



## A Bilevel Optimization Perspective on Meta-Learning, Few-Shot Learning, and Parameter-Efficient Adaptation

Karima Saidi 

Computer Science Department, Mathematics and Computer Science Faculty, Mustapha Benboulaïd Batna 2, Batna University, Batna 05078, Algeria

Corresponding Author Email: [karima.saidi@univ-batna2.dz](mailto:karima.saidi@univ-batna2.dz)

Copyright: ©2026 The author. This article is published by IIETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/isi.310915>

### ABSTRACT

**Received:** 10 July 2026

**Revised:** 25 August 2026

**Accepted:** 1 September 2026

**Available online:** 30 September 2026

#### **Keywords:**

*bilevel optimization, meta-learning, few-shot learning, parameter-efficient fine-tuning, low-rank adaptation, in-context learning, cross-domain few-shot learning*

Meta-learning and few-shot learning address adaptation when only limited labelled data are available, but their relationship to parameter-efficient fine-tuning (PEFT) and in-context learning (ICL) is often discussed across separate literatures. This paper develops a bilevel-optimization perspective for organizing task-adaptation methods. Gradient-based, metric-based, and model-based approaches are represented through a common task-level adaptation operator, while PEFT, low-rank adaptation (LoRA), Bayesian meta-learning, and ICL are examined as extensions or related formulations. For LoRA, we establish a representational result for a single weight matrix: when the adapter rank reaches the matrix's maximum possible rank, the factorized update can represent any update matrix, including an unconstrained one-step gradient update. This result concerns representational capacity and does not establish that optimization of the low-rank factors converges to the same solution. The paper also reviews published few-shot classification results reported on miniImageNet with a ResNet-12 backbone and compares selected cross-domain results across EuroSAT, CropDisease, ISIC, and ChestX. The reported in-domain results favor metric-based methods in the examined sample, whereas cross-domain outcomes vary across methods and target datasets; the evidence does not establish a universal ranking of method families. Finally, five conjectures are formulated with proposed experiments concerning LoRA optimization, implicit adaptation through ICL, task-distribution geometry, curriculum sampling, and feature hallucination. The framework provides a common vocabulary for comparing adaptation mechanisms while identifying theoretical claims that remain to be tested.

## 1. INTRODUCTION

Meta-learning and few-shot learning address the challenges of limited training data and the need for rapid adaptation to new tasks. Unlike conventional machine learning, which relies on large amounts of labelled data, these approaches enable models to learn from small amounts of data and adapt quickly. These approaches have therefore attracted considerable research interest.

Meta-learning (or "learning to learn") creates models that learn how to learn a task [1]. The goal is to create models that quickly adapt to new tasks by leveraging knowledge from previous experiences. This is particularly useful when traditional methods fail due to limited data or the need for rapid adaptation [2].

Meta-learning can be used to train few-shot learning models, which classify new classes from one to five examples per class [3]. This ability is crucial for designing AI systems that adapt and perform well in real-world scenarios where training data is scarce [4].

Humans can easily learn new classes with only a few examples and apply general rules. Researchers seek to embed these principles into machine learning to create AI systems

that learn like humans [5].

Meta-learning is linked to few-shot learning, but the connection remains unclear in the literature. Few-shot learning serves as a testbed for meta-learning algorithms and a distinct field with its own tasks, benchmarks, and methods, including transfer learning and fine-tuning large pretrained models. This ambiguity has led to confusion over training and partitioning queries and supports into episodes, as well as learning at the task-level. In this paper, the few-shot learning problem is treated as a meta-learning problem, a modeling assumption rather than a fact.

Previous research has shown that few-shot learning is a specific instance of task-based learning and that meta-learning offers a principled way of solving tasks [6]. However, current methods are usually method-specific, limited to a single family of methods. This paper includes a single bilevel optimization vocabulary that covers gradient-based, metric-based, and model-based families, as well as parameter-efficient fine-tuning (PEFT), Bayesian meta-learning, and in-context learning (ICL).

The paper makes five main contributions. First, we propose an organizing scheme that describes gradient-based, metric-based, and model-based methods as special cases of a common

adaptation operator with different parameterizations. Second, we formally prove that, at full adapter rank, the set of reachable low-rank adaptation (LoRA) updates is identical to that of an unconstrained one-step gradient update, thereby resolving the representational part of the convergence question raised in Section 6. Third, we extend the framework to ICL in large language models (LLMs), PEFT, including LoRA, and diffusion-based data augmentation as a conceptual parallel. Fourth, we conduct a meta-analysis of published benchmark results from studies using a ResNet-12 backbone. Finally, we formulate five open questions as falsifiable conjectures and propose concrete experiments to test them.

The rest of the paper is organized as follows: the unified view is described in Section 2, and the comparative analysis of method families is presented in Section 3. Section 4 presents a meta-analysis of benchmark results. Section 5 discusses the applications and benchmarks. In addition, unsolved problems and future directions are given in Section 6, while a conclusion is drawn in Section 7.

## 2. UNIFIED FRAMEWORK: LEARNING ON A DISTRIBUTION OF TASKS

### 2.1 Task distribution and episode construction

We assume a distribution  $T$  over tasks. Each task  $\tau \sim T$  is characterized by a support set  $S_\tau = \{(x_i, y_i)\}$  and a query set  $Q_\tau = \{(x_j, y_j)\}$ . The support set serves as a small training set for the task, and the query set is used to test generalization. In the N-way K-shot few-shot learning setting, a task consists of N classes with K examples per class ( $|S_\tau| = NK$ ). Notably, the episodic training process (whereby the learner is trained on a sequence of tasks sampled at meta-training time) is repeated at the meta-level (as shown in Figure 1), following the supervised learning cycle. Vinyals et al. [7] first introduced this idea, which connects few-shot learning to meta-learning. Few-shot learning is a special case of meta-learning with task diversity obtained by sub-sampling classes and examples from a shared pool.

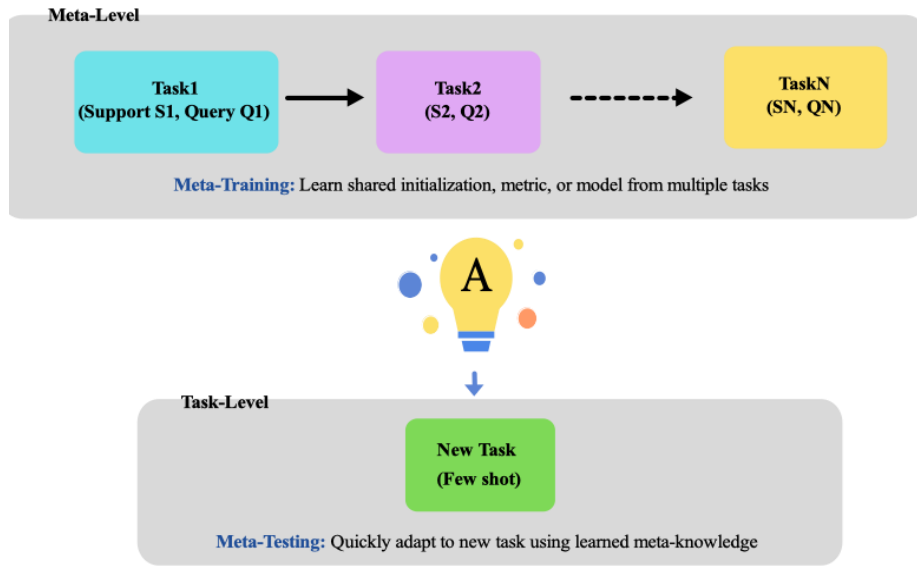


Figure 1. How meta-learning connects to few-shot learning

### 2.2 Bilevel optimization formulation

The meta-learning objective is formulated as a bilevel problem, where  $\theta$  denotes the shared meta-parameters and  $\varphi_\tau$  denotes the task-specific parameters. The meta-level objective is the expected loss over tasks:

