Data-Centric Meta-Learning for Robust Few-Shot Generalization

Jongmin Lim, Soobin Cha, Jaehun Park, Inho Oh, Minho Park, Kwangsu Kim*
Sungkyunkwan University, South Korea
jm.lim@skku.edu, chasoobin99@gmail.com, {pk9403, inho2007, parkminho, kim.kwangsu}@skku.edu

Abstract

Few-shot learning aims to enable rapid adaptation to unseen tasks using limited data. Optimization-based meta-learning addresses this challenge by acquiring shared prior knowledge across diverse tasks. However, its effectiveness degrades in cross-domain scenarios where unseen tasks differ significantly from training tasks. We identify this degradation as a failure to acquire generalizable prior knowledge, which is fundamentally caused by gradient discrepancies—conflicting update directions arising in the meta-training environment with diverse task distributions. To achieve robust few-shot generalization, we propose Data-Centric Meta-Learning (DCML), a novel framework that mitigates gradient discrepancies by aligning task-specific input distributions with shared prior knowledge. DCML accomplishes this alignment through a meta-learnable visual prompt that is integrated into the entire meta-learning process—unlike previous prompt-based methods restricted solely to test-time adaptation. During meta-training, the prompt transforms each task's inputs to induce more consistent gradients, thereby facilitating the learning of generalizable prior knowledge. Leveraging this robust knowledge, DCML enables rapid and parameter-efficient test-time adaptation by updating only the lightweight prompt and classifier while keeping the backbone frozen. Extensive experiments demonstrate that DCML consistently outperforms baselines, particularly in challenging few-shot cross-domain scenarios, establishing a data-centric perspective for robust meta-learning.

1. Introduction

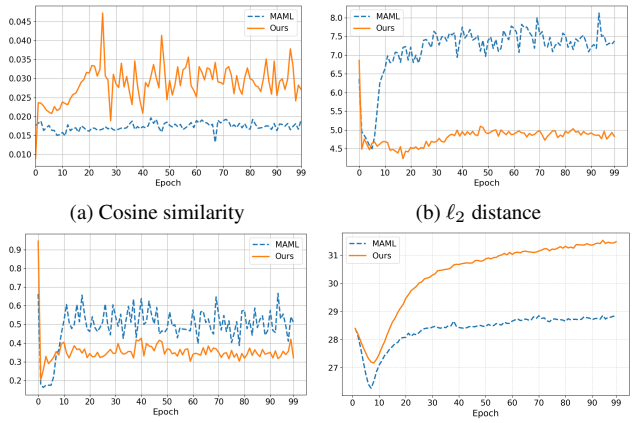
Few-shot learning aims to enable rapid adaptation to unseen tasks using only a limited number of examples, which is particularly valuable in real-world scenarios where data collection is costly or infeasible. Meta-learning, or "learning to learn," addresses this challenge by elevating the learning level from data to tasks [13, 21] and acquiring shared prior knowledge that generalizes across tasks, enabling models to adapt rapidly to previously unseen tasks. Accordingly, the objectives of meta-learning are twofold: (1) to learn shared prior knowledge that is generalizable across diverse tasks, and (2) to enable fast and efficient adaptation to unseen tasks by leveraging the acquired prior knowledge.

Among various meta-learning approaches, optimization-based meta-learning adopts a bi-level optimization framework to acquire generalizable prior knowledge. It consists of an inner-level optimization that adapts the model to each specific task, and an outer-level optimization that aggregates the learning signals across tasks to update the shared prior knowledge. For example, Model-Agnostic Meta-Learning (MAML) [9] learns prior knowledge in the form of initial parameters, enabling rapid adaptation with just a few gradient steps. Subsequent works have improved task adaptation by meta-learning not only the initial parameters but also the update rules (e.g., gradient preconditioners [16], or adaptive learning rates [33]).

Despite its effectiveness, optimization-based meta-learning often struggles in few-shot cross-domain scenarios, where the distribution of unseen tasks significantly deviates from the training tasks. We attribute this critical limitation to the difficulty in acquiring generalizable prior knowledge under meta-training environments with diverse task distributions, primarily due to gradient discrepancies—substantial directional differences among task-specific gradients computed during the inner-level optimization. Since the outer-level optimization aggregates these task-specific gradients to compute the meta-gradient, such discrepancies make the meta-gradient inconsistent, as it reflects mutually conflicting update signals across tasks, ultimately hindering the learning of robust prior knowledge.

