A Complete Survey on Contemporary Methods, Emerging Paradigms and Hybrid Approaches for Few-Shot Learning

Georgios Tsoumplekas[1*], Vladislav Li[2],
Panagiotis Sarigiannidis[1,3], Vasileios Argyriou[2]

[1]MetaMind Innovations P.C., Kozani, 50100, Greece.
[2]Department of Networks and Digital Media, Kingston University, Kingston upon Thames, KT1 2EE Surrey, UK.
[3]Department of Electrical and Computer Engineering, University of Western Macedonia, Kozani, 50100, Greece.

*Corresponding author(s). E-mail(s): gtsoumplekas@metamind.gr;
Contributing authors: v.li@kingston.ac.uk; psarigiannidis@uowm.gr;
vasileios.argyriou@kingston.ac.uk;

Abstract

Despite the widespread success of deep learning, its intense requirements for vast amounts of data and extensive training make it impractical for various real-world applications where data is scarce. In recent years, Few-Shot Learning (FSL) has emerged as a learning paradigm that aims to address these limitations by leveraging prior knowledge to enable rapid adaptation to novel learning tasks. Due to its properties that highly complement deep learning's data-intensive needs, FSL has seen significant growth in the past few years. This survey provides a comprehensive overview of both well-established methods as well as recent advancements in the FSL field. The presented taxonomy extends previously proposed ones by incorporating emerging FSL paradigms, such as in-context learning, along with novel categories within the meta-learning paradigm for FSL, including neural processes and probabilistic meta-learning. Furthermore, a holistic overview of FSL is provided by discussing hybrid FSL approaches that extend FSL beyond the typically examined supervised learning setting. The survey also explores FSL's diverse applications across various domains. Finally, recent trends shaping the field, outstanding challenges, and promising future research directions are discussed.

1 Introduction

In the past decade, there has been significant progress in machine/deep learning fields, mainly due to the abundance of computational resources, the availability of massive datasets, and the development of sophisticated algorithms, like the Transformer architecture (Vaswani et al, 2017). The efficacy of machine and deep learning models has been materialized in diverse fields, including computer vision (He et al, 2016), reinforcement learning (Silver et al, 2017), natural language processing (Brown et al, 2020), and biomedicine (Jumper et al, 2021), while also surpassing human performance in some cases (Silver et al, 2017). A vital aspect of these models, considered the 'golden rule' of modern deep learning, is their training from scratch on a specific task using a large dataset.

Although modern machine learning models have shown significant success, they still face clear limitations. Due to their increasing size, training these models is a very data- and compute-hungry procedure, while at the same time, these models struggle in continual or incremental learning scenarios without forgetting previously acquired knowledge. As a result, real-world application of these models can be challenging, especially in cases where data is scarce, expensive, or time-consuming to obtain or where there is a lack of necessary computational resources.

On the other hand, one of the most critical aspects of human intelligence is the ability to learn quickly from limited examples and adapt to unfamiliar settings. For instance, young children can identify new object categories given only a few examples. Generally, when learning a new task, an individual uses prior knowledge and experiences and adapts them to the new task based on the provided contextual framework. Consequently, they can adapt more quickly to the new task by transferring already acquired knowledge rather than relearning it from scratch, as machine/deep learning models do.

Parallel to the development of large-scale models, there is a growing interest in integrating such aspects of human-like learning processes within the existing machine learning paradigm. As a result, to bridge the gap between machine and human learning, Few-Shot Learning (FSL) has emerged as a promising direction. More precisely, FSL aims to develop models that can generalize within a specific task, even with a limited number of training samples available. This is typically achieved by using prior knowledge extracted from similar tasks. Notably, such models can adapt to novel tasks with small datasets, which is difficult for conventional machine learning models without the risk of overfitting. Additionally, they can significantly reduce the need for collecting and preprocessing extensive amounts of data for training while also minimizing the time required for adapting to new data. On a more practical level, FSL can also enable learning of rare cases, which is crucial in applications such as drug discovery (Altae-Tran et al, 2017).

Table 1 Comparison of existing FSL surveys and our survey. Symbol ✓means the topic is covered while ✗ means the topic is not covered.

Survey	In-Context Learning	Novel Meta-Learning Model Families	Non Meta-Learning FSL Methods	Hybrid Methods Beyond Supervised FSL
Wang et al (2020b)	✗	✗	✓	✗
Parnami and Lee (2022)	✗	✗	✓	✓
Song et al (2023a)	✗	✗	✓	✗
Gharoun et al (2024)	✗	✓	✗	✗
Ours	✓	✓	✓	✓

Leveraging prior knowledge to adapt to novel tasks has been explored since the 1980s via self-referential learning (Schmidhuber, 1987) and learning biologically plausible rules that enable adaptation to new tasks (Bengio et al, 1991). Other early approaches have also included object recognition from a single image using Bayesian models (Fei-Fei et al, 2006), learning data augmentations (Miller et al, 2000), and metric learning (Fink, 2004). However, the proliferation of FSL and its influence on various aspects of modern deep learning over the past decade have necessitated consolidating knowledge on recent advancements in the field.

One of the first surveys on FSL was conducted by Wang et al (2020b), who provided the first formal definition and taxonomy of existing methods and applications in the field. This taxonomy was later extended in Parnami and Lee (2022) to include various hybrid methods beyond the standard supervised setting. More recently, in Song et al (2023a), a novel taxonomy has been proposed based on how prior information is abstracted in FSL and the different challenges arising in each case. Meanwhile, other surveys such as in Gharoun et al (2024) have narrowed the focus to meta-learning approaches and their application in the field of FSL. Additionally, several recent surveys have focused on specific application areas of FSL, such as FSL in medical imaging (Pachetti and Colantonio, 2024), FSL in channel state information-based human sensing (Wang et al, 2024), and few-shot object detection (Köhler et al, 2023), among others.

However, there has not yet been a comprehensive overview of FSL that covers the full range of modern FSL methods and applications. Specifically, certain existing surveys have primarily focused on meta-learning approaches for FSL (Gharoun et al, 2024) while most recent attempts to provide a broader perspective on the field have not captured emerging FSL paradigms such as the trending in-context learning (Song et al, 2023a; Gharoun et al, 2024). Meanwhile, discussions regarding the extensions of FSL beyond the supervised learning setting using hybrid methods have been limited in the past (Parnami and Lee, 2022). Finally, several surveys (Pachetti and Colantonio, 2024; Wang et al, 2024; Köhler et al, 2023) have focused exclusively on specific application fields of FSL.

This paper aims to address the current lack of a broader overview in the field of FSL by providing a comprehensive and up-to-date survey. In particular, it includes an exploration of novel model families under existing FSL paradigms, such as neural processes and probabilistic meta-learning, incorporates novel FSL paradigms, such as in-context learning, and offers a more detailed review of hybrid FSL methods that extend FSL beyond supervised learning settings. Table 1 illustrates how our survey integrates these various elements extending upon previous surveys in the field. Finally, emerging challenges and future directions for FSL in the era of large-scale deep learning models such as human-like learning and Green AI are discussed. Overall, our contributions can be summarized as follows:

- We propose a novel taxonomy that builds upon previously proposed ones and extends classic approaches, such as meta-learning, transfer learning, and data augmentation, by introducing new subcategories, including neural processes and probabilistic extensions of meta-learning. We primarily focus on recent advancements while also including well-established methods.
- We present, for the first time in an FSL-related survey, an analysis of the recently explored in-context learning within the FSL framework, examining its similarities and differences with more traditional FSL approaches as well as various techniques that combine these two paradigms.
- We explore FSL's connections with relevant learning problems and review various existing hybrid approaches that extend the standard supervised FSL framework to different settings, including unsupervised FSL, few-shot federated learning, cross-domain FSL, and class-incremental FSL, among others.
- We highlight the latest trends in FSL and identify some arising challenges that need to be tackled to advance the field. These can bring additional attention to the field of FSL, sparking interest in further advancements.

The remainder of this survey is organized as follows. Section 2 provides an overview of FSL, including its formal mathematical formulation, commonly used datasets, and related fields of interest. Section 3 introduces the proposed taxonomy and provides an overview of various recent and standard approaches within each category. Section 4 extends the presented taxonomy beyond the supervised FSL setting by introducing hybrid methods at the intersection of FSL and different learning paradigms. Section 5 analyzes FSL's fields of application. Section 6 analyzes multiple trends that have emerged in recent years and discusses less explored directions within the field that are promising for future research. Finally, Section 7 concludes the survey.

2 Background

Before delving deeper into the various methods used within the FSL field, it is essential to provide a rigorous mathematical and practical formulation of the problem these methods try to solve. Next, an overview of some standard benchmarks, as well as some emerging ones in novel settings, is provided. Finally, the relationship between FSL and different learning paradigms is examined.

2.1 Problem Formulation

As previously mentioned, FSL refers to a set of learning problems in which machine learning models aim to adapt and be able to generalize given only a limited number of training examples. Wang et al (2020b) provide a more formal definition framing FSL using the same terminology as the one used for defining standard machine learning. In general, given a task T and a performance measure P, a model is said to learn when its performance P on the task T is improved by leveraging experience E related to the task (e.g., training data).

More specifically, the experience E associated with T is a dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$ of N samples, where $x_i \in \mathbb{R}^I$, $y_i \in \mathbb{R}^O$ are the input feature vector and corresponding label, respectively. In general, $(x_i, y_i) \sim p(x, y)$, meaning that there is an underlying joint distribution p that generates the data samples within this task. The aim is to create a model $f_\theta : \mathbb{R}^I \to \mathbb{R}^O$ parameterized by θ, which is optimized according to a performance measure which typically is a loss function $\mathcal{L} : \mathbb{R}^O \times \mathbb{R}^O \to \mathbb{R}$. This leads to an optimal set of model parameters θ^* for which:

$$\theta^* = \operatorname*{argmin}_{\theta} \mathbb{E}_{(x,y) \sim p(x,y)}[\mathcal{L}(f_\theta(x), y)] \tag{1}$$

Since $p(x, y)$ is unknown, θ is optimized through *empirical risk minimization* by approximating the expected loss over the joint distribution by the average loss on the training set samples:

$$\theta^* = \operatorname*{argmin}_{\theta} \left[\frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(f_\theta(x_i), y_i) \right] \tag{2}$$

Based on this definition, FSL is substantially different from typical supervised machine learning in the sense that there is less training data on which the model can be trained and gain experience for the task. Under these learning conditions, a typical learning model would severely overfit the training dataset and thus could not generalize to unseen examples. As a result, standard FSL approaches rely on extracting knowledge from a set of similar tasks, each consisting of only a few training examples.

Specifically, instead of a single learning task T, we now consider a group of tasks $\{T_j\}_{j=1}^{M}$, which are realizations of a task distribution $q(T)$. The derivation of the tasks from the same distribution implies the existence of a common shared structure that can be extracted and leveraged to adapt more quickly to these tasks. As a result, in this case, the aim is to find an optimal set of parameters θ^*, which minimizes the loss across all given tasks. These optimal parameters would satisfy the following:

$$\theta^* = \operatorname*{argmin}_{\theta} \mathbb{E}_{T \sim q(T)}[\mathbb{E}_{(x,y) \sim p(x,y)} \mathcal{L}(f_\theta(x), y)] \tag{3}$$

The existence of the two expectations, one over the task distribution and one over the data generating process within each task, in (3), imposes an additional level of intractability. As a result, similarly to how the expected error was handled in (2), an approximation of the expected error over the task distribution is necessary. Each of the tasks $\{T_j\}_{j=1}^{M}$ corresponds to a dataset $\{D_j\}_{j=1}^{M}$ and therefore, the expected loss over all tasks can be approximated by the average loss on the available tasks used

during training. For clarity, we adhere to the nomenclature commonly used in meta-learning and FSL and refer to this set of datasets as the *meta-train* set. Additionally, a similar set of datasets from the same task distribution, named *meta-test* set, can be used to evaluate the model on novel tasks.

Accordingly, within each task in the *meta-train* set, a set of data points available during training is used for the error approximation within the task. This set is referred to as the *support set*, while a second set of the task's data points is used for evaluation within the task, named the *query set*. Precisely, each dataset $D_j \in \{D_j\}_{j=1}^M$ is split into $D_j = (D_j^S, D_j^Q)$ where D_j^S, D_j^Q are two disjoint sets that correspond to the *support* and *query* sets of the task. Combining all of the above, (3) can be approximated by:

$$\theta^* = \underset{\theta}{\mathrm{argmin}} \left[\frac{1}{M} \sum_{j=1}^M \left[\frac{1}{N} \sum_{(x,y) \in D_j^Q} \mathcal{L}(f_{\theta|D_j^S}(x), y) \right] \right] \tag{4}$$

It is worth noting here that θ is optimized over the *query* set samples of each task. Typically, the model parameters are refined within each task based on the provided *support* set, and the evaluation on the *query* set is performed using these refined parameters. This parameter refinement is visible in (4) on the dependence of θ from the *support* set D_j^S of the examined task. Since the aim of FSL models is to be capable of generalizing within each given task, it makes sense that the loss calculation is performed on samples different than the ones used for refining the model.

2.2 N-way K-shot Formulation

Historically, research around FSL has mainly focused on image classification. In this particular setting, one typically refers to the *N-way-K-shot* classification (Vinyals et al, 2016), in which each support set contains samples from N different classes, with K different examples in each class. As a result, each support set consists of $|S| = NK$ training samples. Similar formulations can also be implemented for different learning problems. For instance, in few-shot semantic segmentation, each task might consist of various images of the same object to be segmented, and in few-shot regression, each task corresponds to a different output function with only a few input-output pairs provided in the support set.

2.3 Datasets

Various datasets suitable for FSL training and evaluation have been proposed throughout the years, with *miniImageNet* (Vinyals et al, 2016) and *Omniglot* (Lake et al, 2015) being the most widely used for few-shot image classification. In particular, *miniImageNet* consists of 60000 images separated into 100 different classes originating from *ImageNet*, and *Omniglot* contains images of 1623 handwritten characters, with 20 different samples included for each character. Other commonly used datasets include *tieredImageNet* (Ren et al, 2018), which alleviates *miniImageNet*'s lack of class diversity, having 608 classes grouped in 34 class categories, and *CIFAR-FS* (Bertinetto et al, 2019). Finally, datasets such as *CUB-200-2011*, *tinyImageNet*, *Aircraft* and *Stanford*

Cars, although not initially developed for FSL purposes, have also been used due to their small size and variety of classes.

However, these datasets lack variability in their number of classes and samples per class, and they are also not suitable for evaluating out-of-domain generalization. To combat these issues, datasets such as *Meta-Dataset* (Triantafillou et al, 2020) and *ORBIT* (Massiceti et al, 2021) have emerged that combine images from different datasets. Extending this framework, there has also been a bloom of other datasets in recent years, aiming to include images from various domains, such as *BSCD-FSL* (Guo et al, 2020), which includes agriculture, satellite, and dermatology. Other recent approaches include *Meta-Album* (Ullah et al, 2022) focusing on a more realistic cross-domain evaluation of FSL models, spanning across ten distinct domains and *HARD-MD++* (Basu et al, 2022) which is the amalgamation of various tasks considered hard to solve from *Meta-Dataset* and other FSL benchmarks, spanning across four domains. Finally, *Meta Omnium* (Bohdal et al, 2023) extends the FSL framework outside of image classification by providing a benchmark for evaluation of generalization across different learning problems such as key-point localization, semantic segmentation, and regression.

Apart from image classification, various other datasets related to specific problems such as semantic segmentation (Li et al, 2020), regression (Gao et al, 2022), pose estimation (Xu et al, 2022b), drug discovery (Stanley et al, 2021) and robotics (Yu et al, 2020) have been proposed, however they usually lack the diversity displayed by modern image classification few-shot benchmarks.

2.4 Relevant Learning Problems

Learning efficiently with few data, transferring knowledge across different datasets, and leveraging auxiliary tasks to enhance learning have been well-known research topics within the machine learning community for years. Under these settings, which are particularly challenging for standard supervised learning, various approaches have been proposed throughout the years. In this section, we compare these types of learning with the FSL paradigm, delineating their similarities and differences. Additionally, Section 4 provides a more detailed discussion, where several hybrid methods that lie at the intersection of FSL and these learning paradigms are presented.

Meta-Learning. The main goal of meta-learning is to enable faster adaptation to novel tasks by utilizing previously acquired knowledge. To achieve this, meta-learning models extract the common structure from a pool of tasks and use it as an inductive bias for rapid adaptation to unseen tasks with scarce data. In general, meta-learning has been widely used to solve FSL tasks (see Section 3.1). However, its goal of generalizing to novel tasks is not necessarily restricted to the assumption of limited data as in FSL.

Weakly Supervised Learning. This type of learning includes incomplete supervision, where both labeled and unlabeled data is utilized during training, inexact supervision that focuses on training with data that has coarser labels than the ones the model wants to predict, and inaccurate supervision that deals with noisy and incorrect data labels. Although both FSL and weakly supervised learning (WSL) rely

on introducing inductive biases by leveraging other data sources, in WSL, these originate from unlabeled data, while in FSL, they usually originate from labeled data from different tasks or pretrained models.

Semi-Supervised Learning.   A typical case of weakly-supervised learning is semi-supervised learning, where the model is trained on a few labeled data and a large number of unlabeled data. Typically, the semi-supervised model aims to infer a general structure, acting as an inductive bias, from the unlabeled data and then leverage it to boost its performance on the few labeled data. This bears great similarity with FSL approaches that infer inductive biases from outer sources. However, FSL uses labeled data from other tasks to extract prior knowledge.

Self-Supervised Learning.   Similarly to FSL, self-supervised models are initially trained on pretext tasks. However, in self-supervised learning (SSL), these tasks are typically self-generated and consist of unlabeled data, while FSL tasks are labeled and predefined. Additionally, in SSL, there is no restriction on whether these downstream tasks contain only a few data. Generally, the two paradigms have a close relationship (Ni et al, 2021) since they aim to extract robust and transferable representations that can be used in downstream tasks. As a result, various recent approaches have incorporated self-supervised techniques such as contrastive learning within the FSL framework (see Section 4.3).

Transfer Learning.   The goal of transfer learning is to transfer knowledge from a source domain where data is widely available to a target domain where data is limited. Regarding FSL, various transfer learning approaches based on model finetuning have successfully been used (see Section 3.2). However, transfer learning deals with the source (abundant data) and target (limited data) tasks only, and its goal is to transfer knowledge from the former to the latter. On the other hand, FSL deals with multiple tasks, each containing only a few data, and its goal is to extract knowledge from the underlying structure shared among tasks.

