SEED: SELF-SUPERVISED DISTILLATION FOR VISUAL REPRESENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper is concerned with self-supervised learning for small models. The problem is motivated by our empirical studies that while the widely used contrastive self-supervised learning method has shown great progress on large model training, it does not work well for small models. To address this problem, we propose a new learning paradigm, named **SE**lf-Sup**E**rvised **D**istillation (SEED), where we leverage a larger network (as Teacher) to transfer its representational knowledge into a smaller architecture (as Student) in a self-supervised fashion. Instead of directly learning from unlabeled data, we train a student encoder to mimic the similarity score distribution inferred by a teacher over a set of instances. We show that SEED dramatically boosts the performance of small networks on downstream tasks. Compared with self-supervised baselines, SEED improves the top-1 accuracy from 42.2% to 67.6% on EfficientNet-B0 and from 36.3% to 68.2% on MobileNet-v3-Large on the ImageNet-1k dataset.

1 INTRODUCTION

"Tell me and I forget, teach me and I may remember, involve me and I learn."

– Benjamin Franklin

The burgeoning studies and success on self-supervised learning (*SSL*) for visual representation are mainly marked by its extraordinary potency of learning from unlabeled data at scale. Accompanying with the *SSL* is its phenomenal benefit of obtaining task-agnostic representations while allowing the training to dispense with prohibitively expensive data labeling. Major ramifications of visual *SSL* include pretext tasks (Noroozi & Favaro, 2016; Zhang et al., 2016; Gidaris et al., 2018; Zhang et al., 2019; Feng et al., 2019), contrastive representation learning (Wu et al., 2018; He et al., 2020; Chen et al., 2020a), online/offline clustering (Yang et al., 2016; Caron et al., 2018; Li et al., 2020; Caron et al., 2020; Grill et al., 2020), etc. Among them, several recent works (He et al., 2020; Chen et al., 2020a; Caron et al., 2020) have achieved comparable or even better accuracy than the supervised pre-training when transferring to downstream tasks, *e.g.* semi-supervised classification, object detection.

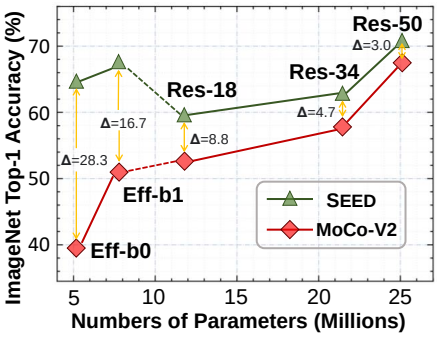

Figure 1: SEED vs. MoCo-v2 (Chen et al., 2020c)) on ImageNet-1K linear probe accuracy. The vertical axis is the top-1 accuracy and the horizontal axis is the number of learnable parameters for different network architectures. Directly applying self-supervised contrastive learning (MoCo-v2) does not work well for smaller architectures, while our method (SEED) leads to dramatic performance boost. Details of the setting can be found in Section 4.

The aforementioned top-performing *SSL* algorithms all involve large networks (*e.g.*, ResNet-50 (He et al., 2016) or larger), with, however, little attention on small networks. Empirically, we find that existing techniques like contrastive learning do not work well on small networks. For instance, the linear probe top-1 accuracy on ImageNet using MoCo-v2 (Chen et al., 2020c) is only 42.2% with MobileNet-v3 (large) (see Figure 1), which is much lower compared with its supervised training accuracy 75.2% (Howard et al., 2019). For EfficientNet-B0, the accuracy is 39.1% compared with its supervised training accuracy 77.1% (Tan & Le, 2019). We conjecture

that this is because smaller models with fewer parameters cannot effectively learn discriminative representation with large amount of data.

To address this challenge, we inject knowledge distillation (KD) (Buciluǎ et al., 2006; Hinton et al., 2015) into self-supervised learning and propose self-supervised distillation (dubbed as SEED) as a new learning paradigm. That is, train the larger, and distill to the smaller both in self-supervised manner. Instead of directly conducting self-supervised training on a smaller model, SEED first trains a large model (as the teacher) in a self-supervised way, and then distills the knowledge to the smaller model (as the student). Note that the conventional distillation is for supervised learning, while the distillation here is in the self-supervised setting without any labeled data. Supervised distillation can be formulated as training a student to mimic the probability mass function over classes predicted by a teacher model. In unsupervised knowledge distillation setting, however, the distribution over classes is not directly attainable. Therefore, we propose a simple yet effective self-supervised distillation method. Similar to the contrastive learning approach, we maintain a queue of negative instances. Given an instance, we first use the teacher network to obtain its similarity scores with all the instances in the queue as well as the instance itself. Then the student encoder is trained to mimic the similarity distribution inferred by the teacher over these instances.

The simplicity and flexibility that SEED brings are self-evident. 1) It does not require any clustering/prototypical computing procedure to retrieve the pseudo-labels or latent classes. 2) The teacher model can be pre-trained with any advanced *SSL* approach, *e.g.* MoCo-V2 (Chen et al., 2020c), SimCLR (Chen et al., 2020a), SWAV (Caron et al., 2020). 3) The knowledge can be distilled to any target small networks (either shallower, thiner, or totally different architectures).

