This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Task-Level Contrastiveness for Cross-Domain Few-Shot Learning

Kristi Topollai Anna Choromanska
New York University
kt2664@nyu.edu, ac5455@nyu.edu

Abstract

Few-shot classification and meta-learning methods typically struggle to generalize across diverse domains, as most approaches focus on a single dataset, failing to transfer knowledge across various seen and unseen domains. Existing solutions often suffer from low accuracy, high computational costs, and rely on restrictive assumptions. In this paper, we introduce the notion of task-level contrastiveness, a novel approach designed to address issues of existing methods. We start by introducing simple ways to define task augmentations, and thereafter define a task-level contrastive loss that encourages unsupervised clustering of task representations. Our method is lightweight and can be easily integrated within existing few-shot/meta-learning algorithms while providing significant benefits. Crucially, it leads to improved generalization and computational efficiency without requiring prior knowledge of task domains. We demonstrate the effectiveness of our approach through different experiments on the MetaDataset benchmark, where it achieves superior performance without additional complexity.

1. Introduction

Deep learning has ignited remarkable progress in large scale image classification [24] to the point of even surpassing humans. However, deep learning approaches are still not able to efficiently learn new visual concepts and generalize to unseen domains or different tasks. This limits their applicability in the real world scenarios, where the number of data classes is infinite and the classes depend on the task itself. This challenge motivates a few-shot classification framework [52], where the objective is to learn a model capable of adapting to new classification tasks given only few labeled samples. The standard approach for this problem is to learn a general case feature extractor, which can then be adapted to new tasks. The meta learning paradigm can be employed to address the few shot learning problem [13, 41, 52], i.e., in this case, episodic training is used to simulate few-shot learning tasks and proper training of the model under both data and task sampling enables it to learn how to generalize on new incoming few-shot learning tasks.

Most of the traditional meta learning methods rely on a single global model which can learn to generalize on similar tasks, but lacks the representational capacity to accommodate vastly different ones. This results in a performance decrease when tasks come from different domains or even modalities, such as in the case when the few-shot tasks are generated from different data sets [48]. In this paper we consider this challenging setting and we propose a simple and effective method to improve algorithms operating in this regime. Our approach is based on contrastive learning, a self-supervised learning [2] paradigm where the pretext task is to project similar samples closer in the feature space while pushing dissimilar ones apart, facilitating representation learning without explicit labels. The main contributions of this paper can be summarized as follows

- We extend the contrastive learning framework for the multi task case, where instead of contrasting data representations, as was commonly done before, we instead contrast task representations in an unsupervised manner.
- We incorporate our method on the top of multi-domain few-shot and meta learning algorithms and evaluate it on the popular MetaDataset benchmark [48].
- We show that our approach, when incorporated into the multi-domain few-shot and meta learning schemes, leads to the following: **(a)** it improves generalization, **(b)** it lowers time and memory requirements, **(c)** it eliminates limiting assumptions, such as knowing the source of each task (i.e. data set of the task, its modality, etc.).

2. Related work

2.1. Meta learning

Meta learning methods are commonly categorized into one of the following three families: metric-based, optimization-based, and model-based methods. Metric-based meta-learning algorithms [34, 46, 52] are approaches that focus on learning a metric or distance function during training, allowing a model to quickly generalize its knowledge from a set of base tasks to novel, unseen tasks. These algorithms

aim to enable efficient adaptation by capturing similarities and differences between examples, making them particularly useful for tasks like few-shot learning. Optimization-based methods, such as MAML [13], seek parameters such that the model can quickly adapt to a new task by running gradient descent starting from these parameters. ANIL [40] is simillar to MAML, but only adapts the last layer, while R2D2 [6] and MetaOptNet [26] replace the last layer with a convex differentiable problem. Finally, model-based methods learn a model that explicitly adapts to new tasks. The most prominent family of methods is Neural Processes (NPs) [14, 16] , which encode a support set and estimate task-specific modulation parameters. All of the above methods adapt well to novel and unseen tasks only if they are similar to the training tasks.

2.2. Multi-domain few-shot learning

Handling heterogeneous cases involves constructing models that are aware of the tasks they are designed to tackle. Two mechanisms commonly used to achieve this are: domain expert routing and domain-aware modulation. Methods employing domain-aware modulation [3, 4, 42, 53, 57, 58] aim to adapt a globally shared model using task-specific information. In this setup, a task embedding network is utilized to extract domain-specific details, which are then used to modulate the model parameters. Obtaining an effective task representation is crucial in these methods. Domain expert routing, on the other hand, involves the mechanism of assigning tasks to specific "experts" for processing. These experts can be either different models or model components. Several methods [21, 25, 35, 51, 61] train the routing component in an unsupervised manner, without prior knowledge of the domain label of each task. Developing a reliable router (or mixer [1]) is crucial in this scenario.

Several cross-domain few shot learning methods employ both aforementioned mechanisms but rely on the knowledge of the domain label of each task. In FLUTE [49] parameters are split into a universal (task agnostic) and task-specific portion. In SUR [12], they pre-train multiple independent networks (one for each domain) and then, in order to learn in the multi-domain setting, for each task the method identifies the most useful features for this task across domain networks. In URL [27, 28] they solve the problem of storing multiple expert networks by distilling their knowledge to a single set of universal deep representations, while in tri-M [31] each domain is assigned its own domain specific modulation parameters. eTT [55] adapts a pre-trained ViT[11] using an attentive prefix tuning strategy. DIPA [36] further improves performance by replacing the standard NCC classifier[42] and introducing a parameter-efficient modulation method. Large foundation models like DINOv2 [33] also perform well on few-shot tasks, while Hu et al. [19] show that a simple pre-training, meta-training, and fine-tuning pipeline can be competitive.

[1] When multiple expert predictions are combined

2.3. Contrastive learning

Contrastive learning [9, 17, 18, 39, 44, 50, 59] has emerged as a successful approach in self-supervised learning for unsupervised representation learning. The contrastive loss operates in the latent space, aiming to pull the anchor sample towards positive samples while pushing it away from negatives [9, 17, 54]. A key aspect of contrastive schemes is the construction of positive and negative samples. Typically, randomly augmented samples from the anchor are considered positives, while all others are treated as negatives. Recent work has delved deeper into this, analyzing optimal augmentation strategies [47] and highlighting the importance of informative negative samples, particularly prioritizing hard negatives for contrastive loss calculation [10, 43, 45].