To validate this, we analyze the relationship between the meta-gradient and task-specific gradients using cosine similarity and ℓ_2 distance. As shown in the top row of Figure 1, MAML, a representative optimization-based meta-learning method, exhibits low cosine similarity and high ℓ_2 distance across epochs, indicating substantial directional discrepancies in gradients across tasks. To further examine how such discrepancies affect the outer-level optimiza-

*Corresponding author.



(a) Cosine similarity (b) ℓ_2 distance

(c) Norm of the meta-gradient (d) Parameter shift

Figure 1. Analysis of gradient discrepancies and meta-optimization dynamics on 5-way 5-shot miniImageNet, comparing MAML and DCML (Ours). Top: Cosine similarity and ℓ_2 distance between the meta-gradient and task-specific gradients, averaged over tasks at each epoch. Bottom: Meta-gradient norm and parameter shift over epochs, where parameter shift denotes the Euclidean distance between the initial and updated parameters.

tion, we follow the analysis protocol of [22] and examine both the norm of the meta-gradient and the parameter shift from initialization over training epochs. As shown in the bottom row of Figure 1, although MAML yields relatively large meta-gradient norms, the parameter shift remains limited. This suggests that gradient discrepancies across tasks cause updates to cancel each other out, resulting in inconsistent outer-level optimization, impeding the acquisition of generalizable knowledge.

To achieve robust few-shot generalization, we propose Data-Centric Meta-Learning (DCML), a novel framework that mitigates gradient discrepancies by aligning the data distributions of training tasks with the shared prior knowledge. Specifically, DCML introduces a meta-learnable visual prompt into the entire meta-learning process. During meta-training, the visual prompt is optimized to adjust each task's input distribution to better match the shared prior knowledge, thereby producing more consistent gradient signals and facilitating the acquisition of generalizable knowledge. Leveraging the robustness of the prior knowledge acquired during training, DCML enables rapid and parameter-efficient adaptation to unseen tasks at meta-test time by updating only the lightweight prompt and classifier while keeping the backbone frozen. Integrating the visual prompt into both meta-training and meta-test phases marks a clear departure from previous prompt-based methods, which typically limit prompt usage to test-time adaptation.

To validate the effectiveness of DCML, we conduct extensive empirical evaluations on standard few-shot classification benchmarks, including both in-domain and challenging cross-domain settings. The results demonstrate that DCML achieves superior and robust performance against strong baselines, while additional analyses further highlight the advantages of our data-centric perspective toward robust few-shot generalization.

In summary, our main contributions are as follows:

- We propose Data-Centric Meta-Learning (DCML), a novel framework that enhances robust few-shot generalization by directly mitigating gradient discrepancies through data space alignment.
- To the best of our knowledge, we are the first to redefine the role of the visual prompt, treating it as a core mechanism for acquiring generalizable prior knowledge rather than a mere test-time auxiliary.
- DCML achieves superior and robust performance, consistently outperforming baselines in challenging few-shot cross-domain scenarios, ensuring parameter-efficient adaptation.

2. Related Works

2.1. Meta-learning for Few-Shot Learning

To enable fast adaptation from limited data, meta-learning approaches are commonly categorized into three types: model-based [24, 30], metric-based [32, 34, 37], and optimization-based methods [10, 11, 45]. Among these, optimization-based meta-learning has attracted considerable attention due to its model-agnostic design, enabling broad applicability to fields such as object tracking [27], viewpoint estimation [36], and medical imaging [8]. A seminal work in this area, MAML [9], proposes a framework that optimizes initial parameters that can be shared across tasks and generalized to diverse task distributions. To enhance the effectiveness and flexibility of task adaptation based on the shared initial parameters, various extensions of MAML have been proposed. ANIL [28] improves computational efficiency by freezing the backbone and updating only the classifier through a feature reuse strategy, while BOIL [25] emphasizes representation change by updating only the backbone during adaptation. Other works have further explored learning update rules that modulate the gradients of the shared initial parameters, enabling more flexible adaptation. For example, GAP [16] learns a geometry-adaptive preconditioner that satisfies the Riemannian metric properties to guide gradient updates more effectively, while Meta-AdaM [33] meta-learns an adaptive optimizer that predicts learning rates from gradient-update history.