To demonstrate the effectiveness, we comprehensively evaluate the learned representations on series of downstream tasks, *e.g.*, fully/semi-supervised classification, object detection, and also asses the transferability on other domains. For example, on ImageNet-1k dataset, SEED improves the linear probe accuracy of Efficientnet-B0 from 42.2% to 67.6% (a gain over 25%), and MobileNet-v3 from 36.3% to 68.2% (a gain over 31%), as shown in Figure 1 and Section 4.

Our contributions can be summarized as follows:

- We are the first to address the problem of self-supervised visual representation learning for small models.
- We propose a self-supervised distillation (SEED) technique to transfer knowledge from a large model to a small model without any labelled data.
- With the proposed distillation technique (SEED), we significantly improve the state-of-the-art *SSL* performance on small models.
- We exhaustively compare a variety of distillation strategies to show the validity of SEED under multiple settings.

2 Related Work

Among the recent literature in self-supervised learning, contrastive based approach stands out as its learned representations show prominent results on downstream tasks. Majority of the techniques along this direction are stemming from noise-contrastive estimation (Gutmann & Hyvärinen, 2010) where the latent distribution is estimated by contrasting with randomly or artificially generated noises. Variations of this idea have been successfully applied to series of tasks, *e.g.*, language embedding (Mnih & Kavukcuoglu, 2013), audio classification (Oord et al., 2018), face recognition (Sun et al., 2014; Zhang et al., 2017), and so forth. Oord et al. (2018) first proposed Info-NCE to learn image representations by predicting the future using an auto-regressive model for unsupervised learning. Follow-up works include improving the efficiency (Hénaff et al., 2019), and using multi-view as positive samples (Tian et al., 2019b). As these approaches can only have the access to limited negative instances, Wu et al. (2018) designed a memory-bank to store the previously seen random representations as negative samples, and treat each of them as independent categories (instance discrimination). However, this approach also comes with a delicacy that the previously stored vectors are inconsistent with the recently computed representations during the earlier stage of pre-training. Chen et al. (2020a) mitigate this issue by sampling negative samples from a large batch with implementations on TPU. Concurrently, He et al. (2020) improve the memory-bank based method and propose to use the

momentum updated encoder for the remission of representation inconsistency. Other techniques include Misra & Maaten (2020) that combines the pretext-invariant objective loss with contrastive learning, and Wang & Isola (2020) that decomposes contrastive loss into alignment and uniformity objectiveness.

Knowledge distillation (Hinton et al., 2015) aims to transfer knowledge from a cumbersome model to a smaller one without losing too much generalization power, which is also well investigated in model compression (Buciluǎ et al., 2006). Instead of mimicking the teacher's output logit, attention transfer (Zagoruyko & Komodakis, 2016) formulates knowledge distillation on attention maps. Similarly, works in (Ahn et al., 2019; Yim et al., 2017; Koratana et al., 2019; Huang & Wang, 2017) have utilized different learning objectives including consistency on feature maps, consistency on probability mass function, and maximizing the mutual information. CRD (Tian et al., 2019a), which is derived from CMC (Tian et al., 2019b), optimizes the student network by a similar objective to Tian et al. (2019b) using a derived lower bound on mutual information. However, the aforementioned efforts all focus on task-specific distillation (*e.g.*, image classification) during the fine-tuning phase rather than a task-agnostic distillation in the pre-training phase. Several works on natural language pre-training, such as DistillBert (Sanh et al., 2019), TinyBert (Jiao et al., 2019), and MobileBert (Sun et al., 2020), have used knowledge distillation for model compression and shown their validity on multiple downstream tasks. Similar works also emphasize the value of smaller and faster models for representation learning by leveraging knowledge distillation (Turc et al., 2019; Sun et al., 2019). SEED closely relates to the above techniques but aims to facilitate **visual representation learning during pre-training phase for small models**, which as far as we know has not been investigated.

3 METHOD

3.1 KNOWLEDGE DISTILLATION IN SUPERVISED CLASSIFICATION

Traditional knowledge distillation is formulated as the process of training a student f_θ^S to mimic the output class-probabilities predicted by a teacher f_θ^T. A commonly used loss function is the cross-entropy loss for classification tasks (Hinton et al., 2015):

$$\hat{\theta}_S = \arg\min_{\theta_S} \sum_i^N \underbrace{-y^T(\mathbf{x}_i;\theta_T)\cdot \log y^S(\mathbf{x}_i;\theta_S)}_{\mathcal{L}_{Distill}} \underbrace{-y\cdot \log y^S(\mathbf{x}_i;\theta_S)}_{\mathcal{L}_{\text{CE}}}, \tag{1}$$

where $y(\mathbf{x};\theta_T)$ denotes the model output for sample $\mathbf{x}$, y denotes the ground-truth label, and θ_T and θ_S are the model parameters of the teacher and the student, respectively. Other variants of $\mathcal{L}_{Distill}$ have also been proposed. Romero et al. (2014) use *l2* distance as an intermediate representation. CRD (Tian et al., 2019a) proposes to adopt contrastive learning to facilitate supervised distillation. Yim et al. (2017) define distillation in terms of flow between layers. Until now, most of these works focus on supervised distillation that requires annotations either for a teacher model's pre-training or distillation.

