Shot in the Dark: Few-Shot Learning with No Base-Class Labels

Zitian Chen Subhransu Maji Erik Learned-Miller
University of Massachusetts Amherst
{zitianchen,smaji,elm}@cs.umass.edu

Abstract

Few-shot learning aims to build classifiers for new classes from a small number of labeled examples and is commonly facilitated by access to examples from a distinct set of 'base classes'. The difference in data distribution between the test set (novel classes) and the base classes used to learn an inductive bias often results in poor generalization on the novel classes. To alleviate problems caused by the distribution shift, previous research has explored the use of unlabeled examples from the novel classes, in addition to labeled examples of the base classes, which is known as the ***transductive setting****. In this work, we show that, surprisingly, off-the-shelf self-supervised learning outperforms transductive few-shot methods by 3.9% for 5-shot accuracy on* mini*ImageNet* ***without using any base class labels****. This motivates us to examine more carefully the role of features learned through self-supervision in few-shot learning. Comprehensive experiments are conducted to compare the transferability, robustness, efficiency, and the complementarity of supervised and self-supervised features.*

1. Introduction

Deep architectures have achieved significant success in various vision tasks including image classification and object detection. Such success have relied heavily on massive numbers of annotated examples. However, in real-world scenarios, we are frequently unable to collect enough labeled examples. This has motivated the study of few-shot learning (FSL), which focuses on building classifiers for novel categories from one or very few labeled examples.

Previous approaches to FSL include meta-learning and metric learning. Meta-learning aims to learn task-agnostic knowledge that improves optimization. Metric learning focuses on learning representations on base categories that can generalize to novel categories. Most previous FSL methods attempt to borrow a strong inductive bias from the *supervised* learning of base classes. However, the challenge of FSL is that a *helpful* inductive bias, i.e., one that improves performance on novel classes, is hard to develop when there is a large difference between the base and novel classes.

To address this challenge, previous research explores using unlabeled examples from novel classes to improve the generalization on novel classes, which is referred to transductive few-shot learning. Typical transductive few-shot learning (TFSL) methods include exploiting unlabeled novel examples that have been classified (by an initial classifier) with high confidence in order to self-train the model [4, 23, 26] or fine-tuning the model on unlabeled novel examples with an auxiliary loss serving as a regularizer [8,24,31]. These methods still focus on **improving** the generalization of inductive bias borrowed from the supervised learning of base classes.

In comparison, our key motivation is that, unlabeled examples from novel classes not only can fine-tune or retrain a pre-trained model, but also can effectively train a new model from scratch. The advantage of doing so is that the model can generalize better on novel classes. In this paper, we demonstrate the effectiveness of **an extremely simple baseline** for transductive few-shot learning. **Our baseline does not use any labels on the base classes**. We conduct self-supervised learning on unlabeled data from both the base and the novel classes to learn a feature embedding. When doing few-shot classification, we directly learn a linear classifier on top of the feature embedding from the few given labeled examples and then classify the testing examples. Surprisingly, this baseline significantly outperforms state-of-the-art transductive few-shot learning methods, which have additional access to base-class labels.

The empirical performance of this baseline should not be "the final solution" for few-shot learning. We believe that meta-learning, metric learning, data augmentation, and transfer learning are also critical for effective few-shot learning. However, this baseline can help us interpret existing results and indicates that using self-supervised learning to learn a generalized representation could be another important tool in addressing few-shot learning.

To investigate the best possible way to use self-supervised learning in few-shot learning, it is necessary to examine more carefully the role of features learned through self-supervision in few-shot learning. For brevity, we refer to these features as 'self-supervised features'. **(1)** In a non-transductive few-shot learning setting, we explore the *complementarity* and *transferability* of supervised and self-supervised features. By directly concatenating self-supervised and supervised features, we get a 2-3% performance boost and achieve new state-of-the-art results. We conduct cross-domain few-shot learning and show that supervised features have better transferability than self-supervised features. However, when more novel labeled examples are given, self-supervised features overtake supervised features. **(2)** In a transductive few-shot learning setting, we show that simple off-the-shelf self-supervised learning significantly outperforms other competitors who have additional access to base-class labels. We confirm the performance gain is not from a better representation but from a representation that better generalizes on the novel classes. The proof is that the self-supervised features achieve the top performance on the novel classes but *not* on other unseen classes. **(3)** For both non-transductive and transductive settings, we conduct comprehensive experiments to explore the effect of different backbone architectures and datasets. We report results using a shallow ResNet, a very deep ResNet, a very wide ResNet, and a specially designed shallow ResNet that is commonly used for few-shot learning. While deeper models generally have significantly better performance for the standard classification task on both large (e.g, ImageNet [7]) and small datasets (e.g., CIFAR-10) as shown in [18], the performance gain is relatively small for supervised features in few-shot learning. In comparison, self-supervised features show a much larger improvement when using a deeper network, especially in the transductive setting. We also conduct experiments on various datasets, including large datasets, small datasets, and datasets that have small or large domain differences between base and novel classes. We show the efficiency and robustness of self-supervised features on all kinds of datasets except for very small datasets.

