SHOT IN THE DARK: FEW-SHOT LEARNING WITH NO BASE-CLASS LABELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-shot learning aims to learn classifiers for new objects from a small number of labeled examples. But it does not do this in a vacuum. Usually, a strong inductive bias is borrowed from the supervised learning of *base classes*. This inductive bias enables more statistically efficient learning of the new classes. In this work, we show that *no labels are needed* to develop such an inductive bias, and that *self-supervised learning* can provide a powerful inductive bias for few-shot learning. This is particularly effective when the unlabeled data for learning such a bias contains not only examples of the base classes, but also examples of the novel classes. The setting in which unlabeled examples of the novel classes are available is known as the *transductive setting.* Our method outperforms state-of-the-art few-shot learning methods, including other transductive learning methods, by 3.9% for 5-shot accuracy on *mini*ImageNet **without using any base class labels**. By benchmarking unlabeled-base-class (UBC) few-shot learning and UBC transductive few-shot learning, we demonstrate the great potential of self-supervised feature learning: self-supervision alone is sufficient to create a remarkably good inductive bias for few-shot learning. This motivates a rethinking of whether base-class labels are necessary at all for few-shot learning. We also explore the relationship between self-supervised features and supervised features, comparing both their transferability and their complementarity in the non-transductive setting. By combining supervised and self-supervised features learned from base classes, we also achieve a new state-of-the-art in the non-transductive setting, outperforming all previous methods.

1 INTRODUCTION

Deep architectures have achieved significant success in various vision tasks including image classification and object detection. Such success has 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.

In this paper, we demonstrate an extremely simple method to address this challenge. By conducting *self-supervised* learning on unlabeled examples from both base and novel classes, our method can develop an inductive bias that directly incorporates information about the novel classes, reducing the sensitivity to the distance between base and novel classes. This method significantly outperforms all previous FSL methods including transductive few-shot learning (TFSL) methods. Moreover, it does this *without using any base class labels*, which motivates a rethinking of whether base-class labels are necessary at all for few-shot learning.

We summarize our contributions as follows.

- We benchmark classical few-shot learning, unlabeled-base-class few-shot learning (UBC-FSL), and unlabeled-base-class transductive few-shot learning (UBC-TFSL). The UBC learning shows competitive performance against labeled-base-class learning in the few-shot setting. In the transductive setting, our simple method easily beats all previous FSL methods, including TFSL. We demonstrate that self-supervision alone is sufficient for strong few-shot learning.
- While most FSL methods focus on optimizing their inductive biases on a validation set of classes, often requiring the addition of complex architectural modifications, we show that similar improvements can be realized simply by using a deeper network. Our FSL, UBC-FSL, and UBC-TFSL methods all benefit from deeper networks, suggesting future FSL methods can achieve high performance on standard architectures.
- In the non-transductive setting, we explore the complementarity between supervised features and unsupervised features. By combining them, we also reach a new state-of-the-art in the non-transductive setting.
- We compare the effectiveness of supervised features and self-supervised features for transfer learning to a new set of classes. We consider this problem for conditions under which the novel classes have both abundant labeled data, and also in the few-shot paradigm, where the novel classes have only a few labeled examples. While some work shows that self-supervised features are better at transferring to novel classes, our results support the conclusion that for few-shot learning, supervised features do better than self-supervised features when transferring to a new set of classes.
- We compare FSL and UBC-FSL across training set sizes. Let N be the number of training examples per novel class. Even though FSL significantly outperforms UBC-FSL for small N, this advantage diminishes as N grows, and UBC-FSL eventually overtakes FSL as N gets large, suggesting that supervised features contain higher-level semantic concepts that is easier to incorporate with a few training instances while self-supervised features are better at transferring with abundant labeled examples from novel classes.