3.2 SELF-SUPERVISED DISTILLATION FOR VISUAL REPRESENTATION

Different from supervised distillation, SEED aims to transfer knowledge from a large model to a small model without requiring labelled data. Since there are no labels, $\mathcal{L}_{Distill}$ is not applicable. Our idea is to inject knowledge distillation into contrastive learning. Similar to the contrastive learning framework[], we also maintain a queue of instances. Given a new sample, we compute its similarity scores with all the instances in the queue using both the teacher and the student models. We require that the similarity score distribution of the student matches with that of the teacher, which is formulated as minimizing the cross entropy of the student and the teacher's similarity score distributions. Specifically, for a randomly augmented view $\mathbf{x}_i$ of image $\mathbf{I}_i$, it is first mapped and normalized into feature vector representations $\mathbf{z}_i^T = f_\theta^T(\mathbf{x}_i)/||f_\theta^T(\mathbf{x}_i)||_2$, and $\mathbf{z}_i^S = f_\theta^S(\mathbf{x}_i)/||f_\theta^S(\mathbf{x}_i)||_2$, where $\mathbf{z}_i^T, \mathbf{z}_i^S \in \mathbb{R}^D$, and f_θ^T and f_θ^S denote the teacher and student encoders, respectively. Let $\mathbf{D} = [\mathbf{d}_1 ... \mathbf{d}_K]$ denote the instance queue where K is the queue length and $\mathbf{d}_j$ is the feature vector obtained from the teacher encoder. Similar to the constrastive learning framework, $\mathbf{D}$ is progressively updated under the "*first-in first-out*" strategy as distillation proceeds. That is, we

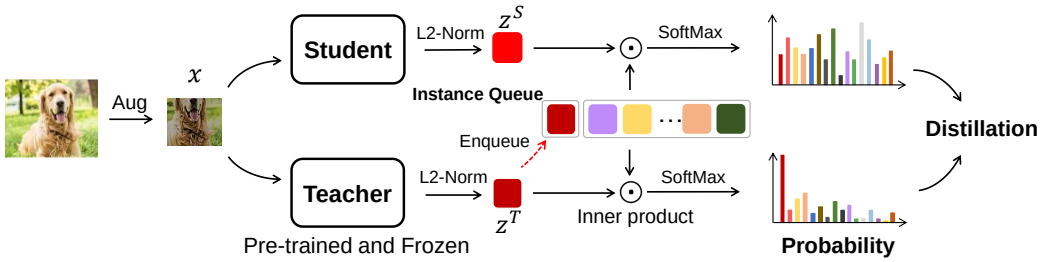

Figure 2: Illustration of our distillation pipeline. The student encoder is trained by minimizing the cross-entropy of probabilities from teacher & student for an identical augmented view of an training image, computed with a dynamically maintained queue. The teacher encoder is pre-trained by *SSL* and kept frozen during the distillation.

de-queue the earliest seen samples and en-queue the visual features of the current batch inferred by the teacher.

Let $\mathbf{p}^T(\mathbf{x}_i;\theta_T;\mathbf{D})$ denotes the similarity scores between $\mathbf{x}_i$ and $\mathbf{d}_j$'s ($j=1,...,K$) computed by the teacher model. $\mathbf{p}^T(\mathbf{x}_i;\theta_T;\mathbf{D})$ is defined as

$$\mathbf{p}^T(x_i;\theta_T,\mathbf{D}) = \left[p_1^T \dots p_K^T\right], \qquad \text{where } p_j^T = \frac{\exp(\mathbf{z}_i^T \cdot \mathbf{d}_j/\tau^T)}{\sum_{\mathbf{d}_j\sim\mathbf{D}}\exp(\mathbf{z}_i^T \cdot \mathbf{d}_j/\tau^T)}, \tag{2}$$

and τ^T is a temperature parameter for the teacher.

Similarly let $\mathbf{p}^S(x_i;\theta_S,\mathbf{D})$ denotes the similarity scores computed by the student model, which is defined as

$$\mathbf{p}^S(x_i;\theta_S,\mathbf{D}) = \left[p_1^S \dots p_K^S\right], \qquad \text{where } p_j^S = \frac{\exp(\mathbf{z}_i^S \cdot \mathbf{d}_j/\tau^S)}{\sum_{\mathbf{d}_j\sim\mathbf{D}}\exp(\mathbf{z}_i^S \cdot \mathbf{d}_j/\tau^S)}, \tag{3}$$

and τ^S is a temperature parameter for the student.

