Label Hallucination for Few-Shot Classification

Yiren Jian, Lorenzo Torresani

Dartmouth College
yiren.jian.gr@dartmouth.edu, LT@dartmouth.edu

Abstract

Few-shot classification requires adapting knowledge learned from a large annotated base dataset to recognize novel unseen classes, each represented by few labeled examples. In such a scenario, pretraining a network with high capacity on the large dataset and then finetuning it on the few examples causes severe overfitting. At the same time, training a simple linear classifier on top of "frozen" features learned from the large labeled dataset fails to adapt the model to the properties of the novel classes, effectively inducing underfitting. In this paper we propose an alternative approach to both of these two popular strategies. First, our method pseudo-labels the entire large dataset using the linear classifier trained on the novel classes. This effectively "hallucinates" the novel classes in the large dataset, despite the novel categories not being present in the base database (novel and base classes are disjoint). Then, it finetunes the entire model with a distillation loss on the pseudo-labeled base examples, in addition to the standard cross-entropy loss on the novel dataset. This step effectively trains the network to recognize contextual and appearance cues that are useful for the novel-category recognition but using the entire large-scale base dataset and thus overcoming the inherent data-scarcity problem of few-shot learning. Despite the simplicity of the approach, we show that that our method outperforms the state-of-the-art on four well-established few-shot classification benchmarks. The code and appendix are available at https://github.com/yiren-jian/LabelHalluc.

Introduction

Deep learning has emerged as the prominent learning paradigm for large data scenarios and it has achieved impressive results in wide range of application domains, including computer vision (Krizhevsky, Sutskever, and Hinton 2012), NLP (Devlin et al. 2019) and bioinformatics (Senior et al. 2020). However, it remains difficult to adapt deep learning models to settings where few labeled examples are available, since large-capacity models are inherently prone to overfitting.

Few-shot learning is usually studied under the episodic learning paradigm, which simulates the few-shot setting during training by repeatedly sampling few examples from a small subset of categories of a large base dataset. Meta-learning algorithms (Finn, Abbeel, and Levine 2017; Ravi and Larochelle 2017; Koch 2015; Vinyals et al. 2016; Snell, Swersky, and Zemel 2017) optimized on these training episodes have advanced the field of few-shot classification. However, recent works (Chen et al. 2019; Dhillon et al. 2020; Tian et al. 2020) have shown that a pure transfer learning strategy is often more competitive. For example, Tian et al. (Tian et al. 2020) proposed to first pretrain a large capacity classification model on the base dataset and then to simply learn a linear classifier on this pretrained representation using the few novel examples. The few-shot performance of the transferred model can be further improved by multiple distillation iterations (Furlanello et al. 2018), or by combining several losses simultaneously, e.g., entropy maximization, rotational self-supervision, and knowledge distillation (Rajasegaran et al. 2020).

In this paper, we follow the transfer learning approach. However, instead of freezing the representation to the features learned from the base classes (Tian et al. 2020; Rajasegaran et al. 2020; Rizve et al. 2021), we finetune the entire model. Since finetuning the network using only the few examples would result in severe overfitting (as evidenced by our ablations), we propose to optimize the model by re-using the entire base dataset but only after having swapped the original labels with pseudo-labels corresponding to the novel classes. This is achieved by running on the base dataset a simple linear classifier trained on the few examples of the novel categories. The classifier effectively "hallucinates" the presence of the novel classes in the base images. Although we empirically evaluate our approach in scenarios where the classes of the base dataset are completely disjoint from the novel categories, we demonstrate that this large-scale pseudo-labeled data enables effective finetuning of the entire model for recognition of the novel classes. The optimization is carried out using a combination of distillation over the pseudo-labeled base dataset and cross-entropy minimization over the few-shot examples. An overview of our proposed approach is provided in Fig. 1.

The intuition is that although the novel classes are not "properly" represented in the base images, many base examples may include objects that resemble those of the novel classes as encoded by the soft pseudo-labels that define the probabilities of belonging to the novel classes. For exam-