2. Related Work

Few-shot Learning. Few-shot learning is a classic problem [27], which refers to learning from one or a few labeled examples for each novel class. Existing FSL methods can be broadly grouped into three categories: data augmentation, meta-learning, and metric learning. Data augmentation methods synthesize [6, 34, 43], hallucinate [16] or deform [5] images to generate additional examples to address the training data scarcity. Meta-learning [10, 21, 28, 32] attempts to learn a parameterized mapping from limited training examples to hidden parameters that accelerate or improve the optimization procedure. Metric learning [1,22,37] aims at learning a transferable metric space (or embedding). MatchingNet [40] and ProtoNet [35] adopt cosine and Euclidean distance to separate instances belonging to different classes. Recently, some works [6,25,39] showed that learning a classifier on top of supervised features can achieve surprisingly competitive performance.

Transductive Few-shot Learning. TFSL methods use the distribution support of unlabeled novel instances to help few-shot learning. Some TFSL methods [23, 26, 42] exploit unlabeled instances with high confidence to train the model. [4] propose a data augmentation method to directly mix base examples and selected novel examples in the image domain to learn generalized features. In addition, previous work [8, 24, 31] seek to take unlabeled testing instances to acquire an auxiliary loss serving as a regularizer to adapt the inductive bias. These methods borrow inductive bias from the supervised learning of the base classes and further utilize unlabeled novel examples to improve it. In comparison, we show that unlabeled novel examples in addition to labeled examples of the base classes can directly develop a very strong inductive bias.

Self-supervised Learning. Self-supervised learning aims to explore the internal data distribution and learns discriminative features without annotations. Some work takes predicting rotation [13], counting [30], predicting the relative position of patches [9], colorization [20, 46], and solving jigsaw puzzles [29] as self-supervised tasks to learn representations. Recently, instance discrimination [2,15,38,44] has attracted much attention. [17] propose a momentum contrast to update models and shows superior performance to supervised learning. In this work, we explore the generalization ability of self-supervised features to new classes in the few-shot setting, i.e., in circumstances where few labeled examples of novel classes are given. Other works that have explored transductive techniques, e.g., [17], have used large training sets for new classes. Gidaris et al. [12] and Su et al. [36] take rotation prediction, solving jigsaw as auxilary tasks to learn better representation on base classes to help few-shot learning. Tian et al. [39] utilize contrastive learning to learn features for non-transductive few-shot learning. In comparison, while previous works only conduct self-supervised learning under the non-transductive, we confirm the effectiveness of self-supervised learning in a transductive few-shot setting. We claim this as our major contribution.

3. Methods

In Fig. 1, we illustrate our few-shot learning settings. We denote the base category set as C_{base} and the novel category set as C_{novel}, in which $C_{base} \cap C_{novel} = \emptyset$. Correspondingly, we denote the labeled base dataset as $D_{base} = \{(I_i, y_i)\}, y_i \in C_{base}$, the labeled novel dataset as $D_{novel} = \{(I_i, y_i)\}, y_i \in C_{novel}$, the unlabeled base dataset as $U_{base} = \{(I_i)\}, y_i \in C_{base}$, and the unlabeled novel dataset as $U_{novel} = \{(I_i)\}, y_i \in C_{novel}$.

In a standard few-shot learning task, we are only given

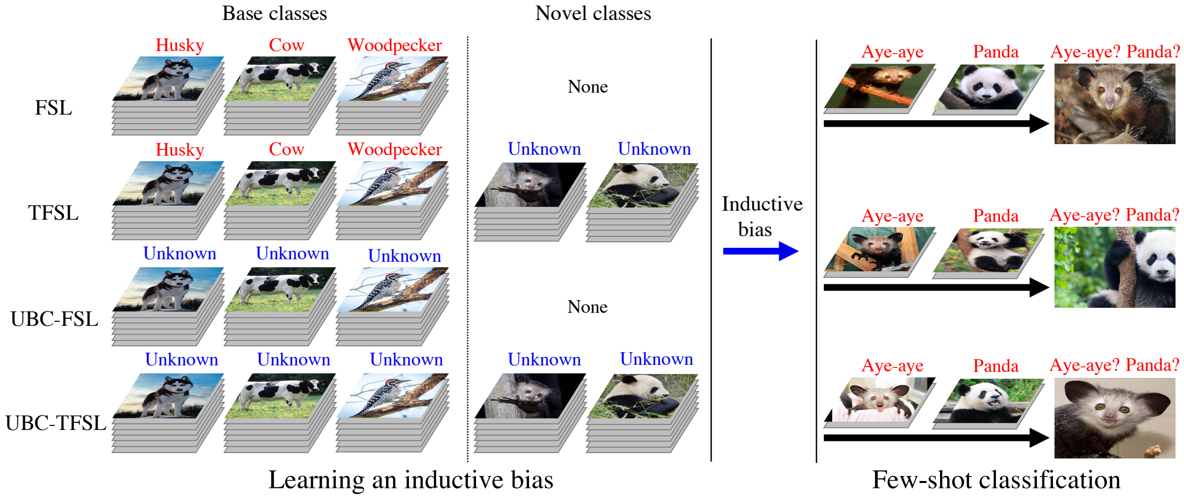


Figure 1: **An illustration of different few-shot learning settings.** There are four few-shot settings, including few-shot learning (FSL), transductive few-shot learning (TFSL), unlabeled-base-class few-shot learning (UBC-FSL), and unlabeled-base-class transductive few-shot learning (UBC-TFSL). The differences between these settings are **whether they have labels for examples from the base classes and unlabeled examples from the novel classes.**

labeled examples from base classes so the training set is $D_{FSL} = D_{base}$. For transductive few-shot learning (TFSL), we are given $D_{TFSL} = D_{base} \cup U_{novel}$. For unlabeled-base-class few-shot learning (UBC-FSL), we have $D_{UBC-FSL} = U_{base}$. For unlabeled-base-class transductive few-shot learning (UBC-TFSL), we denote the training set as $D_{UBC-TFSL} = U_{base} \cup U_{novel}$. Note that UBC-TFSL has strictly less supervision than TFSL and this setting has not been explored before. We claim **we are the first to explore this setting**.