2 Related Work

Few-shot Learning. Few-shot learning is a classic problem (Miller et al., 2000), 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 (Wang et al., 2018; Chen et al., 2019d; Schwartz et al., 2018), hallucinate (Hariharan & Girshick, 2017) or deform (Miller et al., 2000; Chen et al., 2019c) images to generate additional examples to address the training data scarcity. Meta-learning (Finn et al., 2017; Ravi & Larochelle, 2017; Munkhdalai & Yu, 2017; Wang & Hebert, 2016; Lee et al., 2019) attempts to learn a parameterized mapping from limited training examples to hidden parameters that accelerate or improve the optimization procedure. Metric learning (Bateni et al., 2020; Li et al., 2020; Sung et al., 2018) aims at learning a transferable metric space (or embedding). MatchingNet (Vinyals et al., 2016) and ProtoNet (Snell et al., 2017) adopt cosine distance and Euclidean distance to separate instances belonging to different classes. Recently, Chen et al. (2019d); Liu et al. (2020); Tian et al. (2020b) showed that learning a classifier on top of supervised features can achieve surprisingly competitive performance. Most FSL methods use relatively small networks, e.g., a modified version of ResNet-12 (He et al., 2015), and adapt the inductive bias learned from base classes. No previous work presents their best results on very deep networks (e.g. ResNet-101), and Chen et al. (2019a) (in their Table A3) show that deeper networks perform worse for some FSL methods. However, all our methods achieve additional gain and the top results by using a very deep network, which will be further discussed in § 4.2.

Transductive Few-shot Learning. TFSL methods use the distribution support of unlabeled novel instances to help few-shot learning. Liu et al. (2018), Wang et al. (2020b), and Li et al. (2019) exploit unlabeled instances with high confidence to train the model. Chen et al. (2019b) propose a data augmentation method to directly mix base examples and selected novel examples in the image domain to learn generalized features. Dhillon et al. (2019), Pau et al. (2020), and Lichtenstein et al. (2020) take unlabeled testing instances to acquire an auxiliary loss serving as a regularizer to adapt the inductive bias. However, in this procedure, a strong inductive bias is first required for clustering

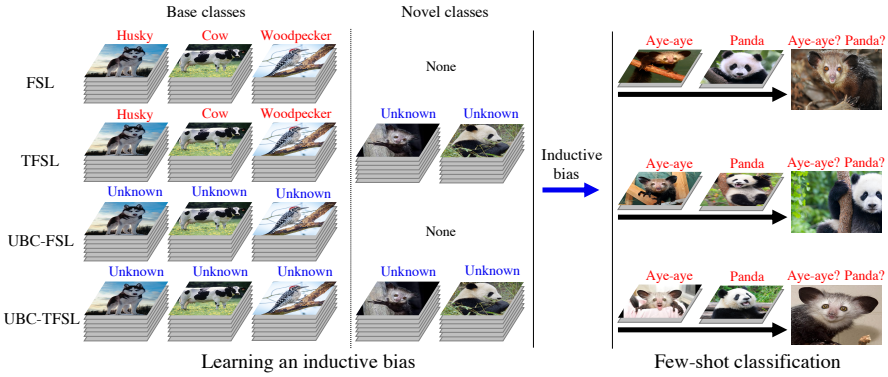

Figure 1: **An illustration of different few-shot learning settings.** The difference between these settings is whether they have labels for examples from base classes and unlabeled examples from novel classes.

or predicting the unlabeled instances. As a result, these previous works still heavily rely on base-class labels and supervised learning, whereas our work can directly develop a strong inductive bias from unlabeled instances.

Self-supervised Learning. Self-supervised learning aims to explore the internal data distribution and learns discriminative features without annotations. Some work takes predicting rotation (Gidaris et al., 2018), counting (Noroozi et al., 2017), predicting the relative position of patches (Doersch et al., 2015), colorization (Zhang et al., 2016; Larsson et al., 2017), and solving jigsaw puzzles (Noroozi & Favaro, 2016) as self-supervised tasks to learn representations. Recently, instance discrimination (Wu et al., 2018) has attracted much attention. Tian et al. (2020a) and Bachman et al. (2019) maximize the mutual information between different views. MoCo (He et al., 2020) proposes a momentum contrast to update models and shows superior performance to supervised learning when transferring to downstream tasks, including detection and segmentation. Su et al. (2020) and Gidaris et al. (2019) use self-supervision to help few-shot learning. However, while Su et al. (2020) claim that self-supervision alone is not enough for few-shot learning, we demonstrate that self-supervision alone is particularly effective when unlabeled examples from not only base classes, but also novel classes are provided.