$$\min_{\theta} \mathbb{E}_{\tau \sim T} [L_Q(\varphi_\tau, Q_\tau)], \varphi_\tau = A(\theta, S_\tau) \quad (1)$$

where,  $A(\theta, S_\tau)$  is the adaptation operator that maps meta-parameters and a support set to task-specific parameters, and  $L_Q$  measures loss on the query set. The adaptation operator  $A$  defines the task-level problem (the learning algorithm).  $A$  can be a family of gradient steps, nearest-neighbor assignment, or a read-out function learned by a neural network.

### 2.3 Method families as special cases of $A$

The three method families differ in their parameterization of  $A$ . These latter are asserted as organizing schemas.

**Organizing scheme** (family unification): The organizing schema follows from the bilevel objective of subsection 2.2.

Therefore, each family can be described as a particular parameterization of the  $A$  in Eq. (1). Table 1 summarizes the general adaptation operator  $A$  and its interpretation for each of the three method families.

**Conceptual parallel** (ICL implicitly as an adaptation operator): The fourth special case of  $A$  as implicit is ICL. The transformer attention mechanism updates task-specific activations without changing any parameters.

This corollary differentiates ICL from the three aforementioned families:

(1) Versus gradient-based ( $A^g$ ). ICL does not need to make any iterative Task-level or gradient steps to adapt; instead, the model must be able to "read" the task from the support set.

(2) ICL is similar to metric-based methods, but it does not tend to compute a closed-form aggregation such as a class prototype or mean embedding.

(3) Versus model-based ( $A^M$ ). The closest relative is ICL; both are based on learned feed-forward functions. However, conventional model-based techniques such as Hyper-networks generate the weights of another network. ICL, on the other hand, does not change the weights of the model, but adjusts the internal state of the model, with the same value of  $\varphi_\tau = \theta$ .

**Table 1.** Adaptation operator  $A$  by method family

| Family         | Adaptation Operator $A(\theta, S_\tau)$ with Interpretation                                                                                                                                                                                                                                                                              | How Published Methods Instantiate It                                                                                                                                                                                                                                                                                                                         |
|----------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Gradient-based | $A^g(\theta, S_\tau) = \theta - \alpha \nabla \theta L_S(\theta)$ ( $T$ gradient steps)<br>Task-level is $T$ steps of stochastic gradient descent (SGD) from the shared meta-initialization $\theta$ ; Meta-level trains $\theta$ via second-order gradients Model-Agnostic Meta-Learning (MAML) or first-order approximation (Reptile). | The MAML objective is the meta-objective, $\min_\theta \mathbb{E}_\tau [L_Q(\theta - \alpha \nabla \theta L_S(\theta))]$ [1]. The approximation of Reptile recovers $T = 1$ [8]. Because differentiating through many task-level steps is costly, implicit differentiation can instead be used to recover implicit Model-Agnostic Meta-Learning (iMAML) [9]. |
| Metric-based   | $A^T(\theta, S_\tau) = c_n = f_\theta(S_\tau^n)$ where, $c_n$ is a closed-form class summary (e.g., mean embedding). $A$ is non-iterative: no gradient step is taken at the task-level. Query classification uses $d(f_\theta(x), c_n)$ for distance function.                                                                           | During the training of the meta-objective, only the shared encoder $\theta$ is trained. By setting $f_\theta(x_i): y_i = n$ the model recovers Prototypical Networks [10], and replacing the distance with a learned comparator recovers Relation Networks [11].                                                                                             |
| Model-based    | $A^M(\theta, S_\tau) = g_\theta(S_\tau)$ where, $g$ is a learned feed-forward function (e.g., memory read-write, hypernetwork weight generation, or amortized posterior). No iterative task-level optimization is required; the architecture internalizes adaptation.                                                                    | The meta-objective trains $g_\theta$ to yield a good $\varphi_\tau = g_\theta(S_\tau)$ on the query set. Hypernetworks and memory-augmented networks [12, 13] are instances.                                                                                                                                                                                 |

## 2.4 Extension to parameter-efficient fine-tuning/low-rank adaptation

PEFT techniques such as LoRA [14] and adapter-based fine-tuning can be extended to the family of gradient-based methods with a limited adaptation operator:

$$A^{LRT}(\theta, S_\tau) = \theta + \Delta\theta \quad (2)$$

where,  $\Delta\theta = BA$ ,  $\theta \in \mathbb{R}^d$  denotes the frozen pretrained meta-parameters; the update is restricted to  $\Delta\theta = BA$  with  $B \in \mathbb{R}^{d \times r}$ ,  $A \in \mathbb{R}^{r \times d}$ , and  $\text{rank}(A) = r \ll d$ . This parameterization assumes that the required task-specific update lies on a lower-dimensional manifold.

LoRA learns a smaller number of parameters for matrices  $A$  and  $B$  compared to updating a full  $d \times d$  matrix, which has  $d^2$  parameters.

The classification shows that PEFT and LoRA are not distinct from meta-learning; however, they are in fact gradient-based meta-learners whose learning algorithm ( $A$ ) has a very efficient low-rank constraint.

All meta-parameters,  $\theta$ , are held fixed and are not updated throughout the task-specific ‘‘Task-level’’; only low-rank matrices  $A$  and  $B$  are adapted with gradients from the support set  $S_\tau$ . This sub-family achieves memory savings of  $O(rd)$  compared to  $O(d^2)$  while performing full fine-tuning and has been the most popular approach for few-shot adaptation of billion-parameter LLMs.

## 2.5 Rank-capacity equivalence between low-rank adaptation and gradient-based adaptation

While subsection 2.4 reinterprets LoRA as an adaptation operator, similar to a gradient, under a rank constraint, it raises a key question: does the LoRA operator match an unconstrained gradient step as the rank  $r$  increases, and at what rank? This question is tied to the practical issue of what adapter rank is ‘‘adequate’’. This subsection provides a brief response to the expressive power question.

We include a single weight matrix  $W_0 \in \mathbb{R}^{d1 \times d2}$  within the meta-parameters, updated at the task-level either by (1) an unconstrained 1-step gradient update,  $\Delta W^g = -\alpha \nabla WLS(W_0)$ , an arbitrary  $d1 \times d2$  matrix, or (2) a LoRA update,  $\Delta W^{LRT} = BA$  with  $B \in \mathbb{R}^{d1 \times r}$  and  $A \in \mathbb{R}^{r \times d2}$ .

