On the Importance of Distractors for Few-Shot Classification

Rajshekhar Das[1] Yu-Xiong Wang[2] José M.F. Moura[1]
[1]Carnegie Mellon University [2]University of Illinois at Urbana-Champaign
rajshekd@andrew.cmu.edu yxw@illinois.edu moura@andrew.cmu.edu

Abstract

Few-shot classification aims at classifying categories of a novel task by learning from just a few (typically, 1 to 5) labelled examples. An effective approach to few-shot classification involves a prior model trained on a large-sample base domain, which is then finetuned over the novel few-shot task to yield generalizable representations. However, task-specific finetuning is prone to overfitting due to the lack of enough training examples. To alleviate this issue, we propose a new finetuning approach based on contrastive learning that reuses unlabelled examples from the base domain in the form of distractors. Unlike the nature of unlabelled data used in prior works, distractors belong to classes that do not overlap with the novel categories. We demonstrate for the first time that inclusion of such distractors can significantly boost few-shot generalization. Our technical novelty includes a stochastic pairing of examples sharing the same category in the few-shot task and a weighting term that controls the relative influence of task-specific negatives and distractors. An important aspect of our finetuning objective is that it is agnostic to distractor labels and hence applicable to various base domain settings. More precisely, compared to state-of-the-art approaches, our method shows accuracy gains of up to 12% *in cross-domain and up to* 5% *in unsupervised prior-learning settings. Our code is available at* https://github.com/quantacode/Contrastive-Finetuning.git

1. Introduction

The ability to learn from very few examples is innate to human intelligence. In contrast, large amounts of labelled examples are required by modern machine learning algorithms to learn a new task. This limits their applicability to domains where data is either expensive to annotate and collect or simply inaccessible due to privacy concerns. To overcome this limitation, few-shot classification has been proposed as a generic framework for learning to classify with very limited supervision [12, 31, 34, 59]. Under this paradigm, most approaches leverage prior knowledge from a (labelled) *base* domain to solve a novel task by either finetuning-based transfer [9, 64] or meta-learning [12, 15, 49, 57, 59, 62, 69]. In particular, when the base and novel domains are related, the hope is that representations learnt in the base domain can be generalized to novel tasks, thus facilitating *positive* knowledge transfer.

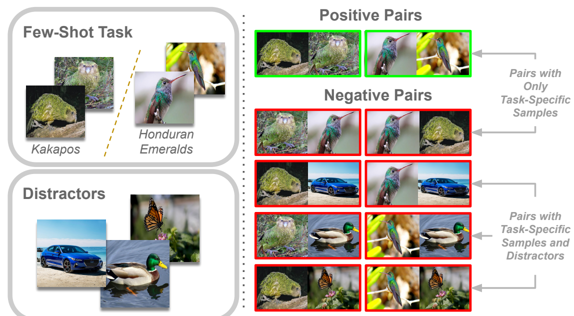

Figure 1. **Classification of Kakapos vs. Honduran Emeralds with just few examples per class and many distractors**: The idea is to leverage unlabelled data in the form of *distractors* that need not be semantically related to the classes in the few-shot task. The hope is that by pairing distractors and task samples as negatives (bottom six red boxes) and encouraging greater dissimilarity between such pairs, image representations of the two classes, Kakapos and Honduran Emeralds, will be pushed farther away. This would ultimately lead to better classification.

While the above paradigm is effective for tasks that can leverage large datasets like ImageNet [52] as the related base domain, for others, such as rare species classification [68] or medical image classification [70], acquiring necessary prior knowledge can be exceedingly difficult due to the absence of a related base domain with labelled data. To relax such data requirements, recent techniques explore alternative ways such as unsupervised learning [25, 28] or cross-domain learning [1, 15, 45, 66] to obtain representations useful for novel tasks. In the absence of labelled base data, approaches like [25, 27, 28] seek to benefit from self-supervised representation learning over unlabelled data in a related domain. In a more challenging scenario where related base data is hard to obtain, cross-domain techniques [10, 64, 66] exploit representations learnt in other domains that do not have the same task characteristics as the novel tasks.

Although the issue of learning a good prior representation remains a core focus in few-shot classification, it addresses only a part of the problem. In this work, we investigate the other important aspect, *i.e.*, *effective finetuning specific to the novel task*. Our main motivation comes from recent findings [1, 9, 16] that demonstrate the outperformance of simple finetuning over more sophisticated prior learning techniques such as meta-learning. Despite its effectiveness, we suspect that finetuning might still suffer from overfitting as a consequence of small training set in a few-shot task. To alleviate this situation, we propose to *leverage additional unlabelled data exclusive to the task*. Such datapoints are referred to as *distractors*. For instance, in the case of classifying Honduran Emeralds and Kakapos (rare species of birds), examples of butterflies, cars or ducks can serve as distractors (Fig. 1). By the virtue of its task-exclusivity, distractors can be obtained from various data-abundant domains with categories that could be semantically unrelated to novel task categories. However, in this work, we restrict ourselves to just the base data as a source for distractors. This allows us to efficiently reuse the data under standard settings and directly compare with prior works.

To this end, we pose the imminent question – *Can distractors improve few-shot generalization?* The answer is, somewhat surprisingly, yes. To elucidate how, we propose *ConFT*, a simple **f**ine**t**uning method based on a **con**trastive loss that contrasts pairs of the same class against those from different classes. We show that with a few simple but crucial modifications to the standard contrastive loss, distractors can be incorporated to boost generalization. We hypothesize that in the absence of extensive in-domain supervision for prior experience, distractor-aware finetuning can yield non-trivial gains. Towards the design of the loss function, we adopt an *asymmetric* construction of similarity pairs to ensure that *distractors contribute only through different-class pairs*. Our key insight here is two-fold – 1) generalization in contrastive learning can be influenced by not only same-class but also different-class pairs; 2) construction of different-class pairs is extremely flexible in that it can include samples from task-specific as well as task-exclusive categories. As a test of generality, we study the effect of our finetuning approach in conjunction with two different prior learning setups, namely, cross-domain and unsupervised prior learning. **Our contributions** are as follows.

- We propose contrastive finetuning, *ConFT*, a novel finetuning method for transfer based few-shot classification.
- We show how distractors can be incorporated in a contrastive objective to improve few-shot generalization.
- The proposed method outperforms state-of-the-art approaches by up to 12 points in the cross-domain few-shot learning and up to 5 points in unsupervised prior learning settings.

2. Related Work

2.1. Few-Shot Classification

Modern algorithms for few-shot classification are predominantly based on meta-learning where the goal is to quickly adapt to novel tasks. These approaches can be broadly classified into three categories: initialization based [12, 38, 39, 50, 53], hallucination based [2, 20, 71], and metric-learning based [3, 31, 57, 59, 62, 69] methods. Despite the growing interest in sophisticated meta-learning techniques, recent works [1, 6, 9, 64] have demonstrated that even simple finetuning based transfer learning [13, 17, 32, 44, 76] can outperform them. Such baselines usually involve cross-entropy training over the base categories followed by finetuning over a disjoint set of novel classes. Following these results, we further the investigation of finetuning for few-shot classification.

