Self-Supervision Can Be a Good
Few-Shot Learner

Yuning Lu[1][*], Liangjian Wen[2], Jianzhuang Liu[2],
Yajing Liu[1], and Xinmei Tian[1,3]

[1]University of Science and Technology of China   [2]Huawei Noah's Ark Lab
[3]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
{lyn0, lyj123}@mail.ustc.edu.cn, xinmei@ustc.edu.cn,
{wenliangjian1, liu.jianzhuang}@huawei.com

Abstract. Existing few-shot learning (FSL) methods rely on training with a large labeled dataset, which prevents them from leveraging abundant unlabeled data. From an information-theoretic perspective, we propose an effective unsupervised FSL method, learning representations with self-supervision. Following the InfoMax principle, our method learns comprehensive representations by capturing the intrinsic structure of the data. Specifically, we maximize the mutual information (MI) of instances and their representations with a low-bias MI estimator to perform self-supervised pre-training. Rather than supervised pre-training focusing on the discriminable features of the seen classes, our self-supervised model has less bias toward the seen classes, resulting in better generalization for unseen classes. We explain that supervised pre-training and self-supervised pre-training are actually maximizing different MI objectives. Extensive experiments are further conducted to analyze their FSL performance with various training settings. Surprisingly, the results show that self-supervised pre-training can outperform supervised pre-training under the appropriate conditions. Compared with state-of-the-art FSL methods, our approach achieves comparable performance on widely used FSL benchmarks without any labels of the base classes.

Keywords: Few-shot image classification, self-supervised learning

1 Introduction

Training a reliable model with limited data, also known as few-shot learning (FSL) [22,43,48,53,59,65,71], remains challenging in computer vision. The core idea of FSL is to learn a prior which can solve unknown downstream tasks. Despite various motivations, most existing methods are *supervised*, requiring a large labeled (base) dataset [61,75] to learn the prior. However, collecting a large-scale base dataset is expensive in practice. Depending on supervision also does not allow the full use of abundant unlabeled data.

---

[*] This work was done during an internship in Huawei Noah's Ark Lab.

Several unsupervised FSL works [3, 35, 37, 38, 57] attempt to solve the problem of *label dependency*. Most of them share a similar motivation of applying existing meta-learning methods (i.e., the popular supervised FSL solutions) to unsupervised data. Instead of leveraging category labels, these approaches generate (meta-)training tasks (or *episodes*) via different unsupervised ways, such as data augmentation [37] or pseudo labels [35]. Despite their worthy attempts, they still have a large performance gap compared with the top supervised FSL methods. Recent work [39] indicates that the episodic training of meta-learning is data-inefficient in that it does not sufficiently exploit the training batch. Several studies [10, 19, 28, 71] of (supervised) FSL also show that a simple pre-training-&-fine-tuning approach outperforms many sophisticated meta-learning methods.

From an information-theoretic perspective, we propose an effective unsupervised FSL method, i.e., learning the representations with self-supervised pre-training. Following the principle of InfoMax [46], the goal of our method is to preserve more information about high-dimensional raw data in the low-dimensional learned representations. In contrast to supervised pre-training [71], self-supervised pre-training focuses on capturing the intrinsic structure of the data. It learns comprehensive representations instead of the most discriminative representations about the base categories. Specifically, our self-supervised pre-training maximizes the mutual information (MI) between the representations of augmented views of the same instance. It is a lower bound of MI between the instance and its representations. Many contrastive learning methods [8, 31, 52] maximize MI by optimizing a loss based on Noise-Contrastive Estimation [29] (also called InfoNCE [52]). However, recent progress [56, 66, 77] shows that the MI estimation based on InfoNCE has *high bias*. We alternatively employ a low-bias MI estimator following the MI neural estimation [4] to address the issue. The experiments in FSL demonstrate the effectiveness of our approach.

To better understand self-supervision and supervision in FSL, we explain that they are maximizing different MI targets. We further construct comprehensive experiments to analyze their different behaviors in FSL across various settings (i.e., backbones, data augmentations, and input sizes). The experiment results surprisingly show that, with appropriate settings, self-supervision *without* any labels of the base dataset can outperform supervision while exhibiting better scalability for network depth. We argue that self-supervision learns less bias toward the base classes than supervision, resulting in better generalization ability for unknown classes. In this manner, extending the network depth can learn more powerful representations without over-fitting to the seen classes.

The scalability of network depth provides an opportunity to use a deep model to guide the learning of a shallow model in FSL. We formulate this problem of unsupervised knowledge distillation as maximizing MI between the representations of different models. Consequently, we propose a simple yet effective loss to perform the knowledge distillation *without* labels. To the best of our knowledge, existing supervised FSL methods [20, 71] only perform the knowledge distillation between shallow models. In summary, our contributions are:

- From an information-theoretic perspective, we propose an effective unsupervised FSL approach that learns representations with self-supervision. Our method maximizes the MI between the instances and their representations with a low-bias MI estimator.
- We indicate that the self-supervised pre-training and supervised pre-training maximize different targets of MI. We construct comprehensive experiments to analyze the difference between them for the FSL problem.
- We present a simple yet effective self-supervised knowledge distillation for unsupervised FSL to improve the performance of a small model.
- Extensive experiments are conducted to demonstrate the advantages of our method. Our *unsupervised* model achieves comparable results with the state-of-the-art *supervised* FSL ones on widely used benchmarks, i.e., *mini*-ImageNet [75] and *tiered*-ImageNet [61], without any labels of the base classes.

2   Related Work

Few-shot learning (FSL). The pioneering works of FSL date back to the Bayesian approach [40, 44]. In recent years, several papers [23, 43, 53, 63, 65, 75] address the problem with a meta-learning paradigm, where the model learns from a series of simulated learning tasks that mimic the real few-shot circumstances. Due to its elegant form and excellent results, it has attracted great interest. However, recent studies [10, 28, 71] show that pre-training an embedding model with the classification loss (cross-entropy) is a simple but tough-to-beat baseline in FSL. Subsequently, many studies [13, 47, 49, 55, 67] focus on how to learn a good embedding instead of designing complex meta-learning strategies. Although considerable progress has been made, the aforementioned approaches rely on the annotation of the base classes, limiting their applications. In addition, most existing supervised methods [10, 22, 43, 51, 53, 65, 71, 75] achieve their best results with a relatively shallow backbone, e.g., ResNet-10/12. Our paper demonstrates that it is possible to build an effective and scalable few-shot learner without any labels of the base classes. It suggests that we should rethink the significance of label information of the base dataset in FSL.