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 labeled examples from base classes so the training set is $D_{FSL} = D_{base}$. For TFSL, we are given $D_{TFSL} = D_{base} \cup U_{novel}$. For UBC-FSL, we have $D_{UBC-FSL} = U_{base}$. For UBC-TFSL, we denote the training set as $D_{UBC-TFSL} = U_{base} \cup U_{novel}$.

These four few-shot learning settings use the same evaluation protocol as in previous works (Vinyals et al., 2016). At inference time, we are given a collection of *N-way-m-shot* classification tasks sampled from D_{novel} to evaluate our method.

3.1 SELF-SUPERVISED LEARNING

Here we use a contrastive loss to do instance discrimination as our self-supervision task. We follow momentum contrast (He et al., 2020), 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):

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

where τ is a temperature hyper-parameter. Since our implementations are based on MoCo-v2 (Chen et al., 2020), please refer to it for further details.

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. We learn a logistic regression classifier on top of the learned feature embedding of $N * m$ training examples and report its accuracy on 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.

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 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. This 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.

Datasets. For single-domain FSL, we conduct experiments on three commonly used datasets: *mini*ImageNet (Vinyals et al., 2016), *tiered*ImageNet (Ren et al., 2018), and Caltech-256 (Griffin et al., 2007). The *mini*ImageNet contains 100 classes randomly selected from ImageNet (Deng et al., 2009) with 600 images per class. We follow Ravi & Larochelle (2017) 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 Chen et al. (2019d), 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 (Wah et al., 2011). CUB is a fine-grained image classification dataset including 200 bird classes and 11788 bird images. We follow Hilliard et al. (2018) 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

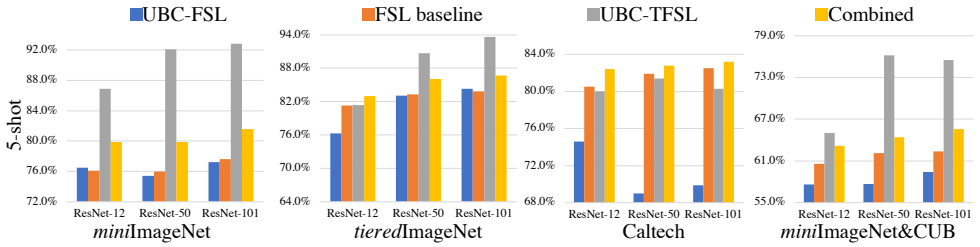

Figure 2: **A comparison between UBC-FSL, FSL baseline, UBC-TFSL, and Combined for 5-shot accuracy on four datasets.** UBC-TFSL significantly outperforms other methods.

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 (Lee et al., 2019), Distill (Tian et al., 2020b), Closer (Chen et al., 2019a), and Neg-Cosine (Liu et al., 2020). We also compare with three transductive few-shot learning methods: ICI (Wang et al., 2020a), TAFSSL (Lichtenstein et al., 2020), and EPNet (Pau et al., 2020). TFSL methods have 100 unlabeled images per novel class by default. EPNet (full)[1] and our UBC-TFSL uses all of the images of novel classes as unlabeled training samples.

Implementation setup. For training details of our methods, please refer to appendix A2.

4.1 Self-supervision alone is enough

Su et al. (2020) shed light on improving few-shot learning with self-supervision and claim that "Self-supervision alone is not enough" for FSL. However, we come to the opposite conclusion: **self-supervised learning alone is enough to develop a strong inductive bias**.

An intuitive comparison of our FSL baseline, UBC-FSL, UBC-TFSL, and Combined methods is shown in Fig. 2. The results on *mini*ImageNet and *tiered*ImageNet are shown in Table 1. (Please refer to Table A1 for results on Caltech-256 and *mini*ImageNet&CUB.) We notice that **(1) UBC-FSL shows some potential.** Even without any base-class labels, it only underperforms the FSL baseline by $4-7\%$ 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) Even without any base-class labels, UBC-TFSL significantly surpasses all other methods, supporting our claim that "self-supervised features alone are enough".** When using the deepest network ResNet-101, it outperforms the FSL baseline by about 10% for both 1-shot and 5-shot accuracy on *mini*ImageNet and *tiered*ImageNet. 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 base and novel classes (cross-domain) whereas such dissimilarity has a relatively minor effect on UBC-TFSL.** In *mini*ImageNet&CUB, 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