In the context of meta-learning and cross-domain few-shot learning, [56] investigates the use of contrastive learning during both pre-training and meta-learning phases to enhance few-shot learning performance. However, their study is limited to single-domain scenarios. In contrast, authors from [15], similar to our approach, leverage contrastive learning to learn a robust representation of the underlying data-generating function (i.e., the data distribution), and validate their method in multi-domain settings. Additionally, in [20, 38] contrastive learning is applied to improve performance in unsupervised few-shot classification tasks.

3. Preliminaries

Few-shot classification is usually formulated as a meta-learning problem. Instead of sampling from the training set, we create a series of learning tasks called episodes, each task consisting of few labeled examples split into a support set and a query set. Specifically, in each episode, a small subset of N classes are sampled from all training classes to construct a support set and a query set. The support set contains K examples (shots) for each of the N classes (i.e., N-way K-shot setting) denoted as $S = \{(X_{1,1}, y_{1,1}), \ldots, (X_{K,N}, y_{K,N})\}$, while the query set Q includes different samples from the same N classes. Under the meta learning formalization, the support set is used to adapt the global model and the query set is used to evaluate the adapted model and subsequently update the global model. This learning setup can be summarized as

$$\min_{\theta} \sum_{i=1}^{|T|} \mathcal{L}(Q_i, \mathcal{A}(\theta, S_i)), \tag{1}$$

where $|T|$ is the number of training tasks/episodes, $\mathcal{L}$ some loss function and $\mathcal{A}$ the adaptation operator, which given a parameter set θ and a support set S, outputs a set of

task adapted parameters $\hat{\theta} = \mathcal{A}(\theta, S)$. In MAML[13], the operator $\mathcal{A}(\theta, S)$ is a standard gradient descent, $\hat{\theta} = \theta - \alpha \nabla_\theta \mathcal{L}(\theta, S)$. In the multi-domain case, the support and query sets $T = (S, Q) \sim D$ are not sampled from a single task distribution D, but may come from one of several such distributions. Lets call the distribution over potential domains $\mathbb{D}$, then the goal is to minimize the expected loss over all domains:

$$\min_\theta \mathbb{E}_{D\sim\mathbb{D}}[\mathbb{E}_{(S,Q)\sim D}[\mathcal{L}(Q, \mathcal{A}(\theta, S))]]. \tag{2}$$

Naturally, the adaptation $\mathcal{A}$ in this case should be more sophisticated, so that it considers the domain information itself in addition to the support set samples. To this end, our focus will be on how to aid the unsupervised domain identification and domain specific adaptation of a task. Our method is based on contrastive learning and in particular SimCLR [9] where authors proposed a simple framework for contrastive learning of visual representations. The process is simple, first, a random minibatch of N samples is drawn and for each sample two different data augmentations are applied, resulting in $2N$ augmented samples in total. This is formalized as follows:

$$\mathbf{x}_1 = t(\mathbf{x}), \quad \mathbf{x}_2 = t'(\mathbf{x}), \quad t, t' \sim \mathcal{T},$$

where t and t' denote two separate data augmentation operators that are sampled from the same family of augmentations $\mathcal{T}$. (x_1, x_2) is one positive pair and the rest $2(N-1)$ data points are treated as negative samples. The contrastive learning loss is based on the NT-Xent loss, using cosine similarity $\mathrm{sim}(.,.)$ on some projection g of the representations $f_\theta(x)$.

$$\mathbf{z}_1 = g\left(f_\theta\left(\mathbf{x}_1\right)\right), \quad \mathbf{z}_2 = g\left(f_\theta\left(\mathbf{x}_2\right)\right).$$

$$\mathcal{L}_{\text{SimCLR}}^{(i,j)} = -\log \frac{\exp\left(\mathrm{sim}\left(\mathbf{z}_i, \mathbf{z}_j\right)/\tau\right)}{\sum_{k=1}^{2N} \mathbb{I}_{[k\neq i]} \exp\left(\mathrm{sim}\left(\mathbf{z}_i, \mathbf{z}_k\right)/\tau\right)} \tag{3}$$

4. Task-level contrastive learning

We propose using task-level contrastiveness to improve the performance of few-shot/meta learning schemes. Our approach can be used on top of most algorithms in the literature and as we will show in the experimental section, provides significant performance boosts. In a manner similar to traditional contrastive learning, the objective is to bring representations of similar tasks close to each other and push away dissimilar ones.

4.1. Task representations

There are several ways to define a task representation, most modulation based few-shot and meta learning algorithms explicitly define a task encoding network g_ϕ parametrized by ϕ to construct task embeddings, which are then used to adapt the model to the target task. Such an encoder should include 1) high representational capacity, and 2) permutational invariance to its inputs. A commonly used encoder that fits the requirements is based on DeepSets [60] and uses convolutional blocks to produce rich representations, while the average pooling operation in both spatial and batch dimension maintains the permutation invariance. The operation of the encoding given a support set S is:

$$z = g_\phi(S) \tag{4}$$

Another possibility is to use the feature extractor f_θ itself and use the individual sample embeddings to construct task embeddings. In addition we can also consider either the pre or post adaptation features, which would allow for label-aware representations. To obtain a permutation invariant task embedding we can aggregate the individual representations with a permutation invariant operator, such as for example mean, or max/min, etc. In case of using the mean, the task embedding is given as:

$$z = \frac{1}{N} \sum_{(x_i, y_i)\in S} f_\theta(x_i). \tag{5}$$

4.2. Task augmentations

Since our main assumption is the lack of domain labels for each task, using tasks from the same domain as positive pairings to define the contrastive loss is not an option. We can however approach this problem similarly to SimCLR by treating two augmented views of the same task as a positive pair and every other task in the batch as negative. Luckily defining task augmentations for contrastive purposes is easier than in the case of instance-level augmentations. We will consider three simple straightforward approaches, illustrated in Figure 1, different from general task-level augmentations found in non contrastive settings such as in [32]:

Relabeling It is exclusive to few shot classification. The idea is to relabel the samples in the task. Define P as the permutation operator which assigns a new label to all samples in the same class, then a positive task pair, given a task $T = (S, Q)$ and two permutation operators P, P', is defined as

$$T_1 = P(S, Q) = \{(X, P(y))\}_{(X,y)\in S\cup Q} \tag{6}$$
$$T_2 = P'(S, Q) = \{(X, P'(y))\}_{(X,y)\in S\cup Q} \tag{7}$$