Domain Adaptation/Generalization.   Domain adaptation aims to train models capable of maintaining a solid performance when evaluated on out-of-distribution samples, while domain generalization extends this framework to create models that exhibit robustness under any distribution shift (Khoee et al, 2024). Recently, domain adaptation/generalization has been studied within the FSL framework, expanding it from creating models capable of adapting to new tasks with only a few data to also being robust to out-of-distribution tasks and data.

Multi-Task Learning.   In multi-task learning (MTL), the goal is to train a model that is jointly optimized on a set of tasks. Each task contributes to the optimization of all tasks, and consequently, performance is increased compared to training a separate model for each task. This type of training is similar to FSL, where shared information is extracted from training tasks and is subsequently used to improve performance on novel tasks. However, MTL focuses on a predetermined set of tasks, while FSL aims to generalize to novel tasks not seen during training.

Zero-Shot Learning. The major difference between zero-shot learning (ZSL) and FSL is the number of available training samples within each task. In particular, FSL tasks contain a small number of training samples, while ZSL contains none. As a result, while FSL relies on adapting to these few samples, ZSL leverages external data, usually from different modalities, to enable learning in these tasks. Overall, both approaches introduce inductive biases that facilitate effective learning in novel tasks, yet their sources differ significantly.

Continual Learning. In continual learning, the goal is to develop models that can incrementally acquire new knowledge from different tasks without their performance deteriorating due to catastrophic forgetting (Khodaee et al, 2024). While both FSL and continual learning incorporate learning across various tasks, FSL methods are trained offline, while continual learning is heavily based on online learning. This shifts the focus of continual learning towards incremental accumulation of knowledge, while FSL focuses on leveraging prior information to generalize within tasks with a limited number of data efficiently.
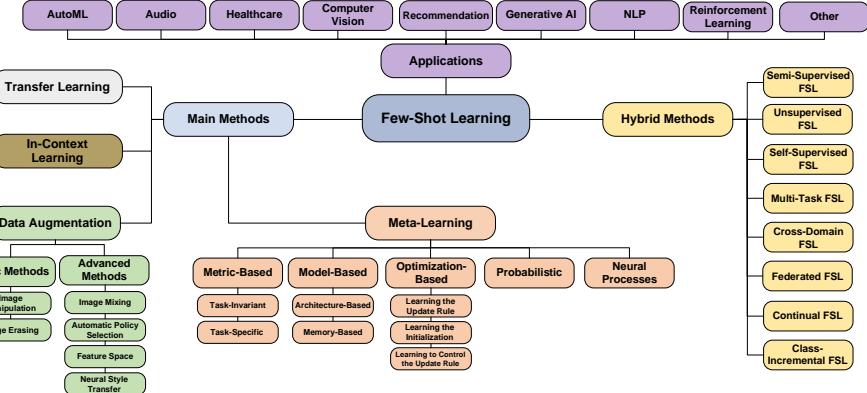
3 Few-Shot Learning Models

In this section, we analyze some of the major approaches proposed over the years in the context of FSL. Most existing surveys have mainly focused on *meta-learning*, *transfer learning*, and *data augmentation* techniques. Yet, due to the recent bloom that FSL has seen, we extend this framework to include additional families of methods such as *neural processes* under the meta-learning methodologies, and the emergent *in-context learning* which is commonly found in the recent large language model (LLM) literature. Our focus is on novel approaches that have surfaced in recent years. Yet, an overview of older but well-established methods is also provided to delineate the background upon which more recent works have been developed. Figure 1 provides an overview of the different types of methods within the FSL paradigm as well as some of their major fields of application.

3.1 Meta-Learning Approaches

Traditionally, meta-learning has been the leading paradigm to deal with FSL. As a result, a wide variety of approaches have been proposed over the years that fall under this category, including *metric-based*, *model-based*, *optimization-based*, and *probabilistic* methods, as well as *neural processes*.

3.1.1 Metric-Based Methods

Due to data scarcity in the FSL setting, obtaining accurate and robust decision boundaries for classification tasks is extremely hard, especially for high-dimensional inputs, such as images. The goal of metric-based methods is to learn an embedding space where it is easier to classify the embedded feature representations of each task's data. At the same time, a suitable metric, either predefined or learned, is used to discriminate the different representations. The main component in these approaches is the feature representation extractor, which can be either *task-invariant*, meaning the

Fig. 1 A taxonomy of FSL methods and their applications. This includes main methods focused on supervised FSL settings (Section 3), hybrid methods for FSL beyond the supervised setting (Section 4), and FSL fields of application (Section 5)
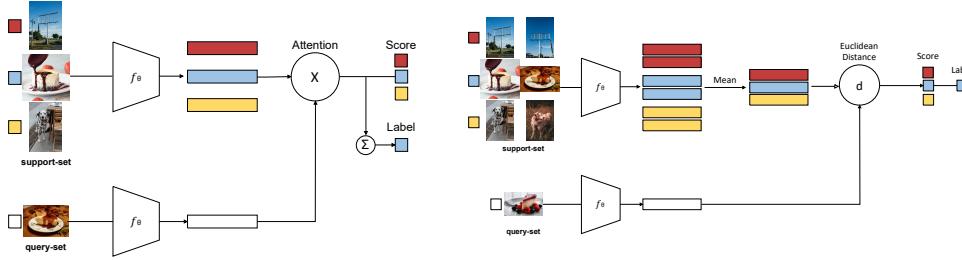
embedding space is the same for all tasks, or *task-specific*, meaning the embedding space is specifically tailored for each task.

Task-Invariant Representations

One of the first FSL models based on metric learning was Siamese Neural Networks (Koch et al, 2015). Specifically, their objective is to learn how to discriminate between two images, similar to contrastive learning. Under this setup, both support and query set samples are encoded using the same Convolutional Neural Network (CNN) feature extractor, and classification is done based on the ℓ_1 distance between each query set sample's representation and the representations of the support set samples.

On the other hand, Matching Networks (Vinyals et al, 2016) employ different CNN embedding networks for support and query sets. The predicted output is the weighted sum of the support set sample labels, and the weights are calculated using an attention mechanism based on cosine similarity between the query and support set samples' representations. Full Context Embeddings are also proposed, where support and query set samples are contextualized on the support set embeddings using an LSTM model. Figure 2 shows a high-level overview of Matching Networks' architecture.

Another popular method, Prototypical Networks (Snell et al, 2017), relies on learning a suitable feature extractor whose feature representations of the same class samples can be clustered around a single prototype representation. A single CNN feature extractor is used for both support and query set samples, and each class prototype is the mean of the representations belonging to that class. Next, a simple nearest class prototype algorithm is employed to classify each query sample. Figure 3 shows a high-level overview of Prototypical Networks' architecture. Recently, Prototypical Networks have been extended in the setting of meta-learning with limited tasks by training them on both real and synthetic tasks. For the creation of synthetic tasks, Conditional Batch Normalization (CNB) is applied to the data of an existing task

Fig. 2 Matching networks architecture (Vinyals et al, 2016). Predictions are made using an attention mechanism to compare the representations of the support and query set samples

Fig. 3 Prototypical networks architecture (Snell et al, 2017). A mean representation is calculated for each class, and the query sample is assigned to the class of the nearest centroid

to create modulated features either using a deterministic objective or via variational inference (Sun et al, 2023).

Similar to Prototypical Networks, Relation Networks (Sung et al, 2018) also use a single feature extractor for both support and query sets and utilize class prototypes. However, in Relation Networks, the nearest class prototype algorithm is replaced by an MLP module that learns with which prototype the query representation is the most similar. In general, this can be seen as a metric learning approach where, instead of using a predefined distance metric (e.g., Euclidean distance in Prototypical Networks), the metric is learned by a neural network. Figure 4 shows a high-level overview of Relation Networks' architecture.

In Garcia and Bruna (2018), FSL is approached by a different perspective where each task is modeled as a partially observed graph. While the feature representations of support and query set samples are still extracted to form the graph, a Graph Neural Network is used to draw inferences, acting as a form of learnable metric. Notably, the proposed method is general enough and includes previous architectures, such as Prototypical Networks and Siamese Networks, as special instances. It can also be applied for semi-supervised and active FSL.

Finally, DeepEMD (Zhang et al, 2020) aims to address the limitations of using global features by leveraging local feature characteristics that provide more discriminative and transferable information across categories. Specifically, instead of comparing representations of whole images, these are broken down into patches, and Earth Mover's Distance (EMD), which is incorporated as a layer, is used to calculate their similarity. Additionally, a cross-reference mechanism is used to discard patches that introduce noise due to cluttered backgrounds and large intra-class appearance variations.
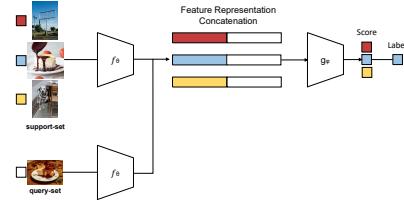
Task-Specific Representations

One of the first works to address the limitations of using task-independent feature extractors was TADAM (Oreshkin et al, 2018), which introduced the use of an extractor contextualized on each task's data. Contrary to Relation Networks' representation contextualization that is done as a post-processing step, in TADAM, the feature extractor is directly modulated via learnable task-specific shift and scale vectors for

each convolutional layer, and a learnable scale parameter is introduced in the model's loss function. Finally, metric scaling is proposed to bridge the gap between using different metrics.
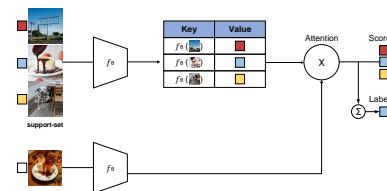
FEAT (Ye et al, 2020) also leverages task-specific representations by contextualizing each support set sample to the rest of the set's samples using a self-attention mechanism, while a contrastive objective is incorporated into the loss function to enhance discrimination between different classes. A transformer-based architecture has also been used in Sparse Spatial Transformers (Chen et al, 2023b) (SSFormers). In particular, SSFormers aim to improve DeepEMD by contextualizing an image's local features to the rest of the image, as well as patches belonging to other support set images, by using sparse attention maps of the extracted task-relevant patches.

Recently, various task-specific adaptations of existing task-invariant methods have also been proposed. For instance, finetuning of Prototypical Networks' backbone feature extractor for each task has been proposed, based on the observation that learned feature representation may fail to generalize to novel tasks due to possible distributional shifts (Hu et al, 2022a). Additionally, in Kim et al (2023), a model that can adapt to different dense prediction tasks by matching similar image patches and using the weighted mean of the output patches as the prediction is proposed. This model can be seen as a generalization of Matching Networks where matching is performed on the latent representations of the labels derived from similar image patches.

Finally, in Hu et al (2022b), the problem of few-shot semantic segmentation is addressed by learning a feature extractor that addresses different sources of heterogeneity. In particular, it incorporates a cross-sample attention module, the fusion of information between different background patches, and a masked image segmentation scheme to reinforce the capacity of contextual inference. Contrary to previously mentioned approaches, this method focuses on learning more separable feature representations by enhancing intra-class compactness and leveraging their distances during training.

Fig. 4 Relation networks architecture (Sung et al, 2018). Support and query set representations are concatenated and classified based on their similarity following a metric learning approach

Fig. 5 Memory-Augmented Neural Networks architecture (Santoro et al, 2016). An external memory module is utilized to store and retrieve refined sample representations within each task

3.1.2 Model-Based Methods

While metric-based meta-learning methods focus on producing easily separable feature representation, model-based methods are generally based on developing specific

model architectures that enable rapid adaptation to novel tasks. Yet, despite their differences, both methods aim to constrain a model's hypothesis space to facilitate its adaptation using less data. Model-based methods can be either *architecture-based*, where the proposed model leverages an internal memory or a specifically designed combination of components that enable FSL, or *memory-based* methods, where an external memory module is used to retain information related to rare class samples.

Architecture-Based Methods

Some of the earliest works in this line of research have been primarily based on the fact that the hidden states of Recurrent Neural Networks (RNNs) can be utilized to model learning algorithms. For instance, in Duan et al (2016), learning a reinforcement learning (RL) algorithm capable of rapidly adapting to novel tasks is also formulated as an RL problem. Based on the concept that in meta-learning, knowledge acquisition takes place in two different time scales, a GRU network's internal state is used as a fast learner to adapt to a new task with few samples, while Trust Region Policy Optimization is used to optimize the slow learner's policy across tasks. Consequently, the model can generalize within different tasks like multi-armed bandits, tabular Markov Decision Processes (MDPs), and visual navigation (partially observable MDPs). On the other hand, in Wang et al (2016), a similar approach based on RL is applied for structured task distributions. The proposed deep meta-RL approach trains an RNN to act as a learning algorithm that can be applied to different MDPs from the same distribution. Using reward information from various tasks, the RNN can quickly adjust to new tasks during testing without finetuning due to its internal state and ability to model different learning algorithms.

On a more general approach, SNAIL (Mishra et al, 2018) can be applied to different types of problems without any assumptions for the task or the optimization procedure. It uses a generic architecture based on three building blocks: dense blocks (temporal 1D causal convolutions), TCBlocks (dense blocks with increasing dilation rates), and a self-attention block. Interestingly, using temporal convolutions interleaved with attention enables selecting relevant information over a larger context window. Overall, this approach allows for learning an optimal strategy without any built-in assumptions, as in gradient-based meta-learning.

Finally, some recent approaches have introduced architecture-based models suitable for or inspired by different domains, such as a meta-learning model for forecasting dynamical systems that generalizes across different time series sharing a common structure (Wang et al, 2022b). This is achieved by extracting time-invariant features of each time series dynamics using a CNN-based encoder and then adapting to the dynamics of each specific task via adaptive modules. On the other hand, inspired by the Lottery Ticket Hypothesis (Frankle and Carbin, 2018), a gradient-based meta-learning method is proposed in Chijiwa et al (2022) to meta-learn a sparse neural network structure suitable for FSL. Specifically, the method aims to learn a parameter mask to derive a sparse subnetwork that demonstrates strong meta-generalization capabilities. This method can be seen as a generalization for architecture-based modeling from hand-crafted architectures to learnable ones.

Memory-Based Methods

External memory modules can store different types of content that facilitate rapid adaptation, such as model weights or sample representations. Meta Networks (Munkhdalai and Yu, 2017) is a model that uses external memory modules to store fast weights for rapid adaptation across tasks. The model consists of a meta-learner (learning across tasks) and a base-learner (task-specific learning) that are both augmented using additional parameters (fast weights) generated on the fly. These fast weights are used to parameterize each module to rapidly adapt to specific tasks and specific task samples. An external memory module is also used to store these fast weights and retrieve their optimal combination based on the task's input representations.

On the other hand, most memory-based models typically rely on external memories to store refined sample representations that enable fast adaptation to classes with few samples. In Santoro et al (2016), the stored values are retrieved using soft attention and writing to memory is controlled by a content-based memory writer that writes memories to either the least used or the most recently used memory location. Figure 5 shows a high-level overview of the model's architecture. Additionally, in Kaiser et al (2017), an independent memory module is proposed that allows neural networks to achieve one-shot and life-long learning. Specifically, it uses a memory that stores class labels, and the main network's layer activations are used to query the memory based on their cosine similarity with the memory keys. An approximate nearest neighbors algorithm is used to scale the memory, and the memory can be combined with various network types, leading to improved classification of rare samples.

Recently, memory-augmented networks have also been utilized in different applications where rare classes are common. In Gkanatsios et al (2023), a memory-augmented 3D object scene segmentation network is proposed. The model consists of a retriever that finds the most relevant representations in the memory and a modulator where the retrieved memories are encoded and enhanced by scene-agnostic queries, while attention modules are also used for contextualization of the query images. Finally, in Schimunek et al (2023), a memory-enhanced model for few-shot drug discovery is proposed. Specifically, it uses a context module based on Modern Hopfield Networks (MHNs) to extract representations of the support set and the query set conditioned on a large external context set of molecules via an attention mechanism. Attention is also used to contextualize the support and query sets, and overall robustness is enhanced by reinforcing the covariance structure and reducing the noise.

3.1.3 Optimization-Based Methods

Optimization algorithms are a crucial component of deep learning. However, they lack guarantees for nonconvex optimization problems like deep neural network optimization and may require many iterations for convergence, leading to suboptimal solutions. Additionally, choosing the best optimizer for different problems is a laborious task. Optimization-based meta-learning aims to learn optimization procedures suitable for optimization across different tasks without multiple iterations or large volumes of data by casting it as a 'learning how to optimize' problem. Subsequently, optimization-based methods can be categorized into three types: learning the *update rule*, learning the *initial parameters*, and learning to *control the update rule*.

Learning the update rule

Based on the principle that learned features are better than hand-crafted ones, in this setting, using hand-crafted update rules such as Stochastic Gradient Descent (SGD) is replaced by learned rules. This is achieved by casting the design of the rule as a learning problem. One of the first approaches towards that direction has used LSTMs to output the updates of the gradient descent step used within each task (Andrychowicz et al, 2016). To control all steps of the optimization procedure, an objective dependent on the entire trajectory of learned optimization is used. LSTMs are suitable for this since they maintain past information in their internal state, allowing for previous updates to inform subsequent ones. Meta-Learner LSTM (Ravi and Larochelle, 2016) extends this framework to the FSL setting by learning both the base learner's initialization and update strategy using only a small, predefined number of steps. Specifically, each LSTM cell state represents the parameters of the learner, and the cell state variations correspond to model updates. However, this method does not scale well with the base learner's size, and both methods fail to capture any dependencies between different model parameters. Finally, finding an optimal optimization algorithm has also been cast as an RL problem (Li and Malik, 2017). Specifically, learning the optimal way to obtain update steps can be seen as selecting an optimal update policy, which is achieved via guided policy search. A specialized cost function is also used that encourages learning update rules that converge fast.

Learning the initial parameters

To address hand-designed optimization algorithms' slow convergence, some optimization-based algorithms focus on learning initial parameters that converge to optimal parameters with only a few gradient descent iterations. Model-Agnostic Meta-Learning (Finn et al, 2017) (MAML) is the first work that aims to learn a model parameter initialization that allows for rapid adaptation to each given task. In particular, meta-learning is cast as a bi-level optimization problem with two learning loops in two different levels. In the inner loop, the base learner is responsible for adapting the initial parameters in the task's support set and evaluating them on the task's query set. In the outer loop, the meta-learner optimizes the initial model parameters using the query set errors across training tasks. As a result, the shared information across tasks is encapsulated in the model's learned initial parameters, which act as a prior that obviates the need for many optimization steps within each task.