These four few-shot learning settings use the same evaluation protocol as in previous works [40]. At inference time, we are given a collection of *N-way-m-shot* classification tasks sampled from D_{novel} to evaluate our methods.

3.1. Self-supervised learning

Here we take instance discrimination as our self-supervision task due to its efficiency. We follow momentum contrast [17], where each training example x_i is augmented twice into x_i^q and x_i^k. x_i^q and x_i^k are then fed into two encoders forming two embeddings $q_i = f_q(x_i^q)$, and $k_i = f_k(x_i^k)$. A standard log-softmax function is used to discriminate a positive pair (2 instances augmented from one image) from several negative pairs (2 instances augmented from 2 images):

$$L(q_i, k_i) = -\log\left(\frac{\exp(q_i^T k_i/\tau)}{\exp(q_i^T k_i/\tau) + \sum_{j\neq i}\exp(q_i^T k_j/\tau)}\right) \tag{1}$$

where τ is a temperature hyper-parameter. Since our implementations are based on MoCo-v2 [3], please refer to it for further details. We also try other self-supervised methods in § 4.6.

3.2. Evaluation Protocols

Here we introduce our protocols for the four different few-shot learning settings. All protocols consist of a training phase and an evaluation phase. In the training phase, we learn a feature embedding on the training sets D_{FSL}, D_{TFSL}, $D_{UBC-FSL}$, and $D_{UBC-TFSL}$. In the evaluation phase, we evaluate the few-shot classification performance. For simplicity and efficiency, we learn a logistic regression classifier on top of the learned feature embedding of $N * m$ training examples and then classify the testing examples. Training and testing examples come from the given *N-way-m-shot* classification task. Such procedures are repeated 1000 times and we report the average few-shot classification accuracies with 95% confidence intervals. Now, we would like to introduce our methods.

Few-shot learning baseline. We learn our embedding network on D_{FSL} using cross-entropy loss under a standard classification process. We use the logit layer as the feature embedding as it is slightly better than the pre-classification layer. This baseline is very simple and achieve the state-of-the-art performance.

Unlabeled-base-class few-shot learning. For UBC-FSL, we learn from self-supervised supervision on $D_{UBC-FSL}$. We follow MoCo-v2 to do instance discrimination. The output of the final layer of the model is used as the feature embedding.

Unlabeled-base-class transductive few-shot learning. For UBC-TFSL, our method is similar to our UBC-FSL

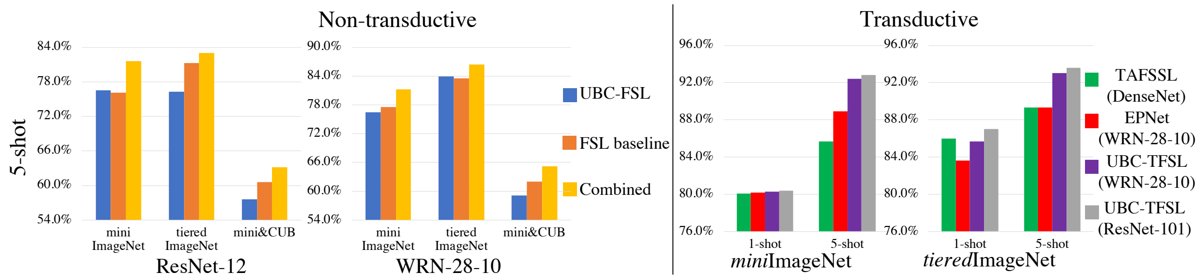


Figure 2: **Comparison between different methods under the non-transductive and transductive few-shot setting.** For the non-transductive few-shot learning, there is great complementarity among supervised and self-supervised features. For the transductive few-shot learning, our UBC-TFSL outperform other competitors even without using base-classes labels.

method. The difference is that we train on $D_{UBC-TFSL}$, which has additional access to unlabeled test instances.

Combination of FSL baseline and UBC-FSL. This method works under standard, non-transductive, few-shot learning setting. We explore the complementarity between supervised features (from the FSL baseline) and self-supervised features (from UBC-FSL). We directly concatenate normalized supervised features and normalized self-supervised features and then do normalization again. T his feature is used as the feature embedding and we refer this method as "Combined".

4. Experiments

We define two types of experiments based upon whether the base and novel classes come from the same dataset or not. We refer to the standard FSL paradigm in which the base and novel classes come from the same dataset (e.g., ImageNet) as *single-domain* FSL. We also perform experiments in which the novel classes are chosen from a separate dataset, which we call *cross-domain* FSL. In cross-domain FSL, the domain differences between the base and novel classes are much larger than the single-domain FSL. For both setting, we report 5-way-1-shot and 5-way-5-shot accuracies.

Datasets. For single-domain FSL, we run experiments on three datasets: *mini*ImageNet [40], *tiered*ImageNet [33], and Caltech-256 [14]. The *mini*ImageNet contains 100 classes randomly selected from ImageNet [7] with 600 images per class. We follow [32] to split the categories into 64 base, 16 validation, and 20 novel classes. The *tiered*ImageNet is another subset of ImageNet but has far more classes (608 classes). These classes are first divided into 34 groups and then further divided into 20 training groups (351 classes), 6 validation groups (97 classes), and 8 testing groups (160 classes), which ensure the distinction between training and testing sets. Caltech-256 (Caltech) has 30607 images from 256 classes. Following [6], we split it into 150, 56, and 50 classes for training, validation, and testing respectively.