**Proposition** (Rank-capacity equivalence; formalizes an

observation of Hu et al. [14]): Target update matrix  $\Delta W \in \mathbb{R}^{d1 \times d2}$  with rank  $r_{\Delta W}$ , and for every adapter rank  $r \geq r_{\Delta W}$ , there exist matrices  $B \in \mathbb{R}^{d1 \times r}$  and  $A \in \mathbb{R}^{r \times d2}$  such that  $BA = \Delta W$  exactly. Since  $r_{\Delta W} \leq \min(d1, d2)$  holds for any  $d1 \times d2$  matrix, it follows that at full rank,  $r = \min(d1, d2)$ , the LoRA parameterization  $\{BA\}$  spans the entire space  $\mathbb{R}^{d1 \times d2}$  of possible weight updates. Consequently, at this rank the reachable set of the LoRA operator  $A^{LRT}$  coincides exactly with the reachable set of the unconstrained one-step gradient operator  $A^g$  restricted to this layer: for any realized gradient update  $\Delta W^g$ , there exist  $B^*, A^*$  with  $B^* A^* = \Delta W^g$ .

**Proof:**  $\Delta W = U\Sigma V^T$  singular value decomposition, where  $U \in \mathbb{R}^{d1 \times d1}$ ,  $V \in \mathbb{R}^{d2 \times d2}$ , and  $\Sigma$  diagonal with the  $r_{\Delta W}$  nonzero singular values  $\sigma_1, \dots, \sigma_{r_{\Delta W}}$ . For  $r \geq r_{\Delta W}$ , take  $U_r$  and  $V_r$  to be the first  $r$  columns of  $U$  and  $V$ , padding with zero singular values if  $r > r_{\Delta W}$ , and set  $B = U_r \Sigma_r^{1/2}$  and  $A = \Sigma_r^{1/2} V_r^T$ . Then  $BA = U_r \Sigma_r V_r^T = \Delta W$ , which proves the first claim. Because  $\text{rank}(\Delta W) \leq \min(d1, d2)$  for any  $d1 \times d2$  matrix, setting  $r = \min(d1, d2)$  satisfies  $r \geq \text{rank}(\Delta W^g)$  for every possible gradient update  $\Delta W^g$ , and the construction above then yields  $B^*, A^*$  with  $B^* A^* = \Delta W^g$ .

The proposition confirms LoRA can represent any single gradient step at full rank, a result previously stated informally in the LoRA contribution [14]. This does not mean gradient descent on low-rank factors can find that representation, since the  $(B, A) \rightarrow BA$  map is non-convex even if the underlying problem is convex. Shuttleworth et al. [15] showed that LoRA and fully fine-tuned models differ structurally at comparable performance levels, and Zindari et al. [16] demonstrated theoretically that rank affects generalization beyond capacity. The proposition only addresses the ‘‘can LoRA represent it’’ part of Conjecture 6.1; training the model remains open.

## 2.6 Extension of the adaptation operator to Bayesian meta-learning

Bayesian meta-learning treats the adaptation operator as posterior inference:

$$A^{BD}(\theta, S_\tau) = p(\varphi_\tau | S_\tau, \theta) \quad (3)$$

where, the meta-parameters  $\theta$  parameterize the prior over task-specific parameters.

Three representative algorithms illustrate this approach. First, Bayesian Model-Agnostic Meta-Learning (BMAML)

[17] extends MAML using Stein variational gradient methods to represent uncertainty. Second, Neural Processes [18] use latent-variable models over functions to estimate predictive uncertainty. Finally, Deep Kernel Transfer [19] combines deep feature extractors with Gaussian process (GP)-based Bayesian inference. The medical imaging field is particularly relevant to the Bayesian approach, where the uncertainty needs to be calibrated for clinical use.

### 2.7 Role of inductive bias

There are two types of inductive bias in the context of meta-learning, including (1) architectural bias: it is the choice of a particular encoder, memory structure, or attention mechanism; (2) algorithmic bias: it is the functional form of A. The latter is assumed by gradient-based methods for new tasks that can be solved by a local change of parameters, starting from a common initialization, as shown in references [1, 8]. Metric-based methods assume a single embedding space offers class-discriminative geometry, directly investigated in prototype- and relation-based approaches of Snell et al. and Sung et al. [10, 11]. Model-based methods assume the architecture can

adapt without iterative task-level optimization, as discussed in references [12, 13]. Evidence for gradient-based and metric-based approaches comes from image-classification benchmarks, where they make their respective assumptions in the visual domain. The additional assumptions such as ICL and PEFT are that (1) the representations are already large, (2) they already contain a lot of prior knowledge, and (3) only light conditioning is required.

## 3. COMPARATIVE ANALYSIS

### 3.1 Meta-learning vs. few-shot learning: Core distinctions

To clarify the conceptual differences mentioned earlier, Table 2 compares meta-learning and few-shot learning across nine key aspects: goal of learning, role, objective, unit of learning, training, adaptation, computational cost, strength, and limitations. It also juxtaposes the meta-learning framing of few-shot learning with the evaluation and benchmark framing used in the few-shot-learning literature.

**Table 2.** The key differences between meta-learning and few-shot learning

| Criteria           | Meta-Learning                                                                        | Few-Shot Learning                                                                                                                    |
|--------------------|--------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------|
| Goal of learning   | Learn an adaptation strategy or a "learning algorithm".                              | Learn a given task from a few examples.                                                                                              |
| The role           | A paradigm for the use of past tasks to enhance new tasks.                           | A setting or evaluation regime is a problem that presents data deficiency (K-shot).                                                  |
| Objective          | Fast adaptation to unseen tasks sampled from a distribution.                         | High accuracy in cases of limited training data.                                                                                     |
| Unit of learning   | Task distribution ( $T$ ); learn over many episodes.                                 | A classification problem that is described by a specific low-data episode; usually a K-shot classification problem involving N ways. |
| Training           | Episodic or task-based training that repeats the supervised cycle at the meta-level. | Episodic training; or applying large pre-trained models via direct inference/prompting.                                              |
| Adaptation         | Task-level optimization, similarity, memory, or learned optimizer.                   | Fine-tuning, nearest-prototype, or prompting.                                                                                        |
| Computational cost | Higher (meta-training with nested optimization).                                     | Lower at inference for metric-based methods.                                                                                         |
| Strength           | Produces transferable priors.                                                        | Practical benchmark and evaluation framework.                                                                                        |
| Limitations        | The complexity and the cost of training.                                             | Not itself a complete algorithmic framework.                                                                                         |

### 3.2 Method family survey

Methods are assigned to a family based on the dominant adaptation operator used at the task-level: iterative gradient updates, closed-form aggregation, or learned feed-forward mapping. Some methods are hybrids, e.g., methods that use a metric head followed by a light gradient fine-tuning step, and

are classified according to the main operator.

#### 3.2.1 Metric-based methods

Table 3 summarizes the major metric-based algorithms reviewed in this subsection, highlighting the type of adaptation operator that they use, the mechanism, and the advantages and limitations of each algorithm.