Cross-Domain Few-Shot Classification: A number of recent works [1, 11, 16, 40, 43, 54, 66, 67] have been proposed to address the cross-domain setup where base and novel classes are not only disjoint but also belong to different domains. Interestingly, [6] demonstrated that in this setup too, finetuning based transfer approaches outperformed popular meta-learning methods by significant margins. Following that, [66] proposed to learn feature-wise transformations via meta-learning to improve few-shot generalization of metric-based approaches. While in standard finetuning, the embedding model is usually frozen to avoid overfitting, recent works like [1, 16] have shown that frozen embeddings can hinder few-shot generalization. In this work, we build upon these developments to propose a more effective finetuning method over the entire embedding model.

In the context of learning from heterogeneous domain, [65] introduced a benchmark for multi-domain few-shot classification. This benchmark has been adopted by some recent works [7, 10, 37, 55]. While multiple base domains can alleviate cross-domain learning, we test our approach on a more challenging setup [66] that only involves a single base domain. Recent works used [66] as a benchmark to evaluate the importance of representation change [40] and spatial contrastive learning [43] in cross-domain few-shot classification. Another related work [67] leveraged unlabelled data from the novel domain in addition to few-shot labelled data to improve the task performance in a similar benchmark [16]. In contrast to [67], we operate under a limited access to novel domain data, *i.e.* only the few-shot labelled data.

Unlabelled Data in Few-Shot Classification: Our use of unlabelled data in the form of distractors is inspired from cognitive neuroscience studies [36] describing the effect of visual distractors on learning and memory. Prior works that use additional unlabelled data for few-shot classification include [4,14,35,51,61,72]. Complementary to [4,14,61] that exploit unlabelled data via self-supervised objectives in the prior learning phase, we use unlabelled data specifically for task-specific finetuning. Nonetheless, combining both perspectives could yield further benefits and is left for future work.

More related approaches [35, 51] combined heterogeneous unlabelled data, *i.e*., task-specific data and distractors, in a semi-supervised framework. Our distractor-aware finetuning differs from these works in two important ways: our few-shot classification is strictly inductive in that we do not use unlabelled data specific to the task, and our method leverages distractors instead of treating them as interference that needs to be masked out. The most relevant methods [1, 13], like us, reused the base (or source) domain as a source for additional data. The key difference, however, is that their success relies on effective alignment of the base and novel classes, whereas we benefit from contrasting the two. While the importance of distractor-aware learning has been investigated in the context of object detection [46, 78], their benefit to few-shot generalization has not been studied before.

Recently, [25,27,28] have studied few-shot classification in the context of unsupervised prior-learning where the base data is unlabelled. In this work, we evaluate the benefit of contrastive finetuning under this setting and and compare it to existing methods.

2.2. Contrastive Learning

Contrastive learning yields a similarity distribution over data by comparing pairs of different samples [58]. Recently, contrastive learning [18, 19, 56, 60] based methods have emerged as the state of the art for supervised [26, 29, 74] and self-supervised [5, 21, 23, 24, 42, 63, 75] representation learning. While the supervised approaches primarily exploit ground-truth labels to construct same-class pairs, self-supervised techniques leverage domain knowledge in the form of data augmentation to generate such pairs. As a special case, [29] maximized the benefit by integrating both forms of contrastive losses into a single objective. In this work, we use a modified version of the supervised contrastive loss when more than one labelled example is available per category. However, in the extreme case of 1-shot classification, it switches to self-supervised contrastive learning. Recent works such as [10, 43] also explored contrastive learning in the context of few-shot classification. While they use contrastive objectives at the *prior-learning* stage to learn a general-purpose representation solely on the base domain, our method uses a contrastive objective at *finetuning* to improve the *downstream-task-specific* representation directly on the target domain task with base domain data as distractors. As a design choice, we adopt the contrastive loss over other losses like cross-entropy since it allows us to leverage distractor data that does not belong to the novel categories but improves generalization.

3. Our Approach

To achieve the goal of few-shot generalization, our contrastive finetuning method, ConFT, optimizes for two simultaneous objectives. First, it aims to bring task-specific samples that share the same class close to each other; and second, it strives to push apart samples that belong to different classes. This two-fold objective can lead to compact clusters that are well separated amongst each other. In the following sections, we first introduce some notations that we then use to formally describe our approach. An overview of our method is presented in Fig. 2.

3.1. Preliminaries

Consider an input space $\mathcal{X}$ and a categorical label set $\mathcal{Y} = \{c_1, \dots, c_M\}$ where each of the M classes is represented via one-hot encoding. A representation space $\mathcal{R} \subset \mathbb{R}^r$ of the input is defined by the composition of an augmentation function $\mathcal{A} : \mathcal{X} \to \mathcal{X}$ and a representation model $\mathcal{M}_\theta : \mathcal{X} \to \mathcal{R}$, parameterized by θ. The augmentation function is a composition of standard image transformations such as random cropping, color jittering, horizontal flipping *etc*. Given a small number K, a few-shot classification task τ can be defined as the collection of a support set, $\tau_{\text{supp}} = \{(x_i, y_i) |\, x_i \in \mathcal{X},\ y_i \in \mathcal{Y}, i \in I_{\text{supp}}\}$ with K examples per class, and a query set, $\tau_q = \{\tilde{x}_j |\, \tilde{x}_j \in \mathcal{X}, j \in I_q\}$ sampled from the same (but unobserved) classes. Here, I_{supp} and I_q are the collection of indices for the support and query sets, respectively. The few-shot classification goal is to leverage the support set to obtain a classifier for the query samples. In this case, the classifier is constructed over the representation model obtained via *contrastive finetuning* of a prior model, $\mathcal{M}_{\theta_0}$ over τ_{supp}.