InfoMax principle in FSL. Some recent studies [5, 19] address the problem of transductive FSL, where unlabeled query samples are utilized in the downstream fine-tuning, from the information-theoretic perspective. The most related work [5] introduces the InfoMax principle [46] to perform transductive fine-tuning. It maximizes the MI between the representations of query samples and their predicted labels during fine-tuning, while ours maximizes the MI between base samples and their representations during pre-training.

Self-supervised learning (SSL). A self-supervised model learns representations in an unsupervised manner via various pretext tasks, such as colorization [41, 82], inpainting [54], and rotation prediction [26]. One of the most competitive methods is contrastive learning [8, 30, 31, 34, 52, 69], which aligns the representation of samples from the same instance (the positive pair, e.g., two augmented views of the same image). A major problem of contrastive learning

is the *representation collapse*, i.e., all outputs are a constant. One solution is the uniformity regularization, which encourages different images (the negative pair) to have dissimilar representations. Recent works [8,31] typically optimize the InfoNCE loss [29,52] to perform both alignment and uniformity, which is considered to maximize the MI between different views. Since InfoNCE can be decomposed into alignment and uniformity terms [9,76], many works introduce new forms of uniformity (and/or alignment) to design new objectives. Barlow Twins [81] encourages the representations to be dissimilar for different channels, not for different samples. Chen and Li [9] propose to explicitly match the distribution of representations to a prior distribution of high entropy as a new uniformity term. Some recent works [12,27,72] introduce asymmetry in the alignment of the positive pair to learn meaningful representations without explicit uniformity.

FSL with SSL. In natural language processing, self-supervised pre-training shows superior performance on few-shot learning [7]. However, the application of SSL in the few-shot image classification is still an open problem. Most works [25, 49, 67] leverage the pretext task of SSL as an auxiliary loss to enhance the representation learning of supervised pre-training. The performance of these methods degrades drastically without supervision. Another way is unsupervised FSL [3, 35, 37, 38, 42, 50, 57], whose setting is the same as ours. Most of these works [3, 35, 37, 38, 50, 57] simply adapt existing supervised meta-learning methods to the unsupervised versions. For example, CACTUs [35] uses a clustering method to obtain pseudo-labels of samples and then applies a meta-learning algorithm. Their performance is still limited by the downstream meta-learning methods, having a large gap with the top supervised FSL methods. In addition, the recent work [21] evaluates existing self-supervised methods on a benchmark [28] of cross-domain few-shot image classification, where there is a large domain shift between the data of base and novel classes. Our approach also obtains the state-of-the-art results on this benchmark [28] compared with other self-supervised and supervised methods (see our supplementary materials). Besides, similar works in continuous [24] and open-world learning [18] also employ SSL to enhance their performances, which can relate to FSL since these fields all aim to generalize the learned representations to the novel distribution. Chen *et al.* [16] suggest that, in the *transductive* setting, the existing SSL method (MoCo v2 [11]) can achieve competitive results with supervised FSL methods. However, their transductive FSL method requires the data of test classes for unsupervised pre-training, which is somewhat contrary to the motivation of FSL.

3   Method

3.1   Preliminaries

FSL setup. In few-shot image classification, given a base dataset $\mathcal{D}_{base} = \{(x_i, y_i)\}$, the goal is to learn a pre-trained (or meta-) model that is capable of effectively solving the downstream few-shot task $\mathcal{T}$, which consists of a support set $\mathcal{S} = \{(x_s, y_s)\}_{s=1}^{N*K}$ for adaptation and a query set $\mathcal{Q} = \{x_q\}_{q=1}^{Q}$ for
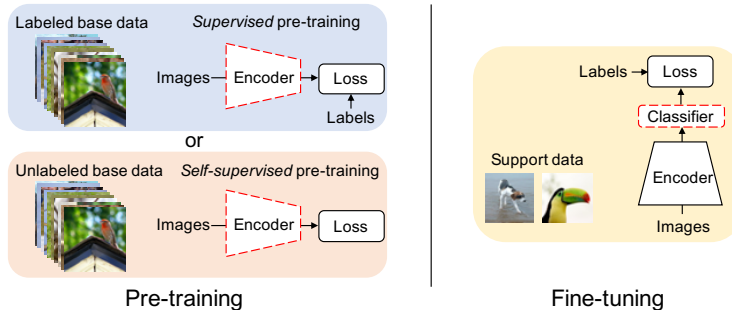
Fig. 1: **The overview of pre-training-&-fine-tuning approach in FSL.**
(**Left**) In the pre-training stage, an encoder network is trained on a labeled (or
unlabeled) base dataset with a supervised (or self-supervised) loss. (**Right**) In
the fine-tuning stage, a linear classifier (e.g., logistic regression) is trained on the
embeddings of a few support samples with the frozen pre-trained encoder.

prediction, where y_s is the class label of image x_s. As an N-way K-shot clas-
sification task $\mathcal{T}$, K is relatively small (e.g., 1 or 5 usually) and the N novel
categories are not in $\mathcal{D}_{base}$.

FSL with supervised pre-training. Recent works [10,71] show that a simple
pre-training-&-fine-tuning approach is a strong baseline for FSL. These methods
pre-train an encoder (e.g., a convolution neural network) on $\mathcal{D}_{base}$ with the stan-
dard classification objective. In downstream FSL tasks, a simple linear classifier
(e.g., logistic regression in our case) is trained on the output features of the fixed
encoder network with the support samples. Finally, the pre-trained encoder with
the adapted classifier is used to infer the query samples (as shown in Fig. 1).

Unsupervised FSL setup. In contrast to supervised FSL where $\mathcal{D}_{base} = \{(x_i, y_i)\}$, only the unlabeled dataset $\mathcal{D}_{base}=\{x_i\}$ is available in the pre-training
(or meta-training) stage for unsupervised FSL. Our self-supervised pre-training
approach follows the standard pre-training-&-fine-tuning strategy discussed above,
except that the base dataset is *unlabeled* (as shown in Fig. 1). Note that, for a
fair comparison, our model is *not* trained on any additional (unlabeled) data.

3.2   Self-Supervised Pre-Training for FSL

**Self-supervised pre-training and supervised pre-training maximize dif-
ferent MI targets.** Supervised pre-training aims to reduce the classification
loss on the base dataset toward zero. A recent study [74] shows that there is
a pervasive phenomenon of *neural collapse* in the supervised training process,
where the representations of within-class samples collapse to the class mean. It
means the conditional entropy $H(Z|Y)$ of hidden representations Z given the
class label Y is small. In fact, Boudiaf *et al.* [6] indicate that minimizing the
cross-entropy loss is equivalent to maximizing the mutual information $I(Z;Y)$
between *representations* Z and *labels* Y. Qin *et al.* [58] also prove a similar result.