Our self-supervised distillation can be formulated as minimizing the cross entropy between the similarity scores of the teacher and the student over all the instances $\mathbf{x}_i$, that is,

$$\mathcal{L}_{SEED} = -\sum_i^N \mathbf{p}^T(\mathbf{x}_i;\theta_T,\mathbf{D}) \cdot \log \mathbf{p}^S(\mathbf{x}_i;\theta_S,\mathbf{D}), \tag{4}$$

Note that the maintained samples in queue $\mathbf{D}$ are mostly irrelevant to the target instance $\mathbf{x}_i$. For example, 4 softly contrasts $\mathbf{x}_i$ with randomly selected samples without directly aligning with the teacher encoder. To address this problem, we add the teacher's embedding ($\mathbf{z}_i^T$) into the queue and form $\mathbf{D}^+ = [\mathbf{z}_i^T, \mathbf{d}_1 ... \mathbf{d}_K]$. Our student encoder is then trained by minimizing the cross entropy between $\mathbf{p}^T(\mathbf{x}_i;\theta_T,\mathbf{D}^+)$ and $\mathbf{p}^S(\mathbf{x}_i;\theta_S,\mathbf{D}^+)$, that is,

$$\begin{aligned}\hat{\theta}_S &= \arg\min_{\theta_S} \sum_i^N -\mathbf{p}^T(\mathbf{x}_i;\theta_T,\mathbf{D}^+) \cdot \log \mathbf{p}^S(\mathbf{x}_i;\theta_S,\mathbf{D}^+) \\ &= \arg\min_{\theta_S} \sum_i^N \sum_j^{K+1} -\frac{\exp(\mathbf{z}_i^T \cdot \mathbf{d}_j/\tau^T)}{\sum_{\mathbf{d}\sim\mathbf{D}^+}\exp(\mathbf{z}_i^S \cdot \mathbf{d}/\tau^T)} \cdot \log \frac{\exp(\mathbf{z}_i^S \cdot \mathbf{d}_j/\tau^S)}{\sum_{\mathbf{d}\sim\mathbf{D}^+}\exp(\mathbf{z}_i^S \cdot \mathbf{d}/\tau^S)}.\end{aligned} \tag{5}$$

Relations with other losses. Our distillation for the student encoder is composed of two objectives: aligning with the embedding computed by the teacher and softly contrasting with samples maintained in the queue. As the cosine similarity for the teacher representation ($\mathbf{z}_i^T$) in $\mathbf{D}^+$ remains constant ($\mathbb{1}$), the weight for the alignment term remains relatively high and can be consolidated by using the temperature τ^T. Specifically, when $\tau^T \to 0$, the softmax computation for $\mathbf{p}^T$ smoothly approaches one-hot, yielding a similar form with *Info-NCE* loss (Oord et al., 2018) which is widely used in contrastive-based *SSL* (see discussion in Appendix):

$$\mathcal{L}_{NCE} = \sum_i^N -\log \frac{\exp(\mathbf{z}_i^T \cdot \mathbf{z}_i^S/\tau)}{\sum_{\mathbf{d}\sim\mathbf{D}^+}\exp(\mathbf{z}_i^S \cdot \mathbf{d}/\tau)} \tag{6}$$

4 EXPERIMENT

4.1 IMPLEMENTATIONS AND SETTINGS

Self-Supervised Pre-training of Teacher Network. We use the adapted version of MoCo (Chen et al., 2020c), MoCo-v2, as our self-supervised pre-training method for the pre-training of teacher architecture. We also show results with several other self-supervised pre-trained methods as the teacher of distillation in ablations, but yields similar conclusions. Following (Chen et al., 2020a), we use ResNet as the network architecture with different depths/widths (*e.g.*, 50, 101 and 152 layers and $\times 1$, $\times 2$ parameters in ResBlock). MoCo-v2 has a multi-layer-perception layer at the end of the encoder after the average pooling, which contains two linear layer and one *ReLU* (Nair & Hinton, 2010) activation layer. We use the produced 128d vector from the *MLP* as our image representation $\mathbf{z}$ for both pre-training and distillation. The teacher networks are pre-trained for 200 epochs with 65,356 negative samples in queue on ImageNet ILSVRC-2012 dataset (Deng et al., 2009). Due to the computational limit, we pre-train our teacher network as well as the distillation process for 200 epochs unless explicitly note.

Unsupervised Distillation on Student Network. In order to compare the effect of distillation on various networks, we chose multiple smaller networks with less learnable parameters as the distillation target: MobileNet-v3-Large (Howard et al., 2017), EfficientNet-b0 (Tan & Le, 2019), and smaller residual neural networks with less layers (ResNet-18, 34 and 50). Similar to MoCo-v2, we add one additional linear layer in the last MLP block on the basis of primitive design of EfficientNet and MobileNet. We use a standard SGD optimizer with momentum 0.9 and a weight decay parameter of 1e-4 for 200 epochs. The initial learning rate is set as 0.03 and updated by a cosine decay scheduler (Loshchilov & Hutter, 2016) with 5 warm-up epochs. The teacher temperature is set at a smaller value $\tau^T = 0.01$ than the student temperature, $\tau^S = 0.2$. During the distillation, we maintain a standard queue with the length to be 65,536 as MoCo. In terms of the training cost, it takes approximately 40 hours for the distillation of efficientNet-b0 from ResNet-50 with 8$\times$NVIDIA V100 GPUs on ImageNet without any special accelerating measure. We further discuss effect of different hyper-parameters in ablations.