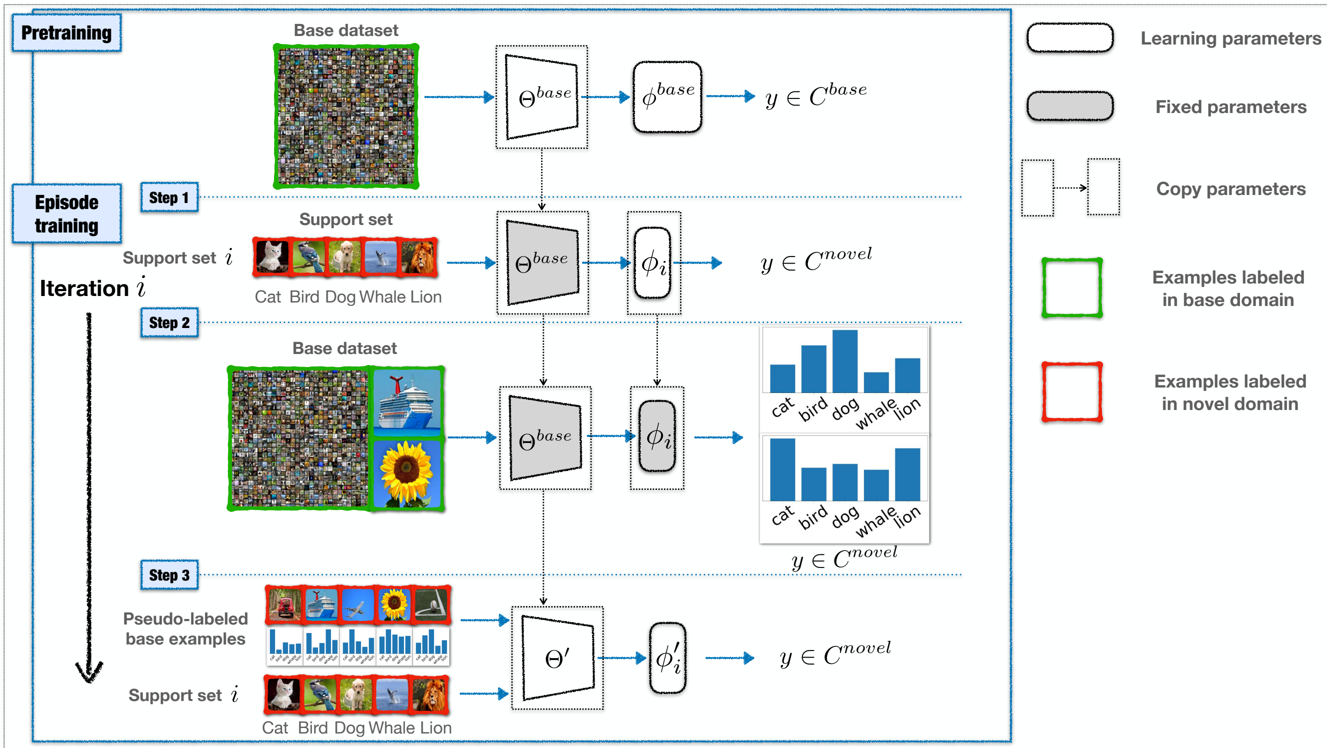

Figure 1: Overview of our proposed approach in an illustrative setting involving 1-shot classification of 5 novel classes. *Pretraining* learns the backbone model Θ and a classification head ϕ_0 from a labeled base dataset. The backbone is used to compute embeddings for the subsequent stages, while the classification head is discarded. During *Episode training*, step 1) learns a linear classifier ϕ_1 in the novel domain using the support set and the fixed embedding Θ. Step 2) pseudo-labels the base dataset with respect to the label space of the novel domain using the fixed embedding Θ and the classifier ϕ_1. Step 3) re-learns both the embedding and the classifier with the support set and the pseudo-labeled base dataset using a combination of distillation and cross-entropy maximization. Note that the base dataset and the support set do not share any classes.

ple, the pseudo-labeling may assign a probability of 0.6 for a base image of a tiger to belong to the novel class "domestic cat" given their appearance similarities. Or it may assign large novel-class pseudo-label probability to a base images because its true base category shares similar contextual background with the novel class, such in the case of "cars" and "pedestrians" which are both likely to appear in street scenes. Fine-tuning the entire model on these soft pseudo-labels using a distillation objective (combined with the cross-entropy loss on the few novel image examples) trains the network to recognize these similar or contextual cues on the base dataset, thus steering the representation towards features that are useful for the recognition of the novel classes. Furthermore, because the base dataset is large-scale, these examples serve the role of massive nonparametric data augmentation yielding a representation that is quite general and does not overfit, thus overcoming the data scarcity problem inherent in few-shot learning. We invite the reader to review the visualizations and the explanation in section Visualizations of Label Hallucination of our Technical Appendix (TechApp) for further insights into the behavior of our system. These visualizations confirm the intuition behind our approach , i.e., the fact that the base images with highest scores tend to be those that contain *contextual elements* that co-occur with the novel-class objects. Examples include foreground objects that have similar appearance to the few-shot images (e.g., the malamute image in Figure 1 of TechApp), or base examples including objects with shape akin to that of the novel class (e.g., the green mamba in Figure 2 of TechApp, which resembles the shape of a nematode), or even examples matching in terms of spatial layout (e.g., the images of tobacco shops and upright pianos have similar spatial layout as the bookshop class in Figure 3 of TechApp).

We note that pseudo-labeling has been widely used before for semi-supervised learning where the unlabeled examples belong to the same classes as the labeled ones (Sohn et al. 2020; Chen et al. 2020; Pham et al. 2021). Pseudo-labeling has also been adapted to the few-shot setting (Lazarou, Avrithis, and Stathaki 2020; Wang et al. 2020) but still under the empirical setting where novel classes are contained in the unlabeled dataset. The novelty of our work lies in showing that the advantages of pseudo-labeling extend even to the extreme setting where the set of base classes and the set of novel classes are completely disjoint. We also note that our work differs from transductive few-shot learning (Wang et al. 2020; Dhillon et al. 2020) which requires the testing set of unlabeled examples used during the training. Instead, our method operates in a pure inductive setting where within each episode only the small set of novel labeled examples and the base dataset are used for finetuning.

Related Work

Meta-Learning or learning to learn is the most common approach for few-shot learning. Meta-learning splits the learning into two phases, i.e., a meta-training phase and a meta-testing phase. During each phase, the meta-training set or meta-testing set is organized into multiple episodes, with each episode sampled from the task distribution. Each episode is further partitioned into a small training set and a testing set. In few-shot classification, the training set within each episode has N classes and K examples per class. Meta-learning methods can be further categorized into metric-based versus optimization-based. Metric-based methods (Koch 2015; Vinyals et al. 2016; Sung et al. 2018; Snell, Swersky, and Zemel 2017) learn embeddings for clustering or comparing examples. Few-shot metric learning methods have also been successfully applied to local descriptors (Li et al. 2019a; Huang et al. 2021; Zhu et al. 2021). Optimization-based methods (Finn, Abbeel, and Levine 2017; Flennerhag et al. 2020; Li et al. 2017; Rusu et al. 2019) learn the parameters of the model or the optimizer for fast adaptation using gradient descent.

Built upon those classical meta-learning methods, DeepEMD (Zhang et al. 2020) learns a new metric using the Earth Mover's Distance. MetaOptNet (Lee et al. 2019) solves a differentiable convex optimization problem to achieve better generalization of linear classifier. MTL (Sun et al. 2019) explores transfer learning in the meta-learning setting. FEAT (Ye et al. 2020) adapts a set-to-set function to learn task-specific and discriminative embeddings. NegCosine (Liu et al. 2020) replaces the softmax loss with a negative margin loss in metric learning. MELR (Fei et al. 2021) adopts an attention module with consistency regularization and explicitly models the relationship between different episodes. Instead of learning unstructured metric, COMET (Cao, Brbic, and Leskovec 2021) proposes to meta-learn along human-interpretable concepts. ConstellationNet (Xu et al. 2021) uses relation learning with self-attention to introduce a cell feature clustering algorithm. PseudoShots (Esfandiarpoor, Hajabdollahi, and Bach 2020) applies a masking module to select useful features from auxiliary labeled data. IEPT (Zhang et al. 2021) devises self-supervised pretext tasks at instance level and episode level for few-shot classification.

Transductive/Semi-Supervised Few-Shot Learning improves the few-shot results by utilizing the information from the query set or extra unlabeled examples for meta-testing episodes. TIM (Boudiaf et al. 2020) maximizes transductive information for few-shot learning. TAFSSL (Lichtenstein et al. 2020) searches discriminative feature sub-spaces for few-shot tasks. ICI (Wang et al. 2020) solves another linear regression hypothesis to filter out less trustworthy instances by pseudo-labeling. The method of Lazarou et al. (Lazarou, Avrithis, and Stathaki 2020) iteratively refines the pseudo-labels on the unlabeled dataset. Our method is an inductive learning approach, which differs from those used in these works. Without having additional data or information on the query set but only with the base dataset and the support set in each episode, inductive learning is the more common formulation of few-shot classification.