3.2. The ConFT Objective

A key component of the ConFT objective is that it includes unlabelled samples, *distractors*, to improve few-shot generalization. Formally, a distractor set, $S_{\text{dt}} = \{x_i | x_i \in \mathcal{X}, i \in I_{\text{dt}}\}$, drawn from a domain $D : \mathcal{X} \times \mathcal{Y}_D$ together with the task-specific support set τ_{supp}, constitutes the training data for few-shot learning. Here, the distractor class set $\mathcal{Y}_D$ is assumed to be *task-exclusive*, $\mathcal{Y}_D \cap \mathcal{Y} = \emptyset$. Starting with a support set example i (a.k.a anchor), we first construct an anchor-negative index set, $N(i) = \{p \in I_{\text{supp}} |\, y_i \neq y_p\}$, and an anchor-positive index set $P(i)$ such that $y_p = y_i, \forall p \in P(i)$. Samples indexed by $N(i)$ are treated as negatives within the task, whereas those indexed

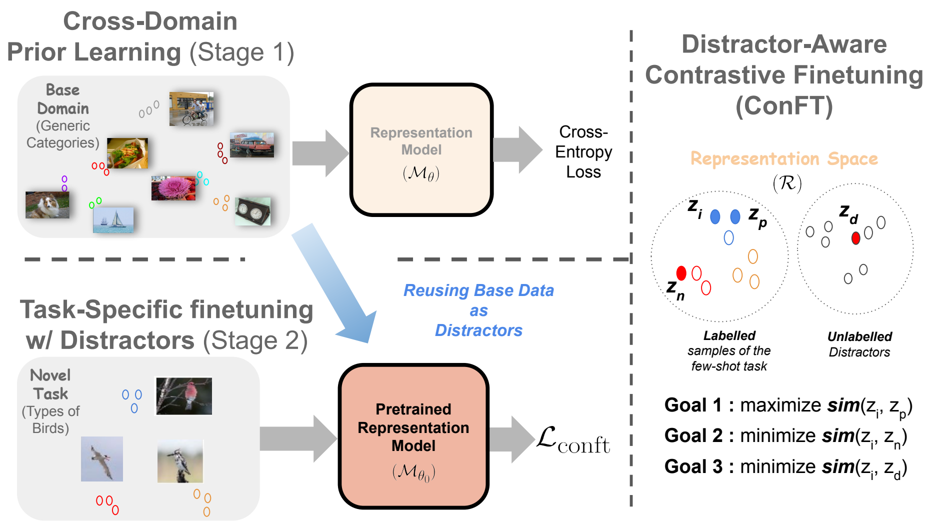

Figure 2. **Contrastive Finetuning in Cross-Domain Few-Shot Learning:** Our contrastive finetuning approach to few-shot classification comprises of two stages: 1) The *prior learning* stage trains a representation model on labelled (under cross-domain settings) base data using a cross-entropy loss; 2) The pretrained representation model is then finetuned over *task-specific* samples as well as *distractors* using a contrastive loss. For each task sample z_i, the contrastive objective (*right*) maximizes a similarity score, *sim*, over same-class pairs while minimizing it over other pair types. In the absence of enough labelled examples, distractors can improve classification by pushing apart task-specific clusters (here, different classes of birds).

by I_{dt} act as negatives exclusive to the task. Finally, we define our contrastive loss that uses a l_2-normalized representation $z \in \mathbb{R}^r$ as follows

$$\mathcal{L}_{\text{conft}}(\theta) = -\frac{1}{|I_{\text{supp}}|}\sum_{i\in I_{\text{supp}}}\frac{1}{|P(i)|}\sum_{p\in P(i)}\log l_{ip},$$
$$l_{ip} = \frac{\exp(\frac{z_i\cdot z_p}{\gamma})}{\exp(\frac{z_i\cdot z_p}{\gamma}) + \sum\limits_{n\in N(i)}\exp(\frac{z_i\cdot z_n}{\gamma}) + \sum\limits_{d\in I_{\text{dt}}}\exp(\frac{z_i\cdot z_d}{\gamma})} \tag{1}$$

where γ is a temperature hyper-parameter. The finetuning objective is simply the minimization of $\mathcal{L}_{\text{conft}}$ to yield optimal parameters θ_τ specific to task τ. To classify the query samples, we construct a nearest-mean classifier [11, 47, 58] atop the updated representation $\mathcal{M}_{\theta_\tau}$. The class-specific weight vectors are computed as an average over the representations of K support examples pertaining to that class. The j^{th} query sample is then assigned to the class whose weight vector has the largest cosine similarity (and hence, nearest in the Euclidean sense) with the query representation. We use the accuracy of this classifier to compare various baselines in the experiment section.

Construction of Anchor-Positive Set $P(i)$: To construct an anchor-positive set, we randomly pair task-samples belonging to the same class with no sample occurring in more than one pair. In each pair, if one is assigned to be the anchor, the other acts as its positive. As an example, in a 5-way 4-shot task, our stochastic pair construction will result in 10 pairs where each of the 5 classes has 2 pairs. In the case when the number of shots is odd, we omit one sample from each class to allow even pairing. The omission is, however, not an issue in the overall scheme of finetuning where multiple steps of gradient descent optimization ensures that eventually each sample gets to participate with equal chance. In the special case 1-shot learning, anchor-positive sets are constructed similar to [5] using augmentation $\mathcal{A}$.

3.3. Relative Importance of Anchor-Negatives

Given the loss formulation of (1), both task-specific (few-shot) and task-exclusive (distractor) anchor-negatives influence the loss proportionate to their respective batch sizes. While the batch size of task-specific negatives $N(i)$ is upper bounded by the number of ways M and the number of shots K, the batch size of distractors can be made as large as that of the domain itself, *i.e.*, $|D|$. In standard contrastive learning paradigms with only task-specific and no task-exclusive training examples, large batch sizes of negatives are known to be beneficial for downstream task performance. However, in our case where both types of negatives exist, naively increasing distractor batch size can be counterproductive (shown in the supplementary). We suspect that too many distractors might overshadow the effect of task-specific negatives that can be more crucial for generalization. Also, the effect might vary according to the proximity of distractors with respect to task samples in the representation space. Nonetheless, there is a need to balance the undue influence of distractors by adjusting the batch sizes. To avoid an extensive search for an optimal batch size specific to the distractor domain, we propose a domain-agnostic weighting scheme for the anchor-negatives proportional to their batch sizes as follows

$$l_{ip} = \frac{\exp(\frac{z_i \cdot z_p}{\gamma})}{\exp(\frac{z_i \cdot z_p}{\gamma}) + \alpha \sum_{n \in N(i)} \exp(\frac{z_i \cdot z_n}{\gamma}) + (2-\alpha) \sum_{d \in I_{\text{dt}}} \exp(\frac{z_i \cdot z_d}{\gamma})}, \tag{2}$$

$$\alpha = 2 \frac{|I_{\text{dt}}|}{|N(i)| + |I_{\text{dt}}|}$$

We found that this simple weighting scheme makes the few-shot performance robust to batch size variations and also improves the overall performance (see the supplementary).

4. Prior Learning and Distractors

Thus far we have assumed the access to a distractor set, S_{dt} and a prior model $\mathcal{M}_\theta$. In this section, we describe how to obtain them and how distractors boost generalization. Recall that our goal is to achieve few-shot generalization by finetuning a prior model over the few-shot task. However, due to the scarcity of task-specific labelled examples, a reasonably strong prior encoded in the model parameters θ_0 is crucial for preventing overfitting, especially when using high-capacity models like neural networks. We next describe two different ways of learning such a prior that can serve as a good initialization for subsequent finetuning.