Following MAML, many different extensions and variants have been proposed, aiming to address its limitations. For instance, MAML++ (Antoniou et al, 2019) is such an extension that introduces various novel features that aim to enhance MAML's performance and ensure its training stability, such as an outer-loop loss that is dependent on all of the inner loop steps' losses, improved batch normalization for meta-learning, and learned inner loop learning rates. The difficulty of choosing optimal learning rates for MAML due to its training instability is also addressed in Behl et al (2019), where the authors propose to learn inner- and outer-loop learning rates using hypergradient descent, which adaptively updates the learning rates through gradient descent. Meanwhile, in Rajeswaran et al (2019), an implicit differentiation approach that removes the need for differentiating through the inner loop's optimization path is proposed to

15

address the high computational cost and memory burden imposed by using higher-order derivatives in MAML. Finally, REPTILE (Nichol et al, 2018) introduces a first-order approximation of MAML that obviates the need to backpropagate through inner loop gradients. Consequently, this mitigates the need to calculate second-order derivatives and significantly reduces computational complexity.

Multiple approaches have also focused on improving MAML's performance and extending its capabilities. In Rusu et al (2019), the proposed method decouples optimization-based meta-learning from the high-dimensional space of the model parameters by embedding the model parameters to a lower dimensional latent space. The produced decoded initial parameters are specific for each task and conditioned in its support set, and MAML uses the latent variables as input. Lastly, by formulating tasks as discriminating between in-distribution and out-of-distribution samples for a specific class, MAML has also been extended in detecting out-of-distribution samples (Jeong and Kim, 2020).

The recent advances in theoretical aspects of meta-learning have also led to the development of various theory-inspired MAML adaptations. In Abbas et al (2022), to avoid the sharp local minima of MAML's loss function, sharpness-aware minimization, a method that simultaneously minimizes the loss value and the loss sharpness, is integrated into MAML. Additionally, in Wang et al (2023c), to improve both meta-generalization and generalization performance within each task after adaptation, standard regularization is applied in the outer loop, and inverted regularization is applied in the inner loop, increasing adaptation difficulty within each task and forcing the model to learn better-generalized hypotheses at the meta-level.

Learning to control the update rule

Learning the update rule in FSL can lead to update rules with non-converging behavior and lacking the inductive bias of gradient descent. To address these limitations, a branch of optimization-based approaches focuses on learning how to control the updates of existing optimizers through a learnable component that modulates the loss gradient (gradient preconditioning). One of the first approaches in that direction extends MAML by additionally meta-learning the learning rates and learning directions for each parameter of the base learner (Li et al, 2017). This allows for more expressive learning but doubles the amount of meta-learned parameters, resulting in higher computational complexity. Gradient preconditioning can also take the form of learning which parameters are shared across tasks and which ones are task-specific by using a learnable mask and transforming the model parameters using meta-learned transformation parameters (Lee and Choi, 2018). This can be seen as extending MAML to learn a parameter subspace for task-specific adaptation using a meta-learned distance metric. In Park and Oliva (2019), MAML is extended to meta-learn curvatures. These curvatures are used to transform gradients before passing through the optimizers, thus constituting a form of preconditioning. Preconditioning can also be induced implicitly by introducing nonlinear layers interleaved with the base learner's layers (Flennerhag et al, 2020). In this case, the occurring preconditioning matrix is meta-learned without backpropagation over the base learner's parameters, leading to increased efficiency and better scaling. Finally, using a preconditioner in MAML's

inner loop to enable more efficient learning by reshaping the geometry of the parameter space has also been proposed (Kang et al, 2023b). The preconditioner is task-specific and formulated as a Riemannian metric, equivalent to natural gradient descent, and it alleviates meta-overfitting by relying more on individual task characteristics.

3.1.4 Probabilistic Extensions

To alleviate deterministic meta-learning's inability to quantify uncertainty, various probabilistic approaches have been developed to improve performance and uncertainty quantification by modeling parameters as random variables and obtaining their distributions. In this direction, Bayesian meta-learning is a popular paradigm that fits expressive generative models to small datasets by incorporating inductive priors learned from related tasks. One of the first works to explore probabilistic FSL is in Edwards and Storkey (2017), where a hierarchical version of Variational Autoencoders (VAEs) is proposed that can be used for FSL by learning the statistics of the training data and then using them to classify new samples. The basic idea is to perform variational inference at two levels using a double variational bound where a higher-level latent variable describes datasets and a lower-level latent variable describes individual examples.

Additionally, many subsequent works have focused on extending MAML to the probabilistic setting. In Grant et al (2018), MAML is reinterpreted as an inference for the parameters of a prior distribution in a hierarchical Bayesian model, and it is shown that the choice of update rule in the inner loop corresponds to a choice of prior over task-specific parameters. To mitigate MAML's inaccurate point approximation, a Bayesian variant of MAML using the Laplace approximation is also proposed. In Finn et al (2018), MAML is extended by learning the joint posterior of initial and task-specific parameters, conditioned on each task's support set, using structured variational inference. An amortized variational inference technique is also proposed to avoid storing separate variational distributions for each task's initial and task-specific parameters. A Bayesian variant of MAML leveraging non-parametric variational inference has also been proposed in Yoon et al (2018), which learns the posterior distributions of task-specific parameters using Stein Variational Gradient Descent to approximate the posterior over the weights in the final layer with a particle ensemble. However, the posterior's expressiveness depends on the number of particles in the ensemble, leading to poor scaling regarding memory and computation costs.

Apart from probabilistic MAML extensions, novel architectures have also been proposed, such as deep kernels where both feature extractor and base kernel parameters are meta-learned across tasks (Patacchiola et al, 2020). The deep kernel's parameters are learned in the outer loop by maximizing the marginal likelihood across all tasks using a Gaussian process approach, obviating the need for an inner loop and leading to an efficient and straightforward closed-form solution. Finally, a novel method is introduced in Pavasovic et al (2023) that meta-learns a Bayesian prior on the function space to avoid overconfident estimations. The problem is formulated as a stochastic process with different realizations, and the prior's score function is estimated using a transformer encoder architecture.
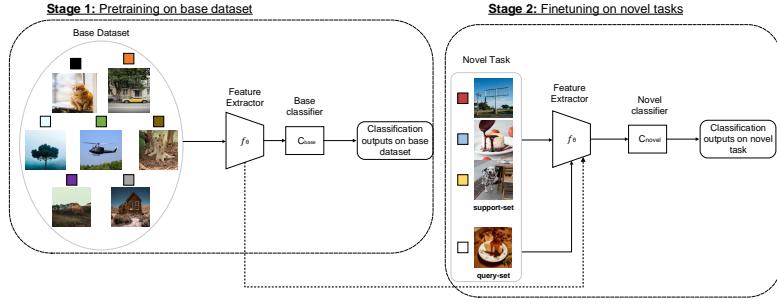
Lastly, probabilistic meta-learning can also mitigate some of metric-based meta-learning's limitations. Specifically, metric-based meta-learning methods typically obtain a single point estimate for each class, leading to biased estimates due to a lack of data. However, a proposed extension in Zhang et al (2019) models each class with its distribution through variational inference, allowing for the calculation of confidence and more accurate classification of novel samples. Expanding on this, in Sun et al (2021), computational complexity is reduced by using fewer random variables for latent class prototypes. A global model is used to learn the posterior distributions of these variables through amortized inference, and task-specific estimates of variational parameters are obtained using initial parameters learned across tasks.

3.1.5 Neural Processes

Parallel to the development of probabilistic models that extend the classic meta-learning framework, significant research effort has revolved around Neural Processes (NPs), another family of probabilistic models suitable for FSL. On a higher level, NPs aim to bridge the gap between deep neural networks (DNNs) and Gaussian Processes (GPs). DNNs excel at function approximation, yet they require retraining for each new function. On the other hand, GPs can quickly infer new functions at test time, but they are computationally expensive and sensitive to the selection of priors (Garnelo et al, 2018a). NPs can make accurate predictions given a few training data points and scale to complex functions and large datasets. NPs were initially developed as function estimators, suitable for modeling stochastic processes. However, their ability to model different functions using only a few samples makes them ideal for few-shot regression, where each function is a different task, and the context data constitutes the support set.

The first proposed model of this family has been Conditional Neural Processes (Garnelo et al, 2018a) (CNPs) that uses a permutation-invariant neural network to parametrize its dependence on the observations. For each task, an aggregated mean representation of the support set is obtained and used by a decoder that models the distribution parameters of the examined target point. The model is trained end-to-end through gradient descent to maximize the conditional likelihood of a random subset of targets given a random observation set. Extending CNPs, NPs (Garnelo et al, 2018b) introduce a more generalized approach by using a global latent variable that allows for better uncertainty modeling. Similar to CNPs, an aggregated mean representation of the support set data is obtained, which conditions the global latent variable. Finally, the decoder receives the target data and a sampled value from the conditioned global latent variable as input and predicts the output values.

Yet, despite their desired properties, these methods lack expressivity power and suffer from underfitting in complex settings. To address these limitations, Attentive Neural Processes (Kim et al, 2019) (ANPs) introduce an attention mechanism to enhance model performance. Specifically, self-attention is used to capture the interactions between support set samples, and cross-attention is used to obtain the aggregated representation. However, these attention modules also lead to increased computational complexity. Convolutional Conditional Neural Processes (Gordon et al, 2019) (CCNPs)

Fig. 6 Two-stage model training in transfer learning. In the first stage, the model is pretrained on a large annotated dataset (base dataset), while in the second stage, it undergoes further finetuning using a smaller target dataset

have also been proposed as an extension of NPs, which introduces translation equivariance as an inductive bias using a convolutional extension of Deep Sets. Additionally, to enable the modeling of temporal dynamics in sequential data, a combination of RNNs and ANPs has been proposed in Qin et al (2019), where the hidden states of an LSTM are used to extract the feature representations. Finally, some recent extensions have also included NPs based on the Transformer architecture (Nguyen and Grover, 2022).

3.2 Transfer Learning Approaches

Although meta-learning approaches have been successfully used in the context of FSL, they are notorious for their architectural and computational complexity. However, dealing with these types of complexity is neither desirable nor affordable in various real-world settings. In contrast, transfer learning approaches are more straightforward and feasible for such cases, and many novel methods tailored for FSL have been proposed in recent years. In general, most of these methods rely on a two-stage learning procedure that includes pretraining a feature extractor backbone on a large base dataset and then performing finetuning in downstream novel tasks with few samples. Typically, most of the proposed approaches differ in the finetuning stage, which can either include finetuning the whole model or only parts of it. Figure 6 illustrates a high-level overview of the two-stage procedure.

One of the first works to explore the efficiency of finetuning techniques for FSL has been in Chen et al (2019), where it is argued that finetuning methods' performance has been underestimated in the past. Consequently, a simple baseline that utilizes a pretrained feature extractor backbone and finetunes only the classification head within each few-shot task is proposed. Classification is then performed based on the cosine similarity between the prototypes extracted from the weight matrix of the classifier head and the feature representations of the given samples. Similarly, in SimpleShot (Wang et al, 2019), a pretrained frozen backbone is also used to extract feature representations. However, a series of simple transformations such as centering and ℓ_2-normalization are applied to the extracted representations, and the final classification is performed using the nearest neighbor/centroid algorithm.

19

On a different approach, in Wang et al (2022a), it is suggested that transferring knowledge from a base dataset to new tasks without retraining the pretrained backbone could lead to biased results due to the extraction of distorted features for out-of-distribution samples. To avoid this and prevent overfitting of the feature extractor due to the small number of samples, the whole model is finetuned using an inverted Firth Bias regularizer that regularizes the distribution of the predicted labels to be more uniform, similar to label smoothing. Full model finetuning is also proposed in Dhillon et al (2020). In particular, the pretrained backbone is combined with a linear classifier head, and a transductive objective is introduced to finetune both the backbone and the classifier within each task, granting it the ability to adapt to different numbers of ways and shots without retraining.

To address both limitations of feature distortion using a frozen pretrained backbone and the possibility of overfitting when finetuning the whole model, partial finetuning, e.g., finetuning only specific layers of the backbone network, has been proposed in Shen et al (2021). To determine the finetuned layers, a genetic programming algorithm is used to learn each layer's learning rates during finetuning, automating the whole procedure. On the other hand, in Shysheya et al (2023), FiLM modules are introduced to modulate the backbone network without retraining all of its parameters, substantially reducing the number of updated parameters. On top of it, a Naive Bayes classifier is used for prediction.

Finally, various other approaches have also emerged, aiming to tackle FSL via model finetuning. In Rizve et al (2021), the effect of invariance and equivariance is explored in the context of learning better feature representation of samples for FSL. Subsequently, the authors propose learning two additional modules alongside the classification head for enforcing equivariance and invariance via self-supervised learning, and the final model is obtained via self-distillation. Transfer learning has also been combined with meta-learning to bridge the gap between the two methods. Specifically, in Demirel et al (2023), meta-learning is used to tune an object detection model's hyperparameters, while finetuning is used to tackle the few-shot object detection objective in each task.

Overall, finetuning methods can be a great asset in few-shot applications where simplicity and efficiency are crucial. However, there is no clear consensus on whether using a frozen backbone or finetuning it either fully or partially leads to better performance (Luo et al, 2023). Additionally, pretraining the backbone network in a large base dataset is not always possible in some domains. Finally, cautiousness is needed when there is a domain gap between the base dataset and the novel tasks since it could hinder learning due to negative transfer.

3.3 Data Augmentation Approaches

Since FSL is characterized by the lack of adequate data to enable accurate adaptation to the problem at hand, a straightforward solution is to augment the given dataset with new samples. This is the main functionality of Data Augmentation (DA) methods, which reduce overfitting in machine learning models either by enforcing a form of regularization through the existing data or by increasing the number of available samples. In general, DA techniques can further be categorized into basic and advanced

DA methods. It is worth noting that FSL has primarily been examined for visual domain tasks, and most of the techniques discussed in this context refer to image input.

3.3.1 Basic Data Augmentation

Basic DA methods include fundamental methods widely used in both FSL as well as standard machine learning and can further be categorized into two groups: (a) *Image Manipulation*, and (b) *Image Erasing*. In general, these methods have been widely adopted due to their simplicity, which makes them easy to implement and adds minimal overhead to the training procedure.

Image Manipulation

This group of DA methods involves altering the visual representation of an image. Some widely used image manipulation techniques include rotation, translation, shearing, cropping, and color jittering. By introducing these geometric and visual modifications, models trained using such data can become more robust by learning feature representations invariant to these alterations. However, hand-picking these transformations beforehand is a challenging task, and simply trying all available methods can add up to a significant computational cost. Consequently, while these methods can be helpful, they can only provide a partial solution in the context of FSL.

Image Erasing

The main logic behind these approaches is to regularize the model by preventing it from overfitting to particular regions of an image and ignoring other parts. In general, this is achieved by removing specific parts of an image and replacing them with new ones, forcing the model to focus on different features. For instance, Cutout randomly removes square regions (cutouts) of the training images, while Random Erasing randomly selects and erases rectangular regions within an image and replaces them with random or predefined pixel values. On the other hand, a more recent approach, Context-Guided Augmentation (Chen et al, 2023a), adds new instances of classes strategically using contextual information to simulate occlusion occurrences, with each segmented instance being selectively positioned onto the object's bounding box in a randomized manner. Overall, image erasing methods can be seen as a form of spatial dropout that induces robustness against object occlusions and different spatial configurations. However, they suffer from the same limitations as image manipulation methods since they must be hand-picked beforehand.

3.3.2 Advanced Data Augmentation

The rapid expansion of the computer vision field has led to the emergence of novel sophisticated DA approaches that aim to overcome the limitations of basic DA methods. Advanced DA techniques typically fall into categories such as image mixing, automatic policy selection, feature space augmentations, and neural style transfer.
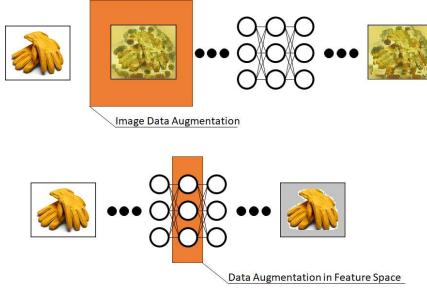
Image Mixing

These methods aim to enhance generalization performance across diverse datasets and mitigate the risk of overfitting by integrating information from multiple sources. In Mixup (Zhang et al, 2018), this is achieved by blending two or more images by taking a convex combination of their pixel values and their corresponding one-hot label encodings to create novel hybrid samples. Notably, these linear interpolations introduce diversity to the training data and assist models in learning more robust and generalized features with minimal computational overhead. CutMix (Yun et al, 2019), on the other hand, extends this concept by cutting and pasting rectangular patches from different images, fostering spatially coherent variations. More recently, Multi-stage Augmented Mixup (Liang et al, 2023) (MiaMix), a novel mixing ratio sampling method, has been proposed, consisting of four stages: random sample pairing, sampling of mixing methods and ratios, mixing mask generation and augmentation, and obtaining the mixed output. MiaMix selects a mixing method from a pool of methods and sets the mixing ratio for each mask. The masks are then augmented and mixed in the final stage. Finally, PatchMix (Xu et al, 2022a) has been proposed to address the existence of biases, such as spurious correlations, in FSL. Specifically, it disentangles causal and non-causal features to improve model invariance and generalizability by augmenting original query images with images from another query while preserving the label of the additional query.
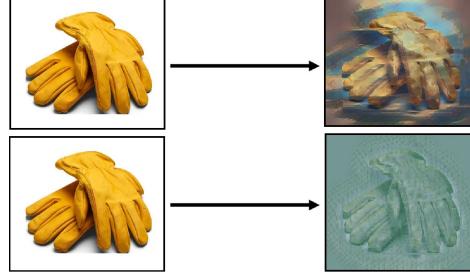
Automatic Augmentation Selection Policies

Although hand-crafted DA methods have been invaluable in dealing with data scarcity and creating models invariant to various transformations, manually selecting the optimal augmentations to apply is a challenging task that often leads to suboptimal results. As a result, novel approaches have been suggested to automate this selection process by systematically finding the optimal selection policy. AutoAugment (Cubuk et al, 2019) formulates a search space within which each policy consists of multiple sub-policies that encapsulate different image transformations. In this setting, the most effective policy is identified through a search algorithm that maximizes validation accuracy on a target dataset based on RL principles. On the other hand, Randaugment (Cubuk et al, 2020) automates this selection by introducing a randomized approach. In particular, it dynamically selects a set of augmentation operations from a transformations pool and applies them with random magnitudes to each image. This approach strikes a balance between variability and consistency while it also obviates using an RL algorithm to find the optimal policy.

Feature Space Augmentations

While traditional DA techniques have focused on applying transformations on the raw input data, feature augmentation techniques have also been used to enhance the learning capabilities of a model by manipulating the extracted feature representations. This can involve methods such as introducing new features, transforming existing ones, or modifying specific feature values. For instance, the work in Wang et al (2023a) proposes the Information Fusion Rectification method to identify relevant base class prototype features from the query set using cosine similarity. The method leverages

Fig. 7 Feature data augmentation techniques extract the image's feature representation and apply the transformation in the feature space

Fig. 8 Neural style transfer methods generate synthetic images by modulating image characteristics while preserving the original semantic information

the relationship between base, support, and query samples and rectifies the support set's distribution using cosine similarity metrics as weights, generating new samples on the extracted feature space relevant to the provided task. Figure 7 shows an example of applying feature data transformations.