Instance augmentation This task augmentation method is a direct extension of sample-level augmentation in SimCLR. Let $t_1, t_2 \sim \mathcal{T}$ be data augmentations, then the task pair, given a task $T = (S, Q)$, is

$$T_1 = (t_1(S), Q) = (\{(t_1(X), y)\}_{(X,y)\in S}, Q) \tag{8}$$
$$T_2 = (t_2(S), Q) = (\{(t_2(X), y)\}_{(X,y)\in S}, Q). \tag{9}$$

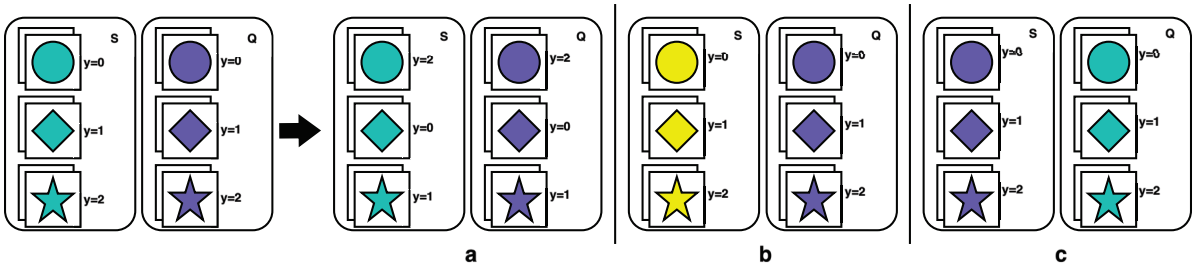


Figure 1. The 3 proposed task augmentation strategies (a) relabeling, (b) instance augmentation , and (c) mixing.

Note that while not necessary, augmentations can be applied on the query samples as well.

Mixing The third task augmentation that we discuss, is mixing of the support and query sets. Intuitively, this augmentation method is the most powerful, as the mutual information between the two task views is smaller compared to the above two augmentation strategies [47]. First we define the number of samples M that we want to substitute from the support set. The most powerful augmentation, meaning the one for which the two task embeddings come from the most distinct support sets, would be the one for which M is set to $M = \min(|S|, |Q|)$. Note that in order to maintain the class balancedness of each set, we substitute M/K samples from each class, where K is the shots of the task. Define the operator $\mathcal{M}_M(\cdot, \cdot)$ that performs this exact mixing and that is drawn from a distribution of such operators, i.e., $\mathcal{M}_M(\cdot, \cdot) \sim \mathbb{M}_M$. For each class in the support set, it substitutes M/K samples from each class with M/K samples from the query set, which comes from the same class. Given $\mathcal{M}_M, \mathcal{M}'_M \sim \mathbb{M}_M$ the task pair, given a task $T = (S, Q)$, is then defined as

$$T_1 = \mathcal{M}_M(T) = \mathcal{M}_M(S, Q) \tag{10}$$

$$T_2 = \mathcal{M}'_M(T) = \mathcal{M}'_M(S, Q). \tag{11}$$

4.3. Enhancing Domain Aware Modulation and Domain Expert Routing Mechanisms

As mentioned in the related work section, we identify two mechanisms utilized by methods operating in this regime, domain-aware modulation and domain expert routing. Both mechanisms can be enhanced within our framework by: **(a)** Contrasting task representations of different domains to encourage domain-aware modulation. **(b)** Contrasting task representations to further refine the implicit clustering of task representations and, consequently, the domain routing mechanism. **(c)** Employing contrastive clustering to enable unsupervised domain routing and eliminate the domain label assumption.

5. Three case studies

For proof of concept, we selected three methods (MMAML, TSA-MAML, Tri-M) due to their simplicity, facilitating a clear demonstration of our task-level contrastive approach. Specifically:**(a)** MMAML: To demonstrate how task-aware modulation can be improved. **(b)** TSA-MAML: To illustrate improvements in unsupervised domain routing. **(c)** Tri-M: To showcase how to enable end-to-end unsupervised domain routing by removing the domain label assumption.

Given that most other methods in this field utilize one or both of these mechanisms, they can be enhanced under our general framework by: **(a)** Identifying which of the aforementioned mechanisms are at play. **(b)** Enhancing the mechanism through the utilization of task-level contrastiveness.

5.1. Enhancing domain aware modulation

5.1.1. MMAML

The first algorithm we consider is MMAML [53], one of the first approaches handling heterogeneous task distributions whose task dependent parameter modulation approach is fundamental and used in a plethora of powerful methods [4, 28]. MMAML and other modulation based methods consist of two complementary neural networks. A modulation network g_ϕ produces a task embedding z , which is used to generate parameters τ_i that modulate the task network. The task network f_θ adapts those modulated parameters (MAML-like gradient based adaptation) to fit to the target task. This extra step of modulation, ensures that the initialization is more domain specific. The MMAML training can be summarized by the following optimization problem:

$$\begin{aligned} \min_{\theta} \quad & \sum_{i=1}^{T} \mathcal{L}(Q_i, \mathcal{A}_\phi(\theta, S_i)) \\ \text{s.t.} \quad & \mathcal{A}_\phi(\theta, S_i) = \hat{\theta}_0 - a \sum_{t=1}^{T} \nabla_\theta \mathcal{L}(Q_i, \hat{\theta}_t) \end{aligned} \tag{12}$$

and the pre adaptation parameters $\hat{\theta}_0$, are obtained by the following task specific modulation

$$\hat{\theta}_0 = \tau \odot \theta, \quad \tau = h_w(g_\phi(S_i)).$$

The modulation parameters τ are produced by a simple network which takes as an input a task embedding $g_\phi(S_i)$. The

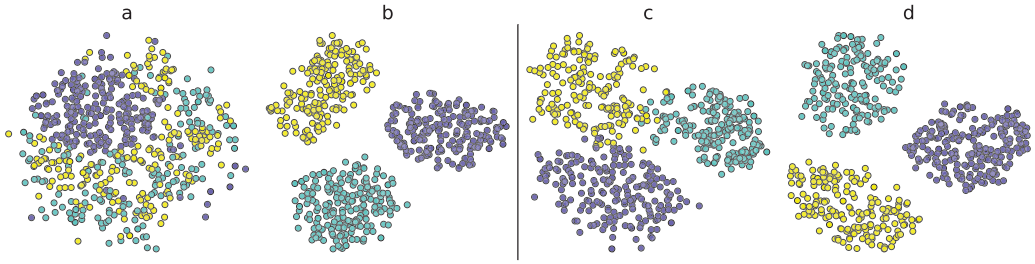


Figure 2. 5 way 1 shot (a,b) and 5 shot (c,d) experiments on fungi, aircraft, birds datasets. We project the adapted parameters for MAML in (a,c) and the task representations for MAML with contrastive loss in (b,d). Unlike task representations, parameters are not a reliable router for TSA-MAML.

modulation operation is performed by the FiLM layer [37]. Minimizing this loss produces reasonable task embeddings, which in turn produce good parameter modulations.