4.1. Types of Prior Learning

Cross-Domain Learning: In the cross-domain setup, we are provided a labelled dataset, $D_l = \{(x_i, y_i) | x_i \in \mathcal{X}_{\text{sc}}, y_i \in \mathcal{Y}_{\text{sc}}\}_{i=1}^{|D_l|}$ drawn from a source domain $\mathcal{X}_{\text{sc}} \times \mathcal{Y}_{\text{sc}}$, such that the categorical label set $\mathcal{Y}_{\text{sc}}$ is disjoint from novel categories $\mathcal{Y}$. The key characteristic of this setup is that the distribution of M-way K-shot tasks, if constructed out of D_l, will be significantly different from novel tasks in the target domain. Such distribution shift could arise due to difference in task granularity (*e.g.*, coarse-grained vs. fine-grained) or shift in input distribution or both. In this work, we consider the case where the shift in task granularity is notably more than the input distribution. Towards the goal of learning a reasonably strong prior, we adopt a simple objective that minimizes cross-entropy loss over all categories in D_l. During finetuning, the distractors are sampled from D_l, thus, naturally satisfying the non-overlapping categories assumption with respect to novel tasks.

Unsupervised Prior Learning: For unsupervised prior learning, we are given an unlabelled dataset, $D_u = \{v_i\}_{i=1}^{|D_u|}$ drawn from a source domain $\mathcal{X}_{\text{su}} \times \mathcal{Y}_{\text{su}}$, such that the corresponding labels in $\mathcal{Y}_{\text{su}}$ are unobserved. While there are no explicit assumptions about the task distribution gap in this setting, the strength of the learnt prior is likely to be more reasonable when the distribution gap is small. To learn a suitable prior using D_u, we use the SimCLR loss [5] as a form of self-supervised objective. Our choice of this objective over others [21] was based on its superior performance found in our preliminary experiments. Priors learnt via self-supervised contrastive objectives on large base datasets (like, ImageNet) have been shown to transfer well to many-shot downstream tasks. In this work we show that such objectives are effective even with smaller base datasets and few-shot downstream tasks. In our experiments, priors learnt in this way already outperform state-of-the-art approaches [25,27,28] that are then further improved by our proposed contrastive finetuning. In this setup, we use D_u as the source for distractors where the assumption of non-overlapping categories is satisfied with high probability, provided the base dataset is relatively large and encapsulates a wide variety of categorical concepts.

4.2. Distractor-Aware Generalization

The most important and perhaps surprising aspect of our method is that distractors, despite being drawn from unrelated (to novel task) categories, can improve generalization. To understand the underlying mechanism, we propose to measure the change in quality of task-specific representation before and after finetuning. Particularly, given a few-shot task with M classes, we define the subset of query samples, $I_q^c \subset I_q$ that share the same class[1] and two other quantities – cluster spread u_{spread}^q and cluster-separation u_{sep}^q that measure the degree of clustering in the representation space. Specifically,

$$u_{\text{spread}}^q(\theta_t) = \frac{1}{M} \sum_{m=1}^{M} \sum_{\substack{i \in I_q^{c_m} \\ j \in I_q^{c_m} \setminus \{i\}}} (1 - z_i \cdot z_j), \tag{3}$$

$$u_{\text{sep}}^q(\theta_t) = \frac{1}{M} \sum_{m=1}^{M} \sum_{\substack{i \in I_q^{c_m}, \\ j \in I_q \setminus I_q^{c_m}}} (1 - z_i \cdot z_j), \tag{4}$$

where θ_t are the parameters of the representation model after t finetuning epochs. For each of the above quantities, we define the change, $\delta_*^q(t) = u_*^q(\theta_t) - u_*^q(\theta_0)$, and relative change, $\delta_*^{\text{rel,q}}(t) = \frac{\delta_*^q(t)}{\kappa(\theta_0)}$ where, the subscript can be *sep* or *spread* and division by a fixed value, $\kappa(\theta_0)$ ensures scale invariance. Finally, to quantify generalization within a given target domain, we define the average relative change, $\mathbb{E}_\tau[\delta_*^{\text{rel,q}}(t)]$ over a large number of tasks sampled from that domain. The average relative change can also be defined for

[1]Note that the query class labels are considered only for analysis purposes. In practice, they are not observed.

support examples by simply swapping superscript 'q' with 's'. Also, in practice, we use $u^s_{\text{sep}}(\theta_0)$ as the fixed value for $\kappa(\theta_0)$ irrespective of the superscript or subscript.

4.3. A Multitask Variant of ConFT

While our original objective (2) is agnostic to distractor supervision, finetuning in the cross-domain setting can further benefit from distractor labels. To that end, we introduce an auxiliary loss $\mathcal{L}_{\text{mtce}}$ during finetuning that minimizes the cross-entropy between predicted probabilities and one-hot encodings of the ground-truth label averaged over the base data, D_l. This leads to a new multitask formulation

$$\mathcal{L}_{\text{mt-conft}} = \mathcal{L}_{\text{conft}} + \lambda \mathcal{L}_{\text{mtce}}, \tag{5}$$

where we fix the relative weighting factor $\lambda = 1$ in our experiments and use a cosine classifier [6] for $\mathcal{L}_{\text{mtce}}$. We found that this simple extension led to significant performance gains in some domains while marginal in others, depending on domain characteristics.

5. Experiments

Following sections first introduce some baselines (§5.1) and present our main results for contrastive finetuning in the cross-domain setup (§5.2). Then, §5.3 elucidates the generalization mechanism of ConFT followed by ablations in §5.4. Finally, §5.5 demonstrates the performance of our approach in the unsupervised prior learning setup.

Datasets and Benchmarks: We evaluate our proposed finetuning method in a variety of novel domains spanning across two different paradigms for prior learning. For cross-domain evaluations, we adopt the benchmark introduced by [66] that comprises of Cars [33], CUB [73], Places [77], and Plantae [68] as the novel domains and *mini*ImageNet [49] as the base domain. Each dataset is split into *train*, *val* and, *test* categories (please refer to the supplementary for details), where tasks sampled from the *test* split are used to evaluate the few-shot performance in respective domains. We use the val splits for cross-validating the hyperparameters and the train split of *mini*ImageNet as our base data. For experiments in unsupervised prior learning, we use the same train split of *mini*ImageNet to learn a self-supervised representation that is then evaluated for few-shot performance on *mini*ImageNet-test. We present additional results on Meta-Dataset [65] in the supplementary.

Backbone (Representation Model): Following best practices in cross-domain few-shot learning, we adopt a ResNet10 [22] model for most of our experiments. In the unsupervised learning case, we use a four-layer CNN consistent with existing works except for a reduced filter size from 64 to 20 in the final layer. This modification was found to improve contrastive finetuning performance.