Neural Style Transfer

A different way to perform DA could be by leveraging modern techniques such as Neural Style Transfer, which offers a unique approach focusing on style-based variations. In Jackson et al (2019), the use of random style transfer is proposed by introducing randomness to texture, contrast, and color while preserving the underlying shape and semantic content of data. This is achieved by adapting an arbitrary style transfer network to perform style randomization, with target style embeddings being sampled from a multivariate normal distribution instead of being computed from a specific style image. Figure 8 illustrates an example of generating synthetic images following different styles. Finally, diffusion models have also been utilized in this context. Specifically, in Feng et al (2023), a pretrained CLIP encoder is used to extract latent features fed to Stable Diffusion to generate synthetic images. Data diversity is ensured by filtering generated images based on cosine similarity, while augmentation is enabled in image and feature spaces.

Overall, DA methods are an effective way of dealing with FSL due to their simplicity, flexibility, and facile combination with approaches such as meta-learning. However, most of these methods leverage domain knowledge and focus on image data, thus lacking transferability between different domains and problems. Consequently, more research towards developing domain-agnostic DA approaches is necessary.

3.4 In-Context Learning

Although not explicitly trained for it, it has been demonstrated that LLMs are capable of generalizing to novel tasks when given just a few demonstrations. This emergent few-shot ability of LLMs was first examined in Brown et al (2020), where scaling the model parameters and the number of training data leads to the creation of models capable of adapting to novel tasks without further parameter finetuning. Ever since, similar FSL capabilities have also been demonstrated under various model settings,

such as bidirectional LLMs (Patel et al, 2022) and different tasks, including machine translation (Garcia et al, 2023). In the literature, this few-shot ability has been referred to as in-context learning (Dong et al, 2022a; Zhou et al, 2024) (ICL) since the model leverages a few examples that form a demonstration context to generalize within it. In-context learning can be seen as a learning mechanism based on analogies where the model draws on its prior knowledge and forms analogies between the provided demonstrations and the query at hand.

Comparing ICL and standard FSL, one of the major differences between the two approaches is their effect on the model parameters. More precisely, the main advantage of ICL is that the model is able to adapt to a new task without having to finetune its parameters through gradient descent. Simply by structuring a prompt that provides the context for the task through input-output pairs of examples, the model can generalize to the task, while in FSL, the model's parameters need further adaptation. As a result, ICL significantly reduces computing power and time associated with novel task adaptation. For instance, based on these few-shot capabilities, an incremental stage-wise training of language models has been proposed that enables faster and more efficient training and adaptation (Reddi et al, 2023). Moreover, another difference is that FSL models are explicitly trained to learn how to rapidly adapt to new tasks with just a few training samples, while ICL naturally emerges in LLMs without the need for explicit optimization. Finally, an additional advantage of ICL is its interpretability since the demonstrations are provided in natural language, thus facilitating the incorporation of human knowledge in them.

Undoubtedly, these desirable characteristics of ICL have rendered LLMs particularly attractive in various applications where data is limited. In Meng et al (2023), an autoregressive LLM is finetuned through prefix-tuning to generate new data. Consequently, the generated data is used to augment the original dataset in an attempt to render supervised learning feasible. In Dai et al (2022), LLMs are used as few-shot query generators to enhance information retrieval in cases where training data is absent. Notably, the proposed method manages to outperform models pretrained on large question-answering datasets by simply using eight examples as demonstration input for the LLM data generator. Furthermore, LLMs have been leveraged as few-shot abductive reasoners (Shi et al, 2023). In that case, given a few annotated examples, their goal is to suggest possible causes for an event as part of an event sequence modeling pipeline. Finally, in recent years, ICL has also been applied to various modalities beyond text, using vision (Foster et al, 2023) and language-vision (Tsimpoukelli et al, 2021) models, demonstrating its potential as a new general form of FSL.

3.4.1 Theoretical Understanding of In-Context Learning

In recent years, there have been several attempts to shed light on the mechanisms that enable ICL in LLMs. On a theoretical level, various studies have analyzed how specific components within LLMs contribute to the emergence of ICL. Specifically, in Olsson et al (2022), the authors explain how certain attention heads in the model's architecture, called induction heads, are responsible for ICL by reproducing previously occurring token sequences for next-token prediction. More recently, this direction has been further explored using a Markov chain modeling task (Edelman

et al, 2024). At the same time, another recent study has examined ICL through the lens of function vectors, which are responsible for transmitting information regarding the demonstration examples across the transformer modules (Todd et al, 2024).

Given the similarities between ICL and meta-learning for FSL, in terms of their ability to adapt to novel tasks using a minimal number of samples, various studies have attempted to frame ICL as a type of gradient-based meta-learning, where the forward pass of the transformer modules in ICL approximates performing finetuning within each task's support set in FSL (Von Oswald et al, 2023). Specifically, in Dai et al (2023), GPT-based models are viewed as meta-optimizers and ICL as implicit model finetuning based on a recasting of transformer attention as a meta-optimization module. However, recent studies have suggested that these similarities may not be sufficient to fully explain ICL due to differences between the inner workings of ICL and gradient descent (Deutch et al, 2024), necessitating further study in that area.

Another line of theoretical research on ICL has focused on analyzing it as a form of function regression learning, where the model learns how to fit a regression function to the provided demonstration examples. In that direction, most existing works have focused on learning linear regression functions from the given examples (Garg et al, 2022; Akyürek et al, 2023). At the same time, a more recent study has also extended this framework to a broader setting by introducing a compositional structure that combines learning a fixed representation function with a linear regression function (Guo et al, 2024).

Lastly, in Xie et al (2021), ICL is presented as a form of implicit Bayesian inference, and the effect of the training data distribution is underlined. The same framing has also been combined with the regression function learning (Panwar et al, 2024) and gradient-based meta-learning (Jeon et al, 2024) frameworks in an attempt to provide a broader and unified perspective of how ICL works.

3.4.2 Empirical Insights on In-Context Learning

Parallel to the works providing a theoretical framework for ICL, various studies have focused on identifying and quantifying the effect of different influencing factors from an empirical perspective. More specifically, model size (Kaplan et al, 2020) has been pinpointed as an influencing factor, yet it is inadequate to illustrate the intrinsic properties that lead to the emergence of ICL. Similarly, other model properties dubbed as possible influencing factors have been computation (Wei et al, 2022a), and the objectives selected during model pretraining (Tay et al, 2023). In terms of the data used during model pretraining, attributes such as their domain (Shin et al, 2022), distribution (Chan et al, 2022), and diversity (Raventós et al, 2024) have been examined, underscoring the importance of the pretraining data for the emergence of ICL.

Since LLMs are not explicitly trained to perform ICL, this emergent property can also be influenced by various factors during inference. Specifically, attributes of the examples used as demonstrations in the ICL setting, such as their diversity (An et al, 2023), their similarity with the test query (Liu et al, 2022), and their ordering (Liu et al, 2024b) have been shown to impact ICL performance. Moreover, another critical factor during inference affecting ICL is the accuracy of the mapping between inputs

and labels in the demonstrations and the samples used during pretraining (Kossen et al, 2023).

However, there are cases where findings seem to be contradictory regarding the actual impact of the various examined factors. For instance, while in Kandpal et al (2023) a correlation between the relevance of the pretraining data and the examined downstream task is identified, in Han et al (2023a), the authors show that this factor does not influence the effectiveness of ICL. Additionally, although initial findings have demonstrated that the input-label mapping does not influence ICL (Min et al, 2022b), subsequent studies have revealed that a consistent input-label mapping between pre-training data and the demonstrations used in ICL can lead to improved performance (Kossen et al, 2023).

3.4.3 Boosting In-Context Learning with Meta-Learning

An interesting aspect of ICL is that despite being itself a form of FSL, it has recently been combined with other FSL paradigms and, most notably, meta-learning to enhance its performance. Specifically, ICL is related to prompt learning, which allows for the use of learnable prompts instead of hand-crafted ones. In this context, meta-learning can be utilized to efficiently learn better prompts across various downstream tasks, thus enhancing the performance of ICL. Although LLMs already demonstrate implicit few-shot capabilities, their performance in FSL scenarios can be further increased through explicit training on few-shot formulated tasks. For instance, in Min et al (2022a), a meta-training scheme based on formulating few-shot tasks is proposed to enhance the ICL capabilities of LLMs. Additionally, in Hou et al (2022), an approach adapting MAML for prompt tuning is proposed, which meta-learns an effective prompt initialization that enables fast adaptation to multiple downstream tasks. Lastly, a novel approach has recently been proposed aiming to improve ICL through meta-ICL (Coda-Forno et al, 2023). More specifically, although ICL allows for rapid adaptation to downstream tasks using a minimal number of demonstrations, different tasks may benefit from different priors and ICL strategies. By introducing an additional level where ICL is performed across sets of downstream tasks, the model is able to adapt its priors and learning strategies and perform ICL more effectively within each group of tasks.
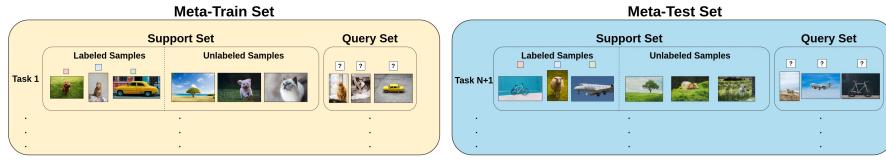
3.4.4 Limitations

Although the benefits of ICL have been both theoretically explored and practically demonstrated, there are still various limitations pertinent to its inherent nature. As discussed earlier, ICL can be unstable, influenced by factors such as the prompting format, the examples provided, and even their ordering (Zhao et al, 2021). To address these issues, different calibration methods have been proposed to make LLMs less sensitive to these parameters. For instance, in Zhao et al (2021), various types of biases in prompts are identified, and a contextual calibration approach is proposed based on content-free test inputs. In Han et al (2023b), prototypical networks are utilized to calibrate the decision boundary of LLMs for classification in the ICL setting. This is achieved by creating prototypical clusters based on the support set samples using a Gaussian Mixture Model. Consequently, when combined with this type of calibration,

LLMs demonstrate increased robustness regarding prompt ordering, class imbalance, and size of estimate set. Lastly, a recent study has posited that ICL can be viewed as a byproduct of using nonlinear or discontinuous metrics rather than an emergent property of LLMs (Schaeffer et al, 2024). Overall, this lack of clear consensus regarding a comprehensive theoretical and empirical justification of ICL highlights the need for further research to shed light on this novel and distinct form of FSL.

4 Few-Shot Learning Beyond Supervised Learning Settings

Something that can be easily inferred from the models presented so far is that almost all of these methods have been developed for use in a supervised setting. However, this greatly limits the applications in which FSL can be used since many ML applications rely on unlabeled data. As a result, various novel approaches have emerged in recent years that aim to address this limitation by extending the FSL paradigm to different types of problems. At the same time, the close relationship of FSL models with techniques developed within other learning paradigms, such as self-supervised learning or multi-task learning, has led to the bridging of these approaches and the development of new hybrid methods that combine the best of both worlds.

4.1 Semi-Supervised FSL

Fig. 9 Overview of Semi-Supervised FSL, where each task's support set includes both labeled and unlabeled samples

The FSL problem setting presented in Section 2.1 assumes that all samples in each task's support set are labeled. Specifically, for any task T in the *meta-train set* and *meta-test set* the standard FSL setting assumes that a task's T support set D_T^S consists of labeled samples, i.e., $D_T^S = \{(x_i, y_i)\}_{i=0}^{N_{T_L}}$, where N_{T_L} is the total number of labeled samples in the support set. In contrast, semi-supervised FSL addresses a more generalized setting that is more eminent under real-world conditions, where the support set includes both labeled samples $D_l = \{(x_{l,i}, y_{l,i})\}_{i=1}^{N_l}$ and unlabeled samples $D_u = \{x_{u,i}\}_{i=1}^{N_u}$. In this context, the support set can be expressed as $D_T^S = D_l \cup D_u$. It is also worth noting that N_l, which is the number of labeled samples in the support set, is typically small, while N_u, which is the number of unlabeled samples in the
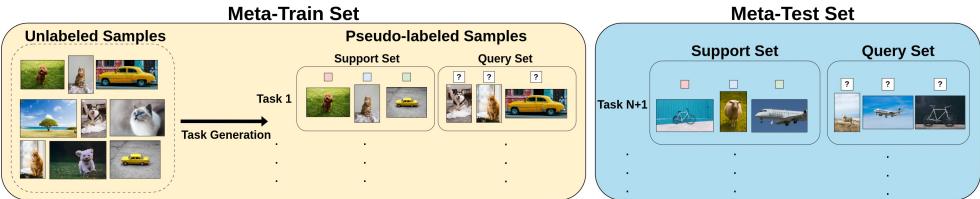
support set, is significantly larger. Figure 9 provides an overview of the above problem formulation. Overall, the objective of semi-supervised FSL is to improve performance for downstream FSL tasks by leveraging both the labeled and unlabeled data of each task.

One of the first works to incorporate unlabeled data within each task's support set was in Ren et al (2018), where the authors propose extending prototypical networks for semi-supervised FSL. Specifically, soft k-means clustering is used, where the initial clusters are the prototypes derived by using the labeled samples, and the refined prototypes are obtained by incorporating the clustered unlabeled samples. The framework is also extended to include unlabeled samples belonging to distractor classes by introducing an additional cluster to deal with distractor samples and proposing a soft masking mechanism in the case of multiple distractor classes.

In general, similar to supervised learning, directly adapting approaches commonly used in standard semi-supervised learning formulations to the FSL setting would yield inferior results due to the limited availability of labeled training samples within each task. In that setting, a common approach to effectively incorporate the available unlabeled data in each task is by propagating labels from the available labeled support set data to the unlabeled data. Specifically, some common approaches have focused on embedding both labeled and unlabeled samples into a graph and leveraging the manifold structure of the data to assign pseudo-labels to the unlabeled samples (Lazarou et al, 2021; Liu et al, 2019), thus increasing the number of labeled samples available for adaptation within each task.

However, predicting pseudo-labels for unlabeled samples can be challenging, especially when only a limited amount of labeled data is available. To tackle this issue, the authors in Wei et al (2022b) proposed reversing the problem of pseudo-labeling prediction to predicting the negative pseudo-labels for each unlabeled sample based on the class that is less probable to belong to. The method is applied iteratively to all unlabeled samples, and the actual positive pseudo-labels are eventually retrieved by eliminating all assigned negative pseudo-labels for each sample.

4.2 Unsupervised FSL

Fig. 10 In the Unsupervised FSL setting, meta-train set tasks are generated using pseudo-labels assigned to the unlabeled samples available during meta-training

A major principle underlying all FSL approaches, regardless of their specific formulation, is their ability to leverage prior knowledge gained from different tasks or data sources to adapt to novel tasks quickly using only a small number of training samples.
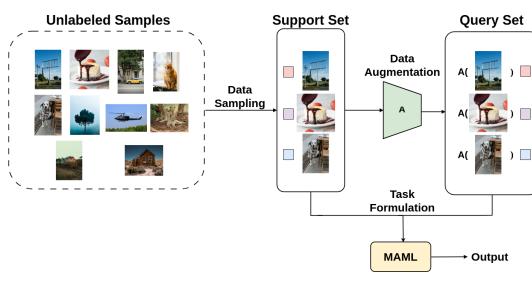
In this context, unsupervised FSL can be seen as a natural extension of supervised and semi-supervised FSL, where a large amount of data is available for meta-training, yet none of it is labeled. More specifically, in unsupervised FSL we consider that the $meta-train$ set contains only unlabeled samples, $D_{meta-train} = \{x_i\}_{i=1}^{N_{total}}$, where N_{total} is the total number of unlabeled samples. Since the classes of the tasks in the $meta-test$ set are unknown during meta-training, it is impossible to propagate pseudo-labels from the labeled samples in the $meta-test$ set to the unlabeled samples. Consequently, techniques from semi-supervised learning cannot be successfully applied in this setting, necessitating a different approach to extracting prior knowledge from the unlabeled samples. Figure 10 provides a high-level overview of the task-generating procedure during meta-training in the Unsupervised FSL problem formulation.

For this specified setting, most approaches combine unsupervised learning and meta-learning to enable rapid learning within different downstream tasks with a few labeled samples. These methods generally utilize information extracted from the unlabeled data and learn how to leverage the underlying shared structure across different tasks. A common approach to dealing with unsupervised FSL is creating synthetic tasks with unlabeled data to enable meta-learning on them. CACTUS (Hsu et al, 2019), one of the first methods in that direction, performs multiple executions of k-means in the feature representation space to obtain partitions that can be used as pseudo-classes, and uses these partitions to construct tasks used in training supervised meta-learning models. In Khodadadeh et al (2021), a generic, domain-independent approach based on generative models is proposed for task generation. These models are trained to model the unlabeled dataset's latent space, and same-class samples are created by sampling in the proximity of a point in the latent space. Similar approaches for unsupervised FSL have also been applied to tabular data (Nam et al, 2023). The proposed method, STUNT, leverages a large unlabeled dataset to create diverse tasks with pseudo-labels using k-means clustering and feature perturbations. Inference is then performed on a small labeled dataset by a Prototypical Network trained on the generated tasks.
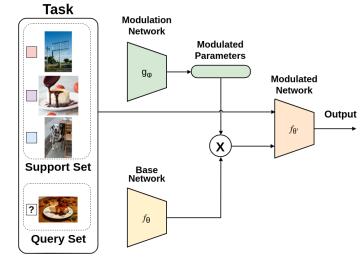
Finally, apart from task generation, other works have also focused on directly learning the multi-modality within each randomly sampled task. In Kong et al (2021), a top-down generative model that learns unknown labels through categorical latent variables is proposed, that consists of an Energy-Based Model (EBL) that couples latent representations with class labels. Interestingly, the label can be inferred from the latent representations, allowing the model to learn the structure of different classes for different tasks, even without access to class labels.

4.3 Self-Supervised FSL

Although meta-learning and self-supervised learning constitute different learning paradigms and are generally used for different problems, they share a common underlying goal: to extract knowledge from upstream tasks and apply it to learning downstream tasks more efficiently. Additionally, similar to the methods for unsupervised FSL, self-supervised learning utilizes unlabeled data during training and uses them to create pretext tasks to train the models and improve their performance in downstream tasks with limited labeled samples. Consequently, various approaches

Fig. 11 In the UMTRA architecture (Khodadadeh et al, 2019), one-shot tasks are generated by matching support set samples with their augmented counterparts in the query set

Fig. 12 In the KML architecture (Abdollahzadeh et al, 2021), the model parameters are modulated based on the concept of transference from multi-task learning

leverage that close relationship (Ni et al, 2021) between these learning paradigms to improve FSL, particularly in the unsupervised setting.