5.1.2. Contrastive MMAML

We build on the top of MMAML and further encourage good task embeddings, which capture the properties of individual domains, by adding a task-level contrastive term and thus changing the MMAML optimization problem into:

$$\min_{\theta,\phi}(\lambda\mathcal{L}_{\text{con}} + \sum_{i=1}^{T}\mathcal{L}(Q_i, \mathcal{A}_\phi(\theta, S_i)). \tag{13}$$

If $B = (S_1, Q_1), \ldots, (S_N, Q_N)$ is the task batch, we apply one of the task augmentations from Section 4.2 on each task to obtain positive pairs $(S'_{2k-1}, Q'_{2k-1}), (S'_{2k}, Q'_{2k})$ for $k = 1, \ldots N$. The augmented batch is then $B' = (S'_1, Q'_1), \ldots, (S'_{2N}, Q'_{2N})$. We can use the MMAML loss on this batch and in addition define the contrastive loss $\mathcal{L}_{\text{con}}$ using the NT-Xent loss 3 $L(i, j)$ on the set of task embeddings $\{z_i\}_{i=1}^{2N} = \{g_\phi(S'_i)\}_{i=1}^{2N}$ as follows:

$$\mathcal{L}_{\text{con}} = \frac{1}{2N}\sum_{k=1}^{2N}(L(2k-1, 2k) + L(2k, 2k-1)). \tag{14}$$

5.2. Improving unsupervised domain routing

5.2.1. TSA-MAML

In TSA-MAML [61], we assume a MAML model pretrained on tasks from all domains. However, instead of relying on this shared global initialization, it obtains domain specific initializations, by first sampling many tasks from all modalities and then clustering their adapted parameters. To fine-tune those domain specific initializations, for each task, a trial adaptation is performed to identify its domain. This domain corresponds to the cluster, whose center lies closest to its trial adapted parameters. One obvious drawback of this method is the need for a trial adaptation, which adds a significant time overhead. Another drawback is the need to store multiple domain specific networks. Furthermore, this method assumes that this parameter-space based routing is reliable. While this is possible in some cases, it also heavily depends on the experiment itself (Fig: 2(a,c)).

5.2.2. Contrastive TSA-MAML

Building on the observation that tasks are not nicely separable in the parameter space, we instead rely on task representations, which under our contrastive framework can form well-separated clusters (see Fig: 2, (b,d)). Instead of using the optimal parameters to perform the routing, we instead use the task representations. We suggest the feature aggregation we introduced in Section 4.1. The main benefit of relying on the features, is that, we can easily encourage contrastive features and in turn task representations by incorporating a contrastive loss. We can therefore improve TSA-MAML by changing the vanilla MAML initalization, to MAML trained with contrastive loss 14. If f_θ is our feature extractor and w are the parameters of the linear classifier layer, then the contrastive MAML loss when using the post or pre-adaptation features $\{z_i\}_{i=1}^{2N} = \{\frac{1}{kN}\sum_{x\in S_i} f_{\mathcal{A}(\theta)}(x)\}_{i=1}^{2N}$ is:

$$\begin{aligned} \min_{w,\theta} \quad & (\lambda\mathcal{L}_{\text{con}} + \sum_{i=1}^{T}\mathcal{L}(Q_i, \mathcal{A}(w, \theta, S_i))) \\ \text{s.t.} \quad & \mathcal{A}(w, \theta, S_i) = (w, \theta) - a\nabla\mathcal{L}(S_i, (w, \theta)) \end{aligned} \tag{15}$$

Our contrastive approach offers another advantage: a reduction in both space and time requirements. Thanks to the reliable expert routing mechanism within our task contrastive framework, we gain greater flexibility in defining the networks for our experts. Instead of each expert constituting an entire network, they can be defined as subsets of parameters. One option we propose is inspired by the findings of Raghu et al [40]. Their work demonstrated that by omitting the inner loop adaptation of the entire feature extractor and only adapting the classification head, we achieve nearly identical results. Therefore, in our case, each expert would consist of a separate classification head. Another possible option, particularly advantageous for large models, involves bias tuning [5, 7]. In this scenario, the experts would be composed of different sets of biases.

Table 1. Task augmentation quality evaluation, MMAML and TSA-MAML are run on 3 datasets, 5 ways, 1 shot

	MetaDataset		MMAML	TSA-MAML
Task Augmentation strategy	**DB**	**Linear Classifier**	**Accuracy**	**Accuracy**
No Augment	1.53	86.6%	56.6%	51.5%
Relabeling	1.79	80.8%	56.5%	53.2%
Mixing	**0.25**	**99.1%**	**57.6%**	**53.6%**
Image Augment	0.62	93.4%	56.8%	53.5%
Mixing+Relabeling	**0.24**	**99.1%**	**57.6%**	**53.7%**
Mixing+Image Augment	0.51	95.8%	56.9%	53.4%
Relabeling+Mixing+Image Augment	0.50	95.8%	56.8%	53.5%

The computational advantages are two. Firstly, our domain-specific experts are smaller subsets of parameters, requiring much less memory to store compared to a full vision model. Secondly, we eliminate the need for trial adaptations to perform routing; a single forward pass is sufficient to identify the task domain.

Table 2. Computational requirements, N is the size of the main model, n the size of the domain expert models ($n < N$) and M the number of domain experts

	Parameters in memory	**Trial Adaptation**
TSA-MAML	N+MN	Yes
our c-TSA-MAML	N+Mn	No

5.3. End-to-end unsupervised domain expert routing

A common assumption in few-shot learning methods that limits their real-world applicability is the requirement for domain labels, indicating the source of each task. This knowledge is crucial for training distinct expert networks, as seen in [27], or for training a domain classifier, as demonstrated in [31, 49]. Notably, in [31], the domain classifier must be trained concurrently with the rest of the architecture, making non-contrastive deep clustering methods impractical [8]. In this section, we introduce task contrastive clustering and its application in Tri-M [31]. In our experimental section, we also present results demonstrating optimal performance in unsupervised task domain classification on MetaDataset. These results serve as evidence of the applicability of our method in FLUTE [49] and URL [27], both of which rely on a domain-router trained with supervision.