Transfer Learning is the de facto approach for many vision tasks (Donahue et al. 2014), when the labeled examples are scarce. But transfer learning has seen success in few-shot learning only very recently. New baseline methods (Chen et al. 2019; Dhillon et al. 2020) show competitive performances of few-shot classification by pretraining on a base training set followed by finetuning on the support set from each episode. RFS (Tian et al. 2020) outperforms all advanced meta-learning methods at its time by learning a fixed embedding model followed by a linear regression. The success of transfer learning methods relies on the high quality of the pretrained feature embeddings. To get more generalized embeddings of examples, SKD (Rajasegaran et al. 2020) proposes to incorporate a rotational self-supervised loss in the pretraining stage. Zhou et al. (Zhou et al. 2020) learn to select a subset of base classes for few-shot classification. Invariant and Equivariant Representations (IER) (Rizve et al. 2021) explore contrastive learning during the embedding learning. Chen et al. (Chen, Maji, and Learned-Miller 2021) have proposed a self-supervised method that pretrains the embedding without using labels for the base dataset. Our work also belongs to the genre of transfer learning. But instead of focusing on improving the embedding representation, we study how to better adapt the base knowledge to each novel task.

Under the transfer learning paradigm, Associative Alignment (AssoAlign) (Afrasiyabi, Lalonde, and Gagné 2020) is the closest to our work. It also exploits the base dataset examples (in the form of a selected subset) to enlarge the novel training set. It aligns the novel examples to the closest base examples in feature space by two strategies: 1) a metric loss to minimize the distance between base examples and the centroids of the novel ones, 2) a conditional Wasserstein adversarial alignment loss. Our method is much simpler than AssoAlign (Afrasiyabi, Lalonde, and Gagné 2020): it does not require complicated losses, selection of base examples or feature alignment of base dataset to novel examples. We show that by simply finetuning the whole network on a pseudo-labeled version of the base dataset, our method achieves stronger results despite using a smaller model (see table 1, table 2).

Pseudo-Labeling (Lee 2013) or self-training (Wei et al. 2021) labels the unlabeled dataset first with the model itself, and then re-trains the model with both the labeled and the pseudo-labeled dataset. It has shown great success in semi-supervised learning (Berthelot et al. 2019; Yu et al. 2020; Sohn et al. 2020), where pseudo-labeling is applied to a large unlabeled dataset with classes overlapping with the labeled set (the unlabeled and labeled dataset have the same or similar distributions). Algorithms are also designed to filter out low-confident pseudo-labeled examples. The latest work (Pham et al. 2021) which combines gradient-based meta learning with pseudo-labeling achieves a new state-of-the-art result on ImageNet benchmark (Deng et al. 2009). The same idea is adapted to semi-supervised few-shot learning as well (Li et al. 2019b; Wang et al. 2020; Lazarou, Avrithis, and Stathaki 2020). Our method has two main differences compared to these approaches. First, the unlabeled and labeled data are from completely disjoint domains. We label the base examples into classes that they do not belong to. Second, we do not have any mechanism to filter the low-confident examples. Our ablation shows that using all the base examples with our pseudo-labels yields better performance than finetuning on only the ones having higher classification probability.

Method

Problem Statement

We now formally define the few-shot classification problem considered in this work. We adopt the common setup which assumes the existence of a large scale labeled base dataset used to discriminatively learn a representation useful for the subsequent novel-class recognition. Let $\mathcal{D}^{base} = \{x_t^{base}, y_t^{base}\}_{t=1}^{N^{base}}$ be the base dataset, with label $y_t^{base} \in \mathcal{C}^{base}$. It is assumed that both the number of classes ($|\mathcal{C}^{base}|$) and the number of examples (N^{base}) are large in order to enable good representation learning. We denote with $\mathcal{D}^{novel} = \{x_t^{novel}, y_t^{novel}\}_{t=1}^{N^{novel}}$ the novel dataset, with $y_t^{novel} \in \mathcal{C}^{novel}$. The base classes and novel classes are disjoint, i.e., $\mathcal{C}^{base} \cap \mathcal{C}^{novel} = \varnothing$. We assume the training and testing of the few-shot classification model to be organized in episodes. At each episode i, the few-shot learner is given a support set $\mathcal{D}_i^{support} = \{x_{i,t}^{support}, y_{i,t}^{support}\}_{t=1}^{NK}$ involving K novel classes and N examples per class sampled from $\mathcal{D}^{novel}$ (with N being very small, typically ranging from 1 to 10). The learner is then evaluated on the query set $\mathcal{D}_i^{query} = \{x_{i,t}^{query}, y_{i,t}^{query}\}_{t=1}^{N'K}$, which contains examples of the same K classes as those in $\mathcal{D}_i^{support}$. Thus, the query/support sets serve as few-shot training/testing sets, respectively. At each episode i, the few-shot learner adapts the representation/model learned from the large-scale $\mathcal{D}^{base}$ to recognize the novel classes given the few training examples in $\mathcal{D}_i^{support}$.

Learning the Embedding Representation on the Base Dataset

We first aim at learning from the base dataset an embedding model that will transfer and generalize well to the downstream few-shot problems. We follow the approach of Tian et al. (Tian et al. 2020) (denoted as RFS) and train discriminatively a convolutional neural network consisting of a backbone f_Θ and a final classification layer g_ϕ. The parameters $\{\Theta, \phi\}$ are optimized jointly for the $|\mathcal{C}^{base}|$-way base classification problem using the dataset $\mathcal{D}^{base}$:

$$\Theta^{base}, \phi^{base} = \text{argmin}_{\Theta,\phi} \mathbb{E}_{\{x,y\} \in \mathcal{D}^{base}} \mathcal{L}_{CE}(g_\phi(f_\Theta(x)), y) \quad (1)$$

where $\mathcal{L}_{CE}$ is the cross-entropy loss. Prior work has shown the quality of the embedding representation encoded by parameters Θ^{base} can be further improved by knowledge distillation (Tian et al. 2020), rotational self-supervision (Rajasegaran et al. 2020) or by enforcing representations equivalent and invariant to sets of image transformations (Rizve et al. 2021). In the experiements presented in this paper, we follow the embedding learning strategies of SKD (Rajasegaran et al. 2020) (using self-supervised distillation) and IER (Rizve et al. 2021) (leveraging invariant and equivariant representations). However, note that our approach is independent of the specific method used for embedding learning.

Hallucinating the Presence of Novel Classes in the Base Dataset

In order to pseudo-label the base dataset according to the novel classes, we first train a classifier on the support set. For each episode i in the meta-learning phase, we learn a linear classifier ϕ_i on top of the *fixed* feature embedding model Θ^{base} using the few-shot support set $\mathcal{D}_i^{support} = \{x_{i,t}^{support}, y_{i,t}^{support}\}_{t=1}^{NK}$.

$$\phi_i = \text{argmin}_\phi \mathbb{E}_{\{x,y\} \in \mathcal{D}_i^{support}} \mathcal{L}_{CE}(g_\phi(f_{\Theta^{base}}(x)), y) \quad (2)$$

Note that in previous works (Tian et al. 2020; Rajasegaran et al. 2020; Rizve et al. 2021), ϕ_i is directly evaluated on query set $\mathcal{D}_i^{query}$ to produce the final few-shot classification results. Instead here we use the resulting model $g_{\phi_i}(f_{\Theta^{base}}(x))$ to re-label the base dataset according to the ontology of the novel classes in episode i. We denote with $\hat{y}_{i,t}^{base}$ the vector of logits (the outputs before the softmax) generated by applying the learned classifier to example x_t^{base}, i.e., $\hat{y}_{i,t}^{base} = g_{\phi_i}(f_{\Theta^{base}}(x_t))$ for $t = 1, \ldots, N^{base}$. These soft pseudo-labels are used to retrain the full model via knowledge distillation, as discussed next.