In general, many of these techniques use an unlabeled dataset to formulate few-shot tasks that are used to train a model that performs FSL in the labeled dataset. For instance, in Khodadadeh et al (2019), one-shot tasks are generated by assuming that each selected sample belongs to a different class, and the corresponding query set is created by applying basic image DA techniques. Figure 11 presents a high-level overview of the model's architecture. Additionally, in Jang et al (2023), each unlabeled sample is considered a query, and the top-k similar samples are selected to be in the support set using a contrastive loss. On the other hand, in Lee et al (2023a), task generation from unlabeled samples has been approached through set representation learning. Specifically, a contrastive loss that considers instance-level, set-level, and cross-similarity between positive pairs and instances is used during meta-training. At test time, a representation is learned for each set, which is then used as initial weights for a linear classifier that is fine-tuned on each task's support set.

Lastly, contrastive learning has been employed in the context of few-shot classification of partial views of images (Jelley et al, 2023). In that case, the partial image views are used as samples within each task, and each image constitutes a different class. Finally, a generalized version of Prototypical Networks that uses a product of experts and leverages contrastive learning is used for classification.
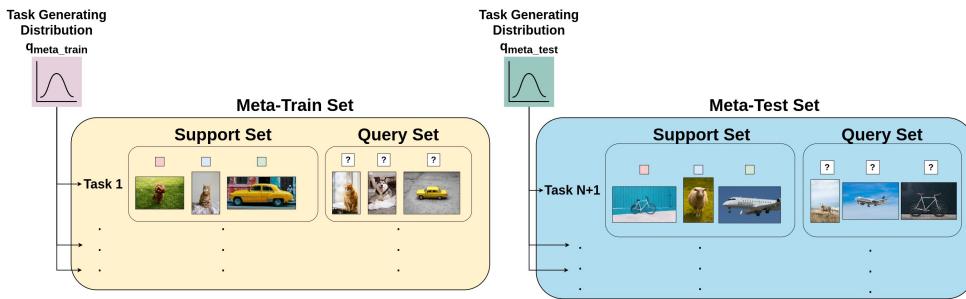
4.4 Multi-Task FSL

Another combination of approaches in FSL includes that of multi-task learning (MTL) and meta-learning. In general, meta-learning can be seen as a more generalized version of MTL, aiming to adapt to previously unknown tasks. In MTL, a specific set of tasks is defined during training, and the model is trained to generalize to these tasks simultaneously. In contrast, in standard FSL and meta-learning settings, the tasks seen during meta-training and meta-testing differ, and the number of tasks in meta-testing may also vary.

Yet, despite these differences, MTL can provide valuable elements and techniques that meta-learning can benefit from. For instance, in Abdollahzadeh et al (2021), the

idea of transference from MTL is extended to quantify the transference of knowledge from training tasks to test tasks in meta-learning. Additionally, a modulation mechanism (KML) based on hard parameter sharing that extends that of Multimodal MAML (Vuorio et al, 2019) to mitigate negative transfer between tasks is proposed. A high-level overview of KML's architecture can be seen in Figure 12.

Moreover, in Ye et al (2021), MAML's bi-level optimization formulation is extended by introducing multiple objective functions at the upper level, and the Multiple Gradient Descent Algorithm (MGDA) is used to tackle the existence of multiple objectives. Finally, in Yu et al (2022), meta-learning is cast as a multi-objective optimization problem, where each task's loss is handled as a different objective. Notably, this method can help mitigate the compromising phenomenon, where the solution focuses on minimizing the losses of only a small subset of tasks and converges faster than MAML.

4.5 Cross-Domain FSL

Fig. 13 Overview of Cross-Domain FSL, where the meta-train set tasks and meta-test set tasks are generated from different task-generating distributions

Cross-Domain FSL (CD-FSL), which combines FSL's knowledge transfer across tasks with different classes with Domain Adaptation's knowledge transfer across different data distributions, has emerged as a new learning paradigm in recent years. More specifically, CD-FSL aims to leverage knowledge extracted from the base tasks to adapt to new tasks that have both novel classes and also follow a different distribution (Wu et al, 2023c).
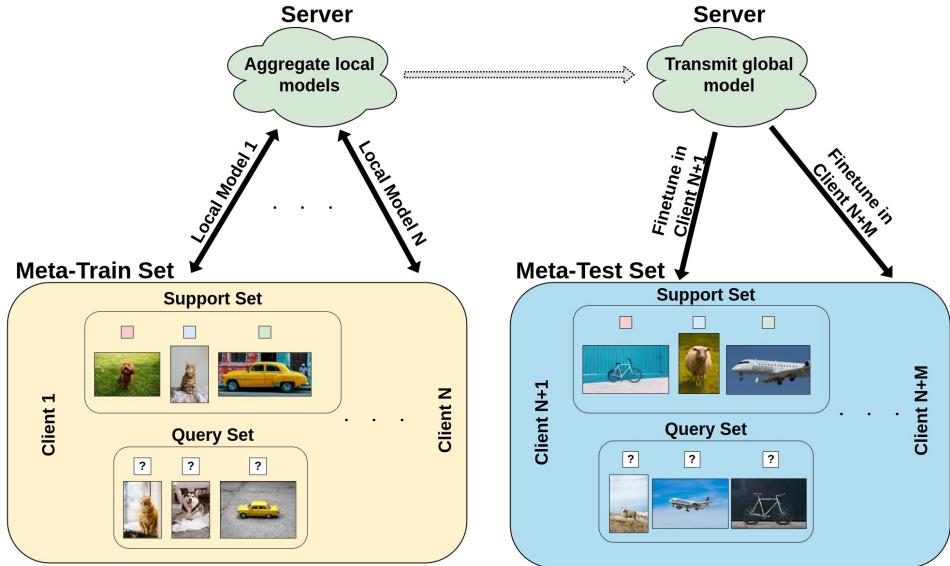
In the problem formulation of FSL in Section 2.1, an assumption was made that all tasks in both *meta-train* and *meta-test* sets are derived from the same task distribution $q(T)$. However, in real-world applications, it is possible for the tasks in the *meta-test* set to originate from a different task distribution compared to those in the *meta-train* set. Consequently, in CD-FSL, the tasks that belong to the *meta-train* set $D_{meta-train} = \{D_i\}_{i=1}^N$, where D_i is the dataset corresponding to a task T_i are drawn from a task distribution $q_{meta-train}(T)$, i.e., $T_i \sim q_{meta-train}(T)$, while the tasks that belong to the *meta-test* set $D_{meta-test} = \{D_j\}_{j=1}^M$ are drawn from a task distribution $q_{meta-test}(T)$, i.e., $T_j \sim q_{meta-test}(T)$, and $q_{meta-train} \neq q_{meta-test}$. Figure 13 illustrates more clearly how tasks in $D_{meta-train}$ and $D_{meta-test}$ originate from different

distributions. Although more challenging, this setting is particularly relevant since, in practice, it is typical for the base and target datasets of FSL to originate from different domains (Zheng et al, 2023b). For instance, a model might be trained on a large dataset of natural images, such as ImageNet, and be adapted and evaluated on a smaller dataset that is fundamentally different, such as medical or satellite images.

A common challenge in CD-FSL is that representations on the target dataset are often biased due to the data volume discrepancy, making it difficult to obtain robust models across different domains. To address this issue, recent approaches have focused on mitigating cross-domain bias during both the development and evaluation stages. At the development level, CD-FSL includes models that are explicitly trained to handle few-shot domain adaptation/generalization, with methods such as finetuning a model's projection and prediction heads (Jang et al, 2023), retraining the feature extractor using an inverted Firth Bias regularizer (Wang et al, 2022a), and using knowledge distillation (Zheng et al, 2023b). The evaluation level includes few-shot cross-domain adaptation of models, even if they were not explicitly trained for this task, which is useful in a wide variety of applications such as visual question answering, temporal point processes, neural architecture search, and drug discovery (Schimunek et al, 2023).

Overall, robustness against domain shifts can be a potential step towards achieving Broad AI (Schimunek et al, 2023). At the same time, cross-domain evaluation in meta-learning could help measure overfitting at the meta-level by attributing generalization to learning or memorization at the task level.

4.6 Few-Shot Federated Learning

Fig. 14 Overview of Few-Shot Federated Learning, where the standard FSL framework is extended to consider each client as a separate task
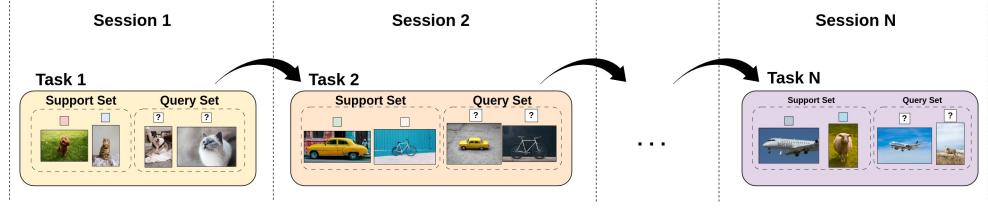
In recent years, decentralized types of learning like Federated Learning (Ayeelyan et al, 2025) (FL) have been extensively researched to leverage scattered data across multiple machines while preserving privacy. In the standard FL setup, there are multiple clients, each having its own private dataset, denoted as D_i for the i-th client. Under this setup, the objective is to learn a global model that performs well across all clients' datasets. To achieve this, various approaches have been developed that aim to aggregate each client's model weights to create a global model that generalizes to all clients' data simultaneously. This way, FL reduces communication costs and enhances data privacy by sharing only model parameters between servers and clients.

Given that there is sufficient data within each client's dataset, it has been demonstrated that these methods can yield strong results. However, FL's performance significantly declines in scenarios where data is limited for each client. To this end, Few-Shot Federated Learning (FS-FL) has been explored as a potential way to address these issues arising in federated applications with limited data. In FS-FL, each client is treated as a separate task with its own associated dataset. Consequently, the problem can be recast as an FSL problem that follows the same formulation as the one outlined in Section 2.1, with the additional consideration that there can be no data exchange between different tasks. Under this formulation, the FL's goal of learning a model that performs well in all clients' data is also recast as learning a model that can rapidly adapt to each client's dataset with minimal training samples. Figure 14 provides an overview of the FS-FL's problem formulation. Based on this recasting of FS-FL as an FSL problem, several approaches combining FL with meta-learning have been proposed, such as adjusting popular meta-learning algorithms for FL (Chen et al, 2018), recasting FedAvg as a meta-learning algorithm for personalized FL (Jiang et al, 2019) and separately executing meta-learning's optimization loops on the server and clients (Fallah et al, 2020).

In the last few years, a plethora of different FS-FL methods have been developed for applications such as facial expression recognition (Shome and Kar, 2021), load forecasting (Tang et al, 2023b), 3D MRI medical image classification (He et al, 2023), fault diagnosis in industrial settings (Chen et al, 2023c), and NLP (Dong et al, 2022b). A common theme across all of these applications is the sensitive nature of the involved data that necessitates privacy preservation, as well as the fact that this data is typically scarce, heterogeneous, and dispersed in different machines. Consequently, FS-FL is an ideal solution for these types of problems.

4.7 Continual FSL

Following the recent trend of bridging the gap between FSL and related fields of study, FSL has also been applied in the context of Continual Learning (De Lange et al, 2021). Continual Learning refers to the ability of a model to learn from an infinite stream of data, acquiring new knowledge along the way while also preserving previously gained information. In general, training in this setting is performed in a session-by-session manner, and one of the main challenges is overcoming catastrophic forgetting where the acquisition of new knowledge during subsequent sessions results in performance degradation on the knowledge accumulated during previous sessions. In the case of Continual FSL, each session consists of a single or a batch of different N-way-K-shot
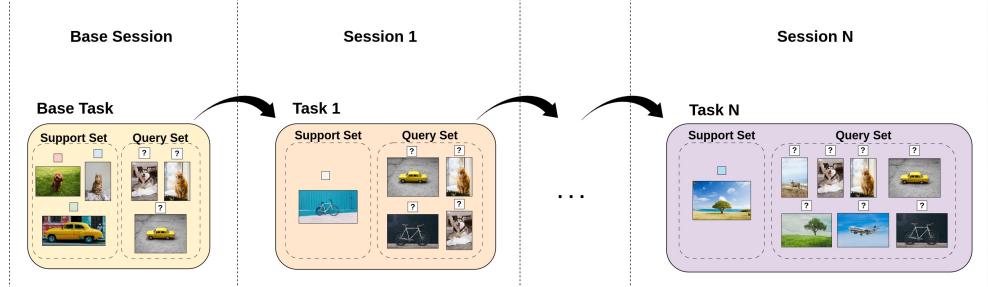
Fig. 15 In the Continual FSL framework, each task appears sequentially in a session-by-session manner

tasks with limited training samples. This scarcity of training data introduces an additional layer of difficulty since typical Continual Learning methods are susceptible to overfitting the new data. Figure 15 illustrates the session-by-session learning procedure that is typical in Continual FSL.

One of the first attempts to address Continual FSL recasts meta-learning in the online learning setting by developing a new algorithm called follow the meta-leader (Finn et al, 2019) (FTML). FTML is the meta-learning version of the widely used follow-the-leader algorithm and extends MAML (Finn et al, 2017) in the online learning setting. However, catastrophic forgetting is not explicitly addressed since all previous data is stored and is available in each session. Overfitting is dealt with by leveraging meta-learning, thus obviating the need to take many gradient steps to refine the model parameters within each task. Other notable approaches in Continual FSL include introducing a mixture of across-tasks shared parameters through a Dirichlet process mixture model to create a meta-learning model capable of incrementally learning heterogeneous tasks (Jerfel et al, 2019) and using Bayesian meta-learning with different meta-knowledge components encapsulating across-tasks knowledge to enable continual learning of sequentially available heterogeneous tasks (Wu et al, 2023a).

4.8 Class-Incremental FSL

Fig. 16 Overview of Class-Incremental FSL where in each session additional classes are added, with only limited support set samples available, and the model is tasked with predicting from all classes seen up to that point

Much of the recent research revolving around Continual FSL has mainly focused on Class Incremental FSL (Zhang et al, 2023a) (CIFSL) where in each session, a new

set of classes is introduced, and the learning model should be able to infer from all classes encountered up to that point. This type of problem differs from the Continual FSL paradigm in that the first session (base session) typically consists of a task with a sufficiently large number of training samples. In contrast, the subsequent tasks in the incremental sessions are structured as N-way-K-shot tasks with only a limited number of samples. Figure 16 illustrates the incremental nature of the Class-Incremental FSL's problem formulation.

In Song et al (2023b), a contrastive supervised loss is used to ensure adequate separation of the base classes and facilitate generalization to novel classes during the incremental sessions. Another common approach in class incremental learning is to leverage knowledge distillation to adapt to new sessions. However, standard knowledge distillation methods fail in the FSL setting as they tend to overfit novel classes with only a few samples. To mitigate this issue, a combination of two teachers has been proposed to generate the student network: one derived from the base session capturing base knowledge and one from the previous session capturing newly acquired knowledge (Zhao et al, 2023). Finally, various recently introduced methods for CIFSL draw inspiration from techniques and concepts in other areas of machine learning, such as recursive analytical learning (Zhuang et al, 2023), neural collapse (Yang et al, 2023), and the Lottery Ticket Hypothesis (Kang et al, 2023a).

5 Few-Shot Learning Applications

Even though academic research has traditionally focused on increasingly large-scale curated datasets for deep learning, collecting, storing, and processing such large volumes of data typically requires a large storing capacity and computational power, which is rarely available in real-world applications. At the same time, considerations such as high data collection costs, data privacy issues, and small datasets scattered across different edge devices have rendered FSL a suitable candidate for a plethora of real-world applications in various fields. In the rest of this section, we examine some of these applications, focusing on recent advancements and innovative approaches.

5.1 Computer Vision

Traditionally, most of the research pertinent to FSL has focused on computer vision problems and specifically image classification (Snell et al, 2017; Finn et al, 2017), primarily due to the existence of appropriate FSL datasets for training and evaluation. However, in recent years, novel FSL approaches have been developed aiming to target different compute vision problems, including image segmentation (Gkanatsios et al, 2023), object detection (Demirel et al, 2023) and 3D reconstruction (Kwak et al, 2023). Finally, the usage of FSL has also been extended to video inputs (Yan et al, 2023).

5.2 Reinforcement Learning

One of the first applications for many of the aforementioned meta-learning techniques has been meta-RL, which involves learning to quickly adapt across various similar environments by finding an optimal policy across tasks, given only a few episodes

in each one of them. This is a particularly useful skill for agents acting in dynamic environments, largely resembling human-like learning, and has been used in the context of imitation learning (Rakelly et al, 2019), visual navigation (Ni et al, 2023), multi-armed bandits (Sung et al, 2017), robot arm manipulation (Wang et al, 2023b) and control (Ni et al, 2023). Novel applications of meta-RL also include online meta-RL (Khattar et al, 2022), black-box optimization (Ma et al, 2024), long-term memory from visual observations (Beck et al, 2024) and implicit language learning (Liu et al, 2023).

5.3 Natural Language Processing

Another field of research where FSL has drawn increasing attention in the last few years is that of NLP, with some standard applications including machine translation (Garcia et al, 2023), and text (Han et al, 2023b) and sentiment (Yu et al, 2018) classification. More recently, FSL has also been extended to novel applications and tasks such as prompt tuning (Jiang et al, 2023b), visual question answering (Najdenkoska et al, 2023), and dense retrieval (Dai et al, 2022). Finally, based on LLMs' ability to handle multiple downstream tasks when given only a few examples in the form of prompts (see Section 3.4), various similar FSL techniques have also been proposed for applications such as natural language inference including multiple downstream tasks like causal reasoning, textual entailment, word sense disambiguation, and question answering (Murty et al, 2021).

5.4 Audio

In recent years, FSL has been successfully applied in the field of acoustic signal processing, with most applications focusing on audio classification (Wang and Anderson, 2022) and its extensions in the continual (Wang et al, 2021) and class-incremental (Li et al, 2023b) learning settings. Additionally, FSL has also been utilized for sound event detection (Wang et al, 2020a), while a promising application is that of speaker (Li et al, 2023b) and sound recognition (Chou et al, 2019), which can be effectively used in mobile devices and can benefit people with hearing difficulties.