For the cross-domain experiments, we construct a dataset that has high dissimilarity between base and novel classes by drawing the base classes from one dataset and the novel classes from another. We denote this dataset as '*mini*ImageNet&CUB', which is a combination of *mini*ImageNet and CUB-200-2011 (CUB) dataset [41]. CUB is a fine-grained image classification dataset including 200 bird classes and 11788 bird images. We follow [19] to split the categories into 100 base, 50 validation, and 50 novel classes. In *mini*ImageNet&CUB, the training set (base classes) contains 64 classes from *mini*ImageNet and the testing set (novel classes) contains 100 classes from CUB. Specifically, the 64 classes in the training set are the 64 base classes in *mini*ImageNet and the 100 classes in the test set are the 100 base classes in CUB.

Competitors. We compare our methods with the top few-shot learning methods: MetaOptNet [21], Distill [39], and Neg-Cosine [25]. We also compare with three transductive few-shot learning methods: ICI [42], TAFSSL [24], and EPNet [31]. TFSL methods have 100 unlabeled images per novel class by default. EPNet (full) and our UBC-TFSL uses all of the images of novel classes as unlabeled training samples.

Implementation details. Most of our settings are the same as [3]. We use a mini-batch size of 256 with 8 GPUs. We set the learning rate as 0.03 and use cosine annealing to decrese the learning rate. The feature dimension for contrastive loss is 128. The momentum for memory update is 0.5 and the temperature is set as 0.07. For *mini*ImageNet, *mini*ImageNet&CUB, and Caltech-256, we sample 2048 negative pairs in our contrastive loss and train 1000 epochs. For *tiered*ImageNet, we sample 20480 negative pairs and train 800 epochs.

Architecture. We use ResNet-12*, ResNet-12, ResNet-50, ResNet-101, and WRN-28-10 as our backbone architecture. ResNet-12* is a modified version of ResNet-12 and will be introduced in Sec.§ 4.2. WRN-28-10 [45] is a very wide version of ResNet-10 and have 36.5M parameters whereas ResNet-50 and ResNet-101 have 25.6M and

Table 1: **Top-1 accuracies(%) on *mini*ImageNet and *tiered*ImageNet.** We report the mean of 1000 randomly generated test episodes as well as the 95% confidence intervals. The top results are highlighted in blue and the second-best results in green. We provide results on Caltech-256 and *mini*ImageNet&CUB in the **supplementary**.

			***mini*ImageNet**		***tiered*ImageNet**	
setting	**method**	**backbone**	**1-shot**	**5-shot**	**1-shot**	**5-shot**
Non-transductive	MetaOptNet	ResNet-12*	62.6±0.6	78.6±0.4	65.9±0.7	81.5±0.5
	Distill	ResNet-12*	64.8±0.6	82.1±0.4	71.5±0.6	86.0±0.4
	Neg-Cosine	ResNet-12*	63.8±0.8	81.5±0.5	-	-
	Neg-Cosine	WRN-28-10	61.7±0.8	81.7±0.5	-	-
	UBC-FSL (Ours)	ResNet-12*	47.8±0.6	68.5±0.5	52.8±0.6	69.8±0.6
	UBC-FSL (Ours)	ResNet-12	56.9±0.6	76.5±0.4	58.0±0.7	76.3±0.5
	UBC-FSL (Ours)	ResNet-50	56.2±0.6	75.4±0.4	66.6±0.7	83.1±0.5
	UBC-FSL (Ours)	ResNet-101	57.5±0.6	77.2±0.4	68.0±0.7	84.3±0.5
	UBC-FSL (Ours)	WRN-28-10	57.1±0.6	76.5±0.4	67.5±0.7	83.9±0.5
	FSL baseline	ResNet-12*	61.7±0.7	79.4±0.5	69.6±0.7	84.2±0.6
	FSL baseline	ResNet-12	61.1±0.6	76.1±0.6	66.4±0.7	81.3±0.5
	FSL baseline	ResNet-50	61.3±0.6	76.0±0.4	69.4±0.7	83.3±0.5
	FSL baseline	ResNet-101	62.7±0.7	77.6±0.5	70.5±0.7	83.8±0.5
	FSL baseline	WRN-28-10	62.4±0.7	77.5±0.5	70.2±0.7	83.5±0.5
	Combined (Ours)	ResNet-12*	59.8±0.8	73.3±0.7	69.2±0.7	82.0±0.6
	Combined (Ours)	ResNet-12	63.8±0.7	79.9±0.6	67.8±0.7	83.0±0.5
	Combined (Ours)	ResNet-50	63.9±0.9	79.9±0.5	72.3±0.7	86.1±0.5
	Combined (Ours)	ResNet-101	65.6±0.6	81.6±0.4	73.5±0.7	86.7±0.5
	Combined (Ours)	WRN-28-10	65.2±0.6	81.2±0.4	73.1±0.7	86.4±0.5
Transductive	ICI	ResNet-12*	66.8±1.1	79.1±0.7	80.7 ±1.1	87.9±0.6
	ICI	ResNet50	60.2±1.1	75.2±0.7	78.6±1.1	86.8±0.6
	ICI	ResNet101	64.3±1.2	78.1±0.7	82.4±1.0	89.4±0.6
	TAFSSL	DenseNet	80.1±0.2	85.7±0.1	86.0±0.2	89.3±0.1
	EPNet	WRN-28-10	79.2±0.9	88.0±0.5	83.6±0.9	89.3±0.5
	EPNet (full)	WRN-28-10	80.2±0.8	88.9±0.5	84.8±0.8	89.9±0.6
	UBC-TFSL (Ours)	ResNet-12*	51.1±0.9	74.6±0.6	57.2±0.6	74.7±0.6
	UBC-TFSL (Ours)	ResNet-12	70.3±0.6	86.9±0.3	65.7±0.7	81.4±0.5
	UBC-TFSL (Ours)	ResNet-50	79.1±0.6	92.1±0.3	81.0±0.6	90.7±0.4
	UBC-TFSL (Ours)	ResNet-101	80.4±0.6	92.8±0.2	87.0±0.6	93.6±0.3
	UBC-TFSL (Ours)	WRN-28-10	80.3±0.6	92.4±0.2	85.7±0.6	93.0±0.3