Maximizing $I(Z;Y)$ is beneficial for recognition on the base classes. However, since FSL requires the representations generalizing on the novel classes, over-fitting to the base classes affects the performance of FSL. In this paper, following the InfoMax principle [46], our method aims to preserve the raw data information as much as possible in the learned representations. Theoretically, we maximize another MI target, i.e., the mutual information $I(Z;X)$ between *representations* Z and *data* X, to learn meaningful representations for FSL. Comparing the two MI objectives $I(Z;Y)$ and $I(Z;X)$, the supervised representations are only required to contain information about the associated labels of the images. In contrast, the representations with self-supervision are encouraged to contain comprehensive information about the data with less bias toward the base labels.

In practice, the calculation of $I(Z;X)$ is intractable. We maximize an alternative MI objective $I(Z^1;Z^2) = I(f(X^1);f(X^2))$, which is a lower bound of $I(Z;X)$ [73], where X^1 and X^2 are two augmented views of X obtained by some data augmentations, and f is the encoder network. In addition, our encoder $f(\cdot) = h_{proj} \circ g(\cdot)$ consists of a backbone $g(\cdot)$ (e.g., ResNet) and an extra *projection* head $h_{proj}(\cdot)$ (e.g., MLP) following contrastive learning methods [8,11], as shown in Fig. 3a. The projection head is only used in the pre-training stage. In the fine-tuning stage, the linear classifier is trained on the representations before the projection head. Next, we introduce two MI estimators for $I(Z^1, Z^2)$ and describe how to perform self-supervised pre-training with them.

Maximizing $I(Z^1; Z^2)$ with I_{NCE} and I_{MINE}. Many contrastive learning methods [8,52] maximize $I(Z^1;Z^2)$ with the *InfoNCE* estimator proposed in [52]:

$$I(Z^1;Z^2) = I(f(X^1);f(X^2)) \tag{1}$$

$$\geq \mathop{\mathbb{E}}_{p(x^1,x^2)}[C(x^1,x^2)] - \mathop{\mathbb{E}}_{p(x^1)}[\log(\mathop{\mathbb{E}}_{p(x^2)}[e^{C(x^1,x^2)}])] \triangleq I_{NCE}(Z^1;Z^2), \tag{2}$$

where $p(x^1,x^2)$ is the joint distribution (i.e., $(x^1,x^2) \sim p(x^1,x^2)$, and (x^1,x^2) is a positive pair) and the critic $C(x^1,x^2)$ is parameterized by the encoder f, e.g., $C(x^1,x^2) = f^T(x^1)f(x^2)/\tau$ with τ being temperature. Given a training batch $\{x_i\}_{i=1}^{2B}$ where x_i and x_{i+B} are positive pair $(i \leq B)$, the well-known method SimCLR [8] minimizes the contrastive loss[1] based on I_{NCE}:

$$\mathcal{L}_{NCE} = \underbrace{-\frac{1}{B}\sum_{i=1}^{B} z_i^T z_{i+B}/\tau}_{Alignment} + \underbrace{\frac{1}{2B}\sum_{i=1}^{2B}\log(\sum_{j \neq i} e^{z_i^T z_j/\tau})}_{Uniformity}, \tag{3}$$

where $z_i = f(x_i)$. Despite the great success of I_{NCE} in contrastive learning, the problem is that I_{NCE} has high bias, especially when the batch size is small and MI is large. For detailed discussions we refer the reader to [56,66].

---

[1] Alignment: the difference between representation of two views of the same sample should be minimized. Uniformity: the difference between representation of two different samples should be maximized.
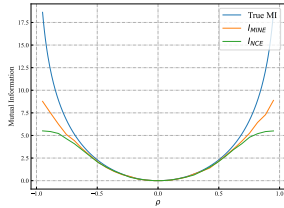
Fig. 2: We estimate MI between two multivariate Gaussians with the component-wise correlation ρ (see the supplementary materials for details). When the true MI is large, I_{NCE} has a high bias compared with I_{MINE}.

Our work employs another MI estimator I_{MINE} following recent progress in the MI neural estimation [4], which has lower bias than I_{NCE} [56,66]:

$$I_{MINE}(Z^1; Z^2) \triangleq \mathbb{E}_{p(x^1,x^2)}[C(x^1, x^2)] - \log(\mathbb{E}_{p(x^1)\otimes p(x^2)}[e^{C(x^1,x^2)}]), \qquad (4)$$

where $p(x^1) \otimes p(x^2)$ is the product of the marginal distributions. We construct a simple experiment on the synthetic data to compare the estimation bias of I_{NCE} and I_{MINE}, as shown in Fig. 2. Based on $I_{MINE}(Z^1; Z^2)$, we can further propose a novel contrastive loss for self-supervised pre-training:

$$\mathcal{L}_{MINE} = \underbrace{-\frac{1}{B}\sum_{i=1}^{B} z_i^T z_{i+B}/\tau}_{Alignment} + \underbrace{\log(\sum_{i=1}^{2B}\sum_{z_j \in Neg(z_i)} e^{z_i^T z_j/\tau})}_{Uniformity}, \qquad (5)$$

where $Neg(z_i)$ denotes the collection of negative samples of z_i.

Improving $\mathcal{L}_{MINE}$ with asymmetric alignment. We can decompose both $\mathcal{L}_{MINE}$ (Eq. 5) and $\mathcal{L}_{NCE}$ (Eq. 3) into two terms: the *alignment* term encourages the positive pair to be close, and the *uniformity* term pushes the negative pair away. In fact, the uniformity term is a regularization used to avoid the *representation collapse*, i.e., the output representations are the same for all samples [76]. Alternatively, without the uniformity term, recent work SimSiam [12] suggests that the Siamese model can learn meaningful representations by introducing asymmetry in the alignment term and obtains better results.