**Table 3.** Survey of metric-based method families

| Algorithm                  | Adaptation Operator $A$ Type | Mechanism                                                                                                                                                      | Advantages                                                                                                                                                                                                    | Limits                                                                                |
|----------------------------|------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|
| Matching Networks [7]      | Closed-form Aggregation      | Uses a soft attention mechanism on the support set using a cosine kernel; uses a bidirectional long short-term memory (LSTM) encoder to encode the samples     | Early method that connected few-shot learning and meta-learning explicitly. Easy to use, efficient, and offers a good inductive bias even when the number of test tasks exceeds the number of training tasks. | Application of the model to different domains may necessitate retraining the encoder. |
| Prototypical Networks [10] | Closed-form Aggregation      | Learn a mapping to a feature space, in which a class prototype will be the mean of support feature vectors; uses squared Euclidean distance for classification |                                                                                                                                                                                                               | The fixed mean causes the model to be sensitive to outliers in the support set.       |
| Relation Network [11]      | Closed-form Aggregation      | Instead of handcrafted distances uses a learned deep comparator (relation                                                                                      | Highly flexible architecture that can                                                                                                                                                                         | The additional flexibility comes at the cost of additional                            |

|                       |                         |                                                                                                                                            |                                                                                                           |                                                                                                                 |
|-----------------------|-------------------------|--------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------|
|                       |                         | module), concatenates query features with prototypes for a learned score head.                                                             | easily be extended to the zero-shot learning case.                                                        | model parameters to be trained, relative to fixed metric approaches.                                            |
| Siamese Networks [20] | Closed-form Aggregation | Compute the pairwise similarity using weight-shared twin networks, compute the L1 distance, then apply a sigmoid activation to the scores. | Decisions are easy to interpret, and pairwise similarity data are relatively simple to collect and label. | By using pairwise comparisons, it can be shown that the complexity is quadratic in the size of the support set. |
| SAPENet [21]          | Closed-form Aggregation | Enrich prototypes with multi-head self-attention and intra-class channel attention to capture complex class structures.                    | Produces more discriminative prototypes than just plain mean embeddings                                   | Adding attention mechanisms results in a higher amount of computation than ProtoNets.                           |

**Table 4.** Survey of gradient/optimization-based method families

| Algorithm                | $\mathcal{A}$ Type                    | Mechanism                                                                                                                                                                                                                                        | Advantages                                                                                                                                                | Limits                                                                                                                                                                                                |
|--------------------------|---------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| MAML [1]                 | Iterative Optimization (2nd-Order)    | Meta-learns an initialization $\theta$ that is sensitive to gradients of the task; uses second-order derivatives along the optimization path.                                                                                                    | Works for all models; applicable to vision, regression, and reinforcement learning.                                                                       | Hessian-vector products are an expensive computation and memory operation.                                                                                                                            |
| Reptile [8]              | Iterative Optimization (1st-Order)    | A scalable variant moving the meta-init towards the parameters learned after multiple SGD steps in a task. Treats the path of the Task-level as an Implicit Function, and removes it from the meta-gradient, aside from the final task solution. | Computationally less expensive than MAML; performs competitively on standard benchmarks.                                                                  | Lacks the convergence guarantees of second-order MAML.                                                                                                                                                |
| iMAML [9]                | Implicit Iterative Optimization       | Learn scale and shift parameters specific to tasks to modulate pre-trained deep network weights.                                                                                                                                                 | Using a constant number of memory accesses per iteration of the loop. Memory access that does not change with the number of iterations of the Task-level. | Needs the Task-level optimization to be fully converted to a local optimum (accurate gradients).                                                                                                      |
| Meta-Transfer [6]        | Scaled/Shifted Iterative Optimization | Compute signSGD instead of standard SGD at the task level, with only the direction (sign) of the gradient.                                                                                                                                       | Freeze the backbone to enable very deep networks (such as ResNet) in meta-learning.                                                                       | It is important to have a large, diverse base set for pre-training that will impact performance. Gradient quantization can sometimes fail to find the optimal convergence in complex loss landscapes. |
| SignMAML [22]            | Sign-based Iterative Optimization     | Reformulate the learning problem as a minimax game so that the model works as a minimax player to guarantee good performance in the worst-case tasks.                                                                                            | Give a first-order method that is more robust than simple approximations, with a theoretical background.                                                  | Robustness constraints can compromise average-case performance for protection against outliers.                                                                                                       |
| Task-Robust MAML [23]    | Minimax Iterative Optimization        | Select easy to hard curriculum for samples; creates challenging tasks with a ‘BrotherNet’.                                                                                                                                                       | Provides constant improvements over standard random episodic sampling.                                                                                    | The curriculum is designed with many complexities related to the domain area.                                                                                                                         |
| CubMeta/SepMeta [24, 25] | Curriculum Iterative Optimization     | Adversarial task augmentation using a Stackelberg game; Explicit generative task-id model.                                                                                                                                                       | Shift in the tasks, and little data are highly robust.                                                                                                    | High training overhead because of an additional generative adversarial loop.                                                                                                                          |
| ARMetaLearn [26]         | Adversarial Iterative Optimization    |                                                                                                                                                                                                                                                  |                                                                                                                                                           |                                                                                                                                                                                                       |

Note: MAML = Model-Agnostic Meta-Learning; iMAML = implicit Model-Agnostic Meta-Learning; MTL = Meta-Transfer Learning.

### 3.2.2 Gradient/optimization-based methods

Table 4 summarizes the covered adaptation-operator types and mechanisms along with the advantages and disadvantages of the different gradient- or optimization-based algorithms presented in this subsection.

### 3.2.3 Model-based/memory methods

Model-based methods, such as memory-augmented networks, hypernetworks, and amortized task posteriors, use a feed-forward function that is learned to model  $\mathcal{A}$ . For example, He et al. [12] and Gharoun et al. [13] are works done by representatives. Lin et al. [27] introduced the Metadropout framework that unifies episodic and pretrain/fine-tune paradigms with structured dropout on transferable knowledge,

and shows a plug-and-play unification with marginal cost over strong baselines.

### 3.2.4 In-context learning and large language models

Table 5 outlines key methods for LLMs. Beyond the four methods described, instruction-tuned and multimodal LLMs apply the in-context adaptation operator to multi-turn and cross-modal prompting.

### 3.2.5 Data augmentation and feature hallucination

Table 6 summarizes the feature-hallucination methods and data augmentation discussed in this subsection, describing the types of adaptation operators, their mechanisms, and limits.

**Table 5.** Survey of in-context learning (ICL) methods in large language models (LLMs)

| Algorithm              | <i>A</i> Type                     | Mechanism                                                                                                                                                                | Advantages                                                                  | Limits                                                                                                                      |
|------------------------|-----------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------|
| MetaICL [28]           | Learned feed-forward conditioning | A language model meta-trained on 142 natural language processing (NLP) tasks makes predictions from in-context examples alone at test time, without further fine-tuning. | Approaches the performance of full fine-tuning with 8× fewer parameters.    | Requires a very diverse multi-task corpus to be effective for meta-training.                                                |
| In-context tuning [29] | Learned Feed-forward              | Recast adapts the task of sequence prediction by adding examples, instructions, and queries together to fine-tune for a variety of tasks.                                | Excellent in the NLP benchmarks and scales well with the size of the model. | The computational overhead increases with the number of shots (and hence with the prompt length), which increases linearly. |
| Meta-in-context [30]   | Implicit attention                | Improve ICL by recursively giving a meta-context of previous ICL episodes when inferring to reshape task priors.                                                         | Does not need any fine-tuning or weight updating at test time.              | Limited context length; currently only examined on toy domains.                                                             |
| CoOp [31]              | Constrained Gradient              | Freezes the backbone while optimizing learnable context tokens by learning them using 1–16 shots.                                                                        | Produces up to 45% improvement over handcrafted prompts.                    | Learned prompts can be specific to the source distribution, adversely impacting out-of-distribution generalization.         |