5.5 Generative AI

One of the most prominent and broadly expanding fields where FSL can be extremely useful is that of Generative AI. The ability to generate novel content given only a few reference samples has been extensively explored in the context of computer vision with recent applications, including image generation (Zheng et al, 2023a), style transfer (Tang et al, 2023a), and semantic image synthesis (Careil et al, 2023). FSL has also been utilized for generative purposes in different modalities such as time series (Jiang et al, 2022), video (Li et al, 2023a), speech (Kumar et al, 2020), and music (Liang et al, 2020) generation. Few-shot generation across different modalities has also gained much popularity recently via applications such as text-to-image (Zhou et al, 2022) and text-to-speech (Huang et al, 2022) generation.

5.6 Automated Machine Learning

Traditionally, Automated Machine Learning (AutoML) has been one of the most used testbeds for novel meta-learning techniques. AutoML aims to automatically find the best model architecture and hyperparameters for a given problem and dataset (Baratchi et al, 2024). This reduces the need for domain-specific knowledge and extensive hyperparameter tuning, making machine learning more accessible and accelerating deployment. As for FSL in the context of AutoML, applications have mainly focused on Neural Architecture Search (Lee et al, 2023b) (NAS), model selection (Park et al, 2022), and hyperparameter optimization (Maraval et al, 2024).

5.7 Healthcare

A natural field of application for FSL is when data collection can be difficult, costly, or even impossible. Consequently, healthcare is a perfect application field for FSL since most patient data is subject to privacy restrictions, yet high performance needs to be maintained under these critical settings. Recently, FSL has been successfully applied in medical image classification (Sun et al, 2023), medical image segmentation (Huang et al, 2023), and 3D medical image segmentation in a federated setting (He et al, 2023). FSL has also been critical in cases where more data acquisition is impossible, such as in rare disease classification (Sun et al, 2023), and drug discovery (Chen et al, 2022). Finally, some novel approaches have focused on developing models that can handle multiple downstream medical tasks such as operational outcomes, anticipating lab test values, assignment of new diagnoses, and anticipating chest X-ray findings (Wornow et al, 2023).

5.8 Recommendation

A field that has particularly benefited from the use of FSL techniques is that of recommendation systems, by helping overcome major challenges, such as the cold-start problem, and improve model performance (Yu et al, 2021). Additionally, personalization is another area that has greatly benefited from FSL (Shysheya et al, 2023). Specifically, by treating each user's data as a separate task, FSL can efficiently learn across all of them, even with limited data. Consequently, FSL has successfully been applied for personalized video retrieval (Yeh et al, 2023), personalized recommendation (Wu et al, 2022a), and personalization in federated learning settings (Zhao et al, 2022).

6 Recent Trends and Future Directions in Few-Shot Learning

The rapid growth of FSL methodologies and applications has established it as a distinct area within the machine learning community. However, FSL is not an isolated field; various methodologies developed within FSL overlap with different areas of machine learning research. Consequently, FSL has been influenced by advancements and new challenges from both its own field and related research areas. For instance,

the emergence of large foundation models has led to their application and evaluation in settings with limited available data. Additionally, more general research directions, such as developing models that exhibit traits of human-like intelligence and advancing AI to be more environmentally friendly, have also penetrated the FSL field in recent years. In this section, we provide a discussion of various trends that have been rising in recent years in the FSL field, as well as the existing and novel challenges that may shape future research directions in FSL.

6.1 Few-Shot Evaluation of Foundation Models

In recent years, the emergence of LLMs and foundation models, primarily in the visual and language domain, has created new opportunities for developing large-scale models capable of handling multiple downstream tasks with little or no adaptation. ICL of LLMs has been critical in that direction, and enabling similar capabilities on vision models via learning more powerful representations is the next logical step.

Based on these observations and the recent bloom of foundation models, one recent trend has been the evaluation and adaptation of these models regarding their FSL capabilities (Liu et al, 2024a). Few-shot and even zero-shot performance is increasingly assessed both in the language (Brown et al, 2020) and visual domain (Dosovitskiy et al, 2020), as well as in multimodal settings (Radford et al, 2021). Additionally, variants of foundation models not explicitly developed for FSL (Kwak et al, 2023) and combinations of foundation models aiming to enhance few-shot performance (Zhang et al, 2023b) have also emerged.

Overall, testing the FSL capabilities of foundation models is a crucial step toward creating general models that can handle multiple tasks with minimal overhead. These foundation models should adapt quickly to novel tasks, alleviate overfitting, and work well in areas with limited data, such as in healthcare, leading to the democratization of AI and the mitigation of environmental damage due to extensive training.

6.2 Meta-Learning in novel settings

Traditionally, most FSL applications have revolved around problems related to computer vision and RL. While FSL has also been applied to different domains, such as NLP, as seen in Section 5, and under different problem formulations, such as unsupervised learning, as seen in Section 4, these constitute well-studied areas within the machine learning community. However, there has been a trend in applying meta-learning in novel and less-explored setups in recent years. It is worth noting that while data efficiency is a central consideration in FSL, time efficiency in terms of fast adaptation to the existing data within each task is critical in many of these settings, elucidating why meta-learning is broadly adopted in these novel scenarios.

Meta-learning has been applied in various games such as zero-sum games, potential games, general sum multi-player games, and Stackelberg security games, showing significant improvements (Harris et al, 2022). Meta-learning an optimal initialization has also been applied in combinatorial optimization (Wang and Li, 2023) by recasting unsupervised learning for combinatorial optimization as a meta-learning problem, as

well as few-shot knowledge graph completion (Wu et al, 2023b). Additionally, meta-learning has been extended to learning cognitive tasks (Miconi, 2023) by introducing another evolutionary level of learning that enables modification of the network in a self-referential manner. Finally, meta-learning has been used to discover optimizing functions without access to gradients efficiently (Lange et al, 2023) and to solve optimal transport problems (Amos et al, 2023) rapidly.

6.3 Green AI via FSL

A central characteristic of modern LLMs and foundation models is their scaling in the regime of billions of parameters. While this has led to breakthrough performance results, it has also led to a commensurate increase in their energy consumption and carbon emissions. The environmental damage associated with training and operating such models has been highlighted in 2023's AI index report (Maslej et al, 2023), showing that their carbon emissions exceed those of an average person's lifetime. Other studies (Patterson et al, 2021) have also focused on measuring the carbon footprint of LLMs, yet more needs to be done to mitigate their impact.

A first step toward this direction has been the development of novel Green AI metrics that quantify these effects. For instance, in Wu et al (2022b), the carbon footprint of AI models is estimated as:

$$CO_2^{embodied} = \sum_i \frac{Time}{Lifetime} CO_2^{embodied}(AI_{system})(i) \tag{5}$$

where $Time$ refers to the time the model is utilized over the specified $Lifetime$. Other tools that enable measurement of model training environmental footprint have also been proposed. On the other hand, in Henderson et al (2020), a formula for calculating the total energy consumption is derived:

$$e_{total} = PUE \sum_p (p_{dram}e_{dram} + p_{cpu}e_{cpu} + p_{gpu}e_{gpu}) \tag{6}$$

where p_x refers to the percentage of each system resource used by process x relative to the total resources used, e_x is the energy usage of that resource, and PUE is the Power Usage Effectiveness index of the data center. However, a unified benchmark and suitable metrics to calculate an AI model's overall carbon footprint are still missing.

Approaches such as FL have been explored in the context of Green AI, yet they face limitations due to hardware constraints, scarcity of renewable energy sources for edge devices, and the need for communication among devices. However, when combined with FSL, they can reduce energy consumption on each device up to 41.2% (Cai et al, 2023). Overall, FSL can alleviate the need for extensive model selection and hyperparameter tuning and reduce the volume of available data needed to achieve strong generalization performance. However, more research is required to establish FSL as a feasible solution for Green AI.

6.4 Meta-Learning vs Transfer Learning debate

While meta-learning has been the leading paradigm for FSL for many years, various recent results have cast doubt on its dominance. Specifically, feature reuse has been

identified as the main factor for fast adaptation in MAML (Raghu et al, 2020) while recent findings show that finetuning with modern pretrained backbones can also form strong baselines (El Baz et al, 2021). Consequently, transfer learning has emerged as a capable competitor of meta-learning, with numerous simple baselines based on finetuning being proposed and demonstrating comparable performance against complex meta-learning architectures.

A shared insight in this line of work is that learning strong, transferable representations is crucial for FSL (Wang et al, 2022a), and the quality of these representations can be affected by various factors such as model size (Tian et al, 2020), invariance and equivariance (Rizve et al, 2021), support set size (Luo et al, 2023), and number of pretraining classes (Dhillon et al, 2020). Overall, further research is required to understand more clearly the importance of expressive and generalizable representations in FSL. An interesting direction would also be to examine FSL's interplay with other learning paradigms with similar objectives, such as self-supervised learning (Ni et al, 2021) and incorporate techniques from these fields to obtain more generalizable representations (Tian et al, 2020).

Moreover, the lack of a unified evaluation protocol has constituted a major issue in comparing meta-learning and finetuning models. It is argued that progress using meta-learning techniques might be illusory, with incongruities in the implementation details of various FSL methods (Dhillon et al, 2020), and lack of evaluation under domain shift (Chen et al, 2019), identified as two main challenges obstructing meaningful comparative assessments. To mitigate these issues, a hardness metric has been proposed (Dhillon et al, 2020), as well as the use of novel datasets (Dumoulin et al, 2021). However, more research is needed to determine more datasets and a unified evaluation protocol for FSL methods.

6.5 Impact of underlying distributions

One of the main differences between FSL and standard machine learning problems is the limited number of support data within each task, which complicates the precise estimation of the underlying data distribution. While meta-learning approaches compensate for it by extracting shared knowledge across different tasks, it is still unclear how to effectively achieve accurate estimates of the underlying task-generating process. Since FSL typically relies on utilizing a set of tasks, this setting introduces another distribution at the meta-level, specifically the distribution of tasks. Consequently, similar to how data augmentation has been extensively utilized in standard machine learning to obtain a better evaluation of the underlying data-generating distribution, similar techniques can also be utilized at the meta-level to increase across-task generalization by learning a more accurate representation of the underlying task-generating distribution. Towards that direction, in recent years, there has been substantial progress in methods that employ interpolation and augmentation techniques at the meta-level.

In the case of task augmentation, various techniques, including Conditional Batch Normalization (Sun et al, 2023), mixing support and query sets with Manifold Mixup to create new query sets (Yao et al, 2021a), and adversarial training (Wang and Deng, 2021) have been employed to generate new tasks. On the other hand, task augmentation techniques typically include leveraging information from two existing

tasks to create a new one. Some of the most recent approaches include using memory banks (Hu et al, 2023), interval bounds (Datta et al, 2023), Manifold Mixup (Yao et al, 2021b), and expressive neural set functions (Lee et al, 2022) to encode and interpolate between tasks.

Yet, despite these practical implementations, we still lack a more enhanced comprehension of the influence of the underlying data distributions on FSL at a theoretical level. Specifically, only a single work (Zuo et al, 2022) that focuses on the theoretical examination of FSL emphasizes the importance of careful modeling of both the task distribution and the data distribution within each task. Consequently, further research is needed to understand more clearly the connection between task and data distributions and how they can efficiently be leveraged in the context of FSL.

6.6 Adoption of FSL in new modalities

Traditionally, most FSL and meta-learning techniques have commonly been applied in the visual domain (Snell et al, 2017) and for RL (Wang et al, 2016). However, since FSL aims to alleviate the adaptation problem in cases where data is scarce, its application could be fruitful in various other domains and data modalities.

A common type of data underexplored in the FSL setting is tabular data. Only recently, some novel approaches have emerged, including using Conditional Batch Normalization to create novel tasks with tabular data (Sun et al, 2023) and applying prototypical networks on tabular data tasks created via unsupervised meta-learning (Nam et al, 2023). On a broader setting focusing on small tabular tasks, in Hollmann et al (2023), the authors have pretrained a transformer-based model on synthetically generated tasks using Bayesian Neural Networks and Structural Causal Models and showed that the model can adapt to downstream tabular tasks without any additional finetuning, similar to the way LLMs perform ICL. Time series data have also received insufficient attention until recently. Some notable first attempts in applying FSL in that modality include a meta-learning approach for few-shot forecasting in high-dimensional time series based on a single latent dynamic function conditioned in each task (Jiang et al, 2022), load forecasting via MAML-based models (Xu et al, 2024), and meta-learning neural processes with application on point processes (Bae et al, 2023).

Overall, one of the main reasons hindering FSL progress outside the visual domain and RL is the limited number of FSL datasets for these domains and modalities. Although some novel FSL datasets, such as FS-MOL (Stanley et al, 2021) for few-shot drug discovery, have emerged, the lack of similar datasets in modalities such as time series necessitates developing and establishing standardized FSL benchmarks in these domains.

6.7 Fully generalizable FSL

The increasing popularity of foundation models on vision and language tasks has also given rise to various multimodal foundation models (Radford et al, 2021). While many of these models demonstrate strong FSL performance or have been adapted to be effective in that setting, their shortcomings in achieving holistic and human-like

generalization capabilities are still notable. Broadly speaking, generalization can be multi-faceted and may be abstracted in different levels based on context:

1. **Single instance of a learning problem.** Better known as domain adaptation/generalization, the model's goal is to generalize strongly to data from a previously unseen distribution for a specific problem.
2. **Multiple instances of the same learning problem.** Generalization at that level could be seen as being able to adapt to different variations of the specific problem, e.g., multi-class classification with a varied number of classes and labels.
3. **Multiple learning problems of the same modality.** The goal at this level is to effectively generalize to different learning problems of the same modality, e.g., from image classification to object detection.
4. **Multiple domains.** At the highest level of abstraction, a model would be able to generalize across different modalities even if the model has never been trained explicitly on them.

While the first two cases have been extensively researched for FSL, only recently has generalization within a specific modality been explored via models that can generalize across different visual prediction tasks (Kim et al, 2023; Bachmann et al, 2024), and the curation of FSL datasets containing multiple visual learning problems (Bohdal et al, 2023). However, current vision learning models struggle to generalize to new learning problems, and overall, FSL methods still face challenges in extrapolating to unseen modalities. Creating a genuine generalist model capable of FSL across novel modalities could be an ambitious but exciting future direction.

6.8 Human-like learning

One of the major attributes of human intelligence is the ability to learn new concepts quickly, master new tasks, and combine them to solve complex problems, given only a few examples, possibly due to humans' inherent inductive biases and their ability to leverage multimodal signals (Lin et al, 2023). However, current machine learning models struggle with fast adaptation to complex tasks and lack human-like intelligence attributes such as compositionality, concept association, relational understanding, and conceptual meaning (Jiang et al, 2023a). Systematicity also remains a challenge for modern models (Lake and Baroni, 2023).

Rapid adaptation is generally prevalent in nature and has been studied for many decades in cognitive science, psychology, and neuroscience, with results highlighting the existence of different and nested timescales of learning. These observations have also been utilized in meta-learning via hierarchical Bayesian models (Lake et al, 2015). The compositionality and expressiveness of natural language have also been leveraged for FSL of concepts, focusing on induction and utilizing a human-like prior extracted from behavioral data from humans (Ellis, 2024). Moreover, meta-learning has been used for human-like learning of algebraic systems, leading to models that learn how to generalize and adopt human inductive biases (Lake and Baroni, 2023) systematically. On a different direction, brain mechanisms such as synapse formation in the human brain have inspired novel approaches in FSL, such as adaptation mechanisms for few-shot generative domain adaptation, enhancing the model with human-like properties

such as memory and domain association ([Wu et al, 2023c](#)). Lastly, it has been shown that increased AI alignment can improve FSL models, adaptation to new domains, and robustness to adversarial examples, possibly by serving as a source of inductive bias that reduces the amount of information needed from data ([Sucholutsky and Griffiths, 2024](#)).

It is worth noting that while cognitive science and neuroscience have focused on discovering the mechanisms of human intelligence, machine learning focuses on replicating these capabilities. In that direction, a new multimodal dataset for FSL has been proposed that enables few-shot multimodal abstract reasoning evaluation of models over various cognitive tasks ([Jiang et al, 2023a](#)). Preliminary results on the dataset suggest significant misalignment between machines and humans regarding FSL capabilities, raising questions on how to align machine and human learning and create agents with human-like FSL capabilities. Consequently, methods combining neuroscience and cognitive science with machine learning might be the way forward.

7 Conclusion

Although deep learning has traditionally relied on large amounts of data to facilitate learning, there is an increasing focus on techniques that can learn with limited data. Similarly, despite the trend of creating large foundation models trained on massive datasets, there is a growing interest in exploring new FSL techniques and applying them in practical settings. Given FSL's recent rapid expansion, this paper aims to provide an overview of this research field, covering its various aspects. The standard FSL formulation is initially provided, based on which different proposed methods are developed. To position FSL within the broader machine learning paradigm, its connection with different learning fields is also examined. Subsequently, we present a comprehensive taxonomy of FSL methods that extends existing taxonomies with recent approaches and model families and also introduces emerging learning paradigms within FSL, such as in-context learning. Additionally, we provide a broad overview of hybrid approaches that extend FSL beyond the presented supervised learning setting. FSL's bloom, which is also reflected in the numerous real-world applications in which it has been applied, is also described. Finally, we discuss various emerging trends, including the few-shot evaluation of foundation models and the application of meta-learning in novel settings, and aim to encourage further research by identifying key areas where FSL can have a significant impact, such as Green AI and bridging the gap between machine learning and human intelligence.