44.4M parameters respectively.

4.1. Self-supervised learning can develop a strong inductive bias with no base-class labels

[36] shed light on improving few-shot learning with self-supervision and claim that "Self-supervision alone is not enough" for FSL. We agree there is still a gap between unlabeled-base-class few-shot learning and few-shot learning. However, in the transductive few-shot classification setting, we present the surprising result that **state-of-the-art performance can be obtained without using any labeled examples from the base classes at all.**

The results on *mini*ImageNet and *tiered*ImageNet are shown in Table 1. A better visualization is shown in Fig. 2. Results on Caltech-256 and *mini*ImageNet&CUB are provided in supplementary material. We notice that **(1) UBC-FSL shows some potential.** Even without any base-class labels, it only underperforms the state-of-the-art few-shot methods by $2-7\%$ in 1-shot and 5-shot accuracy on *mini*ImageNet and *tiered*ImageNet. **(2) There is great complementarity among supervised features and self-supervised features.** Combining supervised and self-supervised features ("Combined") beats the FSL baseline on all four datasets for all backbone networks. Specifically, it gives 4% and 2.9% improvements in 5-shot accuracy on *mini*ImageNet and *tiered*ImageNet when using ResNet-101. Also, it beats all other FSL competitors on *tiered*ImageNet. **(3) For the transductive few-shot classification setting, state-of-the-art can be obtained without actually using any labeled examples at all.** Even without any base-class labels, UBC-TFSL significantly surpasses all other methods. In Table 1, it outperforms all other TFSL methods by 3.5% and 3.9% for 5-shot accuracy on *mini*ImageNet and *tiered*ImageNet respectively. **(4) The FSL baseline struggles to learn a strong inductive bias with high dissimilarity between**

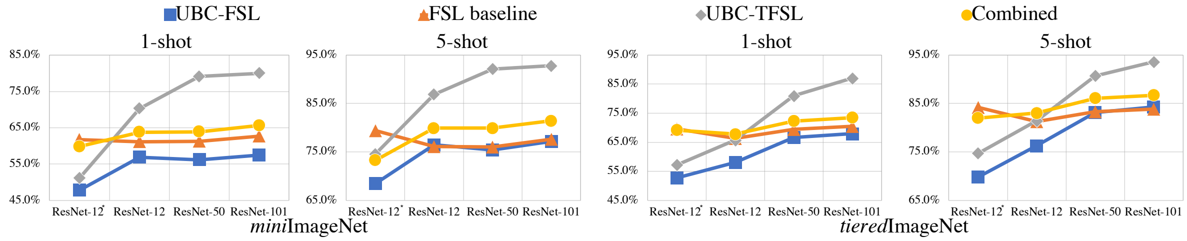


Figure 3: **Few-shot classification accuracy with various depths of backbone architectures.** Our UBC-FSL, FSL baseline, UBC-TFSL, and Combined have better performance with a deeper network. **The performance gain is relatively small for supervised features (FSL baseline) and large for self-supervised features (UBC-FSL), especially in a transductive setting (UBC-TFSL).**