**Table 6.** Survey of data augmentation and feature hallucination methods

| Algorithm                         | <i>A</i> Type                           | Mechanism                                                                                                                                                                          | Advantages                                                                                       | Limits                                                                         |
|-----------------------------------|-----------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------|
| AFHN [32]                         | Learned Feed-forward (Generative)       | Uses a conditional Wasserstein generative adversarial network (cWGAN) with a classification regularize and anti-collapse mechanism to hallucinate various discriminative features. | Explicit control of the diversity and discriminability of features.                              | GAN instability and needs good-quality annotated base data.                    |
| Saliency Hallucination [33]       | Learned Feed-forward (Structural)       | Use a two-stream feature mixer, after automatic foreground/background split, in order to guide the mixing using a saliency detector for foreground/background feature maps.        | Easy to train, low requirements, no need for complex GAN training.                               | The accuracy of the saliency detector plays a big role in feature quality.     |
| Tensor Feature Hallucination [34] | Learned Feed-forward (Spatial)          | Generates spatial tensor features (rather than vectors) using a lightweight generator and a simple loss function to outperform complex GANs.                                       | Simple design beats complex GANs with respect to preserving spatial relations.                   | Limited to visual transfer; multimodal transfer is not yet possible.           |
| Proto Meta-Hallucination [35]     | Learned Feed-forward (Metric-targeting) | Meta-learns prototypes using inter-sample differences, uses a difference-distribution generator and meta-learned re-weighting.                                                     | Adjusts to the unique features of each class dynamically.                                        | Adds more difficulty to the meta-training process than normal metric training. |
| SHKT [36]                         | Learned Feed-forward (Cross-modal)      | Symmetrically hallucinates images and text with cross-modal knowledge transfer (KT) with semantic relations in both modalities.                                                    | Achieves great improvements (+12.84% 1-shot accuracy on Mini ImageNet) via multimodal relations. | Needs certain matched pairs of textual comments to operate.                    |

**Table 7.** Survey of cross-domain few-shot learning (CD-FSL) methods

| Algorithm                                                     | <i>A</i> Type                                 | Mechanism                                                                                                                                                                                    | Advantages                                                                                                                                               | Limits                                                                                              |
|---------------------------------------------------------------|-----------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------|
| ReFine [37]                                                   | Iterative Optimization (Initialization-Reset) | Prior to fine-tuning the target, propose some source parameters randomly to stop the model from "defect inheritance" and avoid being stuck in a local minimum in the source space.           | It is very effective in large domain gaps, where source features may be detrimental to target learning.                                                  | Loses valuable source knowledge when the target domain is close to the source.                      |
| Local-global Distillation Prototypical Network (LDP-Net) [38] | Learned Feed-forward (Local-Global)           | Distill and avoid "simplicity-bias" shortcuts in a two-branch network for local-global distillation, forcing the model to distill knowledge from images of the whole world into local crops. | It does not let the model rely on spurious global correlations.                                                                                          | The dual-branch architecture boosts computational and inference expenses.                           |
| Loss Landscape Flattening [39]                                | Representation-centric Regularization         | Normalizes and interpolated-representation flattens the representation-space loss landscape, which is lightweight and makes feature transfer stable.                                         | Achieves up to 9% gains compared to state of the art (SOTA); can be used with convolutional neural network (CNN) and vision transformer (ViT) backbones. | The flattening radius must be carefully tuned; there is another layer of hyperparameter complexity. |
| StyleAdv [40]                                                 | Iterative Optimization (Adversarial Style)    | Adversarial perturbs the visual styles by using a signed style gradient attack to generate "hard" fictitious domains in the training loop.                                                   | Flexible and powerful; excellent performance at the state-of-the-art level on 8 different target datasets.                                               | The adversarial augmentation loop takes much longer to train and is more expensive.                 |

### 3.2.6 Cross-domain few-shot learning

Table 7 provides a review of few-shot algorithms specifically developed for cross-domain few-shot learning (CD-FSL) that focus on adaptation strategies and known limitations when faced with distribution shifts.

## 4. META-ANALYSIS

The reported accuracy of few-shot tasks has been found to be susceptible to experimental variables like backbone architecture (Conv-4, ResNet-10, ResNet-12, ResNet-18, WRN-28), data-augmentation pipeline, and random seed, making direct comparison across papers unreliable [41]. We applied three inclusion criteria to control variability: benchmarking on standard miniImageNet 5-way, a ResNet-12 backbone, and an episodic evaluation protocol with both 1-shot and 5-shot accuracy reported. For the purposes of a meta-analysis, the 28 published results that met all three criteria were used. The reported results were not re-evaluated under standardized training settings, augmentation pipelines, or optimizers. Therefore, this analysis compares published results under matched benchmark and backbone conditions rather than results from a controlled experiment; differences may still reflect variations in experimental protocols.

### 4.1 In-domain results (miniImageNet, ResNet-12 backbone)

Table 8 shows meta-analysis results on miniImageNet 5-shot with a standard ResNet-12 backbone. We evaluate 1-shot and 5-shot accuracy for each method family.

In this sample, the metric-based methods achieved an average in-domain accuracy approximately 3.7 percentage points (pp) higher than the optimization-based methods (this is not a rule for all samples of papers). SHKT achieved the highest reported accuracy among the methods considered, with data augmentation now a distinct category. The method implicit Model-Agnostic Meta-Learning (iMAML) substantially narrows the gap with metric-based methods.

These latter methods can remove the task-level memory bottleneck.

**Table 8.** In-domain benchmark results (miniImageNet, ResNet-12 backbone)

| Method        | Family        | 1-Shot Accuracy (%) | 5-Shot Accuracy (%) |
|---------------|---------------|---------------------|---------------------|
| ProtoNet [10] | Metric        | 60.37 $\pm$ 0.83    | 78.02 $\pm$ 0.57    |
| MAML [1]      | Optimization  | 58.90 $\pm$ 1.01    | 71.50 $\pm$ 0.72    |
| Reptile [8]   | Optimization  | 57.42 $\pm$ 0.72    | 70.10 $\pm$ 0.60    |
| iMAML [9]     | Optimization  | 60.20 $\pm$ 0.90    | 74.10 $\pm$ 0.71    |
| MTL [6]       | Optimization  | 61.20 $\pm$ 1.80    | 75.50 $\pm$ 0.80    |
| SAPENet [21]  | Metric        | 65.84 $\pm$ 0.71    | 82.35 $\pm$ 0.48    |
| SHKT [36]     | Hallucination | 67.31 $\pm$ 0.65    | 83.41 $\pm$ 0.44    |

Note: SHKT = Symmetric Hallucination with Knowledge Transfer; MTL = Meta-Transfer Learning; MAML = Model-Agnostic Meta-Learning; iMAML = implicit Model-Agnostic Meta-Learning; MTL = Meta-Transfer Learning.

### 4.2 Cross-domain results of cross-domain few-shot learning benchmark

Table 9 reports 1-shot accuracy on four CD-FSL target datasets (EuroSAT, CropDisease, ISIC, and ChestX, with miniImageNet as the source) for representative optimization-based, representation-centric, and metric-based methods, illustrating a reversal of the in-domain ranking under cross-domain shift.