References

Abbas M, Xiao Q, Chen L, et al (2022) Sharp-maml: Sharpness-aware model-agnostic meta learning. In: International conference on machine learning, PMLR, pp 10–32

Abdollahzadeh M, Malekzadeh T, Cheung NMM (2021) Revisit multimodal meta-learning through the lens of multi-task learning. Advances in Neural Information Processing Systems 34:14632–14644

Akyürek E, Schuurmans D, Andreas J, et al (2023) What learning algorithm is in-context learning? investigations with linear models. In: The Eleventh International Conference on Learning Representations

Altae-Tran H, Ramsundar B, Pappu AS, et al (2017) Low data drug discovery with one-shot learning. ACS central science 3(4):283–293

Amos B, Luise G, Cohen S, et al (2023) Meta optimal transport. In: International Conference on Machine Learning, PMLR, pp 791–813

An S, Lin Z, Fu Q, et al (2023) How do in-context examples affect compositional generalization? In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 11027–11052

Andrychowicz M, Denil M, Gomez S, et al (2016) Learning to learn by gradient descent by gradient descent. Advances in neural information processing systems 29

Antoniou A, Edwards H, Storkey A (2019) How to train your maml. In: Seventh International Conference on Learning Representations

Ayeelyan J, Utomo S, Rouniyar A, et al (2025) Federated learning design and functional models: survey. Artificial Intelligence Review 58(1):1–38

Bachmann R, Kar OF, Mizrahi D, et al (2024) 4m-21: An any-to-any vision model for tens of tasks and modalities. arXiv preprint arXiv:240609406

Bae W, Ahmed MO, Tung F, et al (2023) Meta temporal point processes. In: The Eleventh International Conference on Learning Representations

Baratchi M, Wang C, Limmer S, et al (2024) Automated machine learning: past, present and future. Artificial Intelligence Review 57(5):1–88

Basu S, Stanley M, Bronskill JF, et al (2022) Hard-meta-dataset++: Towards understanding few-shot performance on difficult tasks. In: The Eleventh International Conference on Learning Representations

Beck J, Vuorio R, Xiong Z, et al (2024) Recurrent hypernetworks are surprisingly strong in meta-rl. Advances in Neural Information Processing Systems 36

Behl H, Baydin A, Torr P (2019) Alpha maml: Adaptive model-agnostic meta-learning. In: 6th ICML Workshop on Automated Machine Learning, Thirty-Sixth International Conference on Machine Learning (ICML)

Bengio Y, Bengio S, Cloutier J (1991) Learning a synaptic learning rule. In: IJCNN-91-Seattle International Joint Conference on Neural Networks, IEEE, pp 969–vol

Bertinetto L, Henriques J, Torr P, et al (2019) Meta-learning with differentiable closed-form solvers. In: International Conference on Learning Representations (ICLR), 2019, International Conference on Learning Representations

Bohdal O, Tian Y, Zong Y, et al (2023) Meta omnium: A benchmark for general-purpose learning-to-learn. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 7693–7703

Brown T, Mann B, Ryder N, et al (2020) Language models are few-shot learners. Advances in neural information processing systems 33:1877–1901

Cai D, Wang S, Wu Y, et al (2023) Federated few-shot learning for mobile nlp. In: Proceedings of the 29th Annual International Conference on Mobile Computing and Networking, pp 1–17

Careil M, Verbeek J, Lathuilière S (2023) Few-shot semantic image synthesis with class affinity transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 23611–23620

Chan S, Santoro A, Lampinen A, et al (2022) Data distributional properties drive emergent in-context learning in transformers. Advances in Neural Information Processing Systems 35:18878–18891

Chen F, Luo M, Dong Z, et al (2018) Federated meta-learning with fast convergence and efficient communication. arXiv preprint arXiv:180207876

Chen H, Hou L, Zhang GK, et al (2023a) Using context-guided data augmentation, lightweight cnn, and proximity detection techniques to improve site safety monitoring under occlusion conditions. Safety science 158:105958

Chen H, Li H, Li Y, et al (2023b) Sparse spatial transformers for few-shot learning. Science China Information Sciences 66(11):210102

Chen J, Tang J, Li W (2023c) Industrial edge intelligence: Federated-meta learning framework for few-shot fault diagnosis. IEEE Transactions on Network Science and Engineering

Chen W, Tripp A, Hernández-Lobato JM (2022) Meta-learning adaptive deep kernel gaussian processes for molecular property prediction. In: The Eleventh International Conference on Learning Representations

Chen WY, Liu YC, Kira Z, et al (2019) A closer look at few-shot classification. In: International Conference on Learning Representations

Chijiwa D, Yamaguchi S, Kumagai A, et al (2022) Meta-ticket: Finding optimal subnetworks for few-shot learning within randomly initialized neural networks. Advances in Neural Information Processing Systems 35:25264–25277

Chou SY, Cheng KH, Jang JSR, et al (2019) Learning to match transient sound events using attentional similarity for few-shot sound recognition. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp 26–30

Coda-Forno J, Binz M, Akata Z, et al (2023) Meta-in-context learning in large language models. Advances in Neural Information Processing Systems 36:65189–65201

Cubuk ED, Zoph B, Mane D, et al (2019) Autoaugment: Learning augmentation strategies from data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 113–123

Cubuk ED, Zoph B, Shlens J, et al (2020) Randaugment: Practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp 702–703

Dai D, Sun Y, Dong L, et al (2023) Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers. In: ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models

Dai Z, Zhao VY, Ma J, et al (2022) Promptagator: Few-shot dense retrieval from 8 examples. In: The Eleventh International Conference on Learning Representations

Datta S, Mullick SS, Chakrabarty A, et al (2023) Interval bound interpolation for few-shot learning with few tasks. In: International Conference on Machine Learning, PMLR, pp 7141–7166

De Lange M, Aljundi R, Masana M, et al (2021) A continual learning survey: Defying forgetting in classification tasks. IEEE transactions on pattern analysis and machine intelligence 44(7):3366–3385

Demirel B, Baran OB, Cinbis RG (2023) Meta-tuning loss functions and data augmentation for few-shot object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 7339–7349

Deutch G, Magar N, Natan T, et al (2024) In-context learning and gradient descent revisited. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp 1017–1028

Dhillon GS, Chaudhari P, Ravichandran A, et al (2020) A baseline for few-shot image classification. In: International Conference on Learning Representations

Dong Q, Li L, Dai D, et al (2022a) A survey for in-context learning. arXiv preprint arXiv:230100234

Dong W, Wu X, Li J, et al (2022b) Fewfedweight: Few-shot federated learning framework across multiple nlp tasks. arXiv preprint arXiv:221208354

Dosovitskiy A, Beyer L, Kolesnikov A, et al (2020) An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations

Duan Y, Schulman J, Chen X, et al (2016) Rl^2: Fast reinforcement learning via slow reinforcement learning. arXiv preprint arXiv:161102779

Dumoulin V, Houlsby N, Evci U, et al (2021) A unified few-shot classification benchmark to compare transfer and meta learning approaches. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)

Edelman E, Tsilivis N, Edelman BL, et al (2024) The evolution of statistical induction heads: In-context learning markov chains. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems

Edwards H, Storkey A (2017) Towards a neural statistician. In: 5th International Conference on Learning Representations, pp 1–13

El Baz A, Ullah I, Alcobaça E, et al (2021) Lessons learned from the neurips 2021 metadl challenge: Backbone fine-tuning without episodic meta-learning dominates for few-shot learning image classification. In: NeurIPS 2021 Competition and Demonstration Track

Ellis K (2024) Human-like few-shot learning via bayesian reasoning over natural language. Advances in Neural Information Processing Systems 36

Fallah A, Mokhtari A, Ozdaglar A (2020) Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. In: Advances in Neural Information Processing Systems, pp 3557–3568

Fei-Fei L, Fergus R, Perona P (2006) One-shot learning of object categories. IEEE transactions on pattern analysis and machine intelligence 28(4):594–611

Feng CM, Yu K, Liu Y, et al (2023) Diverse data augmentation with diffusions for effective test-time prompt tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 2704–2714

Fink M (2004) Object classification from a single example utilizing class relevance metrics. Advances in neural information processing systems 17

Finn C, Abbeel P, Levine S (2017) Model-agnostic meta-learning for fast adaptation of deep networks. In: International conference on machine learning, PMLR, pp 1126–1135

Finn C, Xu K, Levine S (2018) Probabilistic model-agnostic meta-learning. Advances in neural information processing systems 31

Finn C, Rajeswaran A, Kakade S, et al (2019) Online meta-learning. In: International Conference on Machine Learning, PMLR, pp 1920–1930

Flennerhag S, Rusu A, Pascanu R, et al (2020) Meta-learning with warped gradient descent. In: International Conference on Learning Representations 2020

Foster T, Croitoru I, Dorfman R, et al (2023) Flexible visual prompts for in context learning in computer vision. In: R0-FoMo:Robustness of Few-shot and Zero-shot Learning in Large Foundation Models

Frankle J, Carbin M (2018) The lottery ticket hypothesis: Finding sparse, trainable neural networks. In: International Conference on Learning Representations

Gao N, Ziesche H, Vien NA, et al (2022) What matters for meta-learning vision regression tasks? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 14776–14786

Garcia V, Bruna J (2018) Few-shot learning with graph neural networks. In: 6th International Conference on Learning Representations, ICLR 2018

Garcia X, Bansal Y, Cherry C, et al (2023) The unreasonable effectiveness of few-shot learning for machine translation. In: International Conference on Machine Learning, PMLR, pp 10867–10878

Garg S, Tsipras D, Liang PS, et al (2022) What can transformers learn in-context? a case study of simple function classes. Advances in Neural Information Processing Systems 35:30583–30598

Garnelo M, Rosenbaum D, Maddison C, et al (2018a) Conditional neural processes. In: International conference on machine learning, PMLR, pp 1704–1713

Garnelo M, Schwarz J, Rosenbaum D, et al (2018b) Neural processes. arXiv preprint arXiv:180701622

Gharoun H, Momenifar F, Chen F, et al (2024) Meta-learning approaches for few-shot learning: A survey of recent advances. ACM Computing Surveys 56(12):1–41

Gkanatsios N, Singh M, Fang Z, et al (2023) Analogy-forming transformers for few-shot 3d parsing. In: The Eleventh International Conference on Learning Representations

Gordon J, Bruinsma WP, Foong AY, et al (2019) Convolutional conditional neural processes. In: International Conference on Learning Representations

Grant E, Finn C, Levine S, et al (2018) Recasting gradient-based meta-learning as hierarchical bayes. In: International Conference on Learning Representations

Guo T, Hu W, Mei S, et al (2024) How do transformers learn in-context beyond simple functions? a case study on learning with representations. In: The Twelfth International Conference on Learning Representations

Guo Y, Codella NC, Karlinsky L, et al (2020) A broader study of cross-domain few-shot learning. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16, Springer, pp 124–141

Han X, Simig D, Mihaylov T, et al (2023a) Understanding in-context learning via supportive pretraining data. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 12660–12673

Han Z, Hao Y, Dong L, et al (2023b) Prototypical calibration for few-shot learning of language models. In: The Eleventh International Conference on Learning Representations

Harris K, Anagnostides I, Farina G, et al (2022) Meta-learning in games. In: The Eleventh International Conference on Learning Representations

He K, Zhang X, Ren S, et al (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778

He X, Tan C, Liu B, et al (2023) Dealing with heterogeneous 3d mr knee images: A federated few-shot learning method with dual knowledge distillation. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), IEEE, pp 1–5

Henderson P, Hu J, Romoff J, et al (2020) Towards the systematic reporting of the energy and carbon footprints of machine learning. The Journal of Machine Learning Research 21(1):10039–10081

Hollmann N, Müller S, Eggensperger K, et al (2023) TabPFN: A transformer that solves small tabular classification problems in a second. In: The Eleventh International Conference on Learning Representations

Hou Y, Dong H, Wang X, et al (2022) Metaprompting: Learning to learn better prompts. In: Proceedings of the 29th International Conference on Computational Linguistics, pp 3251–3262

Hsu K, Levine S, Finn C (2019) Unsupervised learning via meta-learning. In: International Conference on Learning Representations

Hu SX, Li D, Stühmer J, et al (2022a) 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, pp 9068–9077

Hu Z, Sun Y, Yang Y (2022b) Suppressing the heterogeneity: A strong feature extractor for few-shot segmentation. In: The Eleventh International Conference on Learning Representations

Hu Z, Shen L, Wang Z, et al (2023) Learning to learn from apis: Black-box data-free meta-learning. arXiv preprint arXiv:230518413

Huang S, Xu T, Shen N, et al (2023) Rethinking few-shot medical segmentation: A vector quantization view. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 3072–3081

Huang SF, Lin CJ, Liu DR, et al (2022) Meta-tts: Meta-learning for few-shot speaker adaptive text-to-speech. IEEE/ACM Transactions on Audio, Speech, and Language Processing 30:1558–1571

Jackson PT, Abarghouei AA, Bonner S, et al (2019) Style augmentation: data augmentation via style randomization. In: CVPR workshops, pp 10–11

Jang H, Lee H, Shin J (2023) Unsupervised meta-learning via few-shot pseudo-supervised contrastive learning. In: The Eleventh International Conference on Learning Representations

Jelley A, Storkey A, Antoniou A, et al (2023) Contrastive meta-learning for partially observable few-shot learning. In: The Eleventh International Conference on Learning Representations

Jeon HJ, Lee JD, Lei Q, et al (2024) An information-theoretic analysis of in-context learning. In: Forty-first International Conference on Machine Learning

Jeong T, Kim H (2020) Ood-maml: Meta-learning for few-shot out-of-distribution detection and classification. Advances in Neural Information Processing Systems 33:3907–3916

Jerfel G, Grant E, Griffiths T, et al (2019) Reconciling meta-learning and continual learning with online mixtures of tasks. Advances in neural information processing systems 32

Jiang G, Xu M, Xin S, et al (2023a) Mewl: Few-shot multimodal word learning with referential uncertainty. In: International Conference on Machine Learning, PMLR,

pp 15144–15169

Jiang W, Zhang Y, Kwok J (2023b) Effective structured prompting by meta-learning and representative verbalizer. In: International Conference on Machine Learning, PMLR, pp 15186–15199

Jiang X, Missel R, Li Z, et al (2022) Sequential latent variable models for few-shot high-dimensional time-series forecasting. In: The Eleventh International Conference on Learning Representations

Jiang Y, Konečnỳ J, Rush K, et al (2019) Improving federated learning personalization via model agnostic meta learning. arXiv preprint arXiv:190912488

Jumper J, Evans R, Pritzel A, et al (2021) Highly accurate protein structure prediction with alphafold. Nature 596(7873):583–589

Kaiser L, Nachum O, Roy A, et al (2017) Learning to remember rare events. In: International Conference on Learning Representations

Kandpal N, Deng H, Roberts A, et al (2023) Large language models struggle to learn long-tail knowledge. In: International Conference on Machine Learning, PMLR, pp 15696–15707

Kang H, Yoon J, Madjid SRH, et al (2023a) On the soft-subnetwork for few-shot class incremental learning. In: The Eleventh International Conference on Learning Representations

Kang S, Hwang D, Eo M, et al (2023b) Meta-learning with a geometry-adaptive preconditioner. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 16080–16090

Kaplan J, McCandlish S, Henighan T, et al (2020) Scaling laws for neural language models. arXiv preprint arXiv:200108361

Khattar V, Ding Y, Sel B, et al (2022) A cmdp-within-online framework for meta-safe reinforcement learning. In: The Eleventh International Conference on Learning Representations

Khodadadeh S, Boloni L, Shah M (2019) Unsupervised meta-learning for few-shot image classification. Advances in neural information processing systems 32

Khodadadeh S, Zehtabian S, Vahidian S, et al (2021) Unsupervised meta-learning through latent-space interpolation in generative models. In: International Conference on Learning Representations

Khodaee P, Viktor HL, Michalowski W (2024) Knowledge transfer in lifelong machine learning: a systematic literature review. Artificial Intelligence Review 57(8):217

Khoee AG, Yu Y, Feldt R (2024) Domain generalization through meta-learning: A survey. Artificial Intelligence Review 57(10):285

Kim D, Kim J, Cho S, et al (2023) Universal few-shot learning of dense prediction tasks with visual token matching. In: The Eleventh International Conference on Learning Representations

Kim H, Mnih A, Schwarz J, et al (2019) Attentive neural processes. In: International Conference on Learning Representations

Koch G, Zemel R, Salakhutdinov R, et al (2015) Siamese neural networks for one-shot image recognition. In: ICML deep learning workshop, Lille

Köhler M, Eisenbach M, Gross HM (2023) Few-shot object detection: A comprehensive survey. IEEE Transactions on Neural Networks and Learning Systems

Kong D, Pang B, Wu YN (2021) Unsupervised meta-learning via latent space energy-based model of symbol vector coupling. In: Fifth Workshop on Meta-Learning at the Conference on Neural Information Processing Systems

Kossen J, Rainforth T, Gal Y (2023) In-context learning in large language models learns label relationships but is not conventional learning. arXiv preprint arXiv:230712375

Kumar N, Goel S, Narang A, et al (2020) Few shot adaptive normalization driven multi-speaker speech synthesis. arXiv preprint arXiv:201207252

Kwak MS, Song J, Kim S (2023) Geconerf: few-shot neural radiance fields via geometric consistency. In: Proceedings of the 40th International Conference on Machine Learning, pp 18023–18036

Lake BM, Baroni M (2023) Human-like systematic generalization through a meta-learning neural network. Nature pp 1–7

Lake BM, Salakhutdinov R, Tenenbaum JB (2015) Human-level concept learning through probabilistic program induction. Science 350(6266):1332–1338

Lange R, Schaul T, Chen Y, et al (2023) Discovering evolution strategies via meta-black-box optimization. In: Proceedings of the Companion Conference on Genetic and Evolutionary Computation, pp 29–30

Lazarou M, Stathaki T, Avrithis Y (2021) Iterative label cleaning for transductive and semi-supervised few-shot learning. In: Proceedings of the ieee/cvf international conference on computer vision, pp 8751–8760

Lee DB, Lee S, Kawaguchi K, et al (2023a) Self-supervised set representation learning for unsupervised meta-learning. In: The Eleventh International Conference on Learning Representations

Lee H, An S, Kim M, et al (2023b) Meta-prediction model for distillation-aware NAS on unseen datasets. In: The Eleventh International Conference on Learning Representations

Lee S, Andreis B, Kawaguchi K, et al (2022) Set-based meta-interpolation for few-task meta-learning. Advances in Neural Information Processing Systems 35:6775–6788

Lee Y, Choi S (2018) Gradient-based meta-learning with learned layerwise metric and subspace. In: International Conference on Machine Learning, PMLR, pp 2927–2936

Li K, Malik J (2017) Learning to optimize. In: International Conference on Learning Representations

Li W, Zhang L, Wang D, et al (2023a) One-shot high-fidelity talking-head synthesis with deformable neural radiance field. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 17969–17978

Li X, Wei T, Chen YP, et al (2020) Fss-1000: A 1000-class dataset for few-shot segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2869–2878

Li Y, Cao W, Xie W, et al (2023b) Few-shot class-incremental audio classification using dynamically expanded classifier with self-attention modified prototypes. IEEE Transactions on Multimedia

Li Z, Zhou F, Chen F, et al (2017) Meta-sgd: Learning to learn quickly for few-shot learning. arXiv preprint arXiv:170709835

Liang W, Liu Z, Liu C (2020) Dawson: A domain adaptive few shot generation framework. arXiv preprint arXiv:200100576

Liang W, Liang Y, Jia J (2023) Miamix: Enhancing image classification through a multi-stage augmented mixed sample data augmentation method. Processes 11(12):3284

Lin Z, Yu S, Kuang Z, et al (2023) Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 19325–19337

Liu EZ, Suri S, Mu T, et al (2023) Simple embodied language learning as a byproduct of meta-reinforcement learning. In: International Conference on Machine Learning, PMLR, pp 21997–22008

Liu F, Zhang T, Dai W, et al (2024a) Few-shot adaptation of multi-modal foundation models: A survey. Artificial Intelligence Review 57(10):268