In our experiments (Table 1), when using the common data augmentation strategy [11, 12], SimSiam is slightly better than models with contrastive loss ($\mathcal{L}_{NCE}$ or $\mathcal{L}_{MINE}$). However, we empirically find that the SimSiam model fails to learn stably in FSL when using *stronger* data augmentation. When the variations in the positive pairs are large, the phenomenon of *dimensional collapse* [36] occurs in SimSiam, i.e., a part of dimensionality of the embedding space vanishes (as shown in Fig. 6). In contrast, models with uniformity regularization do not suffer from significant dimensional collapse. This paper further improve $\mathcal{L}_{MINE}$ with the asymmetric alignment:

$$\mathcal{L}_{AMINE} = \underbrace{-\frac{1}{2B}\sum_{i=1}^{B}(p_i^T SG(z_{i+B}) + p_{i+B}^T SG(z_i))}_{Asymmetric\ Alignment} + \lambda \underbrace{\log(\sum_{i=1}^{2B}\sum_{z_j \in Neg(z_i)} e^{z_i^T z_j/\tau})}_{Uniformity}, \quad (6)$$

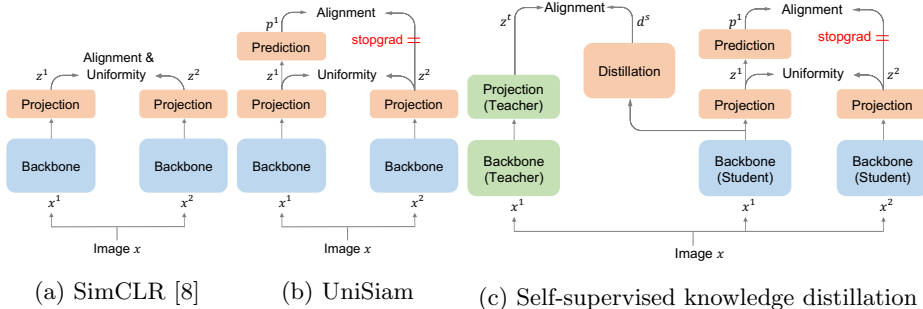(a) SimCLR [8]   (b) UniSiam   (c) Self-supervised knowledge distillation

Fig. 3: (**a**) SimCLR [8] for comparison. (**b**) Our UniSiam for self-supervised pre-training. (**c**) The architecture of our self-supervised knowledge distillation.

where λ is a weighting hyper-parameter, $p_i = h_{pred}(z_i)$ is the output of the additional *prediction* head $h_{pred}(\cdot)$ [12], and the SG (stop gradient) operation indicates that the back-propagation of the gradient stops here. Similar to the projection head, the prediction head is only used in the pre-training stage. Compared with SimSiam, our method can learn with stronger data augmentation to improve the invariance of the representations, resulting in better out-of-distribution generalization for FSL. Since our model can be considered as Sim**Siam** with the **Uni**formity regularization, we term it UniSiam (as shown in Fig. 3b).

Thus, we obtain the final self-supervised pre-training loss $\mathcal{L}_{AMINE}$ (Eq. 6). We can train our UniSiam model by minimizing this objective. After self-supervised pre-training, the pre-trained backbone can be used in FSL tasks by training a classifier on the output embeddings (discussed in Sec. 3.1). Note that the projection head and prediction head are removed in the fine-tuning stage. Next, we introduce how to perform self-supervised knowledge distillation with a pre-trained UniSiam model.

3.3   Self-Supervised Knowledge Distillation for Unsupervised FSL

A large model (teacher) trained with the self-supervised loss (Eq. 6) can be used to guide the learning of a small self-supervised model (student)[2]. In [70], the knowledge transfer from a teacher model to a student model is defined as maximizing the mutual information $I(X^s; X^t)$ between the representations of them. Maximizing the objective is equivalent to minimizing the conditional entropy $H(X^t|X^s)$, since $I(X^s; X^t) = H(X^t) - H(X^t|X^s)$ and the teacher model is fixed. It means the difference between their outputs should be as small as possible. So, simply aligning the outputs of them can achieve the purpose.

Specifically, as shown in Figure 3c, the pre-trained teacher encoder $f^t(\cdot)$ (consisting of the backbone $g^t(\cdot)$ and the projection head $h^t_{proj}(\cdot)$) is used to

---

[2] While larger models have better performance, training a smaller model is also meaningful since it can be more easily deployed in practical scenarios such as edge devices.

guide the training of the student backbone $g^s(\cdot)$ with a distillation head $h_{dist}(\cdot)$. The self-supervised distillation objective can be written as:

$$\mathcal{L}_{dist} = -\frac{1}{2B}\sum_{i=1}^{2B}(d_i^s)^T z_i^t, \tag{7}$$

where $d^s = h_{dist} \circ g^s(x)$ is the output of the distillation head on the student backbone, and $z^t = h_{proj}^t \circ g^t(x)$ is the output of the teacher model. Finally, the total objective that combines both distillation and pre-training is:

$$\mathcal{L} = \alpha\mathcal{L}_{AMINE} + (1-\alpha)\mathcal{L}_{dist}, \tag{8}$$

where α is a hyper-parameter. We set $\alpha = 0.5$ for all our experiments. Given a large UniSiam model pre-trained by Eq. 6, we can employ it as a teacher network to guide the training of a small model (from scratch) by minimizing Eq. 8.

4 Experiments

4.1 Datasets and Settings

Datasets. We perform experiments on two widely used few-shot image classification datasets, *mini*-ImageNet [75] and *tiered*-ImageNet [61]. *mini*-ImageNet [75] is a subset of ImageNet [62], which contains 100 classes with 600 images per class. We follow the split setting used in previous works [60], which randomly select 64, 16, and 20 classes for training, validation, and testing, respectively. *tiered*-ImageNet [61] is a larger subset of ImageNet with 608 classes and about 1300 images per class. These classes are grouped into 34 high-level categories and then divided into 20 categories (351 classes) for training, 6 categories (97 classes) for validation, and 8 categories (160 classes) for testing.

Implementation details. We use the networks of ResNet family [32] as our backbones. The projection and prediction heads of UniSiam are MLPs with the same setting as SimSiam [12], except that the ResNets without bottleneck blocks (e.g., ResNet-18) on *mini*-ImageNet use 512 output dimensions to avoid overfitting. The distillation head is a 5-layer MLP with batch normalization applied to each hidden layer. All the hidden fully-connected layers are 2048-D, except that the penultimate layer is 512-D. We find that this distillation head structure, which is similar to the combination of the projection and the prediction (as shown in Figure 3c), is suited for the knowledge distillation. The output vectors of the projection, prediction, and distillation heads are normalized by their L2-norm [79]. More implementation details can be found in the supplementary materials.