Optimization and Hyperparameters: In this work, we evaluate few-shot performance over 5-way 1-shot and 5-way 5-shot tasks with 15 query samples, irrespective of the prior learning setup. For the contrastive finetuning, we use an ADAM [30] optimizer with a suitable learning rate and early-stopping criteria. Our proposed method has a few hyperparameters such as the temperature (γ), learning rate, early-stopping criteria, and data augmentation ($\mathcal{A}$). However, recent studies [41] have highlighted that excessive hyperparameter tuning on large validation sets can lead to overoptimistic results in limited-labelled data settings like semi-supervised learning. Thus, we keep an extremely small budget for hyperparameter tuning. Among the mentioned hyperparameter, the one with the most number of parameters is the augmentation function $\mathcal{A}$. In this work, we do not tune $\mathcal{A}$ to any specific target domain. Instead, we use a fixed augmentation scheme introduced by [6] for the cross-domain setting and AutoAugment [8] for the unsupervised prior learning case. Please refer to the supplementary for a detailed summary of hyperparameters used in our experiments.

5.1. Baseline Comparisons

We begin our evaluations by comparing various baselines for finetuning in Table 1. These include two simple baselines (introduced in [6]) and two strong baselines (introduced in [16]). While the simple baselines freeze the backbones, the others allow finetuning over the entire embedding model. Another key difference is that the simple baselines are evaluated using standard linear evaluation [5, 6], whereas the rest are evaluated using nearest-mean classifiers. We compare the performance of all these baselines to our vanilla and multi-task (MT) versions of ConFT. Following previous works, the learning rates for the simple baselines are kept at 0.01, whereas for others (including ours), we use smaller learning rates (0.005 or 0.0005). We observe that among the baselines, the cosine classifier based baseline, FT-all (CC), outperforms the linear classifier based FT-all (LC). However, both versions of our finetuning approach significantly outperform all baselines across various dataset and shot settings.

5.2. ConFT for Cross-Domain Prior Learning

In this section, we present our main results on cross-domain few-shot learning (see Table 2). We compare our approach with various prior works on the LFT benchmark [66]. We observe that overall our proposed approaches, ConFT and MT-ConFT, significantly outperform the best previous results in Cars (by 3 **to** 12 **points**), Plantae (by 7 **to** 9 **points**) and CUB (by 1.7 **to** 2 **points**) domains. We also observe higher gains in the 5-shot setting than the 1-shot case, since more labelled examples can improve few-shot generalization. Further, we find that using the auxiliary

Finetuning Method		CUB		Cars		Places		Plantae	
Loss	FT Type	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Cross-Entropy	fixed-BB (LC) [6]	39.77 ± 0.66	51.33 ± 0.70	33.99 ± 0.64	44.14 ± 0.70	44.53 ± 0.75	55.94 ± 0.69	37.07 ± 0.70	46.58 ± 0.69
	fixed-BB (CC) [6]	43.26 ± 0.76	62.87 ± 0.74	25.33 ± 1.85	50.40 ± 0.74	47.70 ± 0.76	69.48 ± 0.69	40.49 ± 0.77	56.64 ± 0.72
	FT-all (LC) [16]	40.81 ± 0.75	61.82 ± 0.72	34.50 ± 0.67	55.63 ± 0.75	45.91 ± 0.77	68.73 ± 0.73	37.51 ± 0.71	58.33 ± 0.68
	FT-all (CC) [16]	44.30 ± 0.73	67.05 ± 0.69	36.79 ± 0.76	57.65 ± 0.76	49.10 ± 0.78	70.32 ± 0.72	40.31 ± 0.76	61.30 ± 0.75
Contrastive	-	43.42 ± 0.75	62.80 ± 0.76	35.19 ± 0.66	51.41 ± 0.72	49.56 ± 0.80	70.71 ± 0.68	40.39 ± 0.79	55.54 ± 0.69
	ConFT (ours)	45.57 ± 0.76	70.53 ± 0.75	$\mathbf{39.11 \pm 0.77}$	61.53 ± 0.75	$\mathbf{49.97 \pm 0.86}$	72.09 ± 0.68	$\mathbf{43.09 \pm 0.78}$	62.54 ± 0.76
	MT-ConFT (ours)	$\mathbf{49.25 \pm 0.83}$	$\mathbf{74.45 \pm 0.71}$	37.36 ± 0.69	$\mathbf{62.54 \pm 0.72}$	$\mathbf{49.94 \pm 0.81}$	$\mathbf{72.71 \pm 0.69}$	41.82 ± 0.75	$\mathbf{63.01 \pm 0.74}$

Table 1. **Baseline Comparisons.** Results on 1-shot and 5-shot tasks on the LFT benchmark [66]. These results are obtained by averaging over 600 novel tasks, each consisting of 5 classes and 15 queries per class. We also present 95% confidence intervals. The train split of the *mini*ImageNet dataset is used as base data. Here, FT-all denotes the case where the entire embedding model is finetuned. Other abbreviations – BB: Backbone model (ResNet-10), LC: Linear Classifier, CC: Cosine Classifier with a multiplication factor of 10.

Method			1-shot			
Prior Learning	Task Specific Finetuning	Backbone	CUB	Cars	Places	Plantae
AAL [1]	arcmax	ResNet18	47.25 ± 0.76	-	-	-
MN [69]	-	ResNet10	35.89 ± 0.51	30.77 ± 0.47	49.86 ± 0.79	32.70 ± 0.60
MN w/ featTx [66]	-	ResNet10	36.61 ± 0.53	29.82 ± 0.44	51.07 ± 0.68	34.48 ± 0.50
RN [62]	-	ResNet10	42.44 ± 0.77	29.11 ± 0.60	48.64 ± 0.85	33.17 ± 0.64
RN w/ featTx [66]	-	ResNet10	44.07 ± 0.77	28.63 ± 0.59	50.68 ± 0.87	33.14 ± 0.62
GNN [57]	-	ResNet10+	45.69 ± 0.68	31.79 ± 0.51	53.10 ± 0.80	35.60 ± 0.56
GNN w/ featTx [66]	-	ResNet10+	47.47 ± 0.75	31.61 ± 0.53	$\mathbf{55.77 \pm 0.79}$	35.95 ± 0.58
MAML [12]	-	Conv4	40.51 ± 0.08	33.57 ± 0.14	-	-
ANIL [48]	-	Conv4	41.12 ± 0.15	34.77 ± 0.31	-	-
BOIL [40]	-	Conv4	44.20 ± 0.15	36.12 ± 0.29	-	-
CE Training	-	ResNet10	43.42 ± 0.75	35.19 ± 0.66	49.56 ± 0.80	40.39 ± 0.79
CE Training	ConFT (ours)	ResNet10	45.57 ± 0.76	$\mathbf{39.11 \pm 0.77}$	49.97 ± 0.86	$\mathbf{43.09 \pm 0.78}$
CE Training	MT-ConFT (ours)	ResNet10	$\mathbf{49.25 \pm 0.83}$	37.36 ± 0.69	49.94 ± 0.81	41.82 ± 0.75