4.2 SCHEME FOR VALIDITY

In order to validate the effectiveness of self-supervised distillation, we choose to asses the performance of representations of the student encoder on several downstream tasks. We first report its performances of linear evaluation and semi-supervised linear evaluation on ImageNet ILSVRC-2012 (Deng et al., 2009) dataset. To measure the feature transferability brought by distillation, we conduct evaluations on other tasks, which includes object detection and segmentation on VOC07 (Everingham et al.) and MS-COCO (Lin et al., 2014) datasets. At the end, we compare the transferability of the features learned by distillation and ordinary self-supervised contrastive learning on the tasks of linear classification on datasets from different domains.

Linear and *K*NN Evaluation on ImageNet. We conduct the supervised linear classification on ImageNet-1M, which contains$\sim$1.3M images for training, and 50,000 images for testing, spanning 1,000 categories. Following previous works in (He et al., 2020; Chen et al., 2020a), we train a single linear layer classifier on the top of frozen network encoder after self-supervised pre-training. SGD optimizer is used to train the linear classifier for 100 epochs with weight decay to be 0. The initial learning rate is set at 30 and is reduced by a factor of 10 at 60 and 80 epochs (similar as Tian et al. (2019a)). Notably, when training the linear classifier for MobileNet and EfficientNet, we reduce the initial learning rate to 3. The results are reported with top-1 and top-5 accuracy. We also perform classification using *K*-Nearest Neighbors (*K*NN) based on the learned 128d vector from the last MLP layer. The sample is classified by taking the most frequent label of its K nearest neighbors, where we set $K = 10$ consistently for all entries in experiment.

Semi-Supervised Evaluation on ImageNet. We then further evaluate the learned representations on the classification task using only a subset of ImageNet. We follow the semi-supervised learning protocol as in (Oord et al., 2018; Kornblith et al., 2019; Kolesnikov et al., 2019), where the fixed 1% and 10% subsets of ImageNet labeled training data (provided by Chen et al. (2020a)) is utilized for linear classification training. We report the Top-1 and Top-5 accuracy on the testing split.

S / T	Eff-b0			Eff-b1			Mob-v3			R-18			R-34		
	K	T-1	T-5	*K*	T-1	T-5	*K*	T-1	T-5	*K*	T-1	T-5	*K*	T-1	T-5
✗	30.0	42.2	68.5	34.4	50.7	74.6	27.5	36.3	62.2	36.7	52.5	77.0	41.5	57.4	81.6
R-50	46.0	61.3	82.7	44.6	58.4	80.3	44.8	55.2	80.3	43.4	57.6	81.8	45.2	58.5	82.6
Δ	+16.0	+19.1	+14.2	+10.2	+7.7	+5.7	+17.3	+18.9	+18.1	+6.7	+5.1	+4.8	+3.7	+1.1	+1.0
R-101	50.1	63.0	83.8	50.3	61.2	84.8	48.8	59.9	83.5	48.6	58.9	82.5	50.5	61.6	84.9
Δ	+20.1	+20.8	+15.3	+5.7	+10.5	+10.2	+21 .3	+23.6	+21.3	+11.9	+6.4	+5.5	+9.0	+4.2	+3.3
R-152	50.7	65.3	86.0	51.2	64.6	85.7	49.5	61.4	84.6	49.1	59.5	83.3	51.4	62.7	85.8
Δ	+20.7	+23.1	+17.5	+16.8	+13.9	+11.1	+22.0	+25.1	+22.4	+12.4	+7.0	+6.3	+9.9	+5.3	+4.2
R50$_{\times 2}$*	57.4	67.6	87.4	60.3	67.8	87.6	55.9	68.2	88.2	55.3	63.0	84.9	58.2	65.7	86.8
Δ	+27.4	+25.4	+28.4	+25.9	+17.1	+13.0	+18.9	+31.9	+26.0	+18.6	+10.5	+7.9	+16.7	+8.3	+5.2

Table 1: ImageNet-1k test *accuracy* (%) under *K*NN and linear classification across multiple students and *deeper*, MoCo-v2 pre-trained teacher architectures. ✗ denotes MoCo-V2 self-supervised learning baselines before distillation. * indicates using a stronger teacher encoder pre-trained by SWAV with additional small-patches during distillation. *K*, T-1 and T-5 denote Top-1 accuracy for *K*NN and Top-1/5 accuracy for linear evaluation.