Finetuning the Whole Model to Recognize Novel Classes

We finally finetune the whole model (i.e., the backbone and the classifier) using mini-batches containing an equal proportion of support and base examples. The loss function for the base examples is knowledge distillation (Hinton, Vinyals, and Dean 2015), while the objective minimized for the support examples is the cross-entropy (CE). In other words, we optimize the parameters of the model on a mixing of the two losses:

$$\Theta_i', \phi_i' = \underset{\Theta,\phi}{\text{argmin}} \alpha \mathbb{E}_{\{x,y\} \in \mathcal{D}^{base}} \mathcal{L}_{KL}(g_\phi(f_\Theta(x)), \hat{y}) + \beta \mathbb{E}_{\{x,y\} \in \mathcal{D}_i^{support}} \mathcal{L}_{CE}(g_\phi(f_\Theta(x)), y) \quad (3)$$

where $\hat{y}$ denotes the hallucinated pseudo-label, $\mathcal{L}_{KL}$ is the KL divergence between the predictions of the model and the pseudo-labels scaled by temperature T, and α, β are hyperparameters trading off the importance of the two losses. Since the support set is quite small (in certain settings, each episode includes five novel classes and only one example for each novel class), we use data augmentation to generate multiple views of each support image, so as to obtain enough examples to fill half of the mini-batch. Specifically, we adopt the standard settings used in prior works (Tian et al. 2020; Rajasegaran et al. 2020; Rizve et al. 2021) and apply random cropping, color jittering and random flipping to generate multiple views.

Finally, the resulting model $g_{\phi'_i}(f_{\Theta'_i}(x))$ is evaluated on the query set $\mathcal{D}_i^{query} = \{x_t^{novel}, y_t^{novel}\}_{t=1}^{N'K}$. The final results are reported by averaging the accuracies of all episodes.

We note that although the operations of pseudo-labeling and finetuning are presented as separate and in sequence, in practice for certain datasets we found more efficient to generate the target pseudo-labels on the fly for the base examples loaded in the mini-batch without having to store them on disk.

Experiments

Datasets

We evaluate our method on four widely used few-shot recognition benchmarks: miniImageNet (Vinyals et al. 2016), tieredImageNet (Ren et al. 2018), CIFAR-FS (Bertinetto et al. 2019), and FC100 (Oreshkin, López, and Lacoste 2018).

Experimental Setup

Network Architecture. To make fair comparison to recent works (Tian et al. 2020; Rajasegaran et al. 2020; Rizve et al. 2021), we adopt the popular ResNet-12 (He et al. 2016) as our backbone.

Further descriptions on the four datasets, the architecture of ResNet-12 we used in experiments and optimization details on embedding learning and our finetuning can be found in the Technical Appendix.

Results on ImageNet-based Few-Shot Benchmarks

Table 1 provides a comparison between our approach and the state-of-the-art in few-show classification on the two ImageNet-based few-shot benchmarks. Our method is denoted as Label-Halluc. On **miniImageNet**, our method using the SKD pretraining of the backbone yields an absolute improvement of 0.96% over SKD-GEN1 in the one-shot setting. The improvement become more substantial under the 5-shot setting, with our method producing a gain of 2.42% over SKD-GEN1. When pretrained with IER (Rizve et al. 2021), our approach achieves one-shot classification accuracy of 68.28 ± 0.77, which is over 1.4% better than all reported results. Under the 5-shot setting, our method improves by 2.04% over IER-distill which had the best reported number, yielding a new state-of-the-art accuracy of 86.54%. On the **tieredImageNet** benchmark, our method pretrained with SKD performs on par with concurrent works (Fei et al. 2021; Zhang et al. 2021) and outperforms SKD (Rajasegaran et al. 2020) by 0.45% under the 1-shot setting and by 0.96% under the 5-shot setting. When pretrained with IER, our approach improves over IER-distill by 0.60% and 1.11% under the 1-shot and 5-shot settings, respectively, yielding a new state-of-the-art even for this benchmark.

Results on CIFAR-based Few-Shot Benchmarks

Table 2 compares our method, Label-Halluc, against the state-of-the-art on the two CIFAR-based few-shot benchmarks. On **CIFAR-FS**, the improvements over SKD-GEN1 (our implementation) for 1-shot and 5-shot are 0.7% and 0.9%, respectively. Note that these gains derive exclusively from the addition of the distillation over pseudo-labeled base examples. When using IER-distill as embedding learning, our method improves the baseline by 0.4% and 0.8% in the 1-shot and the 5-shot settings, respectively. On **FC100**, our method achieves improves over the best reported numbers by 0.8% and 3.0% in the 1-shot and 5-shot setting, respectively, when pretrained with SKD. The improvements are 1.0% and 3.0% when pretrained with IER.

Ablations

The ablation studies are performed to validate improved transfer performances via the use of pseudo-labeled base examples, the use of distillation loss on soft-labels over one-hot encoding labels and finetuning the whole network over learning only the classifier. We further study the effect of different embedding learning methods. Finally, we make an apple-to-apple experimental comparison to AssoAlign (Afrasiyabi, Lalonde, and Gagné 2020), which also exploits the use of the base dataset during finetuning.

Unless otherwise stated, all the experiments in this section (Ablations) use ResNet-12 and the embedding training by SKD-GEN1 (Rajasegaran et al. 2020).