5.3.1. Task Contrastive Clustering

Contrastive Clustering [30] first learns the feature matrix (the matrix of instance representations). After that, the instance- and cluster-level contrastive learning are conducted in the row and column space of the feature matrix by gathering the positive pairs and scattering the negatives. By considering the instance- and cluster-level similarity under their dual contrastive learning framework, Contrastive Clustering is able to simultaneously learn discriminative features and perform clustering in an online and end-to-end manner. This is possible by optimizing the loss:

$$\mathcal{L}_{\text{con}} = \mathcal{L}_{\text{con}}^{\text{task}} + \mathcal{L}_{\text{con}}^{\text{clu}} - \mathcal{H}, \tag{16}$$

where the term $\mathcal{L}_{\text{con}}^{\text{task}}$ is the task embedding contrastive loss we introduced before, while $\mathcal{L}_{\text{con}}^{\text{clu}}$ is again the contrastive loss, applied on the columns of the task embedding matrix after being projected on a space with dimensionality equal to the number of domains. This projection acts as a soft label of the domain. To avoid the trivial solution in which most instances are assigned to the same cluster, Contrastive Clustering also includes the loss term $\mathcal{H}$ that is the entropy of cluster assignment probabilities. Using our task-level contrastive loss, leads to task-level contrastive clustering.

5.3.2. Tri-M Algorithm

The tri-M algorithm is based on task specific modulation, the main difference being, the usage of two sets of modulation parameters. The domain-specific set and the domain-cooperative set, which work complementarily to explore both the intra-domain and inter-domain information. The modulation is then performed again by FiLM layers. The authors treat the problem of choosing the appropriate modulation parameters as a domain classification problem, where a task encoder is trained using the domain labels. By substituting the supervised cross entropy loss of the task encoder of Tri-M with our Task Contrastive Clustering loss, we are able to overcome the need for domain labels. This gives rise to our Contrastive Tri-M.

6. Experiments

6.1. Ablation study on task augmentations

We first conduct a simple ablation study to choose the most promising task augmentation for both routing and modulation based methods. In Table 1, we first evaluate the quality of the clustering achieved. We present the Davies-Bouldin index and the linear classification accuracy, where we use task embeddings as features and their domains as labels. We also present the accuracy for contrastive MMAML and TSA-MAML. Notably, mixing outperforms other methods.

Table 3. The performance (classification accuracy) for the MMAML and its contrastive counterpart

		1 Shot				5 Shot			
		MAML	**MMAML**	**contrastive MMAML**	**supervised MMAML**	**MAML**	**MMAML**	**contrastive MMAML**	**supervised MMAML**
3 datasets	**Omniglot**	$86.1_{\pm 1.1}$	$92.6_{\pm 1.2}$	$93.8_{\pm 0.8}$	$93.8_{\pm 0.5}$	$88.9_{\pm 0.9}$	$91.8_{\pm 0.9}$	$93.1_{\pm 0.7}$	$93.1_{\pm 0.6}$
	MiniImage	$42.4_{\pm 1.3}$	$42.0_{\pm 0.9}$	$42.9_{\pm 0.7}$	$43.4_{\pm 0.5}$	$52.7_{\pm 0.6}$	$54.7_{\pm 0.7}$	$56.6_{\pm 0.5}$	$56.8_{\pm 0.4}$
	FC100	$33.6_{\pm 0.7}$	$35.2_{\pm 1.0}$	$36.1_{\pm 0.7}$	$36.3_{\pm 0.4}$	$47.8_{\pm 0.5}$	$44.0_{\pm 0.6}$	$47.2_{\pm 0.4}$	$47.3_{\pm 0.4}$
	Avg	$54.0_{\pm 1.1}$	$56.6_{\pm 1.0}$	$\mathbf{57.6_{\pm 0.7}}$	$57.8_{\pm 0.5}$	$63.1_{\pm 0.7}$	$63.5_{\pm 0.7}$	$\mathbf{65.6_{\pm 0.5}}$	$65.7_{\pm 0.4}$
5 datasets	**Omniglot**	$84.4_{\pm 1.0}$	$90.7_{\pm 1.2}$	$91.0_{\pm 0.6}$	$91.4_{\pm 0.5}$	$89.4_{\pm 0.9}$	$93.5_{\pm 1.0}$	$91.3_{\pm 0.8}$	$91.2_{\pm 0.8}$
	MiniImage	$39.6_{\pm 1.3}$	$40.0_{\pm 1.1}$	$42.4_{\pm 0.7}$	$42.7_{\pm 0.6}$	$51.3_{\pm 0.7}$	$53.3_{\pm 0.8}$	$55.6_{\pm 0.6}$	$56.0_{\pm 0.5}$
	FC100	$34.1_{\pm 0.9}$	$34.9_{\pm 0.9}$	$34.7_{\pm 1.0}$	$34.7_{\pm 0.8}$	$43.4_{\pm 0.7}$	$43.0_{\pm 0.7}$	$48.2_{\pm 0.5}$	$48.2_{\pm 0.5}$
	Birds	$37.9_{\pm 0.9}$	$43.0_{\pm 1.0}$	$43.8_{\pm 0.8}$	$43.7_{\pm 0.6}$	$47.9_{\pm 0.5}$	$53.2_{\pm 0.6}$	$60.0_{\pm 0.4}$	$60.3_{\pm 0.4}$
	Aircraft	$44.6_{\pm 1.3}$	$43.1_{\pm 1.2}$	$44.5_{\pm 0.7}$	$45.0_{\pm 0.5}$	$56.6_{\pm 1.0}$	$54.1_{\pm 1.0}$	$53.8_{\pm 0.8}$	$54.0_{\pm 0.6}$
	Avg	$48.1_{\pm 1.1}$	$50.3_{\pm 1.1}$	$\mathbf{51.3_{\pm 0.8}}$	$51.5_{\pm 0.6}$	$57.7_{\pm 0.8}$	$59.4_{\pm 0.8}$	$\mathbf{61.8_{\pm 0.6}}$	$61.9_{\pm 0.5}$

Table 4. The performance (classification accuracy) for the TSA-MAML and its contrastive counterpart