Despite these advances, a fundamental limitation remains: it is difficult to acquire generalizable prior knowledge in meta-training environments with diverse task distributions. In response, a line of research has introduced additional network components that operate in the parameter space to adjust the shared initial parameters on a per-

task basis. For instance, MMAML [39] modulates meta-learned prior knowledge using task embeddings derived from each task, and L2F [4] attenuates irrelevant prior knowledge based on gradient information to reduce conflict among tasks. In contrast, we take a data-centric perspective that directly addresses the challenges posed by diverse task distributions. By aligning task-specific inputs with shared prior knowledge through a lightweight meta-learnable visual prompt, our approach enhances robust few-shot generalization without relying on complex parameter modulation or auxiliary networks.

2.2. Prompt-based Learning

As a newly emerging paradigm, prompt learning is a data-centric approach that aligns input data, enabling a frozen model to effectively perform a given task [15, 43]. It was initially introduced in natural language processing (NLP) to enable large-scale language models to perform diverse downstream tasks without fine-tuning [19, 20]. Following their success in NLP, prompt-based methods have been extended to vision-language models [23, 44, 47]. More recently, there has been growing interest in applying prompts to vision-only models that do not rely on textual information. In the vision domain, Visual Prompt [2] has been successfully applied by directly adding learnable patches—such as padding, random patches, or fixed patches—into the pixel space of input images. Subsequent studies on visual prompts [12, 15] have explored multi-prompt strategies, where multiple prompts are learned to address diverse data distributions or domain conditions, improving adaptability across downstream tasks. However, most existing methods [7, 42, 47] employ prompts as lightweight adaptation tools while keeping the backbone frozen. As a result, their potential to contribute to the learning of generalizable prior knowledge during training remains relatively underexplored. In contrast, our approach integrates a meta-learnable visual prompt into both meta-training and adaptation, allowing the prompt to actively participate in the training process and improve the acquisition of robust prior knowledge that generalizes across tasks.

3. Preliminaries

3.1. Problem Setting

We address the few-shot classification problem, where the objective is to enable a model to rapidly adapt to unseen tasks using only a small number of labeled examples. Meta-learning is performed over a task distribution $p(\mathcal{T})$, with training tasks $\{\mathcal{T}_i\}_{i=1}^{\tau}$ sampled to acquire generalizable prior knowledge. Each task $\mathcal{T}_i$ consists of a support set $S_i = \{(\mathbf{x}_i^s, y_i^s)\}_{s=1}^{K}$ and a query set $Q_i = \{(\mathbf{x}_i^q, y_i^q)\}_{q=1}^{M}$, where $\mathbf{x}_i \in \mathcal{X}$ denotes an input sample and $y_i \in \mathcal{Y}$ is the corresponding class label. The support set contains K labeled examples per class, while the query set contains M examples used to evaluate the task performance. Although both sets share the same set of N classes, their samples are disjoint, i.e., $S_i \cap Q_i = \emptyset$. Each task is thus formulated as an N-way K-shot classification problem and can be represented as $\mathcal{T}_i = \{S_i, Q_i\}$. At meta-test time, the model leverages the prior knowledge acquired during meta-training to rapidly adapt to an unseen task using only the support set, and its performance is evaluated on the corresponding query set. Test tasks may be sampled from the same distribution as the training tasks (i.e., few-shot in-domain classification), or from a different distribution $p'(\mathcal{T}) \neq p(\mathcal{T})$ (i.e., few-shot cross-domain classification), where adaptation becomes more challenging if the learned prior knowledge is not sufficiently generalizable.