Benefits of the Pseudo-Labeled Base Dataset In Table 3 we ablate on different strategies to transfer the base knowledge to the novel-class recognition. *Transfer w/ frozen backbone (LR)* is the traditional transfer learning procedure of using the few-shot examples to train a linear regression model on top of the frozen backbone learned from the base dataset. *Transfer w/ finetuning* uses the support set *only* to finetune the entire model (backbone and classifier). *Hard LabelHalluc + finetuning* is a variant of our approach where the base dataset is pseudo-labeled with one-hot hard novel-class labels and the entire model is subsequently finetuned using cross-entropy over the base and support set. Finally, *Soft LabelHalluc + finetuning* is our method using soft pseudo-labels. Note that we train *Transfer w/ finetuning* and **Hard LabelHalluc + finetuning** for 300 steps only for both 1-shot and 5-shot, since we find that training longer in these two settings leads to worse results. From the results in this Table we can infer several findings. First of all, we can observe that *Transfer w/ frozen backbone* performs much better than *Transfer w/ finetuning*. This confirms the findings of previous works (Tian et al. 2020; Afrasiyabi, Lalonde, and Gagné 2020; Chen et al. 2019) which observed that learning a regression model with the fixed embedding leads to better performance over finetuning the whole network (Dhillon et al. 2020). This happens because finetuning the entire network only with the support set results in overfitting. With the combination of distillation loss on the soft-labeled base dataset, our method (*Soft LabelHalluc + finetuning*) eliminates the overfitting problem yielding 5-shot gains of 5.57%, 3.8% and 5.3% on miniImageNet, CIFAR-FS, and FC100, respec-

Model	Net	miniImageNet 5-way 1-shot	miniImageNet 5-way 5-shot	tieredImageNet 5-way 1-shot	tieredImageNet 5-way 5-shot
ProtoNet (Snell, Swersky, and Zemel 2017)	R-12	60.37 ± 0.83	78.02 ± 0.57	65.65 ± 0.92	83.40 ± 0.65
TapNet (Yoon, Seo, and Moon 2019)	R-12	61.65 ± 0.15	76.36 ± 0.10	63.08 ± 0.15	80.26 ± 0.12
MetaOptNet (Lee et al. 2019)	R-12	62.64 ± 0.61	78.63 ± 0.46	65.99 ± 0.72	81.56 ± 0.53
MTL (Sun et al. 2019)	R-12	61.20 ± 1.80	75.50 ± 0.80	65.62 ± 1.80	80.61 ± 0.90
DSN-MR (Simon et al. 2020)	R-12	64.60 ± 0.72	79.51 ± 0.50	67.39 ± 0.83	82.85 ± 0.56
DeepEMD (Zhang et al. 2020)	R-12	65.91 ± 0.82	82.41 ± 0.56	71.16 ± 0.87	86.03 ± 0.58
FEAT (Ye et al. 2020)	R-12	66.78 ± 0.20	82.05 ± 0.14	70.80 ± 0.23	84.79 ± 0.16
Neg-Cosine (Liu et al. 2020)	R-12	63.85 ± 0.81	81.57 ± 0.56	-	-
RFS-simple (Tian et al. 2020)	R-12	62.02 ± 0.63	79.64 ± 0.44	69.74 ± 0.72	84.41 ± 0.55
RFS-distill (Tian et al. 2020)	R-12	64.82 ± 0.82	82.41 ± 0.43	71.52 ± 0.69	86.03 ± 0.49
AssoAlign (Afrasiyabi et al. 2020)	R-18†	59.88 ± 0.67	80.35 ± 0.73	69.29 ± 0.56	85.97 ± 0.49
AssoAlign (Afrasiyabi et al. 2020)	W-28‡	65.92 ± 0.60	82.85 ± 0.55	74.40 ± 0.68	86.61 ± 0.59
SKD-GEN1 (Rajasegaran et al. 2020)	R-12	$66.54 \pm 0.97^{\S}$	$83.18 \pm 0.54^{\S}$	$72.35 \pm 1.23^{\S}$	$85.97 \pm 0.63^{\S}$
P-Transfer (Shen et al. 2021)	R-12	64.21 ± 0.77	80.38 ± 0.59	-	-
InfoPatch (Gao et al. 2021)	R-12	67.67 ± 0.45	82.44 ± 0.31	71.51 ± 0.52	85.44 ± 0.35
MELR (Fei et al. 2021)	R-12	67.40 ± 0.43	83.40 ± 0.28	72.14 ± 0.51	87.01 ± 0.35
IEPT (Zhang et al. 2021)	R-12	67.05 ± 0.44	82.90 ± 0.30	72.24 ± 0.50	86.73 ± 0.34
IER-distill (Rizve et al. 2021)	R-12	$66.85 \pm 0.76^{\S}$	$84.50 \pm 0.53^{\S}$	$72.74 \pm 1.25^{\S}$	$86.57 \pm 0.81^{\S}$
Label-Halluc (pretrained w/ SKD-GEN1)	R-12	67.50 ± 1.01	85.60 ± 0.52	72.80 ± 1.20	86.93 ± 0.60
Label-Halluc (pretrained w/ IER-distill)	R-12	$\mathbf{68.28 \pm 0.77}$	$\mathbf{86.54 \pm 0.46}$	$\mathbf{73.34 \pm 1.25}$	$\mathbf{87.68 \pm 0.83}$

Table 1: Comparison of our method (Label-Halluc) against the state-of-the-art on miniImageNet and tieredImageNet. We report our results with 95% confidence intervals on meta-testing split of miniImageNet and tieredImageNet. Training is done on the training split only. † indicates using a higher resolution of training images. ‡ indicates a larger model than ResNet-12. § indicates our implementations. This makes the fairest comparisons to ours by allowing that those methods are evaluated on exact same episodes.

Model	Net	CIFAR-FS 5-way 1-shot	CIFAR-FS 5-way 5-shot	FC-100 5-way 1-shot	FC-100 5-way 5-shot
ProtoNet (Snell, Swersky, and Zemel 2017)	R-12	72.2 ± 0.7	83.5 ± 0.5	37.5 ± 0.6	52.5 ± 0.6
MetaOptNet (Lee et al. 2019)	R-12	72.6 ± 0.7	84.3 ± 0.5	41.1 ± 0.6	55.5 ± 0.6
MTL (Sun et al. 2019)	R-12	-	-	45.1 ± 1.8	57.6 ± 0.9
DSN-MR (Simon et al. 2020)	R-12	75.6 ± 0.9	86.2 ± 0.6	-	-
DeepEMD (Zhang et al. 2020)	R-12	-	-	46.5 ± 0.8	63.2 ± 0.7
RFS-simple (Tian et al. 2020)	R-12	71.5 ± 0.8	86.0 ± 0.5	42.6 ± 0.7	59.1 ± 0.6
RFS-distill (Tian et al. 2020)	R-12	73.9 ± 0.8	86.9 ± 0.5	44.6 ± 0.7	60.9 ± 0.6
AssoAlign (Afrasiyabi et al. 2020)	R-18‡	-	-	45.8 ± 0.5	59.7 ± 0.6
SKD-GEN1 (Rajasegaran et al. 2020)	R-12	$76.6 \pm 0.9^{\S}$	$88.6 \pm 0.5^{\S}$	$46.5 \pm 0.8^{\S}$	$64.2 \pm 0.8^{\S}$
InfoPatch (Gao et al. 2021)	R-12	-	-	43.8 ± 0.4	58.0 ± 0.4
IER-distill (Rizve et al. 2021)	R-12	$77.6 \pm 1.0^{\S}$	$89.7 \pm 0.6^{\S}$	$48.1 \pm 0.8^{\S}$	$65.0 \pm 0.7^{\S}$
Label-Halluc (pretrained w/ SKD-GEN1)	R-12	77.3 ± 0.9	89.5 ± 0.5	47.3 ± 0.8	67.2 ± 0.8
Label-Halluc (pretrained w/ IER-distill)	R-12	$\mathbf{78.0 \pm 1.0}$	$\mathbf{90.5 \pm 0.6}$	$\mathbf{49.1 \pm 0.8}$	$\mathbf{68.0 \pm 0.7}$

Table 2: Comparison of Label-Halluc (ours) to prior works on CIFAR-FS and FC-100. We report our results with 95% confidence intervals on meta-testing split of CIFAR-FS and FC-100. Training is done on the training split only. ‡ indicates a different model. § indicates our implementations.

tively, over finetuning with the episode examples only. It can be observed that our improves also over the simple strategy of *Transfer w/ frozen backbone* since it provides the advantage of adapting the backbone representation to the specific characteristics of the novel classes.