Method			5-shot			
Prior Learning	Task Specific Finetuning	Backbone	CUB	Cars	Places	Plantae
Baseline [6]	-	ResNet18	65.57 ± 0.70			
Baseline ++ [6]	-	ResNet18	62.04 ± 0.76			
DiversityNCoop [11]	-	ResNet18	66.17 ± 0.55	-	-	-
AAL [1]	arcmax	ResNet18	72.37 ± 0.89	-	-	-
BOIL [40]	-	ResNet12	-	49.71 ± 0.28	-	-
MN [69]	-	ResNet10	51.37 ± 0.77	38.99 ± 0.64	63.16 ± 0.77	46.53 ± 0.68
MN w/ featTx [66]	-	ResNet10	55.23 ± 0.83	41.24 ± 0.65	64.55 ± 0.75	41.69 ± 0.63
RN [62]	-	ResNet10	57.77 ± 0.69	37.33 ± 0.68	63.32 ± 0.76	44.00 ± 0.60
RN w/ featTx [66]	-	ResNet10	59.46 ± 0.71	39.91 ± 0.69	66.28 ± 0.72	45.08 ± 0.59
GNN [57] [57]	-	ResNet10+	62.25 ± 0.65	44.28 ± 0.63	70.84 ± 0.65	52.53 ± 0.59
GNN w/ featTx [66]	-	ResNet10+	66.98 ± 0.68	44.90 ± 0.64	$\mathbf{73.94 \pm 0.67}$	53.85 ± 0.62
MAML [12]	-	Conv4	53.09 ± 0.16	44.56 ± 0.21	-	-
ANIL [48]	-	Conv4	55.82 ± 0.21	46.55 ± 0.29	-	-
BOIL [40]	-	Conv4	60.92 ± 0.11	50.64 ± 0.22	-	-
CE Training	-	ResNet10	62.80 ± 0.76	51.41 ± 0.72	70.71 ± 0.68	55.54 ± 0.69
CE Training	ConFT (ours)	ResNet10	70.53 ± 0.75	61.53 ± 0.75	72.09 ± 0.68	62.54 ± 0.76
CE Training	MT-ConFT (ours)	ResNet10	$\mathbf{74.45 \pm 0.71}$	$\mathbf{62.54 \pm 0.72}$	72.71 ± 0.69	$\mathbf{63.01 \pm 0.74}$)

Table 2. **Cross-Domain Few-Shot Classification Results.** We present the results with 95% confidence intervals and highlight the best performing methods. The results are an average over 600 tasks. Here, '-' denotes numbers not reported by previous works.

loss (MT-ConFT) is more beneficial in the 5-shot case. In fact, it performs worse than ConFT in the 1-shot cases for Cars, Places, and Plantae. Such a degradation could be due to a misalignment between the self-supervised objective (to which ConFT boils down in the 1-shot case) and the auxiliary cross-entropy loss. In the Places domain, "GNN w/ featTx" yields the best performance, whereas our approach outperforms the rest for the 5-shot case. We suspect that the use of a more sophisticated model in"GNN w/ featTx", namely, graph neural net [57] built on top of a ResNet-10 model, leads to a better cross-domain generalization when the domain gap is smaller.

5.3. Effect of Distractors on Generalization

In this section, we investigate the central question – *How do distractors improve generalization?* We present two sets of plots in Figure 3 that track the change in cluster-spread and cluster-separation as the finetuning progresses. In the first set, we plot the average relative change, $\mathbb{E}[\delta^{\text{rel}}_{\text{spread}}(t)]$ and $\mathbb{E}[\delta^{\text{rel}}_{\text{sep}}(t)]$ (see §4.2) as a function of finetuning epochs, t for both support and query samples in 2 different settings, namely, CUB (5-shot) and Cars (5-shot). We observe that for support examples (yellow lines), cluster-spread decreases with increasing epochs while the cluster-separation increases. This is indeed what is expected for training datapoints (here, support examples) and serves as a sanity check. For query samples (green plot), on the other hand, both cluster-spread and separation increase with the progress in finetuning epochs. The key observation, however, is that cluster-separation increases to a much greater extent than the cluster-spread, thus improving overall discriminability between classes represented by these clusters.

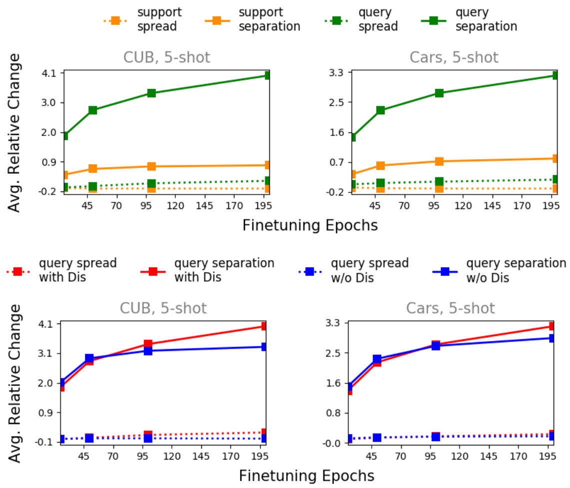

Figure 3. **Understanding Generalization in ConFT.** *Top:* Average relative change in cluster-spread and cluster-separation of support and query samples as a function of finetuning epochs. *Bottom:* Comparing the average relative change in cluster-spread and cluster-separation of only query samples under the presence and absence of distractors. The spread and separation quantities are averaged over 600 tasks for both *top* and *bottom*.

Anchor-Positives		Anchor-Negatives			Accuracy	
SPC	T-Pos	T-Neg	D-Neg	W	CUB, 5-shot	Cars, 5-shot
	✓	✓	✓	✓	$69.61_{\pm 0.68}$	$61.01_{\pm 0.74}$
✓			✓		$70.16_{\pm 0.70}$	$57.42_{\pm 0.80}$
✓		✓			$67.44_{\pm 0.71}$	$59.00_{\pm 0.73}$
✓		✓	✓		$70.26_{\pm 0.68}$	$60.58_{\pm 0.77}$
✓		✓	✓	✓	$70.53_{\pm 0.75}$	$61.53_{\pm 0.75}$

Table 3. **Ablation 1.** Novel task performance with various types of anchor-positives and anchor-negatives. Here, SPC: Stochastic Pair Construction, T: Task, D: Distractor, W: relative weighting.

