, S.; Swersky, K.; Norouzi, M.; and Hinton, G. 2020b. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*.

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

Chen, X.; Xie, S.; and He, K. 2021. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fang, Z.; Wang, J.; Wang, L.; Zhang, L.; Yang, Y.; and Liu, Z. 2021. Seed: Self-supervised distillation for visual representation. *arXiv preprint arXiv:2101.04731*.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, 178–178. IEEE.

Gao, T.; Yao, X.; and Chen, D. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. *arXiv preprint arXiv:2104.08821*.

Gao, Y.; Zhuang, J.-X.; Li, K.; Cheng, H.; Guo, X.; Huang, F.; Ji, R.; and Sun, X. 2021. DisCo: Remedy Self-supervised Learning on Lightweight Models with Distilled Contrastive Learning. *arXiv preprint arXiv:2104.09124*.

Gu, J.; Liu, W.; and Tian, Y. 2021. Simple Distillation Baselines for Improving Small Self-supervised Models. *arXiv preprint arXiv:2106.11304*.

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.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Howard, A.; Sandler, M.; Chu, G.; Chen, L.-C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. 2019. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1314–1324.

Howard, A. G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; and Adam, H. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Huynh, T.; Kornblith, S.; Walter, M. R.; Maire, M.; and Khademi, M. 2020. Boosting Contrastive Self-Supervised Learning with False Negative Cancellation. *arXiv preprint arXiv:2011.11765*.

Koohpayegani, S. A.; Tejankar, A.; and Pirsiavash, H. 2020. Compress: Self-supervised learning by compressing representations. *arXiv preprint arXiv:2010.14713*.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Nakkiran, P.; Kaplun, G.; Bansal, Y.; Yang, T.; Barak, B.; and Sutskever, I. 2019. Deep double descent: Where bigger models and more data hurt. *arXiv preprint arXiv:1912.02292*.

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

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, 6105–6114. PMLR.

Tian, Y.; Krishnan, D.; and Isola, P. 2019. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*.

Tian, Y.; Sun, C.; Poole, B.; Krishnan, D.; Schmid, C.; and Isola, P. 2020. What makes for good views for contrastive learning? *arXiv preprint arXiv:2005.10243*.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, 10347–10357. PMLR.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, F.; and Liu, H. 2021. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2495–2504.

Wang, G.; Wang, K.; Wang, G.; Torr, P. H.; and Lin, L. 2021. Solving Inefficiency of Self-supervised Representation Learning. *arXiv preprint arXiv:2104.08760*.

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.

Wu, M.; Mosse, M.; Zhuang, C.; Yamins, D.; and Goodman, N. 2020. Conditional Negative Sampling for Contrastive Learning of Visual Representations. *arXiv preprint arXiv:2010.02037*.

Wu, Z.; Xiong, Y.; Yu, S.; and Lin, D. 2018. Unsupervised feature learning via non-parametric instance-level discrimination. *arXiv preprint arXiv:1805.01978*.

Xu, H.; Fang, J.; Zhang, X.; Xie, L.; Wang, X.; Dai, W.; Xiong, H.; and Tian, Q. 2021. Bag of Instances Aggregation Boosts Self-supervised Learning. *arXiv preprint arXiv:2107.01691*.

Ye, M.; Zhang, X.; Yuen, P. C.; and Chang, S.-F. 2019. Unsupervised embedding learning via invariant and spreading instance feature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6210–6219.

Zhang, R.; Isola, P.; and Efros, A. A. 2016. Colorful image colorization. In *European conference on computer vision*, 649–666. Springer.