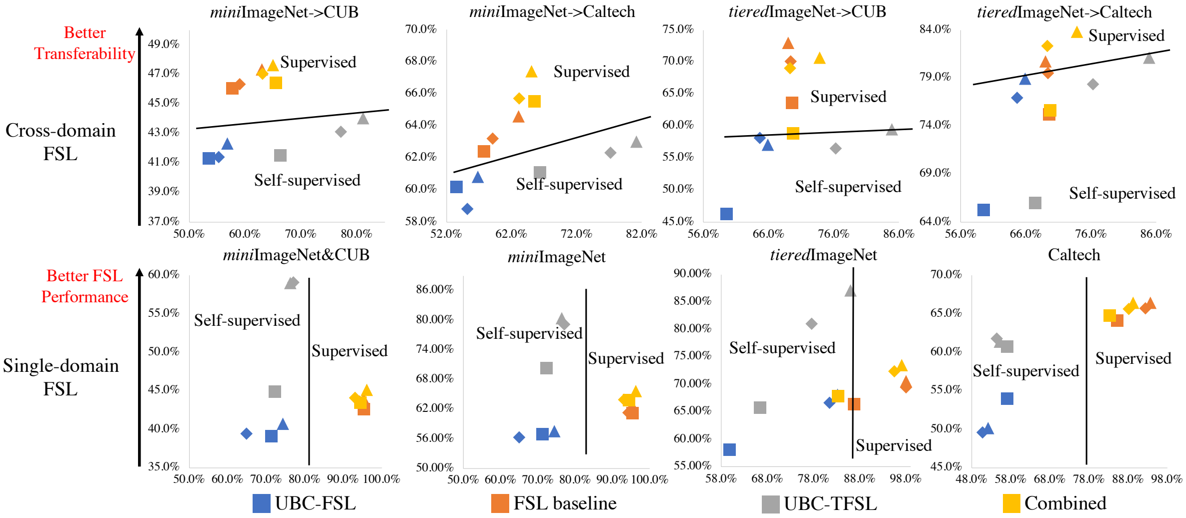


Figure 4: **Accuracy of 1-shot cross-domain FSL (first row) or single-domain FSL (second row).** First row: we visualize 1-shot test accuracy on the source dataset (x-axis) and the target dataset (y-axis). Second row: we visualize 1-shot accuracy on the base classes (x-axis) and the novel classes (y-axis). Squares, diamonds, and triangles denote ResNet-12, ResNet-50, and ResNet-101 respectively. We provide detailed statistics in the supplementary. From the first row, the results suggest that **supervised features are better when transferring to a new target dataset** even if self-supervised features (UBC-TFSL) have more training data and better performance on the source dataset. From the second row, the results suggest that in few-shot learning, even if supervised features have better performance on base classes, it underperforms self-supervised features (UBC-TFSL) on novel classes. It confirms that **UBC-TFSL benefit from the representation that better generalized on the novel classes**.

base and novel classes (cross-domain) whereas such dissimilarity has a relatively minor effect on UBC-TFSL. In *mini*ImageNet&CUB (please refer to supplementary), UBC-TFSL outperforms the FSL baseline by 15% and 13% for 1-shot and 5-shot accuracy respectively.

4.2. A deeper network is better

While deeper models generally have better performance for the standard classification task on both large (e.g, ImageNet [7]) and small dataset(e.g., CIFAR-10) as shown in [18], most previous few-shot methods [21, 39] report the best results with a modified version (ResNet-12*) of ResNet-12 [18]. ResNet-12* employs several modifications, including making it 1.25× wider, changing the input size from 224×224 to 84×84, using Leaky ReLU's instead of ReLU's, adding additional Dropblock layers [11], and removing the global pooling layer after the last residual block. We feel that the effect of different backbone architecture is not very clear in few-shot learning literature. We want to know if using a very deep network (e.g., ResNet-101) can bring significant improvement in few-shot classification as in the standard classification task. More importantly, we want to explore the differences between supervised and self-supervised features when using various backbone architec-

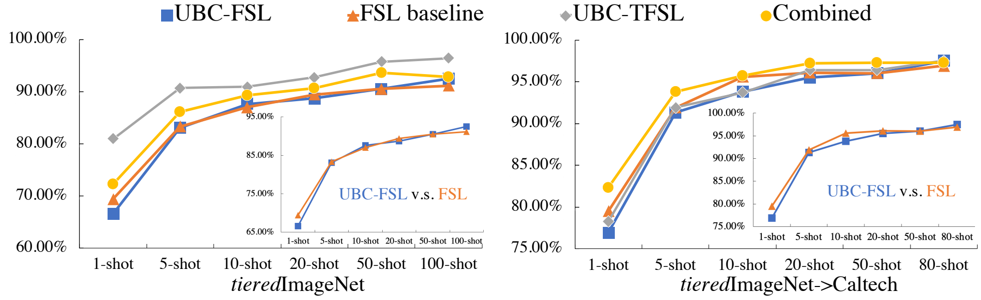


Figure 5: **Few-shot classification accuracy with larger shots.** We use ResNet-50 as our backbone architecture and evaluate on *tiered*ImageNet and Caltech (transferred from *tiered*ImageNet). Training on same data, supervised features (FSL baseline) outperform self-supervised features (UBC-FSL) in few-shot setting. However, self-supervised features are better when large numbers (100-shot) of labeled novel examples are given.

tures in few-shot learning.

As shown in Table 1, we report results using ResNet-12*, ResNet-12, ResNet-50, ResNet-101, and WRN-28-10. To better compare the effect of depth of backbone architecture, we visualize the performance in Fig. 3. We notice that (1) ResNet-101 have the best performance and all our baselines benefit from deeper network in most cases. (2) The commonly used ResNet-12* work well for FSL baseline but do not suit for self-supervised learning based baselines. (3) The wide network WRN-28-10 has very good performance on all our baselines, only slightly underperform ResNet-101. We confirm that few-shot learning can actually benefit from a deeper or wider backbone architecture. **The performance gain is small for supervised features (FSL baseline) and large for self-supervised features (UBC-FSL), especially in a transductive setting (UBC-TFSL).**

4.3. Supervised vs. self-supervised features in cross-domain FSL

Another interesting question is whether models learned in a single domain can perform well in a new domain (with highly dissimilar classes). To study this, we conduct cross-domain FSL, in which we learn models on *mini*ImageNet or *tiered*ImageNet and evaluate our models on Caltech-256 and CUB. Specifically, the FSL baseline and UBC-FSL are trained on base classes of the source dataset, and UBC-TFSL are trained on both base and novel classes of the source dataset. Then, we evaluate our methods on the testing set of target datasets (Caltech-256 and CUB).