Liu J, Shen D, Zhang Y, et al (2022) What makes good in-context examples for gpt-3? In: Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on

Knowledge Extraction and Integration for Deep Learning Architectures, pp 100–114

Liu NF, Lin K, Hewitt J, et al (2024b) Lost in the middle: How language models use long contexts. Transactions of the Association for Computational Linguistics 12:157–173

Liu Y, Lee J, Park M, et al (2019) Learning to propagate labels: Transductive propagation network for few-shot learning. In: International Conference on Learning Representations

Luo X, Wu H, Zhang J, et al (2023) A closer look at few-shot classification again. In: International Conference on Machine Learning, PMLR, pp 23103–23123

Ma Z, Guo H, Chen J, et al (2024) Metabox: a benchmark platform for meta-black-box optimization with reinforcement learning. Advances in Neural Information Processing Systems 36

Maraval A, Zimmer M, Grosnit A, et al (2024) End-to-end meta-bayesian optimisation with transformer neural processes. Advances in Neural Information Processing Systems 36

Maslej N, Fattorini L, Brynjolfsson E, et al (2023) Artificial intelligence index report 2023. arXiv preprint arXiv:231003715

Massiceti D, Zintgraf L, Bronskill J, et al (2021) Orbit: A real-world few-shot dataset for teachable object recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 10818–10828

Meng Y, Michalski M, Huang J, et al (2023) Tuning language models as training data generators for augmentation-enhanced few-shot learning. In: International Conference on Machine Learning, PMLR, pp 24457–24477

Miconi T (2023) Learning to acquire novel cognitive tasks with evolution, plasticity and meta-meta-learning. In: International Conference on Machine Learning, PMLR, pp 24756–24774

Miller EG, Matsakis NE, Viola PA (2000) Learning from one example through shared densities on transforms. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662), IEEE, pp 464–471

Min S, Lewis M, Zettlemoyer L, et al (2022a) Metaicl: Learning to learn in context. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp 2791–2809

Min S, Lyu X, Holtzman A, et al (2022b) Rethinking the role of demonstrations: What makes in-context learning work? In: EMNLP

Mishra N, Rohaninejad M, Chen X, et al (2018) A simple neural attentive meta-learner. In: International Conference on Learning Representations

Munkhdalai T, Yu H (2017) Meta networks. In: International conference on machine learning, PMLR, pp 2554–2563

Murty S, Hashimoto TB, Manning CD (2021) Dreca: A general task augmentation strategy for few-shot natural language inference. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp 1113–1125

Najdenkoska I, Zhen X, Worring M (2023) Meta learning to bridge vision and language models for multimodal few-shot learning. In: The Eleventh International Conference on Learning Representations

Nam J, Tack J, Lee K, et al (2023) STUNT: Few-shot tabular learning with self-generated tasks from unlabeled tables. In: The Eleventh International Conference on Learning Representations

Nguyen T, Grover A (2022) Transformer neural processes: Uncertainty-aware meta learning via sequence modeling. In: International Conference on Machine Learning, PMLR, pp 16569–16594

Ni F, Hao J, Mu Y, et al (2023) Metadiffuser: Diffusion model as conditional planner for offline meta-rl. In: International Conference on Machine Learning

Ni R, Shu M, Souri H, et al (2021) The close relationship between contrastive learning and meta-learning. In: International Conference on Learning Representations

Nichol A, Achiam J, Schulman J (2018) On first-order meta-learning algorithms. arXiv preprint arXiv:180302999

Olsson C, Elhage N, Nanda N, et al (2022) In-context learning and induction heads. arXiv preprint arXiv:220911895

Oreshkin B, Rodríguez López P, Lacoste A (2018) Tadam: Task dependent adaptive metric for improved few-shot learning. Advances in neural information processing systems 31

Pachetti E, Colantonio S (2024) A systematic review of few-shot learning in medical imaging. Artificial intelligence in medicine p 102949

Panwar M, Ahuja K, Goyal N (2024) In-context learning through the bayesian prism. In: The Twelfth International Conference on Learning Representations

Park E, Oliva JB (2019) Meta-curvature. Advances in Neural Information Processing Systems 32

Park N, Rossi RA, Ahmed N, et al (2022) Metagl: Evaluation-free selection of graph learning models via meta-learning. In: The Eleventh International Conference on Learning Representations

Parnami A, Lee M (2022) Learning from few examples: A summary of approaches to few-shot learning. arXiv preprint arXiv:220304291

Patacchiola M, Turner J, Crowley EJ, et al (2020) Bayesian meta-learning for the few-shot setting via deep kernels. Advances in Neural Information Processing Systems 33:16108–16118

Patel A, Li B, Rasooli MS, et al (2022) Bidirectional language models are also few-shot learners. arXiv preprint arXiv:220914500

Patterson D, Gonzalez J, Le Q, et al (2021) Carbon emissions and large neural network training. arXiv preprint arXiv:210410350

Pavasovic KL, Rothfuss J, Krause A (2023) MARS: Meta-learning as score matching in the function space. In: The Eleventh International Conference on Learning Representations

Qin S, Zhu J, Qin J, et al (2019) Recurrent attentive neural process for sequential data. arXiv preprint arXiv:191009323

Radford A, Kim JW, Hallacy C, et al (2021) Learning transferable visual models from natural language supervision. In: International conference on machine learning, PMLR, pp 8748–8763

Raghu A, Raghu M, Bengio S, et al (2020) Rapid learning or feature reuse? towards understanding the effectiveness of maml. In: International Conference on Learning Representations

Rajeswaran A, Finn C, Kakade SM, et al (2019) Meta-learning with implicit gradients. Advances in neural information processing systems 32

Rakelly K, Zhou A, Finn C, et al (2019) Efficient off-policy meta-reinforcement learning via probabilistic context variables. In: International conference on machine learning, PMLR, pp 5331–5340

Raventós A, Paul M, Chen F, et al (2024) Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. Advances in Neural Information Processing Systems 36

Ravi S, Larochelle H (2016) Optimization as a model for few-shot learning. In: International conference on learning representations

Reddi SJ, Miryoosefi S, Karp S, et al (2023) Efficient training of language models using few-shot learning. In: International Conference on Machine Learning, PMLR,

pp 14553–14568

Ren M, Ravi S, Triantafillou E, et al (2018) Meta-learning for semi-supervised few-shot classification. In: International Conference on Learning Representations

Rizve MN, Khan S, Khan FS, et al (2021) Exploring complementary strengths of invariant and equivariant representations for few-shot learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10836–10846

Rusu AA, Rao D, Sygnowski J, et al (2019) Meta-learning with latent embedding optimization. In: International Conference on Learning Representations

Santoro A, Bartunov S, Botvinick M, et al (2016) Meta-learning with memory-augmented neural networks. In: International conference on machine learning, PMLR, pp 1842–1850

Schaeffer R, Miranda B, Koyejo S (2024) Are emergent abilities of large language models a mirage? Advances in Neural Information Processing Systems 36

Schimunek J, Seidl P, Friedrich L, et al (2023) Context-enriched molecule representations improve few-shot drug discovery. In: The Eleventh International Conference on Learning Representations

Schmidhuber J (1987) Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook. PhD thesis, Technische Universität München

Shen Z, Liu Z, Qin J, et al (2021) Partial is better than all: revisiting fine-tuning strategy for few-shot learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 9594–9602

Shi X, Xue S, Wang K, et al (2023) Language models can improve event prediction by few-shot abductive reasoning. arXiv preprint arXiv:230516646

Shin S, Lee SW, Ahn H, et al (2022) On the effect of pretraining corpora on in-context learning by a large-scale language model. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp 5168–5186

Shome D, Kar T (2021) Fedaffect: Few-shot federated learning for facial expression recognition. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 4168–4175

Shysheya A, Bronskill JF, Patacchiola M, et al (2023) Fit: Parameter efficient few-shot transfer learning for personalized and federated image classification. In: The Eleventh International Conference on Learning Representations

Silver D, Schrittwieser J, Simonyan K, et al (2017) Mastering the game of go without human knowledge. nature 550(7676):354–359

Snell J, Swersky K, Zemel R (2017) Prototypical networks for few-shot learning. Advances in neural information processing systems 30

Song Y, Wang T, Cai P, et al (2023a) A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. ACM Computing Surveys

Song Z, Zhao Y, Shi Y, et al (2023b) Learning with fantasy: Semantic-aware virtual contrastive constraint for few-shot class-incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 24183–24192

Stanley M, Bronskill JF, Maziarz K, et al (2021) Fs-mol: A few-shot learning dataset of molecules. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)

Sucholutsky I, Griffiths T (2024) Alignment with human representations supports robust few-shot learning. Advances in Neural Information Processing Systems 36

Sun W, Du Y, Zhen X, et al (2023) Metamodulation: Learning variational feature hierarchies for few-shot learning with fewer tasks. In: International Conference on Machine Learning, PMLR, pp 32847–32858

Sun Z, Wu J, Li X, et al (2021) Amortized bayesian prototype meta-learning: A new probabilistic meta-learning approach to few-shot image classification. In: International Conference on Artificial Intelligence and Statistics, PMLR, pp 1414–1422

Sung F, Zhang L, Xiang T, et al (2017) Learning to learn: Meta-critic networks for sample efficient learning. arXiv preprint arXiv:170609529

Sung F, Yang Y, Zhang L, et al (2018) Learning to compare: Relation network for few-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1199–1208

Tang H, Liu S, Lin T, et al (2023a) Master: Meta style transformer for controllable zero-shot and few-shot artistic style transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 18329–18338

Tang L, Xie H, Wang X, et al (2023b) Privacy-preserving knowledge sharing for few-shot building energy prediction: A federated learning approach. Applied Energy 337:120860

Tay Y, Dehghani M, Tran VQ, et al (2023) UL2: Unifying language learning paradigms. In: The Eleventh International Conference on Learning Representations

Tian Y, Wang Y, Krishnan D, et al (2020) Rethinking few-shot image classification: a good embedding is all you need? In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16, Springer, pp 266–282

Todd E, Li M, Sharma AS, et al (2024) Function vectors in large language models. In: The Twelfth International Conference on Learning Representations

Triantafillou E, Zhu T, Dumoulin V, et al (2020) Meta-dataset: A dataset of datasets for learning to learn from few examples. In: International Conference on Learning Representations

Tsimpoukelli M, Menick JL, Cabi S, et al (2021) Multimodal few-shot learning with frozen language models. Advances in Neural Information Processing Systems 34:200–212

Ullah I, Carrión-Ojeda D, Escalera S, et al (2022) Meta-album: Multi-domain meta-dataset for few-shot image classification. Advances in Neural Information Processing Systems 35:3232–3247

Vaswani A, Shazeer N, Parmar N, et al (2017) Attention is all you need. Advances in neural information processing systems 30

Vinyals O, Blundell C, Lillicrap T, et al (2016) Matching networks for one shot learning. Advances in neural information processing systems 29

Von Oswald J, Niklasson E, Randazzo E, et al (2023) Transformers learn in-context by gradient descent. In: International Conference on Machine Learning, PMLR, pp 35151–35174

Vuorio R, Sun SH, Hu H, et al (2019) Multimodal model-agnostic meta-learning via task-aware modulation. Advances in neural information processing systems 32

Wang H, Deng ZH (2021) Cross-domain few-shot classification via adversarial task augmentation. arXiv preprint arXiv:210414385

Wang H, Yue T, Ye X, et al (2022a) Revisit finetuning strategy for few-shot learning to transfer the emdeddings. In: The Eleventh International Conference on Learning Representations

Wang H, Tian S, Fu Y, et al (2023a) Feature augmentation based on information fusion rectification for few-shot image classification. Scientific Reports 13(1):3607

Wang HP, Li P (2023) Unsupervised learning for combinatorial optimization needs meta learning. In: The Eleventh International Conference on Learning Representations

Wang J, Zhang J, Jiang H, et al (2023b) Offline meta reinforcement learning with in-distribution online adaptation. In: Proceedings of the 40th International Conference on Machine Learning

Wang JX, Kurth-Nelson Z, Tirumala D, et al (2016) Learning to reinforcement learn. arXiv preprint arXiv:161105763

Wang L, Zhou S, Zhang S, et al (2023c) Improving generalization of meta-learning with inverted regularization at inner-level. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 7826–7835

Wang R, Walters R, Yu R (2022b) Meta-learning dynamics forecasting using task inference. Advances in Neural Information Processing Systems 35:21640–21653

Wang Y, Anderson DV (2022) Hybrid attention-based prototypical networks for few-shot sound classification. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp 651–655

Wang Y, Chao WL, Weinberger KQ, et al (2019) Simpleshot: Revisiting nearest-neighbor classification for few-shot learning. arXiv preprint arXiv:191104623

Wang Y, Salamon J, Bryan NJ, et al (2020a) Few-shot sound event detection. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp 81–85

Wang Y, Yao Q, Kwok JT, et al (2020b) Generalizing from a few examples: A survey on few-shot learning. ACM computing surveys (csur) 53(3):1–34

Wang Y, Bryan NJ, Cartwright M, et al (2021) Few-shot continual learning for audio classification. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp 321–325

Wang Z, Li J, Wang W, et al (2024) Review of few-shot learning application in csi human sensing. Artificial Intelligence Review 57(8):195

Wei J, Tay Y, Bommasani R, et al (2022a) Emergent abilities of large language models. Transactions on Machine Learning Research

Wei XS, Xu HY, Zhang F, et al (2022b) An embarrassingly simple approach to semi-supervised few-shot learning. Advances in Neural Information Processing Systems 35:14489–14500

Wornow M, Thapa R, Steinberg E, et al (2023) EHRSHOT: An EHR benchmark for few-shot evaluation of foundation models. In: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track

Wu B, Meng Z, Zhang Q, et al (2022a) Meta-learning helps personalized product search. In: Proceedings of the ACM Web Conference 2022, pp 2277–2287

Wu B, Fang J, Zeng X, et al (2023a) Adaptive compositional continual meta-learning. In: International Conference on Machine Learning, PMLR, pp 37358–37378

Wu CJ, Raghavendra R, Gupta U, et al (2022b) Sustainable ai: Environmental implications, challenges and opportunities. Proceedings of Machine Learning and Systems 4:795–813

Wu H, Yin J, Rajaratnam B, et al (2023b) Hierarchical relational learning for few-shot knowledge graph completion. In: The Eleventh International Conference on Learning Representations

Wu Y, Li Z, Wang C, et al (2023c) Domain re-modulation for few-shot generative domain adaptation. In: Thirty-seventh Conference on Neural Information Processing Systems

Xie SM, Raghunathan A, Liang P, et al (2021) An explanation of in-context learning as implicit bayesian inference. arXiv preprint arXiv:211102080

Xu C, Liu C, Sun X, et al (2022a) Patchmix augmentation to identify causal features in few-shot learning. IEEE Transactions on Pattern Analysis and Machine Intelligence

Xu J, Li K, Li D (2024) An automated few-shot learning for time-series forecasting in smart grid under data scarcity. IEEE Transactions on Artificial Intelligence 5(6):2482–2492

Xu L, Jin S, Zeng W, et al (2022b) Pose for everything: Towards category-agnostic pose estimation. In: European Conference on Computer Vision, Springer, pp 398–416

Yan K, Li X, Wei F, et al (2023) Two-shot video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 2257–2267

Yang Y, Yuan H, Li X, et al (2023) Neural collapse inspired feature-classifier alignment for few-shot class-incremental learning. In: The Eleventh International Conference on Learning Representations

Yao H, Huang LK, Zhang L, et al (2021a) Improving generalization in meta-learning via task augmentation. In: International conference on machine learning, PMLR, pp 11887–11897

Yao H, Zhang L, Finn C (2021b) Meta-learning with fewer tasks through task interpolation. arXiv preprint arXiv:210602695

Ye F, Lin B, Yue Z, et al (2021) Multi-objective meta learning. Advances in Neural Information Processing Systems 34:21338–21351

Ye HJ, Hu H, Zhan DC, et al (2020) Few-shot learning via embedding adaptation with set-to-set functions. In: Proceedings of the IEEE/CVF conference on computer

vision and pattern recognition, pp 8808–8817

Yeh CH, Russell B, Sivic J, et al (2023) Meta-personalizing vision-language models to find named instances in video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 19123–19132

Yoon J, Kim T, Dia O, et al (2018) Bayesian model-agnostic meta-learning. Advances in neural information processing systems 31

Yu M, Guo X, Yi J, et al (2018) Diverse few-shot text classification with multiple metrics. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)

Yu R, Gong Y, He X, et al (2021) Personalized adaptive meta learning for cold-start user preference prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 10772–10780

Yu R, Chen W, Wang X, et al (2022) Enhancing meta learning via multi-objective soft improvement functions. In: The Eleventh International Conference on Learning Representations

Yu T, Quillen D, He Z, et al (2020) Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In: Conference on robot learning, PMLR, pp 1094–1100

Yun S, Han D, Oh SJ, et al (2019) Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 6023–6032

Zhang C, Cai Y, Lin G, et al (2020) Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12203–12213

Zhang H, Cisse M, Dauphin YN, et al (2018) mixup: Beyond empirical risk minimization. In: International Conference on Learning Representations

Zhang J, Zhao C, Ni B, et al (2019) Variational few-shot learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 1685–1694

Zhang J, Liu L, Silven O, et al (2023a) Few-shot class-incremental learning: A survey. arXiv preprint arXiv:230806764

Zhang R, Hu X, Li B, et al (2023b) Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 15211–15222

Zhao L, Lu J, Xu Y, et al (2023) Few-shot class-incremental learning via class-aware bilateral distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 11838–11847

Zhao Y, Yu G, Wang J, et al (2022) Personalized federated few-shot learning. IEEE Transactions on Neural Networks and Learning Systems

Zhao Z, Wallace E, Feng S, et al (2021) Calibrate before use: Improving few-shot performance of language models. In: International Conference on Machine Learning, PMLR, pp 12697–12706

Zheng C, Liu B, Zhang H, et al (2023a) Where is my spot? few-shot image generation via latent subspace optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 3272–3281

Zheng H, Wang R, Liu J, et al (2023b) Cross-level distillation and feature denoising for cross-domain few-shot classification. In: The Eleventh International Conference on Learning Representations

Zhou Y, Li C, Chen C, et al (2022) Lafite2: Few-shot text-to-image generation. arXiv preprint arXiv:221014124

Zhou Y, Li J, Xiang Y, et al (2024) The mystery of in-context learning: A comprehensive survey on interpretation and analysis. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp 14365–14378

Zhuang H, Weng Z, He R, et al (2023) Gkeal: Gaussian kernel embedded analytic learning for few-shot class incremental task. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 7746–7755

Zuo X, Chen Z, Yao H, et al (2022) Understanding train-validation split in meta-learning with neural networks. In: The Eleventh International Conference on Learning Representations