While the increase in cluster-separation hints towards the possible reason behind improved generalization, it is not clear how much of the improvement, if any, is a consequence of incorporating distractors. To delineate the effects of distractors from the contrastive loss itself, we present the second set of plots that compare the average relative change in *query* cluster spread and separation under the presence (red line) and absence (blue line) of distractors for the same data settings. We observe that with increasing finetuning epochs the gap between cluster-separation and cluster-spread widens to a larger extent in the presence of distractors than in their absence. This leads to our conclusion that **distractors help generalization by increasing task-specific cluster separation.**

5.4. Ablations

In this section, we introduce a few important ablations that help deconstruct the ConFT and MT-ConFT objectives. In Table 3, we compare our stochastic anchor-positive construction with naive inclusion of all positives for every anchor. We also, ablate the contribution of each type of

Distractor Domain Size				Accuracy	
512	1024	2048	38400	CUB 5-shot	Cars 5-shot
✓				70.34 ± 0.71	61.16 ± 0.76
	✓			70.28 ± 0.70	61.11 ± 0.77
		✓		70.92 ± 0.69	61.31 ± 0.76
			✓	70.53 ± 0.75	61.53 ± 0.75

Table 4. **Ablation 2.** Novel task performance with varying sizes of the distractor domain, *i.e.*, *mini*ImageNet-train. Note that this is *different* from distractor batch size $|S_{\text{dt}}|$.

Method			5-way *mini*ImageNet	
Prior Learning	Finetuning	BB	1-shot	5-shot
Sup. MML	MFT	Conv4	$46.81_{\pm 0.77}$	$62.13_{\pm 0.72}$
Sup. PN	-	Conv4	$46.56_{\pm 0.76}$	$62.29_{\pm 0.71}$
-	RandInit	Conv4	$27.59_{\pm 0.59}$	$38.48_{\pm 0.66}$
BG-MML [25]	MFT	Conv4	$36.24_{\pm 0.74}$	$51.28_{\pm 0.6}$
BG-PN [25]	-	Conv4	$36.62_{\pm 0.70}$	$50.16_{\pm 0.7}$
DC-MML [25]	MFT	Conv4	$39.90_{\pm 0.74}$	$53.97_{\pm 0.70}$
DC-PN [25]	-	Conv4	$39.18_{\pm 0.71}$	$53.36_{\pm 0.70}$
U-MML [27]	MFT	Conv4	39.93	50.73
LG-MML [28]	MFT	Conv4	$40.19_{\pm 0.58}$	$54.56_{\pm 0.55}$
LG-PN [28]	-	Conv4	$40.05_{\pm 0.60}$	$52.53_{\pm 0.51}$
SimCLR [5]	-	Conv4	$41.54_{\pm 0.61}$	$56.57_{\pm 0.59}$
SimCLR	ConFT	Conv4	$\mathbf{43.45}_{\pm \mathbf{0.60}}$	$\mathbf{60.02}_{\pm \mathbf{0.57}}$

Table 5. **Unsupervised Prior Learning.** The results are averaged over 1000 novel tasks and are presented with 95% confidence intervals. Here, MFT refers to meta-style fintuning [12]. MML:Maml, PN: ProtoNet, DC: DeepCluster, U: Umtra, LG: Lasium-Gan, BG: BiGAN, Sup.:Supervised, BB: Backbone.

anchor-negatives: task-specific and distractors and compare that to relative weighting of the two. In Table 4, we studied the importance of distractor domain size and found that novel task performance is fairly robust to the size of the distractor domain. This is particularly encouraging since we need not store the entire base data during finetuning.

5.5. ConFT for Unsupervised Prior Learning

In Table 5, we demonstrate the generality of contrastive finetuning by evaluating on the unsupervised prior learning benchmark *mini*ImageNet. The key distinction from cross-domain settings is that we do not have labelled base data to learn from. So, we leverage self-supervised contrastive learning [5] on the unlabelled base data and show that it outperforms state of the art by 1 to 2 points. Finetuning the resultant representation with our ConFT objective further improves the accuracy by 2to 4 points. This is particularly significant, as the results come very close to supervised baselines that serve as performance upper bound in this setting [28].

6. Conclusion

We introduce a novel contrastive finetuning approach to few-shot classification. Specifically, our method leverages distractors to improve generalization by encouraging cluster separation of the novel task samples. We show that our method leads to significant performance gains in both cross-domain and unsupervised prior learning setups.

References

[1] Arman Afrasiyabi, Jean-François Lalonde, and Christian Gagn'e. Associative alignment for few-shot image classification. In *ECCV*, 2020. 1, 2, 3, 7

[2] Antreas Antoniou, Amos J. Storkey, and Harrison A Edwards. Data augmentation generative adversarial networks. *arXiv*, 2017. 2

[3] Jane Bromley, James Bentz, Leon Bottou, Isabelle Guyon, Yann Lecun, Cliff Moore, Eduard Sackinger, and Rookpak Shah. Signature verification using a "siamese" time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 7:25, 1993. 2

[4] Guy Bukchin, Eli Schwartz, Kate Saenko, Ori Shahar, R. Feris, Raja Giryes, and Leonid Karlinsky. Fine-grained angular contrastive learning with coarse labels. *arXiv*, 2020. 3

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv*, 2020. 3, 4, 5, 6, 8

[6] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *ICLR*, 2019. 2, 6, 7

[7] Yinbo Chen, Xiaolong Wang, Zhuang Liu, Huijuan Xu, and Trevor Darrell. A new meta-baseline for few-shot learning. *arXiv*, 2020. 2

[8] E. Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation policies from data. *arXiv*, 2018. 6

[9] Guneet Singh Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. A baseline for few-shot image classification. In *ICLR*, 2020. 1, 2

[10] C. Doersch, A. Gupta, and Andrew Zisserman. Crosstransformers: Spatially-aware few-shot transfer. In *NeurIPS*, 2020. 1, 2, 3

[11] Nikita Dvornik, Cordelia Schmid, and Julien Mairal. Diversity with cooperation: Ensemble methods for few-shot classification. In *ICCV*, 2019. 2, 4, 7

[12] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017. 1, 2, 7, 8

[13] Weifeng Ge and Yizhou Yu. Borrowing treasures from the wealthy: Deep transfer learning through selective joint fine-tuning. In *CVPR*, 2017. 2, 3

[14] Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Pérez, and Matthieu Cord. Boosting few-shot visual learning with self-supervision. In *CVPR*, 2019. 3

[15] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. *CVPR*, 2018. 1

[16] Yunhui Guo, Noel C Codella, Leonid Karlinsky, James V Codella, John R Smith, Kate Saenko, Tajana Rosing, and Rogerio Feris. A broader study of cross-domain few-shot learning. ECCV, 2020. 2, 6, 7

[17] Yunhui Guo, Honghui Shi, Abhishek Kumar, Kristen Grauman, Tajana Rosing, and Rogerio Feris. Spottune: Transfer learning through adaptive fine-tuning. *arXiv*, 2018. 2