Soft or Hard Labels The distillation loss on the pseudo-labeled base dataset is the KL-divergence between the predictions of the linear classifier (soft-labels) and the predic-

	mini-IN		CIFAR-FS		FC100	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Frozen (LR)	66.54	83.18	76.6	88.6	46.5	64.2
Finetuning	61.43	80.03	68.8	85.7	43.1	61.9
Hard-halluc	65.04	80.68	75.3	85.3	44.6	62.4
Soft-halluc	**67.50**	**85.60**	**77.3**	**89.5**	**47.3**	**67.2**

Table 3: Ablation study on different strategy to transfer the base knowledge to the novel-class recognition. Our approach in its complete form (*Soft-halluc*) achieves gains over traditional transfer learning (*Frozen (LR)* and *Finetuning*) as well as a variant of our label hallucination method using hard pseudo-labeling (*Hard-hallluc*).

Support		Base		miniImageNet	
Net	**Clf**	**Net**	**Clf**	1-shot	5-shot
✓	✓			61.43	80.03
✓	✓		✓	63.59	81.53
✓	✓	✓		66.18	84.36
✓	✓	✓	✓	**67.50**	**85.60**

Table 4: Ablation study on gradients from base examples. We study which part of the model (the embedding or the linear classifier) benefits from the use of pseudo-labeled base examples. Having no gradient on either *Net* or *Clf* is equal to finetuning with the novel support set only. And Having gradients on both parts is the default setting of our method.

	miniImageNet		CIFAR-FS		FC100	
	LR	ours	LR	ours	LR	ours
RFS_0	79.33	**81.75**	86.6	**87.3**	58.1	**61.2**
RFS_1	81.15	**82.74**	86.5	**87.1**	61.0	**63.9**
SKD_0	82.31	**84.14**	87.8	**88.8**	62.8	**66.5**
SKD_1	83.18	**85.60**	88.6	**89.5**	64.2	**67.2**
IER_0	83.88	**85.86**	89.5	**90.2**	63.8	**67.2**
IER_1	84.50	**86.54**	89.7	**90.5**	65.0	**68.0**
		+2.05		**+0.8**		**+3.2**

Table 5: Ablation study on different embedding learning methods in 5-shot classification. *LR* is Linear Regression with a fixed embedding. RFS_0 denotes RFS-simple model and RFS_1 is RFS-distill, etc.

tions of the current model. We adopt soft-labeling because hard-labeling examples with novel categories that are not truly represented in those images has a negative effect. This can be clearly observed by the poor performance of *Hard LabelHalluc + finetuning* in Table 3. The use of soft labels over hard labels contributes 5-shot classification gains of 4.92% in miniImageNet, 4.2% in CIFAR-FS, and 4.8% in FC100. Using the distillation loss with soft-labels is crucial for the success of our method.

Finetuning the Backbone vs the Classifier with Hallucinated Pseudo-Labels In this section, we aim at studying whether it is beneficial to finetune the whole network with the hallucinated pseudo-labels, as opposed to just the classifier. We expect that training only the linear classifier (which has few parameters) would not reap the full benefits of the extended re-labeling of the large-scale base dataset, whereas unfreezing the high capacity backbone network may yield further gains. Table 4 shows the comparisons between 1) applying no gradient from hallucinated pseudo-labels, 2) applying gradients from hallucinated pseudo-labels to the final classifier only, 3) applying gradients from hallucinated pseudo-labels to the backbone feature network only and 4) applying gradients from hallucinated pseudo-labels to the whole network, which is our default setting. As we can see, applying gradients from hallucinated pseudo-labels to the final classifier only leads to small grains (2.16% and 1.50% in 1-shot and 5-shot respectively). While learning the backbone feature network with gradients from hallucinated pseudo-labels solely already increases by 4.65% and 4.33% in 1-shot and 5-shot. Most of the improvements of our method are coming from learning the backbone feature network with pseudo-labeled base examples.

	CA	Sub	KD	mini 5-shot	FS 5-shot	FC 5-shot
finetune				80.03	85.7	61.9
Asso †		✓		81.21	85.3	60.6
Asso †	✓	✓		82.38	86.3	62.6
Asso	✓	✓		83.47	87.6	63.6
Ours		✓	✓	85.18	89.3	66.8
Ours			✓	**85.60**	**89.5**	**67.2**

Table 6: We compare our method to AssoAlign (Asso) on ResNet-12 in miniImageNet (*mini*) with 84×84 image resolution, CIFAR-FS (*FS*) and FC100 (*FC*). *CA* is the centroid alignment introduced by AssoAlign. *Sub* indicates the use of the scoring matrix in AssoAlign for selection of base examples. *KD* is our method which uses distillation loss on pseudo-labeled base dataset. † indicates using arcMax activation over softMax.

Different Embedding Learning Methods Our finetuning approach can be used with different embedding learning strategies for pretraining the backbone from the base dataset. We experiment with six different pretraining methods proposed in RFS (Tian et al. 2020), SKD (Rajasegaran et al. 2020) and IER (Rizve et al. 2021). Table 5 shows that our approach constantly improves the classification accuracies over the linear regression (LR) with fixed embeddings. In miniImageNet 5-shot classification, our method has an average 2.05% improvement over *LR*. In CIFAR-FS and FC100 5-shot, the average improvement are 0.8% and 3.2% respectively.

Comparing to AssoAlign AssoAlign (Afrasiyabi, Lalonde, and Gagné 2020) also exploits examples from the base dataset to extend the finetuning dataset. The key differences to our method are: 1) AssoAlign experiments with both arcMax and softMax cross-entropy loss; 2) AssoAlign uses a similarity matrix for selecting a subset of base dataset (this requires searching another hyper-parameter to control how many examples to select), while our method uses the entire base dataset; 3) AssoAlign uses centroid alignments on feature space between novel examples and the selected base examples, while our approach uses a distillation loss on the pseudo-labeled base dataset.