Across this sample, there is a 1.4 pp difference between the methods that add robustness to task-distribution shift, like the optimization-based one, and the standard baseline (metric-based) methods shown here. This aligns with the hypothesis that gradient-based adaptation generalizes better when meta-training and meta-test distributions differ. For loss landscape flattening, Local-global Distillation Prototypical Network (LDP-Net) and loss landscape flattening [39] perform best across domains, indicating that representation-level interventions have more influence than algorithmic interventions on distribution shift.

**Table 9.** Cross-domain 1-shot accuracy on the cross-domain few-shot learning (CD-FSL) benchmark

| Methods                                                       | EuroSAT 1-Shot Accuracy (%) | CropDisease 1-Shot Accuracy (%) | ISIC 1-Shot Accuracy (%) | ChestX 1-Shot Accuracy (%) |
|---------------------------------------------------------------|-----------------------------|---------------------------------|--------------------------|----------------------------|
| Local-global Distillation Prototypical Network (LDP-Net) [38] | 70.1 $\pm$ 0.8              | 88.5 $\pm$ 0.5                  | 51.2 $\pm$ 0.9           | 25.4 $\pm$ 0.6             |
| Loss Landscape Flattening [39]                                | 71.3 $\pm$ 0.7              | 89.1 $\pm$ 0.4                  | 52.4 $\pm$ 0.8           | 26.1 $\pm$ 0.5             |
| StyleAdv [40]                                                 | 69.8 $\pm$ 0.9              | 87.3 $\pm$ 0.6                  | 50.7 $\pm$ 0.9           | 24.9 $\pm$ 0.7             |
| Task-Robust Model-Agnostic Meta-Learning (MAML) [23]          | 68.4 $\pm$ 1.1              | 85.9 $\pm$ 0.7                  | 50.1 $\pm$ 1.0           | 24.3 $\pm$ 0.8             |
| ProtoNet [10]                                                 | 66.8 $\pm$ 0.9              | 84.7 $\pm$ 0.6                  | 48.6 $\pm$ 0.9           | 23.1 $\pm$ 0.6             |

## 5. APPLICATIONS AND BENCHMARKS

### 5.1 Standard benchmarks

Since 2016, few-shot learning research has relied on a small number of standard benchmarks. One is miniImageNet, with 600 classes and 100 images from each class in ImageNet. TieredImageNet and Omniglot, with 20 categories and 20 images for each class, are also used for 1-shot benchmarks. A new benchmark was introduced at the Conference on Computer Vision and Pattern Recognition (CVPR) in 2020,

allowing evaluations across different domains. This includes 8 datasets with different types of images, such as natural images, satellite images, medical images, and images of crop diseases.

### 5.2 Medical imaging

Alsaleh et al. [42] evaluated MAML on liver, spleen, and kidney segmentation, achieving mean Dice coefficients of 93.70%, 85.98%, and 89.58% in the 10-shot setting. Khadka et al. [43] used polyp and skin lesion segmentation, achieving 2–4% higher Dice than standard MAML. Bayesian meta-

learning methods are promising for clinical decision-making.

### 5.3 Natural language processing

Few-shot learning and meta-learning in natural language processing (NLP) have merged. Few-shot relation extraction of meta-learning and prompt learning is explored in the survey of Hang et al. [44]. Pan et al. [45] introduced prototype tuning-document-level relation extraction using episodic training with prototype-based in-context prompts. In addition, a Meta-Training model across 12 relation extraction datasets for enhanced zero- and few-shot relation extraction is proposed by Li et al. [46] and benefits from scaling with the model size.

### 5.4 Few-shot class-incremental learning

Few-shot class-incremental learning (FSCIL) is a few-shot learning approach with the goal of learning new classes while preserving previously learned knowledge. Tian et al. [47] organized FSCIL methods into five categories, including traditional machine learning, meta-learning, feature space methods, replay-based techniques, and dynamic network structures. It is used in a wide range of applications, such as image classification, object detection and segmentation, NLP, and graph learning.

## 6. OPEN PROBLEMS AND FUTURE DIRECTIONS

We formulate five open problems as conjectures along with suggested research directions with the description of gaps.

**Conjecture 1** (Model-Agnostic Meta-Learning-low-rank adaptation convergence at high rank): We hypothesize that when the LoRA rank  $r$  approaches the full rank,  $r = \min(d_1, d_2)$ , LoRA and unconstrained one-step gradient descent converge to similar task-specific solutions. This conjecture is divided into two claims: representational and optimization. In Section 2.5, the representational claim is demonstrated at full rank: LoRA's updates are identical to those of an unconstrained one-step gradient update. The conjecture is partially solved, but it remains unknown whether gradient descent on low-rank factors  $(B, A)$  converges to a solution equal to full gradient descent fine-tuning (MAML/Reptile-style) or merely represents it. We propose testing the remaining by systematically varying  $r$  in miniImageNet and comparing convergence speed and final accuracy of trained LoRA factors against full gradient-based fine-tuning.

**Conjecture 2** (ICL as implicit meta-learning): We conjecture, but do not claim to have shown, that ICL with a large transformer is a special case of bilevel learning, where  $A = \text{Attn}(S_r)$ . It remains to be answered in theory [48]. We propose comparing gradient flows of examples in-context with explicit MAML updates on matched tasks, and comparing the correlation of task-specific representations from each.

**Conjecture 3** (Geometry of task distribution): The Wasserstein-2 distance  $W_2(T_{\text{train}}, T_{\text{test}})$  in task-embedding space upper-bounds the generalization gap between the meta-train and meta-test datasets. We propose to create controlled distribution shifts of known  $W_2$  distance and evaluate the empirical error in accuracy for all four method families.

**Conjecture 4** (Curriculum ordering and convergence): Under mild smoothness assumptions on the loss landscape, under the same conditions of convergence rate, the curriculum-based meta-learning (CubMeta/SepMeta)

outperforms random task sampling by an error rate of  $O(\log T/T)$ . We propose that convergence curves are similar for three curriculum schedules (easy-to-hard, hard-only, and random) when both compute budgets are equal.

**Conjecture 5** (Hallucination bias amplification): The feature hallucination techniques exacerbate the original class imbalance in the support set, making the prototypes of minority sub-populations more concentrated for the modes of the generator, by a factor proportional to the inverse of the latent space variance of the generator. We propose to compute class accuracy variation before and after a hallucination for artificially balanced support sets; use disagreement between the features in an ensemble to give a measure of uncertainty in the generated features.

There are further directions including (a) Online/continual meta-learning: the meta-learner itself adapts to a non-stationary task stream [3]; (b) scalable gradient-based meta-learning for billion-parameter models: second-order MAML is impractical; however, the gap between it and other alternative approaches such as implicit differentiation (iMAML) and first-order approximations on complex tasks is not characterized; (c) diffusion-model data augmentation for few-shot settings, fusing the synthesis quality of modern generative models with the diversity control of feature hallucination.