		MAML	**ANIL**	**TSA-MAML**	**TSA-ANIL**	**contrastive TSA-MAML**	**contrastive TSA-ANIL**	**supervised TSA-MAML**
3 datasets	**Fungi**	$40.9_{\pm 1.0}$	$40.8_{\pm 1.1}$	$41.1_{\pm 1.3}$	$41.0_{\pm 1.2}$	$42.7_{\pm 0.9}$	$42.4_{\pm 0.8}$	$42.6_{\pm 0.8}$
	Birds	$52.9_{\pm 0.9}$	$52.7_{\pm 1.0}$	$53.2_{\pm 1.2}$	$52.8_{\pm 1.1}$	$54.4_{\pm 0.8}$	$53.6_{\pm 0.8}$	$54.6_{\pm 0.7}$
	Aircraft	$60.8_{\pm 1.2}$	$60.8_{\pm 1.3}$	$61.0_{\pm 1.3}$	$61.1_{\pm 1.3}$	$63.5_{\pm 0.8}$	$62.7_{\pm 0.9}$	$63.8_{\pm 0.8}$
	Avg	$51.5_{\pm 1.0}$	$51.4_{\pm 1.1}$	$51.8_{\pm 1.3}$	$51.6_{\pm 1.2}$	$\mathbf{53.6_{\pm 0.8}}$	$52.9_{\pm 0.8}$	$53.7_{\pm 0.8}$
5 datasets	**Omniglot**	$93.2_{\pm 1.0}$	$94.0_{\pm 1.3}$	$93.8_{\pm 1.1}$	$93.8_{\pm 1.3}$	$94.5_{\pm 0.9}$	$94.2_{\pm 1.0}$	$94.4_{\pm 0.8}$
	MiniImage	$39.9_{\pm 1.1}$	$39.1_{\pm 1.2}$	$40.0_{\pm 1.2}$	$39.7_{\pm 1.2}$	$44.0_{\pm 0.8}$	$42.8_{\pm 0.9}$	$44.3_{\pm 0.7}$
	Fungi	$40.7_{\pm 0.8}$	$40.9_{\pm 1.0}$	$40.6_{\pm 0.9}$	$40.9_{\pm 1.1}$	$42.1_{\pm 0.7}$	$41.7_{\pm 0.9}$	$42.1_{\pm 0.8}$
	Birds	$53.1_{\pm 0.9}$	$52.5_{\pm 1.0}$	$53.4_{\pm 1.0}$	$52.8_{\pm 1.0}$	$55.1_{\pm 0.8}$	$53.9_{\pm 0.8}$	$54.9_{\pm 0.8}$
	Aircraft	$60.5_{\pm 1.1}$	$59.9_{\pm 1.1}$	$60.7_{\pm 1.2}$	$60.5_{\pm 1.1}$	$61.9_{\pm 0.9}$	$61.9_{\pm 1.0}$	$62.5_{\pm 0.8}$
	Avg	$57.5_{\pm 1.0}$	$57.3_{\pm 1.1}$	$57.7_{\pm 1.1}$	$57.5_{\pm 1.1}$	$\mathbf{59.5_{\pm 0.8}}$	$58.9_{\pm 0.9}$	$59.6_{\pm 0.8}$

Relabeling, while beneficial for TSA-MAML with post-adaptation embeddings, performs poorly with MMAML (and Tri-M) due to collapsed representations. The collapse occurs because the label is only included in the input of the task encoder, which does not ensure sufficiently distinct embeddings for a positive pair of tasks. Instance augmentation demonstrates strong performance. However, we hypothesize that the composition of augmentations in SimCLR may not be optimal for our own datasets, resulting in a suboptimal feature extractor. Lastly, achieving optimal performance on the unsupervised task domain classification in MetaDataset serves as direct evidence of our method's applicability in FLUTE [49] and URL [27], both of which rely on a domain-router trained in a supervised manner.

6.2. Experimental results

For all three methods and their contrastive counterparts we use the exact same experimental setting as in the respective papers. We refer to the original papers and our Supplement for the detailed experimental settings. We also refer to our Supplement for more experimental results, in particular 5-shot experiments for the TSA-MAML family of methods. In the case of MMAML, we see in Table 3 that adding contrastiveness results in better performance. This is because it improves the ability to distinguish between different domains and in turn results in better modulation parameters. TSA-MAML, Table 4, also benefits from contrastiveness due to proper domain assignment, while, contrastive TSA-ANIL, which is our proposed method, where experts are different classification heads instead of full networks, exhibits favorable performance compared to TSA-MAML at a fraction of the computational requirements of TSA-MAML.

Both non-contrastive MMAML and TSA-MAML rely on implicit clustering of task representations and parameters, which cannot be reliably achieved without an explicit loss. Consequently, non-contrastive experiments exhibit higher standard deviation in the final accuracy. Proper discrimination between domains is further demonstrated in Figure 2 (b, d). It can be observed that relying on clustering in the parameter space may cause all centers to collapse to a single point, as all tasks can potentially update all domain-specific centers. When such implicit clustering occurs (Figure 2 (c)), the benefits of the added contrastiveness are less significant. We also compare with supervised versions of MMAML and TSA-MAML, where we assume access to each task's source. For MMAML we include an extra cross-entropy loss term for the task encoder in conjunction with an additional classification head (similar to tri-M). For supervised TSA-MAML we train the router with supervised contrastive learning [22]. As seen in Tables 3,4, the results with domain labels are better than their unsupervised counterparts, and only slightly better than our con-

Table 5. Computational benefits of (a) TSA-MAML variants on 3 datasets 5 ways 1 shot and (b) URL on Metadataset

Frozen Layers	Avg Acc	#Params	Trial Adaptation
0 (TSA-MAML)	51.7	484372/100%	Yes
0 (c-TSA-MAML)	53.6	484372/100%	No
3 (c-TSA-ANIL)	52.9	256276/52.9%	No

	Avg Acc	#Nets
Single task networks	76.4	8
Random grouping	76.3	3
Task similarity grouping	76.6	3

Table 6. The performance (accuracy) for the Tri-M and its contrastive counterpart. We see that our contrastive approach perfroms just as well, **without the need for domain labels.**

	Tri-M	con-Tri-M
ImageNet	54.5 ± 1.1	54.3 ± 1.1
Omniglot	92.2 ± 0.6	92.0 ± 0.5
Aircraft	81.8 ± 0.6	82.8 ± 0.7
Birds	75.9 ± 0.9	75.6 ± 0.9
Textures	69.5 ± 0.8	70.2 ± 0.7
QuickDraw	76.9 ± 0.8	77.3 ± 0.8
Fungi	48.1 ± 1.2	47.1 ± 1.2
VGGFlower	91.5 ± 0.5	90.5 ± 0.6
TrafficSigns	57.5 ± 1.1	58.5 ± 1.1
MSCOCO	49.7 ± 1.1	49.8 ± 1.1
MNIST	94.3 ± 0.4	95.0 ± 0.4
CIFAR10	74.0 ± 0.8	73.1 ± 0.8
CIFAR100	62.3 ± 1.0	62.1 ± 1.0
In domain Average	73.8	73.7
Out of domain Average	67.6	67.7
Overall Average	71.4	71.4

trastive method. This indicates that our approach gets close to what can be perceived as optimal in these experimental settings.