4.2 Self-Supervised vs. Supervised Pre-Training in FSL

In this subsection, we explore how several factors (network depth, image size, and data augmentation) affect the FSL performance of self-supervised and supervised pre-training. On *mini*-ImageNet, we compare supervised pre-training (training with the cross-entropy loss [71]) with our self-supervised UniSiam and two recent
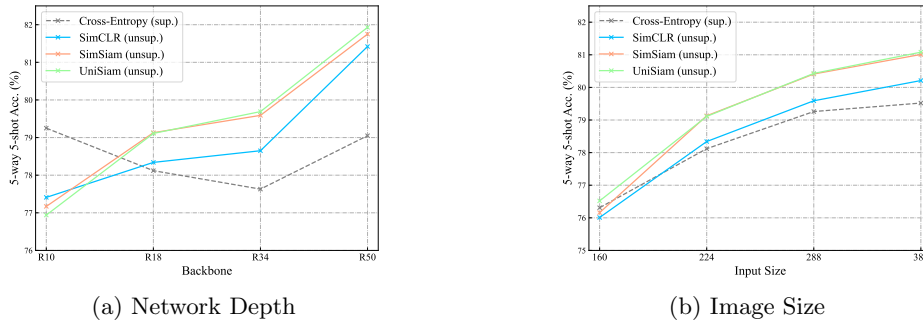
(a) Network Depth   (b) Image Size

Fig. 4: **Effect of network depth and image size.** (**a**) Self-supervised methods have better scalability for network depth compared to supervised pre-training in FSL. (**b**) A larger image size improves the FSL performance of self-supervised methods. Note that unsupervised (unsup.) approaches perform pre-training on the base dataset *without* any labels.

SSL models SimCLR [8] and SimSiam [12]. SimCLR is an well-known contrastive learning method that optimizes L_{NCE} (Eq. 3), and SimSiam is a relevant baseline to our UniSiam (i.e., $\lambda = 0$ in Eq. 6). More detailed comparison among the self-supervised methods is in Sec 4.3.

For a fair comparison, all methods use the same SGD optimization with cosine learning decay for 400 epochs, with batch size 256. Other hyper-parameters in each algorithm are chosen optimally using the grid search. To evaluate their performances in FSL, after pre-training on the base dataset of *mini*-ImageNet (i.e., the data of the training classes), we train a logistic regression classifier (with their fixed representations) for each few-shot classification task, which is sampled from the testing classes of *mini*-ImageNet. The reported results are the average of the accuracies on 3000 tasks for each method. More details about the baselines and evaluation can be found in the supplementary materials. Note that our self-supervised knowledge distillation is *not* used in this experiment.

Network depth. Figure 4a compares the performances of different methods with various depths of ResNet (i.e., ResNet-10/18/34/50). The input image size is 224×224. We use the data augmentation (DA) strategy, widely used in self-supervised learning [11, 12], termed *default DA*. The details of the default DA are described in the supplementary materials.

We can find that when the backbone is shallow (i.e., ResNet-10), supervised pre-training has an advantage compared to self-supervised methods. However, as the network deepens, the self-supervised methods gradually outperform the supervised one. When the backbone changes from ResNet10 to ResNet50, the performance improvement of the self-supervised approach is larger than 4%. In contrast, the performance of supervised pre-training is decreased by 0.2%.
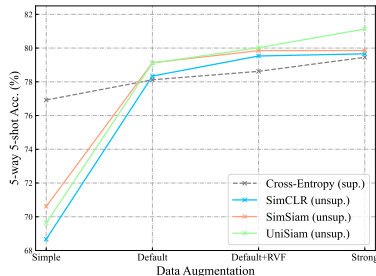
Fig. 5: **Effect of data augmentation.** Stronger data augmentation can substantially improve the performances of the self-supervised pre-training compared to supervised pre-training.

Image size. Fig. 4b shows the performances of different approaches with various input sizes (160×160, 224×224, 288×288, and 384×384). All methods use ResNet-18 as the backbone with the default DA strategy. We find that a larger image size is more important for self-supervised methods. When the image size is small (i.e., 160×160), the performances of different methods are close. However, when the image size increases, self-supervised methods have larger performance gains compared with supervised pre-training. Although a larger image size can bring significant performance improvement, we still use the image size of 224×224 in other experiments following the typical setting in the community.

Data augmentation. Fig. 5 shows the performances of various pre-training methods with different levels of data augmentation. All mehtods use the ResNet-18 backbone with input size 224×224. Here we introduces two effective DA for FSL: *RandomVerticalFlip* (RVF) and *RandAugment* (RA) [17]. We set 4 levels of DA (from slight to heavy) as follows: (1) "Simple" denotes the strategy used for traditional supervised pre-training (including *RandomResizedCrop*, *ColorJitter*, and *RandomHorizontalFlip*), (2) "Default" is the same as the default DA mentioned above, (3) "Default+RVT" denotes the default DA plus the RVF, and (4) "Strong" represents the default DA plus RVF and RA.

Supervised pre-training can bring more information than self-supervised methods in the case of simple DA. However, default DA substantially improves the performances of the self-supervised methods, but it has a limited gain for supervised pre-training. In addition, RVF can further improve the performances of all methods. RA improves the performances of most methods, except for SimSiam. We consider that the strong data augmentation leads to the dimensional collapse of SimSiam, as shown in the next subsection.

4.3 Self-Supervised Pre-Training with Strong Augmentation

We compare SimCLR, SimSiam, and the variants of our UniSiam under default and strong DA (in Table 1). We observe that self-supervised pre-training with the uniformity term obtains a larger improvement from strong DA compared with SimSiam. In addition, the uniformity term of $\mathcal{L}_{MINE}$ has a more significant improvement than the uniformity term of $\mathcal{L}_{NCE}$. Asymmetric alignment can also improve the FSL performance than the symmetric alignment.
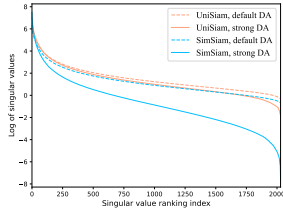
Fig. 6: **Singular value spectrum of embedding space.** The uniformity regularization alleviates the dimensional collapse under strong DA.

Method	Align	Uniform	R18		R50	
			DefaultDA	StrongDA	DefaultDA	StrongDA
SimCLR	symm.	NCE (Eq. 3)	78.34±0.27	79.66±0.27	81.42±0.25	81.51±0.26
SimSiam	asymm.	-	**79.13±0.26**	79.85±0.27	81.75±0.24	79.66±0.27
	symm.	MINE (Eq. 5)	78.04±0.27	80.72±0.26	81.45±0.24	82.84±0.24
	asymm.	NCE (Eq. 3)	78.95±0.26	80.66±0.26	81.51±0.24	82.54±0.24
UniSiam	asymm.	MINE (Eq. 5)	79.11±0.25	**81.13±0.26**	**81.93±0.24**	**83.18±0.24**

Table 1: **Comparison of self-supervised methods under default and strong data augmentations.** We report their 5-way 5-shot accuracy (%) on *mini*-ImageNet. "symm." and "asymm." denote using the symmetric alignment (Eq. 3 or Eq. 5) and the asymmetric alignment (Eq. 6) respectively.