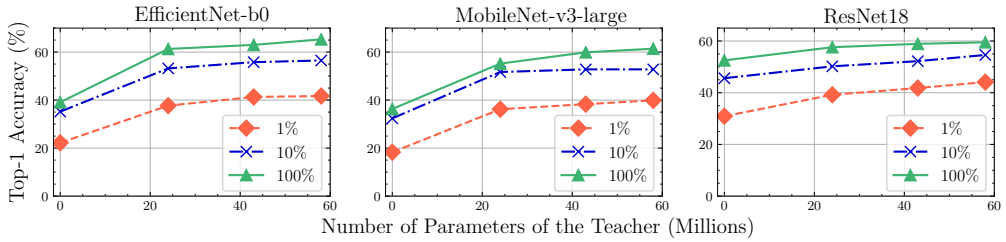

Figure 3: ImageNet-1k Top-1 test *accuracy* for semi-supervised evaluations using 1% 10% label fractions w/o distillation under different Teachers.

Table 1 summarizes the performances of representations from various networks by SEED distillation under *K*NN and linear evaluation on ImageNet. We list the baseline of contrastive self-supervised pre-training using MoCo-v2 (Chen et al., 2020c) in the first row per each student architecture. We can see clearly that smaller networks tend to perform rather worse, that MobileNet-v3 can only reach 36.3% using contrastive pre-training with ∼5M learnable parameters. This aligns well with previous conclusions from (Chen et al., 2020a;b), that the bigger models are more likely to perform better in contrastive based self-supervised pre-training. We conjecture that this is mainly caused by the inability of smaller network handling large-scale dataset. The results clearly demonstrate that the distillation from a larger network help boosting the performances on all tasks, and show obvious improvement compared with ordinary contrastive self-supervised pre-training. Concretely, by leveraging a larger network like ResNet-152, EfficientNet-b0 (with only∼5M learnable parameters) achieves 65.3%@Top-1 Acc., leading MoCo-v2 using ResNet-34 (4×larger than EfficientNet-b0) by a large margin of 8.0%, and even approaching the performances of MoCo-v2 using ResNet-50 on the linear evaluation task. This is also consistent with the results on *K*NN and semi-supervised evaluating protocols (see Figure 3). We list results of distillation from a stronger *SSL* pre-trained ResNet-50×2 model with additional small patches utilized and trained for 800 epochs (see last row of Table 1), where EfficientNet-b0 further reaches 67.6% Top-1 Acc. on ImageNet-1k. We also report the distillation results using ResNet-50 as larger student encoder (see Appendix), and observe same improvement trend: it reaches 74.3% Top-1 Acc. with a ResNet-50×2 as Teacher trained for 800 epochs. We note that the gain benefited from distillation becomes more distinct on smaller architectures and we further study the effect of various distilling sources in ablations.

Comparisons with Different Teachers. We compare the performances of distillation from different teachers. In particular, Figure 4 summarizes the testing accuracy of ResNet-18 and EfficientNet-b0 distilling from wider ResNet architectures (×1, ×2 parameters in ResBlock) when no small patches are used. We see clear performance improvement as depth and width increase: comparing to ResNet-50, deeper (ResNet-101) and wider (ResNet-50×2) substantially improve the testing accuracy. However, further architectural enlargement has relatively limited effects: ResNet-152×2 does not obviously affect the performances. We additionally show the distillation results of a ResNet-18 from various methods of self-supervised ResNet-50 in Table 2. To be specific, we compare the distillation results of MoCo-v1 (He et al., 2020), MoCo-v2 (Chen et al., 2020c) trained 200 and 800 epochs, SimCLR (Chen et al., 2020a), and SWAV (Caron et al., 2020). We observe that both longer Teacher *SSL* pre-training and distillation epochs can yield beneficial effects. Notably, the aforementioned

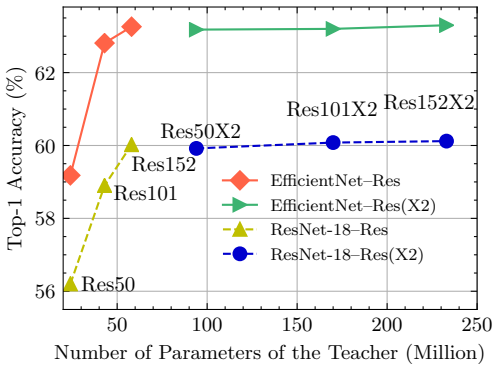

Figure 4: *Accuracy* (%) of student networks (EfficientNet-b0 and ResNet-18) on ImageNet distilled from *wider* MoCo-v2 pre-trained ResNet (ResNet-50/101/152×2).

Teacher	P-E	D-E	T. Top-1	S. Top-1	S. Top-5
✗	✗	✗	✗	52.5	77.0
MoCo	200	200	60.6	52.1	77.0
SimCLR	200	200	65.6	57.5	81.7
MoCo-v2	200	200	67.4	57.6	81.8
	800	200	71.1	60.5	83.5
SWAV	800	100	75.3	61.1	83.8
	800	200	75.3	61.7	84.2
	800	400	75.3	62.0	84.4
SWAV*	800	200	75.3	**62.6**	**84.8**

Table 2: ImageNet-1k test *Accuracy* (%) of student network (ResNet-18) distilled from variants of self-supervised ResNet-50. P-E/D-E represent pre-training and the distillation epochs. T./S.-Top represent testing accuracy of Teacher and Student. * represents distillation using additional small patches. First row is the ResNet-18 *SSL* baseline using MoCo-v2 trained for 200 epochs.