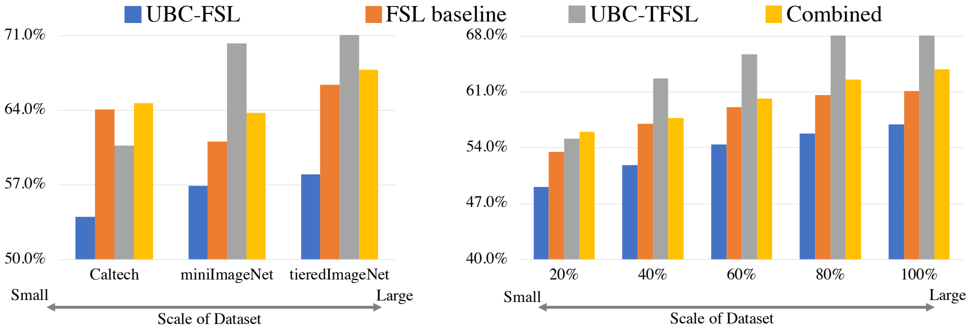
Notice that the way we are applying the UBC-TFSL model, it does not qualify as a true transductive setting, since the model does not have access to unlabeled data from the testing set. Instead, we are testing whether this model can improve its performance on cross-domain classes with unlabeled data from **additional** classes in the source data set.

Previous work [17] compares supervised and self-supervised features when transferring to a new domain for classification, object detection, and instance segmentation. It shows that self-supervised features have better transferability for these tasks. However, the conclusion is based on when large numbers of labeled examples are used to learn the final linear classifier. In few-shot setting, we show that **supervised features do better than self-supervised features**.

In the first row of Fig. 4, we compare UBC-FSL, FSL baseline, UBC-TFSL, and Combined in cross-domain FSL. The x-axis and y-axis denote the 1-shot testing accuracy on the source and target dataset respectively. Surprisingly, supervised features (FSL baseline, Combined) significantly outperform self-supervised features (UBC-FSL, UBC-TFSL) on the target dataset even if they have lower accuracy on the source dataset. In the second row of Fig. 4, we visualize the performance of our methods on base and novel classes in single-domain FSL. The x-axis and y-axis denote the 1-shot accuracy on base and novel classes respectively. As you can see, UBC-TFSL (gray points) outperforms FSL baseline (orange) on novel classes but underperforms on base classes. These experiments show that UBC-TFSL has mediocre performance when it does **not** have access to unlabeled data from the test classes, but performs extremely well when it does. In other words, it is not simply access to additional unlabeled data that helps, but rather, data from the test classes themselves.

4.4. Supervised vs. self-supervised features with larger shots

In Fig. 5, we compare UBC-FSL, the FSL baseline, UBC-TFSL and Combined with larger shots using ResNet-50 on *tiered*ImageNet and *tiered*ImageNet-Caltech (cross-domain FSL). For 1-shot learning, there is a large gap around 5% between UBC-FSL and the FSL baseline. How-



(a) (b)

Figure 6: **1-shot testing accuracy under various scales of dataset size.** ResNet-12 is our backbone architecture. In (a), we compare UBC-FSL, FSL baseline, UBC-TFSL, and Combined on three datasets of different sizes (30607, 60000, and 779165 images). In (b), we randomly select part of *mini*ImageNet (e.g., 20% of the whole dataset) and compare our methods.

ever, as the shots become larger, this gap gradually diminishes. For 100-shot on *tiered*ImageNet and 80-shot on Caltech, UBC-FSL even outperforms the FSL baseline by 1.3% and 0.6% respectively.

We suggest that **supervised features may contain higher-level semantic concepts that are easier to digest with a few training instances** while self-supervised features have better transferability with abundant training data. This statement is compatible with previous work [17], which use abundant labeled data to learn the final classification layer and claims that self-supervised features have better transferability.

4.5. Supervised vs. self-supervised features and dataset size

In this section, we compare supervised and self-supervised features under various dataset sizes. We conduct experiments on Caltech, *mini*ImageNet, and *tiered*ImageNet, which have 30607, 60000, and 779165 images respectively. We also randomly select subsets of *mini*ImageNet (20%, 40%, 60%, 80%, and 100%) and report the 1-shot accuracy. An equal portion of examples from each class are randomly selected. As shown in Fig. 6, self-supervised features (UBC-TFSL) significantly outperform other methods with a big dataset. However, when the dataset is small (e.g., Caltech-256 and 20% of *mini*ImageNet), it is overtaken by the FSL baseline. This result suggests that **supervised features are more robust to dataset size.**

4.6. Comparing different self-supervised methods

As shown in Table 2, we compare three different instance discrimination methods to learn the feature embedding. Here we compare MoCo-v2 [3], CMC [38], and SimCLR [2]. From the results, we can see that all these self-supervised methods can learn a powerful inductive bias, especially in the transductive setting, suggesting that most self-supervised methods can be generalized to learn a good embedding for few-shot learning.

Table 2: **Few-shot classification accuracy with different self-supervised methods.** We run experiments using MoCo-v2 [3], CMC [38], and SimCLR [2] as our self-supervised methods to learn the feature embedding. The top results are highlighted in blue and the second-best results in green.

		***mini*ImageNet**	
method	**backbone**	**1-shot**	**5-shot**
UBC-FSL (MoCo-v2)	ResNet-101	57.5±0.6	77.2±0.4
UBC-FSL (CMC)	ResNet-101	56.9±0.6	76.9±0.5
UBC-FSL (SimCLR)	ResNet-101	57.6±0.7	76.7±0.6
UBC-TFSL (MoCo-v2)	ResNet-101	80.4±0.6	92.8±0.2
UBC-TFSL (CMC)	ResNet-101	79.7±0.6	92.1±0.3
UBC-TFSL (SimCLR)	ResNet-101	79.5±0.7	92.2±0.3

5. Conclusion

Most previous FSL methods borrow a strong inductive bias from the supervised learning of base classes. In this paper, we show that no base class labels are needed to develop such an inductive bias and that self-supervised learning can provide a powerful inductive bias for few-shot learning. We examine the role of features learned through self-supervision in few-shot learning through comprehensive experiments.