Most top FSL methods (Lee et al., 2019; Liu et al., 2020; Tian et al., 2020b) use shallow networks with low input resolution as they achieves the best performance. They manually modify the plain ResNet-12 with several tricks, including making it $1.25\times$ wider, changing the input size from 224×224 to 84×84, using Leaky ReLU's instead of ReLU's, adding additional Dropblock layers (Ghiasi et al., 2018), and removing the global pooling layer after the last residual block. This modified architecture is referred to as 'ResNet-12*'. What's more, Chen et al. (2019a) show that ResNet-10

[1]We implement EPNet (full) using code available at https://github.com/ElementAI/embedding-propagation.

setting	method	backbone	*mini*ImageNet 1-shot	*mini*ImageNet 5-shot	*tiered*ImageNet 1-shot	*tiered*ImageNet 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
	Closer	ResNet-10	53.9±0.7	75.9±0.6	-	-
	Closer	ResNet-18	51.8±0.7	75.6±0.6	-	-
	Closer	ResNet-34	52.6±0.8	76.1±0.6	-	-
	Neg-Cosine	ResNet-12*	63.8±0.8	81.5±0.5	-	-
	Neg-Cosine	ResNet-18	62.3±0.8	80.9±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
	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
	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
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

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. Please refer to Table A1 for results on Caltech-256 and *mini*ImageNet&CUB. The top results are highlighted in blue and the second-best results in green.

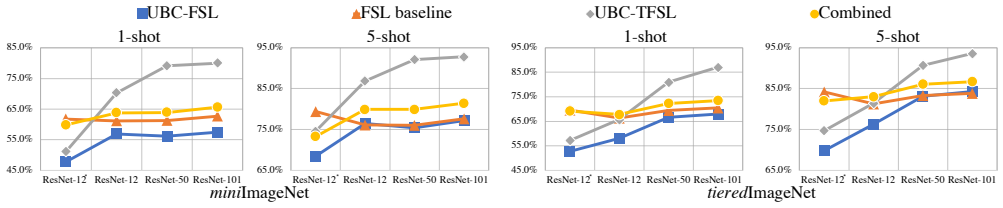

Figure 3: **Few-shot classification accuracy with various depths of backbone architectures.**

outperforms ResNet-34 for their methods and other methods (Finn et al., 2017; Vinyals et al., 2016; Sung et al., 2018) (in their Table A3). Also, Closer and ICI are worse for deeper networks (Table 1).

However, unlike previous methods based on smaller networks with bags of tricks, **we show that a plain deep network (ResNet-101) achieves top performance without bells and whistles**. Our UBC-FSL, FSL baseline, UBC-TFSL, and Combined all benefit from using deeper networks. As shown in Fig. 3 and Table 1, for UBC-FSL, FSL baseline, UBC-TFSL, and Combined, ResNet-101 significantly outperforms ResNet-12 by 10.0%, 4.1%, 21.3%, 5.7% respectively for 1-shot accuracy on *tiered*ImageNet. The plain ResNet-101 also beats ResNet-12* by 1.0% and 0.9% for 1-shot accuracy on *mini*ImageNet and *tiered*ImageNet. Since deeper networks provide better inductive bias, **we hope our result will encourage future work on standard deep networks. This can save effort in modifying architectures and avoid developing methods that are limited to shallow networks.**