To further demonstrate the importance of uniformity, we visualize the singular value spectrum of the embedding space of SimSiam and our UniSiam under different DAs in Fig. 6. The backbone is ResNet-50. Both SimSiam and UniSiam have a flat singular value spectrum when using the default DA. However, when DA is strong, some singular values of SimSiam are reduced. It means the features of SimSiam fall into a lower-dimensional subspace. This phenomenon is termed dimensional collapse by [36]. In contrast, the singular value spectrum of UniSiam is flat even with strong DA, which indicates the significance of the uniformity.

4.4   Our Self-Supervisied Knowledge Distillation

The previous work RFS [71] employs the standard knowledge distillation [33] to improve the supervised pre-training model in FSL. However, it is based on the logits that cannot be applied in unsupervised FSL. We use the standard knowledge distillation to transfer knowledge from a large supervised pre-training model to small ones, being a compared baseline to our self-supervised knowledge distillation (as shown in Tabel 2). Note that our method does not use any labels in the pre-training and the distillation stage. All methods use the default DA and the image size of 224×224. We can see that our knowledge distillation approach improves the performances of the smaller networks. Although the distillation loss allows supervised pre-training models to capture the relationships between classes to learn information beyond labels, our model after distillation still outperforms them when the backbones are larger (ResNet-18 and ResNet-34).

4.5   Comparison with the State-of-the-Art

We compare with state-of-the-art FSL approaches in Table 3 and Table 4. Our method uses the strong DA and the image size of 224×224. In addition, we reim-

	Teacher ResNet-50	distillation		Student	
			ResNet-10	ResNet-18	ResNet-34
RFS [71] (sup.)	79.05±0.26	N	79.25±0.26	78.12±0.26	77.63±0.27
		Y	79.44±0.25	80.15±0.25	80.55±0.26
UniSiam (unsup.)	81.93±0.24	N	76.94±0.27	79.11±0.25	79.69±0.26
		Y	78.58±0.26	80.35±0.26	81.39±0.25

Table 2: **Effect of our self-supervised knowledge distillation.** We report the 5-way 5-shot classification accuracy (%) on the *mini*-ImageNet dataset.

Backbone	Method	Size		1-shot	5-shot
ResNet-18	Δ-Encoder [64]	224	sup.	59.9	69.7
	SNCA [78]	224	sup.	57.8±0.8	72.8±0.7
	iDeMe-Net [15]	224	sup.	59.14±0.86	74.63±0.74
	Robust+dist [20]	224	sup.	63.73±0.62	81.19±0.43
	AFHN [45]	224	sup.	62.38±0.72	78.16±0.56
	ProtoNet+SSL [67]	224	sup.+ssl	-	76.6
	Neg-Cosine [47]	224	sup.	62.33±0.82	80.94±0.59
	Centroid Alignment [2]	224	sup.	59.88±0.67	80.35±0.73
	PSST [14]	224	sup.+ssl	59.52±0.46	77.43±0.46
	UMTRA‡ [37]	224	unsup.	43.09±0.35	53.42±0.31
	ProtoCLR‡ [50]	224	unsup.	50.90±0.36	71.59±0.29
	SimCLR‡ [8]	224	unsup.	62.58±0.37	79.66±0.27
	SimSiam‡ [12]	224	unsup.	62.80±0.37	79.85±0.27
	UniSiam (Ours)	224	unsup.	63.26±0.36	81.13±0.26
	UniSiam+dist (Ours)	224	unsup.	**64.10±0.36**	**82.26±0.25**
ResNet-34	MatchingNet† [75]	224	sup.	53.20±0.78	68.32±0.66
	ProtoNet† [65]	224	sup.	53.90±0.83	74.65±0.64
	MAML† [22]	224	sup.	51.46±0.90	65.90±0.79
	RelationNet† [68]	224	sup.	51.74±0.83	69.61±0.67
	Baseline [10]	224	sup.	49.82±0.73	73.45±0.65
	Baseline++ [10]	224	sup.	52.65±0.83	76.16±0.63
	SimCLR‡ [8]	224	unsup.	63.98±0.37	79.80±0.28
	SimSiam‡ [12]	224	unsup.	63.77±0.38	80.44±0.28
	UniSiam (Ours)	224	unsup.	64.77±0.37	81.75±0.26
	UniSiam+dist (Ours)	224	unsup.	**65.55±0.36**	**83.40±0.24**

Table 3: **Comparison to previous works on *mini*-ImageNet**, using the averaged 5-way classification accuracy (%) with the 95% confidence interval on the testing split. Note that UniSiam+dist is trained by our self-supervised knowledge distillation (Fig. 3c) with ResNet-50 being the teacher's backbone. †: the results obtained from [10]. ‡: the results are from our implementations. Models that use knowledge distillation are tagged with the suffix "+*dist*".

plement two unsupervised FSL methods (ProtoCLR [50] and UMTRA [37]) with the same DA strategy (strong DA) on *mini*-ImageNet. More baseline details are in the supplementary materials. On *mini*-ImageNet, our unsupervised UniSiam achieves the state-of-the-art results compared to other supervised methods with the ResNet-18 and ResNet-34 backbones. UniSiam also has a significant improvement than some methods that incorporate self-supervised objective and supervised pre-training ("sup.+ssl"). In addition, our method outperforms previous unsupervised FSL methods [37,50] by a larger margin.

On *tiered*-ImageNet, since only a few studies use standard ResNet [32] as their backbones, we also compare with some methods that use other backbones. For a fair comparison, we count the number of parameters and MACs of different backbones. Note that ResNet-12 modifies the original architecture of ResNet (e.g., larger channel dimensions). It has a larger computation overhead than standard

Method	Backbone (#Params)	Size	MACs		1-shot	5-shot
MetaOptNet [43]	ResNet-12 (8.0M)	84	3.5G	sup.	65.99±0.72	81.56±0.53
RFS+dist [71]	ResNet-12 (8.0M)	84	3.5G	sup.	**71.52±0.72**	**86.03±0.49**
BML [83]	ResNet-12 (8.0M)	84	3.5G	sup.	68.99±0.50	85.49±0.34
Roubst+dist [20]	ResNet-18 (11.2M)	224	1.8G	sup.	70.44±0.32	85.43±0.21
Centroid Alignment [2]	ResNet-18 (11.2M)	224	1.8G	sup.	69.29±0.56	85.97±0.49
SimCLR‡ [8]	ResNet-18 (11.2M)	224	1.8G	unsup.	63.38±0.42	79.17±0.34
SimSiam‡ [12]	ResNet-18 (11.2M)	224	1.8G	unsup.	64.05±0.40	81.40±0.30
UniSiam (Ours)	ResNet-18 (11.2M)	224	1.8G	unsup.	65.18±0.39	82.28±0.29
UniSiam+dist (Ours)	ResNet-18 (11.2M)	224	1.8G	unsup.	67.01±0.39	84.47±0.28
LEO [63]	WRN-28-10 (36.5M)	84	41G	sup.	66.33±0.05	81.44±0.09
CC+Rot [25]	WRN-28-10 (36.5M)	84	41G	sup.+ssl	**70.53±0.51**	84.98±0.36
FEAT [80]	WRN-28-10 (36.5M)	84	41G	sup.	70.41±0.23	84.38±0.16
UniSiam (Ours)	ResNet-34 (21.3M)	224	3.6G	unsup.	67.57±0.39	84.12±0.28
UniSiam+dist (Ours)	ResNet-34 (21.3M)	224	3.6G	unsup.	68.65±0.39	85.70±0.27
UniSiam (Ours)	ResNet-50 (23.5M)	224	4.1G	unsup.	69.11±0.38	85.82±0.27
UniSiam+dist (Ours)	ResNet-50 (23.5M)	224	4.1G	unsup.	69.60±0.38	**86.51±0.26**