References

[1] Peyman Bateni, Raghav Goyal, Vaden Masrani, Frank Wood, and Leonid Sigal. Improved few-shot visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning

of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 2, 8

[3] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 3, 4, 8

[4] Zitian Chen, Yanwei Fu, Kaiyu Chen, and Yu-Gang Jiang. Image block augmentation for one-shot learning. In *Association for the Advancement of Artificial Intelligence (AAAI)*, volume 33, pages 3379–3386, 2019. 1, 2

[5] Zitian Chen, Yanwei Fu, Yu-Xiong Wang, Lin Ma, Wei Liu, and Martial Hebert. Image deformation meta-networks for one-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8680–8689, 2019. 2

[6] Zitian Chen, Yanwei Fu, Yinda Zhang, Yu-Gang Jiang, Xiangyang Xue, and Leonid Sigal. Multi-level semantic feature augmentation for one-shot learning. *IEEE Transactions on Image Processing*, 28(9):4594–4605, 2019. 2, 4

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 2, 4, 6

[8] Guneet Singh Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. A baseline for few-shot image classification. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019. 1, 2

[9] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1422–1430, 2015. 2

[10] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1126–1135, 2017. 2

[11] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Dropblock: A regularization method for convolutional networks. In *Advances in Neural Information Processing Systems (NeurIps)*, 2018. 6

[12] Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Pérez, and Matthieu Cord. Boosting few-shot visual learning with self-supervision. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 8059–8068, 2019. 2

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

[14] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007. 4

[15] 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. 2

[16] Bharath Hariharan and Ross Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3018–3027, 2017. 2

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2020. 2, 3, 7, 8

[18] 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 (CVPR)*, pages 770–778, 2015. 2, 6

[19] Nathan Hilliard, Lawrence Phillips, Scott Howland, Artëm Yankov, Courtney D Corley, and Nathan O Hodas. Few-shot learning with metric-agnostic conditional embeddings. *arXiv preprint arXiv:1802.04376*, 2018. 4

[20] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6874–6883, 2017. 2

[21] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10657–10665, 2019. 2, 4, 6

[22] Aoxue Li, Weiran Huang, Xu Lan, Jiashi Feng, Zhenguo Li, and Liwei Wang. Boosting few-shot learning with adaptive margin loss. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12576–12584, 2020. 2

[23] Xinzhe Li, Qianru Sun, Yaoyao Liu, Qin Zhou, Shibao Zheng, Tat-Seng Chua, and Bernt Schiele. Learning to self-train for semi-supervised few-shot classification. In *Advances in Neural Information Processing Systems (NeurIps)*, 2019. 1, 2

[24] Moshe Lichtenstein, Prasanna Sattigeri, Rogerio Feris, Raja Giryes, and Leonid Karlinsky. Tafssl: Task-adaptive feature sub-space learning for few-shot classification. *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 522–539, 2020. 1, 2, 4

[25] Bin Liu, Yue Cao, Yutong Lin, Qi Li, Zheng Zhang, Mingsheng Long, and Han Hu. Negative margin matters: Understanding margin in few-shot classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 438–455, 2020. 2, 4

[26] Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sung Ju Hwang, and Yi Yang. Learning to propagate labels: Transductive propagation network for few-shot learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. 1, 2

[27] Erik G. Miller, Nicholas E. Matsakis, and Paul A. Viola. Learning from one example through shared densities on transforms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 464–471, 2000. 2

[28] Tsendsuren Munkhdalai and Hong Yu. Meta networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 2554–2563, 2017. 2

[29] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 69–84, 2016. 2

[30] Mehdi Noroozi, Hamed Pirsiavash, and Paolo Favaro. Representation learning by learning to count. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, pages 5898–5906, 2017. 2

[31] Rodríguez Pau, Laradji Issam, Drouin Alexandre, and Lacoste Alexandre. Embedding propagation: Smoother manifold for few-shot classification. *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 121–138, 2020. 1, 2, 4

[32] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017. 2, 4

[33] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. 4

[34] Eli Schwartz, Leonid Karlinsky, Joseph Shtok, Sivan Harary, Mattias Marder, Abhishek Kumar, Rogerio Feris, Raja Giryes, and Alex Bronstein. Delta-encoder: An effective sample synthesis method for few-shot object recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2

[35] Jake Snell, Kevin Swersky, and Richard S. Zemeln. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2

[36] Jong-Chyi Su, Subhransu Maji, and Bharath Hariharan. When does self-supervision improve few-shot learning? *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 645–666, 2020. 2, 5

[37] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H.S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1199–1208, 2018. 2

[38] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2, 8

[39] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? *arXiv preprint arXiv:2003.11539*, 2020. 2, 4, 6

[40] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 2, 3, 4

[41] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 4

[42] Yikai Wang, Chengming Xu, Chen Liu, Li Zhang, and Yanwei Fu. Instance credibility inference for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12836–12845, 2020. 2, 4

[43] Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7278–7286, 2018. 2

[44] 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 (CVPR)*, pages 3733–3742, 2018. 2

[45] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2016. 4

[46] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 649–666, 2016. 2