The original AssoAlign is implemented with a different optimizer (Adam) and a different image resolution (224×224 in miniImageNet with ResNet-18). Its backbones (Conv4$\times$4, RestNet-18 or WRN-28-10) differ from those used in recent works (84×84 resolution with ResNet-12). The published results of AssoAlign can be found in table 1 and table 2. However, in order to assess AssoAlign and our method on equal ground, we apply the public implementation of AssoAlign to our setting. Starting from the same data augmentations and the same learning policy of ours with their suggested hyper-parameters for ResNet. We report results for AssoAlign finetuned for 100 steps with SGD, since we found this to be the optimal number of steps and the optimizer for this method (we tried both SGD and Adam with 50, 100, 150, 200 250, 300, 350 and 400 steps). Both our method and AssoAlign here use the same embedding model pretrained with SKD-GEN1 (Rajasegaran et al. 2020).

As shown in table 6, AssoAlign alleviates the overfitting issue of finetuning on the support set only (first row) for all three benchmarks. But our method achieves superior results over AssoAlign in all three datasets by only using a simple distillation loss with pseudo-labeled base examples. Though using the arcMax alone yieds an improvement of 1.18% over *finetune* in miniImageNet, we find that combining arcMax with centroid alignment leads to inferior results in our experimental setup based on ResNet-12 and SGD. AssoAlign with softMax and centroid alignment outperforms *finetune* by 3.44%, 1.9% and 2.7%, whereas our method outperforms AssoAlign by 2.13%, 1.9% and 3.6% in miniImageNet, CIFAR-FS and FC100 respectively.

Computational Cost Finetuning methods have a high latency, due to the fact that it requires optimizing all parameters in a large deep neural network for each episode. This is true for our method, AssoAlign and prior works (Dhillon et al. 2020; Liu et al. 2019). However, due to the simplicity of our method, the average running time for each training step is 0.35 second only (a total of 105 seconds for each episode), compared to AssoAlign's 0.87 second (87 seconds for each episode) in miniImageNet 5-shot. Furthermore, we invite the reader to review the section "Speeding up the training" in the Technical Appendix, where we present and quantitatively evaluate strategies to lower the total training cost of our approach by more than 66% while still maintaining significantly higher accuracy compared to the state-of-the-art.

Additional Experiments

We refer the reader to the Technical Appendix for several additional experiments, including:

- Several visualizations of label hallucination providing useful insights into the effectiveness of our system.
- Discussion and evaluation of strategies to reduce the computational cost of the training procedure.
- Performance as a function of the base dataset size.
- Stochastic finetuning using random mini-batches of the support set.
- Results for large number of novel classes (10-way and 20-way experiments) in each episode.
- Results when the base classes and the novel categories are far apart.
- Experiments with imbalanced base classes.
- Simultaneous recognition of base and novel classes.
- Ablation on knowledge distillation.
- Experiment with feature distillation.

Conclusion

We propose the simple strategy of label hallucination to enable effective finetuning of large-capacity models from few-shot examples of the novel classes. Results on four well-established few-shot classification benchmarks show that even in the extreme scenario where the labels of the base dataset and the labels of the novel examples are completely disjoint, our procedure achieves state-of-the-art accuracy and consistently improves over popular strategies of transfer learning via finetuning or methods that perform linear classification on top of pretrained representations.

References

Afrasiyabi, A.; Lalonde, J.-F.; and Gagné, C. 2020. Associative Alignment for Few-shot Image Classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; and Raffel, C. A. 2019. MixMatch: A Holistic Approach to Semi-Supervised Learning. In *Advances in Neural Information Processing Systems*.

Bertinetto, L.; Henriques, J. F.; Torr, P.; and Vedaldi, A. 2019. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations*.

Boudiaf, M.; Ziko, I.; Rony, J.; Dolz, J.; Piantanida, P.; and Ben Ayed, I. 2020. Information Maximization for Few-Shot Learning. In *Advances in Neural Information Processing Systems*.

Cao, K.; Brbic, M.; and Leskovec, J. 2021. Concept Learners for Few-Shot Learning. In *International Conference on Learning Representations*.

Chen, T.; Kornblith, S.; Swersky, K.; Norouzi, M.; and Hinton, G. E. 2020. Big Self-Supervised Models are Strong Semi-Supervised Learners. In *Advances in Neural Information Processing Systems*.

Chen, W.-Y.; Liu, Y.-C.; Kira, Z.; Wang, Y.-C. F.; and Huang, J.-B. 2019. A Closer Look at Few-shot Classification. In *International Conference on Learning Representations*.

Chen, Z.; Maji, S.; and Learned-Miller, E. 2021. Shot in the Dark: Few-Shot Learning With No Base-Class Labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2668–2677.

Deng, J.; Dong, W.; Socher, R.; Li, L.; Kai Li; and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 4171–4186. Association for Computational Linguistics.

Dhillon, G. S.; Chaudhari, P.; Ravichandran, A.; and Soatto, S. 2020. A Baseline for Few-Shot Image Classification. In *International Conference on Learning Representations*.

Donahue, J.; Jia, Y.; Vinyals, O.; Hoffman, J.; Zhang, N.; Tzeng, E.; and Darrell, T. 2014. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. In *Proceedings of the 31st International Conference on Machine Learning*, 647–655. PMLR.

Esfandiarpoor, R.; Hajabdollahi, M.; and Bach, S. H. 2020. Pseudo Shots: Few-Shot Learning with Auxiliary Data. *arXiv preprint arXiv:2012.07176*.

Fei, N.; Lu, Z.; Xiang, T.; and Huang, S. 2021. {MELR}: Meta-Learning via Modeling Episode-Level Relationships for Few-Shot Learning. In *International Conference on Learning Representations*.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, 1126–1135.

Flennerhag, S.; Rusu, A. A.; Pascanu, R.; Visin, F.; Yin, H.; and Hadsell, R. 2020. Meta-Learning with Warped Gradient Descent. In *International Conference on Learning Representations*.

Furlanello, T.; Lipton, Z.; Tschannen, M.; Itti, L.; and Anandkumar, A. 2018. Born Again Neural Networks. In *Proceedings of the 35th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 1607–1616. PMLR.

Gao, Y.; Fei, N.; Liu, G.; Lu, Z.; Xiang, T.; and Huang, S. 2021. Contrastive Prototype Learning with Augmented Embeddings for Few-Shot Learning. *arXiv preprint arXiv:2101.09499*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. In *NIPS Deep Learning and Representation Learning Workshop*.

Huang, H.; Wu, Z.; Li, W.; Huo, J.; and Gao, Y. 2021. Local descriptor-based multi-prototype network for few-shot Learning. *Pattern Recognition*, 116: 107935.

Koch, G. R. 2015. Siamese Neural Networks for One-Shot Image Recognition. In *ICML deep learning workshop*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*.

Lazarou, M.; Avrithis, Y.; and Stathaki, T. 2020. Iterative label cleaning for transductive and semi-supervised few-shot learning. *arXiv preprint arXiv:2012.07962*.