Table 4: **Comparison to previous FSL works on *tiered*-ImageNet**, using the averaged 5-way classification accuracy (%) on the testing split. ‡: the results are from our implementations. ResNet-50 is the teacher's backbone.

ResNet-18, even with a smaller input size. Our method with a shallow backbone ResNet-18 is slightly worse than top supervised FSL methods on *tiered*-ImageNet. The main reasons are twofold. One is that increasing the number of classes alleviates the overfitting problem of supervised methods on the *tiered*-ImageNet dataset. The more important reason is that existing FSL methods utilize a variety of techniques to implicitly alleviate the problem of over-fitting to the base classes. For example, *Robust+dist* [20] trains 20 different networks to learn diverse information for avoiding overfitting. *RFS+dist* [71] repeats self-distillation many times, which can capture the relation between the classes to learn more information beyond the labels. However, these methods require complicated processes and troublesome human designs, which limit their application and scalability. In contrast, our self-supervised UniSiam is a concise and effective approach that fundamentally avoids bias. When the backbone (i.e., ResNet-34) has similar computational overhead, UniSiam also achieves comparable results with the state-of-the-art supervised FSL methods on *tiered*-ImageNet.

5   Conclusion

This paper proposes an effective few-shot learner without using any labels of the base dataset. From a unified information-theoretic perspective, our self-supervised pre-training learns good embeddings with less bias toward the base classes for FSL by maximizing the MI of the instances and their representations. Compared with state-of-the-art supervised FSL methods, our UniSiam achieves comparable results on two popular FSL benchmarks. Considering the simplicity and effectiveness of the proposed approach, we believe it would motivate other researchers to rethink the role of label information of the base dataset in FSL.

References

1. MindSpore. `https://www.mindspore.cn/`
2. Afrasiyabi, A., Lalonde, J.F., Gagné, C.: Associative alignment for few-shot image classification. In: ECCV (2019)
3. Antoniou, A., Storkey, A.: Assume, augment and learn: Unsupervised few-shot meta-learning via random labels and data augmentation. arxiv:1902.09884 (2019)
4. Belghazi, M.I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., Hjelm, D.: Mutual information neural estimation. In: ICML (2018)
5. Boudiaf, M., Masud, Z.I., Rony, J., Dolz, J., Piantanida, P., Ayed, I.B.: Transductive information maximization for few-shot learning. In: NeurIPS (2020)
6. Boudiaf, M., Rony, J., Masud, Z.I., Granger, E., Pedersoli, M., Piantanida, P., Ayed, I.B.: A unifying mutual information view of metric learning: Cross-entropy vs. pairwise losses. In: ECCV (2020)
7. Brown, T.B., Mann, B., et al.: Language models are few-shot learners. In: NeurIPS (2020)
8. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020)
9. Chen, T., Li, L.: Intriguing properties of contrastive losses. arXiv:2011.02803 (2020)
10. Chen, W.Y., Liu, Y.C., Kira, Z., Wang, Y.C.F., Huang, J.B.: A closer look at few-shot classification. In: ICLR (2019)
11. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv:2003.04297 (2020)
12. Chen, X., He, K.: Exploring simple siamese representation learning. In: CVPR (2021)
13. Chen, Y., Wang, X., Liu, Z., Xu, H., Darrell, T.: A new meta-baseline for few-shot learning. arXiv:2003.04390 (2020)
14. Chen, Z., Ge, J., Zhan, H., Huang, S., Wang, D.: Pareto self-supervised training for few-shot learning. In: CVPR (2021)
15. Chen, Z., Fu, Y., Wang, Y.X., Ma, L., Liu, W., Hebert, M.: Image deformation meta-networks for one-shot learning. In: CVPR (2019)
16. Chen, Z., Maji, S., Learned-Miller, E.: Shot in the dark: Few-shot learning with no base-class labels. In: CVPRW (2021)
17. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.: Randaugment: Practical automated data augmentation with a reduced search space. In: NeurIPS (2020)
18. Dhamija, A.R., Ahmad, T., Schwan, J., Jafarzadeh, M., Li, C., Boult, T.E.: Self-supervised features improve open-world learning. arXiv:2102.07848 (2021)
19. Dhillon, G.S., Chaudhari, P., Ravichandran, A., Soatto, S.: A baseline for few-shot image classification. In: ICLR (2020)
20. Dvornik, N., Mairal, J., Schmid, C.: Diversity with cooperation: Ensemble methods for few-shot classification. In: ICCV (2019)
21. Ericsson, L., Gouk, H., Hospedales, T.M.: How well do self-supervised models transfer? arXiv:2011.13377 (2020)
22. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: ICML (2017)
23. Flennerhag, S., Rusu, A.A., Pascanu, R., Visin, F., Yin, H., Hadsell, R.: Meta-learning with warped gradient descent. In: ICLR (2020)
24. Gallardo, J., Hayes, T.L., Kanan, C.: Self-supervised training enhances online continual learning. In: BMVC (2021)