S	T	VOC Obj. Det.			COCO Obj. Det.			COCO Inst. Segm.		
		AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
	✗	46.1	74.5	48.6	35.2	54.1	37.9	31.2	51.3	33.3
R-18	**R-50**	46.1(0.0)	74.8(**+0.3**)	49.1(**+0.5**)	35.5(**+0.3**)	54.3(+0.2)	37.9(0.0)	31.4(+0.2)	51.3(0.0)	33.4(+0.1)
	R-101	46.8(**+0.7**)	75.8(**+1.3**)	49.3(**+0.7**)	35.4(+0.2)	54.2(+0.1)	38.0(+0.1)	31.3(+0.1)	51.4(+0.1)	33.5(+0.2)
	R-152	46.8(**+0.7**)	75.9(**+1.4**)	50.2(**+1.6**)	35.4(+0.2)	54.3(+0.2)	38.0(+0.1)	31.4(+0.2)	51.4(+0.1)	33.5(+0.2)

Table 3: Object detection and instance segmentation results using contrastive self-supervised learning and SSD distillation using ResNet-18 as backbone: bounding-box AP (AP^{bb}) and mask AP (AP^{mk}) evaluated on VOC07-val and COCO testing split. More results on different backbones can be found in the Appendix. Subscript in green represents improvement is larger than 0.3.

methods all unanimously adopt contrastive based pre-training except SWAV, which is based upon online clustering. We find that our SEED is pre-training agnostic that the distillation can be indeed conducted on clustering based self-supervised model.

Transferring to Other Tasks. In order to comprehensively assess the transferability of the representations from SEED, we test it on different downstream tasks, *e.g.*, object detection and segmentation. Following He et al. (2020), we fine-tune all the layers of a Faster R-CNN (Ren et al., 2015) with C4-backbone on the VOC-07+12 train+val set, and evaluate it on VOC-07 test split. Specifically, the backbone in the detector contains only the convolutional layers from ResNet ends at conv4 stage, and the mask/box prediction head consists of the conv5 stage (with global pooling layer). Similarly, we use Mask R-CNN (He et al., 2017) with the C4 backbone for COCO segmentation. During training, we tune the Batch-Normalization layer as (He et al., 2020). In Table 3, we exhibit both the detection and segmentation results evaluated using AP_{50} and AP_{75} metric (threshold of IoU = 50, 75) and bounding box AP (AP^{bb}), mask AP (AP^{mk}). By distilling from the larger architecture, SEED improves the performances on the detection task: **+0.7** on AP, **+1.4** on AP_{50} and **+1.6** on AP_{75} than the MoCo-v2 baseline. Consistently, we observe same trend in segmentation results. Comparisons of performances using different backbones and the tine-tuning details can be found in the Appendix.

Additional Classification Results on Other Domains. To further study whether the improvement of learned representation of distillation is confined to ImageNet dataset, we evaluate on additional classification datasets to study the generalization and transferability of the features. We strictly follow the linear evaluation and fine-tuning settings from (Kornblith et al., 2019; Chen et al., 2020a; Grill et al., 2020), that a linear layer is trained on the basis of frozen features. We report Top-1 Accuracy of models before and after distillation from various architectures on CIFAR-10, CIFAR-100 (Krizhevsky et al., 2009), SUN-397 (Xiao et al., 2010) datasets. More details regards the pre-processing and training can be found in Appendix. Notably, we observe that our distillation surpass the contrastive self-supervised pre-training consistently on all benchmarks, verifying the effectiveness of SEED. This also proves the representations from distillation is generically beneficial to a wide-span of downstream tasks.

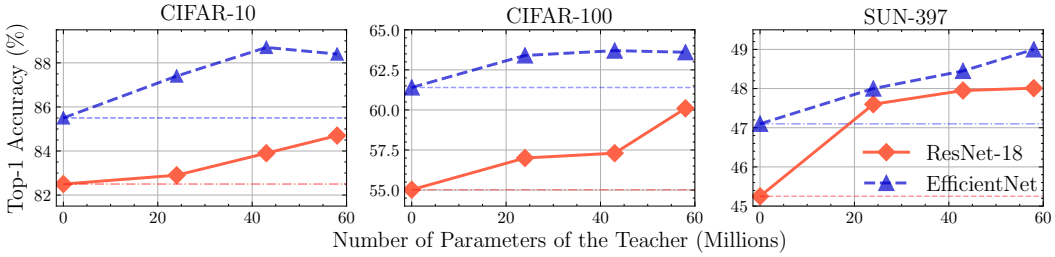

Figure 5: ImageNet-1k test *Accuracy* (%) of student network (EfficientNet-b0 and ResNet-18) transferred to other domains (CIFAR-10, CIFAR-100, SUN-397 datasets) w/o distillation from lager architectures (ResNet-50/101/152).

Method	Top-1 Acc.	Top-5 Acc.
L2 Distance	55.3	80.3
K-Means (4k)	51.0	75.8
Online Clustering	56.4	81.2
Binary Contr. Loss	57.1	81.5
SSD	**57.9**	**82.0**

Table 4: Top-1/5 accuracy of linear classification results on ImageNet using different distillation strategies on ResNet-18 (student) and ResNet-50 (teacher) architectures.