3.2. Model-agnostic Meta-Learning

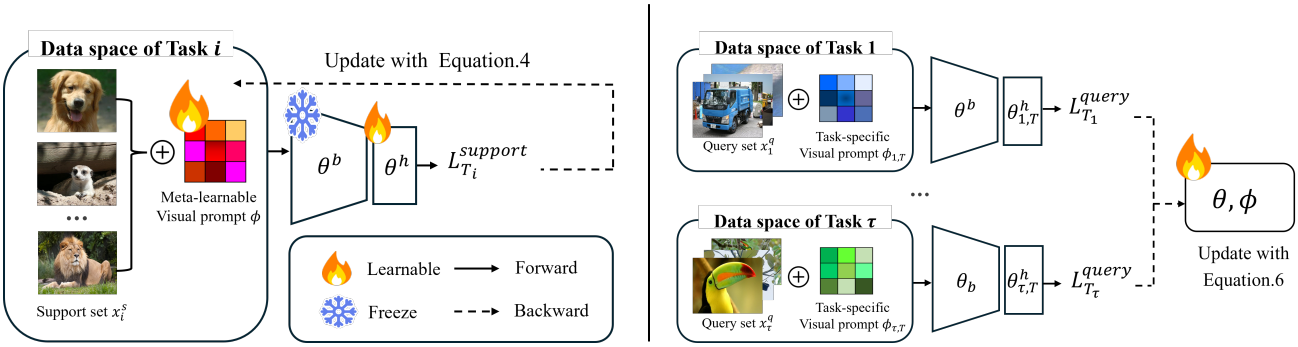
Model-agnostic Meta-Learning (MAML) offers a general framework by formulating meta-learning as a bi-level optimization problem, where the objective is to learn initial parameters θ that serve as prior knowledge. Bi-level optimization consists of two stages: an inner-level optimization for task adaptation and an outer-level optimization for updating the shared prior knowledge across tasks [17]. In the inner-level optimization phase, the model $f_\theta(\cdot)$ is initialized with θ. For each task $\mathcal{T}_i = \{S_i, Q_i\}$, the model is adapted by minimizing the support loss $\mathcal{L}_{\mathcal{T}_i}^{\text{support}}$, as follows:

$$\theta_{i,t+1} = \theta_{i,t} - \alpha \nabla_{\theta_{i,t}} \mathcal{L}_{\mathcal{T}_i}^{\text{support}}(f_{\theta_{i,t}}(\mathbf{x}_i^s), y_i^s). \quad (1)$$

where α is the inner-level learning rate and $t = 0, \ldots, T-1$ denotes the step index, with initial condition $\theta_{i,0} = \theta$. The final updated parameters $\theta_{i,T}$ are then evaluated on the query set using the query loss $\mathcal{L}_{\mathcal{T}_i}^{\text{query}}$. In the outer-level optimization phase, the initial parameters θ are updated to minimize the expected query loss across tasks:

$$\theta \leftarrow \theta - \beta \nabla_\theta \, \mathbb{E}_{\mathcal{T}_i \sim p(\mathcal{T})} \left[\mathcal{L}_{\mathcal{T}_i}^{\text{query}}(f_{\theta_{i,T}}(\mathbf{x}_i^q), y_i^q) \right]. \quad (2)$$

where β is the outer-level learning rate. While the bi-level optimization framework is designed to produce initial parameters that serve as a good starting point for fast adaptation to unseen tasks [5], the meta-gradient $\nabla_\theta \mathbb{E}_{\mathcal{T}_i \sim p(\mathcal{T})} \left[\mathcal{L}_{\mathcal{T}_i}^{\text{query}}(f_{\theta_{i,T}}(\mathbf{x}_i^q), y_i^q) \right]$ is computed by averaging task-specific gradients based on their query losses. However, when the directions of task-specific gradients differ substantially across tasks, gradient discrepancies emerge, as empirically shown in Figure 1. These discrepancies lead to an inconsistent meta-gradient that aggregates conflicting update signals across tasks. As a result, the outer-level optimization process becomes less effective in optimizing the shared initial parameters, ultimately hindering the acquisition of robust prior knowledge that generalizes well across diverse tasks.