25. Gidaris, S., Bursuc, A., Komodakis, N., Perez, P.P., Cord, M.: Boosting few-shot visual learning with self-supervision. In: ICCV (2019)
26. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. In: ICLR (2018)
27. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M.: Bootstrap your own latent: A new approach to self-supervised learning. In: NeurIPS (2020)
28. Guo, Y., Codella, N.C., Karlinsky, L., Codella, J.V., Smith, J.R., Saenko, K., Rosing, T., Feris, R.: A broader study of cross-domain few-shot learning. In: ECCV (2020)
29. Gutmann, M., Hyvärinen, A.: Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In: AISTATS (2010)
30. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: CVPR (2006)
31. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020)
32. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
33. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv:1503.02531 (2015)
34. Hjelm, R.D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., Bengio, Y.: Learning deep representations by mutual information estimation and maximization. In: ICLR (2018)
35. Hsu, K., Levine, S., Finn, C.: Unsupervised learning via meta-learning. In: ICLR (2018)
36. Jing, L., Vincent, P., LeCun, Y., Tian, Y.: Understanding dimensional collapse in contrastive self-supervised learning. In: ICLR (2021)
37. Khodadadeh, S., Boloni, L., Shah, M.: Unsupervised meta-learning for few-shot image classification. In: NeurIPS (2019)
38. Khodadadeh, S., Zehtabian, S., Vahidian, S., Wang, W., Lin, B., Boloni, L.: Unsupervised meta-learning through latent-space interpolation in generative models. In: ICLR (2021)
39. Laenen, S., Bertinetto, L.: On episodes, prototypical networks, and few-shot learning. In: NeurIPS (2021)
40. Lake, B.M., Salakhutdinov, R., Tenenbaum, J.B.: Human-level concept learning through probabilistic program induction. Science (2015)
41. Larsson, G., Maire, M., Shakhnarovich, G.: Learning representations for automatic colorization. In: ECCV (2016)
42. Lee, D.B., Min, D., Lee, S., Hwang, S.J.: Meta-gmvae: Mixture of gaussian vae for unsupervised meta-learning. In: ICLR (2021)
43. Lee, K., Maji, S., Ravichandran, A., Soatto, S.: Meta-learning with differentiable convex optimization. In: CVPR (2019)
44. Li, F.F., Rob, F., Pietro, P.: One-shot learning of object categories. TPAMI (2006)
45. Li, K., Zhang, Y., Li, K., Fu, Y.: Adversarial feature hallucination networks for few-shot learning. In: CVPR (2020)
46. Linsker, R.: Self-organization in a perceptual network. Computer (1988)
47. Liu, B., Cao, Y., Lin, Y., Li, Q., Zhang, Z., Long, M., Hu, H.: Negative margin matters: Understanding margin in few-shot classification. In: ECCV (2020)
48. Lu, Y., Liu, J., Zhang, Y., Liu, Y., Tian, X.: Prompt distribution learning. In: CVPR (2022)

49. Mangla, P., Kumari, N., Sinha, A., Singh, M., Krishnamurthy, B., Balasubrama-nian, V.N.: Charting the right manifold: Manifold mixup for few-shot learning. In: WACV (2020)
50. Medina, C., Devos, A., Grossglauser, M.: Self-supervised prototypical transfer learning for few-shot classification. In: ICMLW (2020)
51. Mishra, N., Rohaninejad, M., Chen, X., Abbeel, P.: A simple neural attentive meta-learner. In: ICLR (2018)
52. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv:1807.03748 (2018)
53. Oreshkin, B.N., López, P.R., Lacoste, A.: Tadam: Task dependent adaptive metric for improved few-shot learning. In: NeurIPS (2018)
54. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context en-coders: Feature learning by inpainting. In: CVPR (2016)
55. Phoo, C.P., Hariharan, B.: Self-training for few-shot transfer across extreme task differences. arXiv:2010.07734 (2020)
56. Poole, B., Ozair, S., van den Oord, A., Alemi, A.A., Tucker, G.: On variational bounds of mutual information. In: ICML (2019)
57. Qin, T., Li, W., Shi, Y., Yang, G.: Unsupervised few-shot learning via distribution shift-based augmentation. arxiv:2004.05805 (2020)
58. Qin, Z., Kim, D., Gedeon, T.: Neural network classifiers as mutual information evaluators. In: ICMLW (2021)
59. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021)
60. Ravi, S., Larochelle, H.: Optimization as a model for few-shot learning. In: ICLR (2017)
61. Ren, M., Triantafillou, E., Ravi, S., Snell, J., Swersky, K., Tenenbaum, J.B., Larochelle, H., Zemel, R.S.: Meta-learning for semi-supervised few-shot classifi-cation. arXiv:1803.00676 (2018)
62. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. IJCV (2015)
63. Rusu, A.A., Rao, D., Sygnowski, J., Vinyals, O., Pascanu, R., Osindero, S., Hadsell, R.: Meta-learning with latent embedding optimization. In: ICLR (2018)
64. Schwartz, E., Karlinsky, L., Shtok, J., Harary, S., Marder, M., Kumar, A., Feris, R.S., Giryes, R., Bronstein, A.M.: Delta-encoder: an effective sample synthesis method for few-shot object recognition. In: NeurIPS (2018)
65. Snell, J., Swersky, K., Zemel, R.S.: Prototypical networks for few-shot learning. In: NeurIPS (2017)
66. Song, J., Ermon, S.: Understanding the limitations of variational mutual informa-tion estimators. In: ICLR (2020)
67. Su, J.C., Maji, S., Hariharan, B.: When does self-supervision improve few-shot learning? In: ECCV (2020)
68. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H., Hospedales, T.M.: Learning to compare: Relation network for few-shot learning. In: CVPR (2018)
69. Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. In: ECCV (2019)
70. Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation. In: ICLR (2020)
71. Tian, Y., Wang, Y., Krishnan, D., Tenenbaum, J.B., Isola, P.: Rethinking few-shot image classification: a good embedding is all you need? In: ECCV (2020)

72. Tian, Y., Chen, X., Ganguli, S.: Understanding self-supervised learning dynamics without contrastive pairs. arXiv:2102.06810 (2021)
73. Tschannen, M., Djolonga, J., Rubenstein, P.K., Gelly, S., Lucic, M.: On mutual information maximization for representation learning. In: ICLR (2020)
74. Vardan, P., Han, X.Y., Donoho, D.L.: Prevalence of neural collapse during the terminal phase of deep learning training. PNAS (2020)
75. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al.: Matching networks for one shot learning. In: NeurIPS (2016)
76. Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: ICML (2020)
77. Wen, L., Zhou, Y., He, L., Zhou, M., Xu, Z.: Mutual information gradient estimation for representation learning. In: ICLR (2020)
78. Wu, Z., Efros1, A.A., Yu, S.X.: Improving generalization via scalable neighborhood component analysis. In: ECCV (2018)
79. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: CVPR (2018)
80. Ye, H.J., Hu, H., Zhan, D.C., Sha, F.: Few-shot learning via embedding adaptation with set-to-set functions. In: CVPR (2020)
81. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: Self-supervised learning via redundancy reduction. arXiv:2103.03230 (2021)
82. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: ECCV (2016)
83. Zhou, Z., Qiu, X., Xie, J., Wu, J., Zhang, C.: Binocular mutual learning for improving few-shot classification. In: ICCV (2021)