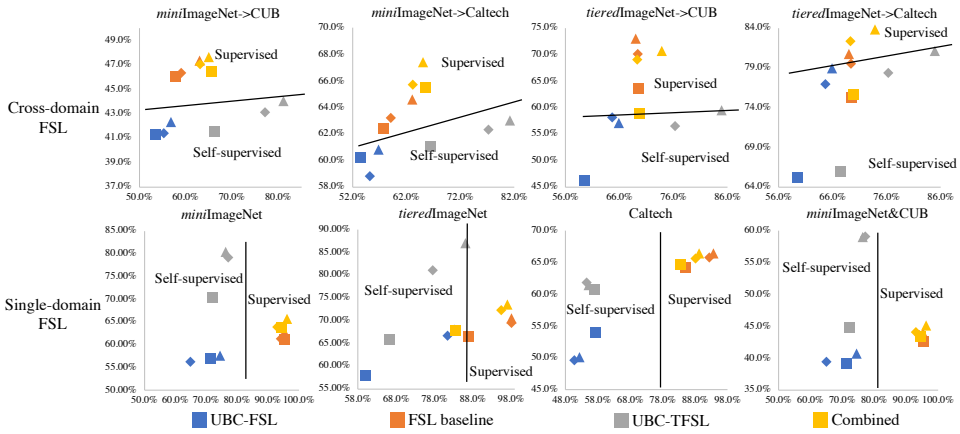

Figure 4: **Accuracy of 1-shot cross-domain FSL (first row) or single-domain FSL (second row).** First row: we visualize 1-shot test accuracy on the source dataset (x-axis) and the target dataset (y-axis). Second row: we visualize 1-shot accuracy on base categories (x-axis) and novel categories (y-axis). Squares, diamonds, and triangles denote ResNet-12, ResNet-50, and ResNet-101 respectively. Please refer to Table A2, A3 for detailed statistics.

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 in this case, 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 (He et al., 2020) 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, in this section, we show that **supervised features do better than self-supervised features in cross-domain FSL**.

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. However, 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.

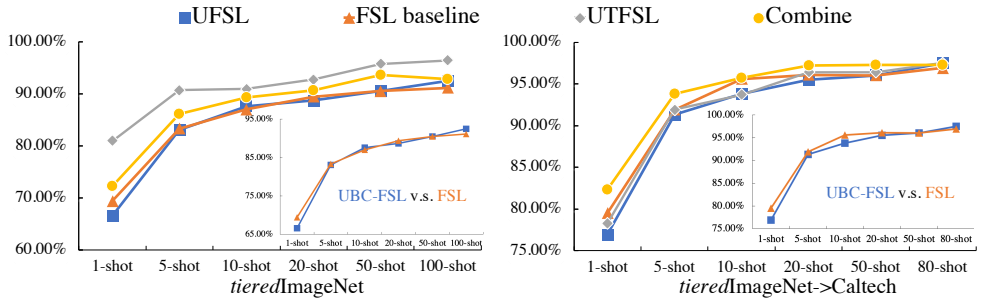

Figure 5: **Few-shot classification accuracy with larger shots.** We use ResNet-50 as our backbone architecture and evaluate on *tiered*ImageNet and Caltech (transferred from *tiered*ImageNet).

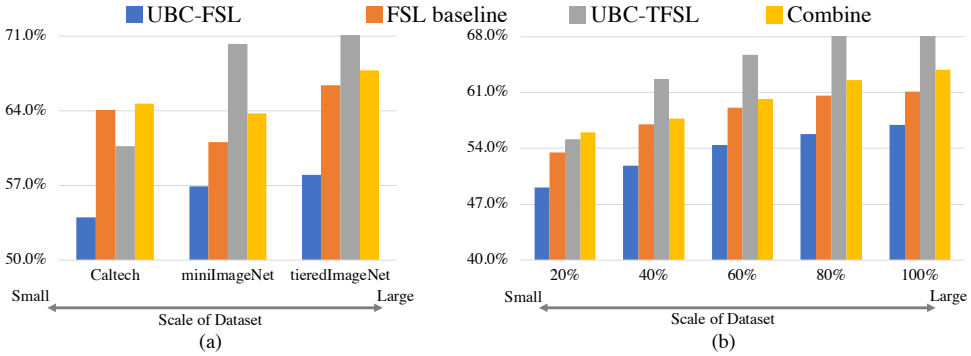

Figure 6: **1-shot testing accuracy under various scales of dataset size.** ResNet-12 is our backbone architecture. In (a), we compare UBC-FSL, FSL baseline, UBC-TFSL, and Combined on three datasets of different sizes (30607, 60000, and 779165 images). In (b), we randomly select part of the *mini*ImageNet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

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 (He et al., 2020), which claims that self-supervised features have better transferability and motivates us to further rethink what supervised and self-supervised features learn.