Lee, D.-H. 2013. Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks. *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*.

Lee, K.; Maji, S.; Ravichandran, A.; and Soatto, S. 2019. Meta-Learning With Differentiable Convex Optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, W.; Wang, L.; Xu, J.; Huo, J.; Yang, G.; and Luo, J. 2019a. Revisiting Local Descriptor based Image-to-Class Measure for Few-shot Learning. In *CVPR*.

Li, X.; Sun, Q.; Liu, Y.; Zhou, Q.; Zheng, S.; Chua, T.-S.; and Schiele, B. 2019b. Learning to Self-Train for Semi-Supervised Few-Shot Classification. In *Advances in Neural Information Processing Systems*.

Li, Z.; Zhou, F.; Chen, F.; and Li, H. 2017. Meta-SGD: Learning to Learn Quickly for Few-Shot Learning. *arXiv preprint arXiv:1707.09835*.

Lichtenstein, M.; Sattigeri, P.; Feris, R.; Giryes, R.; and Karlinsky, L. 2020. TAFSSL: Task-Adaptive Feature Sub-Space Learning for Few-Shot Classification. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J.-M., eds., *Computer Vision – ECCV 2020*.

Liu, B.; Cao, Y.; Lin, Y.; Li, Q.; Zhang, Z.; Long, M.; and Hu, H. 2020. Negative Margin Matters: Understanding Margin in Few-Shot Classification. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J.-M., eds., *Computer Vision – ECCV 2020*.

Liu, Y.; Lee, J.; Park, M.; Kim, S.; Yang, E.; Hwang, S.; and Yang, Y. 2019. Learning to Propagate Labels: Transductive Propagation Network for Few-shot Learning. In *International Conference on Learning Representations*.

Oreshkin, B. N.; López, P. R.; and Lacoste, A. 2018. TADAM: Task dependent adaptive metric for improved few-shot learning. In *NeurIPS*.

Pham, H.; Dai, Z.; Xie, Q.; Luong, M.-T.; and Le, Q. V. 2021. Meta Pseudo Labels. *arXiv preprint arXiv:2003.10580*.

Rajasegaran, J.; Khan, S.; Hayat, M.; Khan, F. S.; and Shah, M. 2020. Self-supervised Knowledge Distillation for Few-shot Learning. *arXiv preprint arXiv:2006.09785*.

Ravi, S.; and Larochelle, H. 2017. Optimization as a Model for Few-Shot Learning. In *ICLR*.

Ren, M.; Ravi, S.; Triantafillou, E.; Snell, J.; Swersky, K.; Tenenbaum, J. B.; Larochelle, H.; and Zemel, R. S. 2018. Meta-Learning for Semi-Supervised Few-Shot Classification. In *International Conference on Learning Representations*.

Rizve, M. N.; Khan, S.; Khan, F. S.; and Shah, M. 2021. Exploring Complementary Strengths of Invariant and Equivariant Representations for Few-Shot Learning. *arXiv preprint arXiv:2103.01315*.

Rusu, A. A.; Rao, D.; Sygnowski, J.; Vinyals, O.; Pascanu, R.; Osindero, S.; and Hadsell, R. 2019. Meta-Learning with Latent Embedding Optimization. In *International Conference on Learning Representations*.

Senior, A.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Žídek, A.; Nelson, A.; Bridgland, A.; Penedones, H.; Petersen, S.; Simonyan, K.; Crossan, S.; Kohli, P.; Jones, D.; Silver, D.; Kavukcuoglu, K.; and Hassabis, D. 2020. Improved protein structure prediction using potentials from deep learning. *Nature*, 577: 1–5.

Shen, Z.; Liu, Z.; Qin, J.; Savvides, M.; and Cheng, K. 2021. Partial Is Better Than All: Revisiting Fine-tuning Strategy for Few-shot Learning. *CoRR*, abs/2102.03983.

Simon, C.; Koniusz, P.; Nock, R.; and Harandi, M. 2020. Adaptive Subspaces for Few-Shot Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical Networks for Few-shot Learning. In *Advances in Neural Information Processing Systems*.

Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C. A.; Cubuk, E. D.; Kurakin, A.; and Li, C.-L. 2020. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. In *Advances in Neural Information Processing Systems*, 596–608.

Sun, Q.; Liu, Y.; Chua, T.-S.; and Schiele, B. 2019. Meta-Transfer Learning for Few-Shot Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H.; and Hospedales, T. M. 2018. Learning to Compare: Relation Network for Few-Shot Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tian, Y.; Wang, Y.; Krishnan, D.; Tenenbaum, J. B.; and Isola, P. 2020. Rethinking Few-Shot Image Classification: A Good Embedding is All You Need? In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J.-M., eds., *Computer Vision – ECCV 2020*.

Vinyals, O.; Blundell, C.; Lillicrap, T.; kavukcuoglu, k.; and Wierstra, D. 2016. Matching Networks for One Shot Learning. In *Advances in Neural Information Processing Systems*.

Wang, Y.; Xu, C.; Liu, C.; Zhang, L.; and Fu, Y. 2020. Instance Credibility Inference for Few-Shot Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wei, C.; Shen, K.; Chen, Y.; and Ma, T. 2021. Theoretical Analysis of Self-Training with Deep Networks on Unlabeled Data. In *International Conference on Learning Representations*.

Xu, W.; yifan xu; Wang, H.; and Tu, Z. 2021. Attentional Constellation Nets for Few-Shot Learning. In *International Conference on Learning Representations*.

Ye, H.-J.; Hu, H.; Zhan, D.-C.; and Sha, F. 2020. Few-Shot Learning via Embedding Adaptation With Set-to-Set Functions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yoon, S. W.; Seo, J.; and Moon, J. 2019. TapNet: Neural Network Augmented with Task-Adaptive Projection for Few-Shot Learning. In *Proceedings of the 36th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 7115–7123. PMLR.

Yu, Z.; Chen, L.; Cheng, Z.; and Luo, J. 2020. TransMatch: A Transfer-Learning Scheme for Semi-Supervised Few-Shot Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, C.; Cai, Y.; Lin, G.; and Shen, C. 2020. DeepEMD: Few-Shot Image Classification With Differentiable Earth Mover's Distance and Structured Classifiers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, M.; Zhang, J.; Lu, Z.; Xiang, T.; Ding, M.; and Huang, S. 2021. {IEPT}: Instance-Level and Episode-Level Pretext Tasks for Few-Shot Learning. In *International Conference on Learning Representations*.

Zhou, L.; Cui, P.; Jia, X.; Yang, S.; and Tian, Q. 2020. Learning to Select Base Classes for Few-Shot Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhu, W.; Li, W.; Liao, H.; and Luo, J. 2021. Temperature network for few-shot learning with distribution-aware large-margin metric. *Pattern Recognition*, 112: 107797.