## 7. CONCLUSION

The paper derives that all three families of meta-learning are instantiations of a single bilevel objective with different adaptation operators. We prove a formal proof result: at full adapter rank, the reachable set of LoRA updates coincides exactly with that of an unconstrained one-step gradient update, formally resolving the representational half of the MAML-LoRA convergence question. In addition, we extend the framework to include ICL, PEFT/LoRA, Bayesian meta-learning, and diffusion-based augmentation, broadening the taxonomy of adaptation operators. We also perform a meta-analysis of the results of the benchmarks based on a ResNet-12 backbone, indicating a 3.7 pp advantage for metric-based methods in domain and a 1.4 pp advantage for optimization-based methods under cross-domain shift. Furthermore, we present five open problems as testable statements with designed experiments.

Finally, the alignment between the distribution of tasks in the meta-training task and the target task distribution determines the quality of the generalization in all paradigms. This latter can be enhanced in curriculum learning by adding adversarial tasks and adaptive sampling. LLMs introduce a new instantiation of the bilevel framework (adaptation by attention) whose relationship to classical meta-learning remains an open direction for future theoretical work.

## REFERENCES

- [1] Finn, C., Abbeel, P., Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In Proceedings of the 34th International Conference on Machine Learning (ICML), PMLR, Sydney, Australia, pp. 1126-1135. <https://proceedings.mlr.press/v70/finn17a.html>.
- [2] Hospedales, T.M., Antoniou, A., Micaelli, P., Storkey, A.J. (2022). Meta-learning in neural networks: A survey.

- IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(9): 5149-5169. <https://doi.org/10.1109/TPAMI.2021.3079209>
- [3] Vettoruzzo, A., Bouguelia, M.R., Vanschoren, J., Rögnvaldsson, T., Santosh, K. (2024). Advances and challenges in meta-learning: A technical review. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(7): 4763-4779. <https://doi.org/10.1109/TPAMI.2024.3357847>
  - [4] Wang, Y.Q., Yao, Q.M., Kwok, J.T., Ni, L.M. (2020). Generalizing from a few examples: A survey on few-shot learning. ACM Computing Surveys, 53(3): 1-34. <https://doi.org/10.1145/3386252>
  - [5] Lake, B.M., Ullman, T.D., Tenenbaum, J.B., Gershman, S.J. (2017). Building machines that learn and think like people. Behavioral and Brain Sciences, 40: e253. <https://doi.org/10.1017/S0140525X16001837>
  - [6] Sun, Q., Liu, Y., Chua, T.S., Schiele, B. (2019). Meta-transfer learning for few-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, pp. 403-412. <https://doi.org/10.1109/CVPR.2019.00049>
  - [7] Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., Wierstra, D. (2016). Matching networks for one shot learning. Advances in Neural Information Processing Systems (NeurIPS), 29. [https://proceedings.neurips.cc/paper\\_files/paper/2016/hash/90e1357833654983612fb05e3ec9148c-Abstract.html](https://proceedings.neurips.cc/paper_files/paper/2016/hash/90e1357833654983612fb05e3ec9148c-Abstract.html)
  - [8] Nichol, A., Achiam, J., Schulman, J. (2018). On first-order meta-learning algorithms. arXiv preprint arXiv:1803.02999. <https://doi.org/10.48550/arXiv.1803.02999>
  - [9] Rajeswaran, A., Finn, C., Kakade, S.M., Levine, S. (2019). Meta-learning with implicit gradients. Advances in Neural Information Processing Systems, 32. [https://proceedings.neurips.cc/paper\\_files/paper/2019/hash/072b030ba126b2f4b2374f342be9ed44-Abstract.html](https://proceedings.neurips.cc/paper_files/paper/2019/hash/072b030ba126b2f4b2374f342be9ed44-Abstract.html)
  - [10] Snell, J., Swersky, K., Zemel, R. (2017). Prototypical networks for few-shot learning. Advances in Neural Information Processing Systems, 30. [https://proceedings.neurips.cc/paper\\_files/paper/2017/hash/cb8da6767461f2812ae4290eac7cbc42-Abstract.html](https://proceedings.neurips.cc/paper_files/paper/2017/hash/cb8da6767461f2812ae4290eac7cbc42-Abstract.html)
  - [11] Sung, F., Yang, Y.X., Zhang, L., Xiang, T., Torr, P.H.S., Hospedales, T.M. (2018). Learning to compare: Relation network for few-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, pp. 1199-1208. <https://doi.org/10.1109/CVPR.2018.00131>
  - [12] He, K., Pu, N., Lao, M.R., Lew, M.S. (2023). Few-shot and meta-learning methods for image understanding: A survey. International Journal of Multimedia Information Retrieval, 12(2): 14. <https://doi.org/10.1007/s13735-023-00279-4>
  - [13] Gharoun, H., Momenifar, F., Chen, F., Gandomi, A.H. (2023). Meta-learning approaches for few-shot learning: A survey of recent advances. arXiv preprint arXiv:2303.07502. <https://doi.org/10.48550/arXiv.2303.07502>
  - [14] Hu, E.J., Shen, Y.L., Wallis, P., et al. (2022). LoRA: Low-rank adaptation of large language models. In Proceedings of the International Conference on Learning Representations (ICLR), 4(1).
  - [15] Shuttlesworth, R., Andreas, J., Torralba, A., Sharma, P. (2025). LoRA vs full fine-tuning: An illusion of equivalence. Advances in Neural Information Processing Systems, 38: 193421-193456. <https://doi.org/10.52202/085713-5808>
  - [16] Zindari, A., Mulayoff, R., Stich, S.U. (2026). LoRA vs. full fine-tuning: A theoretical perspective. arXiv preprint arXiv:2605.19018. <https://doi.org/10.48550/arXiv.2605.19018>
  - [17] Kim, T., Yoon, J., Dia, O., Kim, S., Bengio, Y., Ahn, S. (2018). Bayesian model-agnostic meta-learning. arXiv preprint arXiv:1806.03836. <https://doi.org/10.48550/arXiv.1806.03836>
  - [18] Le, T.A., Kim, H., Garnelo, M., Rosenbaum, D., Schwarz, J., Teh, Y.W. (2018). Empirical evaluation of neural process objectives. In NeurIPS Workshop on Bayesian Deep Learning, Montreal, Canada, pp. 1-9. <https://openreview.net/pdf?id=8DEYng45Ts>
  - [19] Patacchiola, M., Turner, J., Crowley, E.J., O'Boyle, M., Storkey, A.J. (2020). Bayesian meta-learning for the few-shot setting via deep kernels. Advances in Neural Information Processing Systems, 33: 16108-16118. [https://proceedings.neurips.cc/paper\\_files/paper/2020/hash/b9cfe8b6042cf759dc4c0cccb27a6737-Abstract.html](https://proceedings.neurips.cc/paper_files/paper/2020/hash/b9cfe8b6042cf759dc4c0cccb27a6737-Abstract.html)
  - [20] Koch, G., Zemel, R., Salakhutdinov, R. (2015). Siamese neural networks for one-shot image recognition. In ICML Deep Learning Workshop, Lille, France, pp. 1-30. <https://www.cs.utoronto.ca/~rsalakhu/papers/oneshot1.pdf>
  - [21] Huang, X., Choi, S.H. (2023). SAPENet: Self-attention based prototype enhancement network for few-shot learning. Pattern Recognition, 135: 109170. <https://doi.org/10.1016/j.patcog.2022.109170>
  - [22] Fan, C., Ram, P., Liu, S.J. (2021). Sign-MAML: Efficient model-agnostic meta-learning by signSGD. arXiv preprint arXiv:2109.07497. <https://doi.org/10.48550/arXiv.2109.07497>
  - [23] Collins, L., Mokhtari, A., Shakkottai, S. (2020). Task-robust model-agnostic meta-learning. Advances in Neural Information Processing Systems, 33: 18860-18871. [https://proceedings.neurips.cc/paper\\_files/paper/2020/hash/da8ce53cf0240070ce6c69c48cd588ee-Abstract.html](https://proceedings.neurips.cc/paper_files/paper/2020/hash/da8ce53cf0240070ce6c69c48cd588ee-Abstract.html)
  - [24] Zhang, J., Song, J.K., Yao, Y.Z., Gao, L.L. (2021). Curriculum-based meta-learning. In Proceedings of the 29th ACM International Conference on Multimedia (MM '21), New York, NY, USA, pp. 1838-1846. <https://doi.org/10.1145/3474085.3475335>
  - [25] Zhang, J., Song, J.K., Gao, L.L., Liu, Y., Shen, H.T. (2022). Progressive meta-learning with curriculum. IEEE Transactions on Circuits and Systems for Video Technology, 32(9): 5916-5930. <https://doi.org/10.1109/TCSVT.2022.3164190>
  - [26] Wang, Q., Lv, Y.Q., Mao, Y.X., Qu, Y., Xu, Y., Ji, X.Y. (2025). Robust fast adaptation from adversarially explicit task distribution generation. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1 (KDD '25), New York, NY, USA, pp. 1481-1491. <https://doi.org/10.1145/3690624.3709337>
  - [27] Lin, S.B., Zeng, X.Y., Zhao, R. (2022). A unified