Finally, for Tri-M, we see in Table 6 that we get no deterioration in performance even after eliminating the need for task domain labels. In Figure 3 we plot the cluster/domain heatmap and we see that even in our unsupervised setting, there are distinct cluster-domain pairs being formed due to our contrastive framework.

6.3. Computational Benefits

Our task-level contrastiveness can also reduce the memory and time requirements of algorithms. Here, we highlight these benefits for two cases: the aforementioned parameter-efficient TSA-MAML variants and Hierarchical Distillation in URL [29]. URL involves aligning the representations of the universal network with multiple single-task networks, a process that can become memory and compute-intensive as the number of tasks (T) increases. By dividing all tasks into different task groups, it is possible to learn a single network for each group g by distilling the single-task representations of the tasks within that group. The universal network can then be obtained by distilling these group networks. The construction of these groups can be done randomly or by utilizing contrastive task representations (see Table 1) and performing K-means to obtain centers, which will be used to classify tasks into each group. Groups therefore are created by taking into account a notion of task similarity. This leads to lower memory requirements and slightly better final average accuracy(see Table 5).

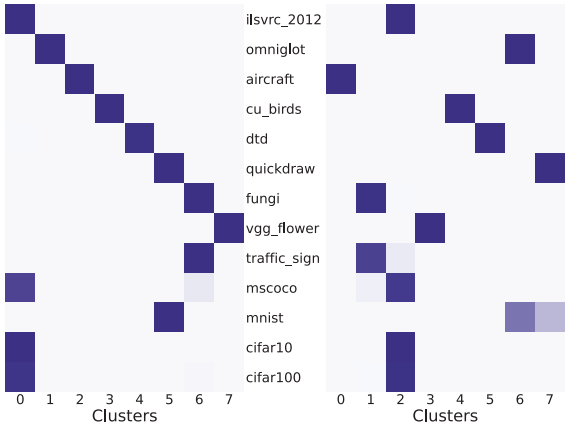


Figure 3. Cluster/domain heatmap. We plot the cluster assignment of tasks for Tri-M (left) and our con-Tri-M (right). Even in the absence of domain labels (right) we are able to get assignements that closely resemble the ones obtained in a supervised manner (left)

7. Conclusion

We introduced task-level contrastiveness to enhance meta-learning and few-shot learning performance in heterogeneous task settings. Our tool seamlessly integrates with most existing methods, demonstrating improvements in generalization, reduced computational and memory requirements, and the elimination of certain modeling assumptions. By selecting key algorithms from the literature, we showed how our approach effectively enhances and enables domain-aware modulation and unsupervised domain-expert routing—the two mechanisms relied upon by state of the art methods. In future work, we plan to explore various settings, such as few-shot dense prediction [23], which represents another interesting application of few-shot learning. Additionally, multi-task settings or mixture of experts approaches could potentially benefit from our task-level contrastiveness. One major limitation of our approach is our choice of task augmentations. While mixing is a powerful technique, several studies have noted that a support/query split, which is necessary for our method, is not always required for meta-learning algorithms [1]. Exploring alternatives to our current augmentation strategies could further enhance the applicability and effectiveness of our approach.

References

[1] Yu Bai, Minshuo Chen, Pan Zhou, Tuo Zhao, Jason Lee, Sham Kakade, Huan Wang, and Caiming Xiong. How important is the train-validation split in meta-learning? In *International Conference on Machine Learning*, pages 543–553. PMLR, 2021. 8

[2] Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, Avi Schwarzschild, Andrew Gordon Wilson, Jonas Geiping, Quentin Garrido, Pierre Fernandez, Amir Bar, Hamed Pirsiavash, Yann LeCun, and Micah Goldblum. A cookbook of self-supervised learning, 2023. 1

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

[4] Peyman Bateni, Jarred Barber, Jan-Willem Van de Meent, and Frank Wood. Enhancing few-shot image classification with unlabelled examples. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2796–2805, 2022. 2, 4

[5] Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9, Dublin, Ireland, 2022. Association for Computational Linguistics. 5

[6] L Bertinetto, J Henriques, P Torr, and A Vedaldi. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations (ICLR), 2019*. International Conference on Learning Representations, 2019. 2

[7] Han Cai, Chuang Gan, Ligeng Zhu Massachusetts Institute of Technology, and Song Han Massachusetts Institute of Technology. Tinytl: reduce memory, not parameters for efficient on-device learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2020. Curran Associates Inc. 5

[8] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 132–149, 2018. 6

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2, 3

[10] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. In *Advances in Neural Information Processing Systems*, pages 8765–8775. Curran Associates, Inc., 2020. 2

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 2

[12] Nikita Dvornik, Cordelia Schmid, and Julien Mairal. Selecting relevant features from a multi-domain representation for few-shot classification. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X*, page 769–786, Berlin, Heidelberg, 2020. Springer-Verlag. 2

[13] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017. 1, 2, 3

[14] Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and SM Ali Eslami. Conditional neural processes. In *International conference on machine learning*, pages 1704–1713. PMLR, 2018. 2

[15] Muhammad Waleed Gondal, Shruti Joshi, Nasim Rahaman, Stefan Bauer, Manuel Wuthrich, and Bernhard Schölkopf. Function contrastive learning of transferable meta-representations. In *International Conference on Machine Learning*, pages 3755–3765. PMLR, 2021. 2

[16] Jonathan Gordon, Wessel P Bruinsma, Andrew YK Foong, James Requeima, Yann Dubois, and Richard E Turner. Convolutional conditional neural processes. In *International Conference on Learning Representations*, 2019. 2

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2

[18] 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 *International Conference on Learning Representations*, 2018. 2

[19] Shell Xu Hu, Da Li, Jan Stühmer, Minyoung Kim, and Timothy M Hospedales. Pushing the limits of simple pipelines for few-shot learning: External data and fine-tuning make a difference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9068–9077, 2022. 2