τ^T	**ImageNet**		**CIFAR-10**	**CIFAR-100**
	Top-1	Top-5	Top-1	Top-1
0.3	54.8	80.0	78.7	46.6
0.1	54.9	80.1	83.0	50.1
0.05	56.5	81.3	84.4	56.2
0.01	**57.9**	**82.0**	**87.5**	60.6
1e-3	57.6	81.8	86.9	**60.8**

Table 5: Effect of τ^T for the distillation of ResNet-18 (student), ResNet-50 (teacher) on multiple datasets.

Ablations on Strategies of Distillation and Hyper-Parameters. We finally compare our SEED, with other several distillation strategies, which include minimizing the *l2*-distance of the embedding (Romero et al., 2014), pseudo-label classification using *K*-Means Clustering (Li et al., 2017), online clustering (Snell et al., 2017), and the contrastive-akin distillation method CRD (Tian et al., 2019a) in Table 4. It's worth noting that, by simply minimizing the *l2*-distance of the student as teacher's visual embedding yields a decent performance: 55.3% on ImageNet test split using ResNet-18 as the student and a pre-trained ResNet-50 as the teacher. This observation aligns with the conclusion in (Grill et al., 2020), where the visual representation is trained by learning to mimic the representations of just augmented views of positive samples. *K*-Means Clustering constructs pseudo-labels as cross-entropy targets by clustering features from the teacher encoder in an offline manner. During practice, we find that when a smaller number of *K* is set, the assigned categorical center for a sample can be less aligned: samples assigned with identical label are discriminating with each others. Thus, in contrary to *l2*-distance, *K*-Means clustering innately lays particular emphasis on discriminating with different samples than the alignment of different views from identical images. In particular, we note that CDR is indeed a variation of contrastive distillation by minimizing an upper-bound of Info-NCE objectiveness (Oord et al., 2018), but was designed to facilitate the supervised distillation. We apply it on our self-supervised distillation task as a binary form of Info-NCE loss, therefore it is expected to produce a close result with SEED. Table 5 summarizes the distillation performances on multiple datasets under different temperature τ^T when τ^S is fixed at 0.2. We observe a better performance with the decreasing of τ^T to 0.01 for ImageNet-1k, CIFAR-10 and CIFAR-100 datasets. We further discuss the detailed configurations for each baseline in the Appendix.

5 CONCLUSIONS

Contrastive based Self-Supervised Learning is established upon instance discrimination, while a critical impedance for the pre-training on smaller architecture comes from its incapacity of dealing enormous number of instances. Instead of directly learn from the un-labeled data, we propose SEED, that learns its representation by just distillation from a bigger, and self-supervised model. We show that as a novel self-supervised learning paradigm, SEED achieves state-of-the-art results on various benchmarks of small architectures.

REFERENCES

Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9163–9171, 2019.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 535–541, 2006.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 132–149, 2018.

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

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

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

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

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

Zeyu Feng, Chang Xu, and Dacheng Tao. Self-supervised representation learning by rotation feature decoupling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10364–10374, 2019.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

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

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 297–304, 2010.

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

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

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

Olivier J Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, SM Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*, 2019.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1314–1324, 2019.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*, 2017.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*, 2019.

Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1920–1929, 2019.

Animesh Koratana, Daniel Kang, Peter Bailis, and Matei Zaharia. Lit: Learned intermediate representation training for model compression. In *International Conference on Machine Learning*, pp. 3509–3518, 2019.

Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2661–2671, 2019.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

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

Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. Learning from noisy labels with distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1910–1918, 2017.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

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

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

Andriy Mnih and Koray Kavukcuoglu. Learning word embeddings efficiently with noise-contrastive estimation. In *Advances in neural information processing systems*, pp. 2265–2273, 2013.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010.

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

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

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

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. 2014.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pp. 4077–4087, 2017.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient knowledge distillation for bert model compression. *arXiv preprint arXiv:1908.09355*, 2019.

Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *Advances in neural information processing systems*, pp. 1988–1996, 2014.

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. Mobilebert: a compact task-agnostic bert for resource-limited devices. *arXiv preprint arXiv:2004.02984*, 2020.

Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, 2019a.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019b.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*, 2019.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. *arXiv preprint arXiv:2005.10242*, 2020.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3733–3742, 2018.

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

Jianwei Yang, Devi Parikh, and Dhruv Batra. Joint unsupervised learning of deep representations and image clusters. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5147–5156, 2016.

Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4133–4141, 2017.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

Liheng Zhang, Guo-Jun Qi, Liqiang Wang, and Jiebo Luo. Aet vs. aed: Unsupervised representation learning by auto-encoding transformations rather than data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2547–2555, 2019.

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

Xiao Zhang, Zhiyuan Fang, Yandong Wen, Zhifeng Li, and Yu Qiao. Range loss for deep face recognition with long-tailed training data. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5409–5418, 2017.