- framework with meta-dropout for few-shot learning. arXiv preprint arXiv:2210.06409. <https://doi.org/10.48550/arXiv.2210.06409>
- [28] Min, S., Lewis, M., Zettlemoyer, L., Hajishirzi, H. (2022). 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, Association for Computational Linguistics, Seattle, WA, USA: pp. 2791-2809. <https://doi.org/10.18653/v1/2022.naacl-main.201>
- [29] Chen, Y.D., Zhong, R.Q., Zha, S., Karypis, G., He, H. (2022). Meta-learning via language model in-context tuning. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Dublin, Ireland, pp. 719-730. <https://doi.org/10.18653/v1/2022.acl-long.53>
- [30] Coda-Forno, J., Binz, M., Akata, Z., Botvinick, M., Wang, J., Schulz, E. (2023). Meta-in-context learning in large language models. *Advances in Neural Information Processing Systems*, 36: 65189-65201. <https://doi.org/10.52202/075280-2844>
- [31] Zhou, K.Y., Yang, J.K., Loy, C.C., Liu, Z.W. (2022). Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337-2348. <https://doi.org/10.1007/s11263-022-01653-1>
- [32] Li, K., Zhang, Y.L., Li, K.P., Fu, Y. (2020). Adversarial feature hallucination networks for few-shot learning. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp. 13467-13476. <https://doi.org/10.1109/CVPR42600.2020.01348>
- [33] Zhang, H.G., Zhang, J., Koniusz, P. (2019). Few-shot learning via saliency-guided hallucination of samples. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, pp. 2765-2774. <https://doi.org/10.1109/CVPR.2019.00288>
- [34] Lazarou, M., Stathaki, T., Avrithis, Y. (2022). Tensor feature hallucination for few-shot learning. In 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, pp. 2050-2060. <https://doi.org/10.1109/WACV51458.2022.00211>
- [35] Zhang, L., Zhou, F., Wei, W., Zhang, Y.N. (2023). Meta-hallucinating prototype for few-shot learning promotion. *Pattern Recognition*, 136: 109235. <https://doi.org/10.1016/j.patcog.2022.109235>
- [36] Wang, S., Zhang, X.Y., Wang, M., He, X.N. (2025). Symmetric hallucination with knowledge transfer for few-shot learning. *IEEE Transactions on Multimedia*, 27: 1797-1807. <https://doi.org/10.1109/TMM.2024.3521802>
- [37] Oh, J., Kim, S., Ho, N., Kim, J.H., Song, H., Yun, S.Y. (2022). ReFine: Re-randomization before fine-tuning for cross-domain few-shot learning. In Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM '22), New York, NY, USA, pp. 4359-4363. <https://doi.org/10.1145/3511808.3557681>
- [38] Zhou, F., Wang, P., Zhang, L., Wei, W., Zhang, Y.N. (2023). Revisiting prototypical network for cross domain few-shot learning. In 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, pp. 20061-20070. <https://doi.org/10.1109/CVPR52729.2023.01921>
- [39] Zou, Y.X., Liu, Y.C., Hu, Y.M., Li, Y.H., Li, R.X. (2024). Flatten long-range loss landscapes for cross-domain few-shot learning. In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp. 23575-23584. <https://doi.org/10.1109/CVPR52733.2024.02225>
- [40] Fu, Y.Q., Xie, Y., Fu, Y.W., Jiang, Y.G. (2023). StyleAdv: Meta style adversarial training for cross-domain few-shot learning. In 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, pp. 24575-24584. <https://doi.org/10.1109/CVPR52729.2023.02354>
- [41] Chen, W.Y., Liu, Y.C., Kira, Z., Wang, Y.C.F., Huang, J.B. (2019). A closer look at few-shot classification. In International Conference on Learning Representations, New Orleans, USA. <https://openreview.net/pdf?id=HkxLXnAcFQ>
- [42] Alsaleh, A.M., Albalawi, E., Algosaihi, A., Albakheet, S.S., Khan, S.B. (2024). Few-shot learning for medical image segmentation using 3D U-Net and model-agnostic meta-learning (MAML). *Diagnostics*, 14(12): 1213. <https://doi.org/10.3390/diagnostics14121213>
- [43] Khadka, R., Jha, D., Hicks, S., et al. (2022). Meta-learning with implicit gradients in a few-shot setting for medical image segmentation. *Computers in Biology and Medicine*, 143: 105227. <https://doi.org/10.1016/j.compbiomed.2022.105227>
- [44] Hang, T.T., Wu, W., Feng, J., Djigal, H., Huang, J. (2025). A survey of few-shot relation extraction combining meta-learning with prompt learning. *Neurocomputing*, 647: 130534. <https://doi.org/10.1016/j.neucom.2025.130534>
- [45] Pan, D.H., Sun, Y.Y., Xu, B., et al. (2025). Prototype tuning: A meta-learning approach for few-shot document-level relation extraction with large language models. In Findings of the Association for Computational Linguistics: NAACL 2025, pp. 1112-1128. <https://doi.org/10.18653/v1/2025.findings-naacl.62>
- [46] Li, G.Z., Wang, P., Liu, J.J., et al. (2024). Meta in-context learning makes large language models better zero and few-shot relation extractors. arXiv preprint arXiv:2404.17807. <https://doi.org/10.48550/arXiv.2404.17807>
- [47] Tian, S.S., Li, L.S., Li, W.J., Ran, H., Ning, X., Tiwari, P. (2024). A survey on few-shot class-incremental learning. *Neural Networks*, 169: 307-324. <https://doi.org/10.1016/j.neunet.2023.10.039>
- [48] Reynolds, L., McDonell, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm. In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, Yokohama Japan, pp. 1-7. <https://doi.org/10.1145/3411763.3451760>