[20] Huiwon Jang, Hankook Lee, and Jinwoo Shin. Unsupervised meta-learning via few-shot pseudo-supervised contrastive learning. *arXiv preprint arXiv:2303.00996*, 2023. 2

[21] Weisen Jiang, James Kwok, and Yu Zhang. Subspace learning for effective meta-learning. In *Proceedings of the 39th International Conference on Machine Learning*, pages 10177–10194. PMLR, 2022. 2

[22] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. 7

[23] Donggyun Kim, Jinwoo Kim, Seongwoong Cho, Chong Luo, and Seunghoon Hong. Universal few-shot learning of

dense prediction tasks with visual token matching. *arXiv preprint arXiv:2303.14969*, 2023. 8

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2012. 1

[25] Jae-Jun Lee and Sung Whan Yoon. XB-MAML: Learning expandable basis parameters for effective meta-learning with wide task coverage. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, pages 3196–3204. PMLR, 2024. 2

[26] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10657–10665, 2019. 2

[27] Wei-Hong Li, Xialei Liu, and Hakan Bilen. Universal representation learning from multiple domains for few-shot classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9526–9535, 2021. 2, 6, 7

[28] Wei-Hong Li, Xialei Liu, and Hakan Bilen. Cross-domain few-shot learning with task-specific adapters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7161–7170, 2022. 2, 4

[29] Wei-Hong Li, Xialei Liu, and Hakan Bilen. Universal representations: A unified look at multiple task and domain learning. *International Journal of Computer Vision*, pages 1–25, 2023. This work was partly supported by the EPSRC programme grant Visual AI EP/T028572/1 and a Huawei Technologies RD (UK) project. 8

[30] Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. Contrastive clustering. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8547–8555, 2021. 6, 4

[31] Yanbin Liu, Juho Lee, Linchao Zhu, Ling Chen, Humphrey Shi, and Yi Yang. A multi-mode modulator for multi-domain few-shot classification. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8433–8442, 2021. 2, 6

[32] Renkun Ni, Micah Goldblum, Amr Sharaf, Kezhi Kong, and Tom Goldstein. Data augmentation for meta-learning. In *International Conference on Machine Learning*, pages 8152–8161. PMLR, 2021. 3

[33] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 2

[34] Boris N. Oreshkin, Pau Rodriguez, and Alexandre Lacoste. Tadam: task dependent adaptive metric for improved few-shot learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, page 719–729, Red Hook, NY, USA, 2018. Curran Associates Inc. 1

[35] Danni Peng and Sinno Jialin Pan. Clustered task-aware meta-learning by learning from learning paths. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9426–9438, 2023. 2

[36] Rashindrie Perera and Saman Halgamuge. Discriminative sample-guided and parameter-efficient feature space adaptation for cross-domain few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23794–23804, 2024. 2

[37] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 5

[38] Stylianos Poulakakis-Daktylidis and Hadi Jamali-Rad. BECLR: Batch enhanced contrastive unsupervised few-shot learning. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[40] Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of maml. In *International Conference on Learning Representations*, 2019. 2, 5

[41] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *In International Conference on Learning Representations (ICLR)*, 2017. 1

[42] James Requeima, Jonathan Gordon, John Bronskill, Sebastian Nowozin, and Richard E Turner. Fast and flexible multi-task classification using conditional neural adaptive processes. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[43] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *International Conference on Learning Representations (ICLR)*, 2021. 2

[44] Amitis Shidani, Devon Hjelm, Jason Ramapuram, Russ Webb, Eeshan Gunesh Dhekane, and Dan Busbridge. Polyview contrastive learning. *arXiv preprint arXiv:2403.05490*, 2024. 2

[45] Abhishek Sinha, Kumar Ayush, Jiaming Song, Burak Uzkent, Hongxia Jin, and Stefano Ermon. Negative data augmentation. In *International Conference on Learning Representations*, 2021. 2

[46] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 4080–4090, Red Hook, NY, USA, 2017. Curran Associates Inc. 1

[47] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning?, 2020. 2, 4

[48] Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Jordan Swersky, Pierre-Antoine Manzagol, and Hugo Larochelle. Meta-dataset: A dataset of datasets for learning to learn from few examples. In *International Conference on Learning Representations (submission)*, 2020. 1

[49] Eleni Triantafillou, Hugo Larochelle, Richard Zemel, and Vincent Dumoulin. Learning a universal template for few-shot dataset generalization. In *International Conference on Machine Learning*, pages 10424–10433. PMLR, 2021. 2, 6, 7

[50] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. 2

[51] Anna Vettoruzzo, Mohamed-Rafik Bouguelia, and Thorsteinn Rögnvaldsson. Meta-learning from multimodal task distributions using multiple sets of meta-parameters. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2023. 2

[52] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning, 2017. 1

[53] Risto Vuorio, Shao-Hua Sun, Hexiang Hu, and Joseph J Lim. Multimodal model-agnostic meta-learning via task-aware modulation. *Advances in neural information processing systems*, 32, 2019. 2, 4

[54] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 2

[55] C. Xu. Exploring efficient few-shot adaptation for vision transformers. *Trans. Mach. Learn. Res.*, 2022, 2023. 2

[56] Zhanyuan Yang, Jinghua Wang, and Yingying Zhu. Few-shot classification with contrastive learning. In *European conference on computer vision*, pages 293–309. Springer, 2022. 2

[57] Huaxiu Yao, Ying Wei, Junzhou Huang, and Zhenhui Li. Hierarchically structured meta-learning. In *International Conference on Machine Learning*, pages 7045–7054. PMLR, 2019. 2

[58] Huaxiu Yao, Xian Wu, Zhiqiang Tao, Yaliang Li, Bolin Ding, Ruirui Li, and Zhenhui Li. Automated relational meta-learning. In *International Conference on Learning Representations*, 2019. 2

[59] Chun-Hsiao Yeh, Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu, Yubei Chen, and Yann LeCun. Decoupled contrastive learning. In *European Conference on Computer Vision*, pages 668–684. Springer, 2022. 2

[60] Manzil Zaheer, Satwik Kottur, Siamak Ravanbhakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J Smola. Deep sets. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 3394–3404, Red Hook, NY, USA, 2017. Curran Associates Inc. 3

[61] Pan Zhou, Yingtian Zou, Xiaotong Yuan, Jiashi Feng, Caiming Xiong, and SC Hoi. Task similarity aware meta learning: Theory-inspired improvement on maml. In *4th Workshop on Meta-Learning at NeurIPS*, 2020. 2, 5