4.5 Supervised vs. self-supervised features and dataset size

In this section, we want to compare supervised and self-supervised features under various scales of dataset size. We conduct experiments on Caltech, *mini*ImageNet, and *tiered*ImageNet, which have 30607, 60000, and 779165 images respectively. We also randomly select only part of the *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 may be more robust to dataset size.**

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. This work lays out important directions for the next few years: using self-supervised learning to develop strong inductive biases and improving few-shot learning based on this strong inductive bias.

REFERENCES

Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

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.

Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *International Conference on Learning Representations (ICLR)*, 2019a.

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

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)*, 2019b.

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)*, 2019c.

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*, 2019d.

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)*, 2009.

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.

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)*, 2015.

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)*, 2017.

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.

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.

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)*, 2019.

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

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)*, 2017.

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)*, 2015.

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)*, 2020.

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.

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)*, 2017.

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)*, 2019.

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)*, 2020.

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.

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)*, 2020.

Bin Liu, Yue Cao, Yutong Lin, Qi Li, Zheng Zhang, Mingsheng Long, and Han Hu. Negative margin matters: Understanding margin in few-shot classification. 2020.

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. 2018.

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)*, 2000.

Tsendsuren Munkhdalai and Hong Yu. Meta networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

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

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)*, 2020.

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

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.

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.

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.

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)*, 2020.

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)*, 2018.

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

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*, 2020b.

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.

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

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)*, 2020a.

Yikai Wang, Chengming Xu, Chen Liu, Li Zhang, and Yanwei Fu. Instance credibility inference for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020b.

Yu-Xiong Wang and Martial Hebert. Learning to learn: Model regression networks for easy small sample learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

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)*, 2018.

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)*, 2018.

Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

APPENDIX

A1 RESULTS ON CALTECH-256 AND MINIIMAGENET&CUB

We report our results on Caltech-256 and miniImageNet&CUB in Table A1.

setting	method	backbone	Caltech 1-shot	Caltech 5-shot	*mini*ImageNet&CUB 1-shot	*mini*ImageNet&CUB 5-shot
Non-transductive	UBC-FSL (Ours)	ResNet-12*	48.7±0.6	68.9±0.6	36.0±0.5	54.3±0.5
	UBC-FSL (Ours)	ResNet-12	54.0±0.6	74.6±0.5	39.1±0.6	57.6±0.5
	UBC-FSL (Ours)	ResNet-50	49.6±0.7	69.0±0.5	39.4±0.6	57.7±0.5
	UBC-FSL (Ours)	ResNet-101	50.1±0.6	69.9±0.5	40.7±0.6	59.4±0.6
	FSL baseline	ResNet-12*	65.7±0.6	81.5±0.5	42.8±0.5	60.9±0.6
	FSL baseline	ResNet-12	64.1±0.6	80.5±0.6	42.6±0.6	60.6±0.5
	FSL baseline	ResNet-50	65.7±0.7	81.9±0.3	43.6±0.6	62.1±0.5
	FSL baseline	ResNet-101	66.4±0.6	82.5±0.4	43.9±0.6	62.4±0.6
	Combined (Ours)	ResNet-12*	65.4±0.6	82.7±0.5	42.9±0.5	61.7±0.7
	Combined (Ours)	ResNet-12	64.7±0.6	82.4±0.4	43.4±0.6	63.2±0.5
	Combined (Ours)	ResNet-50	65.6±0.6	82.8±0.4	44.1±0.6	64.4±0.5
	Combined (Ours)	ResNet-101	66.5±0.5	83.2±0.4	45.1±0.6	65.6±0.5
Transductive	UBC-TFSL (Ours)	ResNet-12*	56.4±0.6	74.8±0.6	39.7±0.4	58.9±0.5
	UBC-TFSL (Ours)	ResNet-12	60.7±0.7	80.0±0.5	44.9±0.6	65.0±0.6
	UBC-TFSL (Ours)	ResNet-50	61.8±0.6	81.4±0.5	59.1±0.8	76.2±0.6
	UBC-TFSL (Ours)	ResNet-101	61.4±0.6	80.3±0.5	59.0±0.8	75.5±0.6

Table A1: **Top-1 accuracies(%) on Caltech-256 and *mini*ImageNet&CUB.** 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.