[18] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*, 2010. 3

[19] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006. 3

[20] B. Hariharan and R. Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *ICCV*, 2017. 2

[21] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 3, 5

[22] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[23] Olivier J. Hénaff, A. Srinivas, J. Fauw, Ali Razavi, C. Doersch, S. Eslami, and A. Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv*, 2020. 3

[24] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR*, 2019. 3

[25] Kyle Hsu, Sergey Levine, and Chelsea Finn. Unsupervised learning via meta-learning. In *ICLR*, 2019. 1, 3, 5, 8

[26] Konstantinos Kamnitsas, Daniel C. Castro, Loïc Le Folgoc, Ian Walker, Ryutaro Tanno, Daniel Rueckert, Ben Glocker, Antonio Criminisi, and Aditya V. Nori. Semi-supervised learning via compact latent space clustering. In *ICML*, 2018. 3

[27] Siavash Khodadadeh, Ladislau Boloni, and Mubarak Shah. Unsupervised meta-learning for few-shot image classification. In *NeurIPS*, 2019. 1, 3, 5, 8

[28] Siavash Khodadadeh, Sharare Zehtabian, Saeed Vahidian, Weijia Wang, Bill Lin, and Ladislau Boloni. Unsupervised meta-learning through latent-space interpolation in generative models. In *ICLR*, 2021. 1, 3, 5, 8

[29] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *NeurIPS*, 2020. 3

[30] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, 2015. 6

[31] Gregory R. Koch. Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop. Vol. 2.*, 2015. 1, 2

[32] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better? *CVPR*, 2019. 2

[33] J. Krause, M. Stark, J. Deng, and L. Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV Workshops*, 2013. 6

[34] B. Lake, R. Salakhutdinov, and J. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350:1332 – 1338, 2015. 1

[35] 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 *NeurIPS*, 2019. 3

[36] Chen Liang, Xiao Yang, Neisarg Dave, Drew Wham, Bart Pursel, and C. Lee Giles. Distractor generation for multiple choice questions using learning to rank. In *Proceedings*

of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, 2018. 3

[37] Orchid Majumder, A. Ravichandran, Subhransu Maji, M. Polito, Rahul Bhotika, and Stefano Soatto. Revisiting contrastive learning for few-shot classification. *arXiv*, 2021. 2

[38] Tsendsuren Munkhdalai and Hong Yu. Meta networks. In *ICML*, 2017. 2

[39] Alex Nichol and John Schulman. Reptile: A scalable metalearning algorithm. *arXiv*, 2018. 2

[40] Jaehoon Oh, Hyungjun Yoo, ChangHwan Kim, and Se-Young Yun. {BOIL}: Towards representation change for few-shot learning. In *ICLR*, 2021. 2, 7

[41] Avital Oliver, Augustus Odena, Colin Raffel, Ekin D Cubuk, and Ian J Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. *arXiv*, 2018. 6

[42] A. Oord, Y. Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv*, 2018. 3

[43] Yassine Ouali, Céline Hudelot, and Myriam Tami. Spatial contrastive learning for few-shot classification. *arXiv*, 2020. 2, 3

[44] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 2010. 2

[45] Cheng Perng Phoo and Bharath Hariharan. Self-training for few-shot transfer across extreme task differences. In *ICLR*, 2021. 1

[46] Horst Possegger, Thomas Mauthner, and Horst Bischof. In defense of color-based model-free tracking. In *CVPR*, June 2015. 3

[47] Hang Qi, David Lowe, and Matthew Brown. Low-shot learning with imprinted weights. In *CVPR*, 2018. 4

[48] Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of maml. In *ICLR*, 2020. 7

[49] S. Ravi and H. Larochelle. Optimization as a model for few-shot learning. In *ICLR*, 2017. 1, 6

[50] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *ICLR*, 2017. 2

[51] Mengye Ren, Sachin Ravi, Eleni Triantafillou, Jake Snell, Kevin Swersky, Josh B. Tenenbaum, Hugo Larochelle, and Richard S. Zemel. Meta-learning for semi-supervised few-shot classification. In *ICLR*, 2018. 3

[52] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015. 1

[53] Andrei A. Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *ICLR*, 2019. 2

[54] Jeongun Ryu, Jaewoong Shin, H. B. Lee, and Sung Ju Hwang. Metaperturb: Transferable regularizer for heterogeneous tasks and architectures. *arXiv*, 2020. 2

[55] Tonmoy Saikia, T. Brox, and C. Schmid. Optimized generic feature learning for few-shot classification across domains. *arXiv*, 2020. 2

[56] Ruslan Salakhutdinov and Geoff Hinton. Learning a nonlinear embedding by preserving class neighbourhood structure. In *AISTATS*, 2007. 3

[57] Victor Garcia Satorras and Joan Bruna Estrach. Few-shot learning with graph neural networks. In *ICLR*, 2018. 1, 2, 7

[58] Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *ICML*, 2019. 3, 4

[59] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017. 1, 2

[60] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *NeurIPS*, 2016. 3

[61] Jong-Chyi Su, Subhransu Maji, and Bharath Hariharan. When does self-supervision improve few-shot learning? In *ECCV*, 2020. 3

[62] 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 *CVPR*, 2018. 1, 2, 7

[63] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv*, 2019. 3

[64] 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*, 2020. 1, 2

[65] Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, and Hugo Larochelle. Meta-dataset: A dataset of datasets for learning to learn from few examples. In *ICLR*, 2020. 2, 6

[66] Hung-Yu Tseng, Hsin-Ying Lee, Jia-Bin Huang, and Ming-Hsuan Yang. Cross-domain few-shot classification via learned feature-wise transformation. In *ICLR*, 2020. 1, 2, 6, 7

[67] Hung-Yu Tseng, Hsin-Ying Lee, Jia-Bin Huang, and Ming-Hsuan Yang. Cross-domain few-shot classification via learned feature-wise transformation. In *ICLR*, 2020. 2

[68] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *CVPR*, 2016. 1, 6

[69] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *NIPS*. 2016. 1, 2, 7

[70] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[71] Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. In *CVPR*, 2018. 2

[72] Yu-Xiong Wang and Martial Hebert. Learning from small sample sets by combining unsupervised meta-training with cnns. In *NeurIPS*, 2016. 3

[73] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010. 6

[74] Zhirong Wu, Alexei A Efros, and Stella Yu. Improving generalization via scalable neighborhood component analysis. In *ECCV*, 2018. 3

[75] Zhirong Wu, Yuanjun Xiong, X Yu Stella, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. 3

[76] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *NeurIPS*, 2014. 2

[77] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *TPAMI*, 40(6):1452–1464, 2018. 6

[78] Zheng Zhu, Qiang Wang, Li Bo, Wei Wu, Junjie Yan, and Weiming Hu. Distractor-aware siamese networks for visual object tracking. In *ECCV*, 2018. 3