A2 IMPLEMENTATION DETAILS

Most of our settings are the same as Chen et al. (2020). 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. For *tiered*ImageNet, we sample 20480 negative pairs. We train 1000, 300, 1000, and 800 epoches for *mini*ImageNet, *tiered*ImageNet, *mini*ImageNet&CUB, and Caltech-256 respectively.

A3 RESULTS FOR CROSS-DOMAIN FSL

method	backbone	*mini*ImageNet→Caltech 1-shot	*mini*ImageNet→Caltech 5-shot	*mini*ImageNet→CUB 1-shot	*mini*ImageNet→CUB 5-shot
UBC-FSL (Ours)	ResNet-12	41.3±0.5	59.1±0.6	60.2±0.7	80.1±0.4
UBC-FSL (Ours)	ResNet-50	41.4±0.6	58.5±0.6	58.8±0.6	79.0±0.5
UBC-FSL (Ours)	ResNet-101	42.3±0.5	59.9±0.6	60.8±0.6	80.7±0.4
FSL baseline	ResNet-12	46.0±0.6	63.7±0.5	62.4±0.6	79.1±0.4
FSL baseline	ResNet-50	46.3±0.6	64.9±0.5	63.2±0.8	79.9±0.5
FSL baseline	ResNet-101	47.3±0.6	65.6±0.5	64.6±0.7	81.1±0.5
Combined (Ours)	ResNet-12	46.4±0.6	65.1±0.5	65.5±0.6	83.0±0.4
Combined (Ours)	ResNet-50	47.0±0.4	66.3±0.5	65.7±0.8	83.2±0.4
Combined (Ours)	ResNet-101	47.6±0.6	67.3±0.5	67.4±0.5	84.5±0.4
UBC-TFSL (Ours)	ResNet-12	41.5±0.5	59.2±0.6	61.1±0.6	81.1±0.5
UBC-TFSL (Ours)	ResNet-50	43.1±0.5	61.0±0.7	62.3±0.6	82.8±0.4
UBC-TFSL (Ours)	ResNet-101	44.0±0.6	61.7±0.6	63.0±0.6	83.3±0.4

Table A2: **Top-1 accuracies(%) for cross-domain FSL.** 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.

method	backbone	*tiered*ImageNet→Caltech 1-shot	*tiered*ImageNet→Caltech 5-shot	*tiered*ImageNet→CUB 1-shot	*tiered*ImageNet→CUB 5-shot
UBC-FSL (Ours)	ResNet-12	46.3±0.5	64.3±0.6	65.2±0.7	84.1±0.5
UBC-FSL (Ours)	ResNet-50	58.1±0.7	76.3±0.6	76.9±0.5	91.3±0.4
UBC-FSL (Ours)	ResNet-101	57.0±0.7	75.4±0.6	78.9±0.8	92.5±0.4
FSL baseline	ResNet-12	63.6±0.7	82.4±0.5	75.2±0.7	90.0±0.4
FSL baseline	ResNet-50	70.0±0.5	85.5±0.5	79.5±0.7	91.9±0.5
FSL baseline	ResNet-101	72.9±0.7	87.2±0.5	80.7±0.7	92.6±0.3
Combined (Ours)	ResNet-12	58.8 ±0.8	79.2±0.6	75.6±0.6	90.6±0.3
Combined (Ours)	ResNet-50	69.0±0.7	86.2±0.4	82.3±0.6	93.8±0.4
Combined (Ours)	ResNet-101	70.6±0.7	87.3±0.3	83.8±0.5	94.6±0.3
UBC-TFSL (Ours)	ResNet-12	44.8±0.6	62.7±0.7	66.0±0.7	84.6±0.7
UBC-TFSL (Ours)	ResNet-50	56.5±0.6	74.5±0.6	78.3±0.6	91.9±0.4
UBC-TFSL (Ours)	ResNet-101	59.4±0.7	76.3±0.7	81.1±0.7	93.3±0.5

Table A3: **Top-1 accuracies(%) for cross-domain FSL.** 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 report our results for cross-domain FSL in Table A2 and Table A3. In Table A2, we show results of learning models on *mini*ImageNet and evaluating them on Caltech-256 and CUB. In Table A3, we show results of learning models on *tiered*ImageNet and evaluating them on Caltech-256 and CUB.