(a) Inner-level Optimization (b) Outer-level Optimization

Figure 2. Overview of the proposed Data-Centric Meta-Learning (DCML) framework. (a) During inner-level optimization, a meta-learnable visual prompt ϕ is element-wise added to the support images of each task to produce aligned inputs. The prompt is updated together with the classifier θ^{h} in a task-specific manner by minimizing the support loss $\mathcal{L}_{\mathcal{T}_i}^{\text{support}}$, while the backbone θ^{b} remains frozen, to better align each task's input distribution with the shared prior knowledge. (b) During outer-level optimization, the resulting task-specific prompt $\phi_{i,T}$ is added to the query images. This transformation mitigates gradient discrepancies arising from diverse task distributions and promotes consistent outer-level optimization. The prior knowledge θ and the visual prompt ϕ are then jointly optimized using the aggregated query losses $\mathcal{L}_{\mathcal{T}_i}^{\text{query}}$.

4. Proposed Method: DCML

4.1. Overview

The goal of DCML is to improve the generalization of prior knowledge, represented by the initial parameters θ, by mitigating gradient discrepancies, which often arise from mismatches in input distributions across tasks. To this end, we adopt a data-centric perspective that adjusts the data space of each task to better align with the shared prior knowledge. Figure 2 illustrates the overall structure of our framework. It begins with the design of a meta-learnable visual prompt, which can be flexibly implemented using various templates—such as patch-based or padding-based structures—without restriction on its form. During inner-level optimization, the backbone is kept frozen to encourage task-specific inputs—modulated by the visual prompt— to align with the shared prior knowledge. The prompt is then element-wise added to the support set of each task and is updated in a task-specific manner to improve compatibility with the shared prior knowledge. During outer-level optimization, the task-specific updated prompt is applied to the query set of each task to align its inputs with the shared prior knowledge. This alignment reduces gradient discrepancies across tasks, enabling more consistent meta-gradient updates to the prior knowledge θ, which improves its generalization ability across tasks. The following subsections detail the inner-level and outer-level optimization procedures.

4.2. Inner-Level Optimization

The inner-level optimization begins by initializing the model $f_\theta(\cdot)$ with the shared initial parameters $\theta = \{\theta^{\text{b}}, \theta^{\text{h}}\}$, where θ^{b} and θ^{h} denote the backbone and classifier parameters, respectively, such that $f_\theta(\cdot) = f_{\theta^{\text{h}}}(f_{\theta^{\text{b}}}(\cdot))$. In conventional optimization-based meta-learning, the entire parameter set θ is adapted to each task $\mathcal{T}_i$ using the fixed support set $S_i = \{(\mathbf{x}_i^s, y_i^s)\}_{s=1}^K$, as formalized in Equation 1.

In contrast, DCML freezes the backbone θ^{b} and instead injects a meta-learnable visual prompt ϕ directly into the data space. The prompt is added element-wise to each input sample in the support set, transforming task-specific inputs to better match the shared initial parameters, as illustrated in Figure 2a. The support loss $\mathcal{L}_{\mathcal{T}_i}^{\text{support}}$ is defined as the cross-entropy loss on the prompt-aligned support set:

$$\mathcal{L}_{\mathcal{T}_i}^{\text{support}} := -\log p\left(y_i^s \mid f_{\theta_{i,t}^{\text{h}}}\left(f_{\theta^{\text{b}}}(\mathbf{x}_i^s + \phi_{i,t})\right)\right), \quad (3)$$

where t denotes the step index in the inner-level optimization, with initial conditions $\theta_{i,0} = \theta$ and $\phi_{i,0} = \phi$. At each step t, both the visual prompt ϕ and classifier θ^{h} are updated by minimizing the support loss:

$$(\phi_{i,t+1}, \theta_{i,t+1}^{\text{h}}) = (\phi_{i,t}, \theta_{i,t}^{\text{h}}) - \alpha_t \nabla_{(\phi_{i,t}, \theta_{i,t}^{\text{h}})} \mathcal{L}_{\mathcal{T}_{i,t}}^{\text{support}}. \quad (4)$$

After T inner steps, the visual prompt and classifier are adapted to task $\mathcal{T}_i$, yielding $\phi_{i,T}$ and $\theta_{i,T}^{\text{h}}$, while the backbone θ^{b} remains frozen throughout the inner-level optimization. Note that we adopt per-step learnable learning rates α_t for the visual prompt and classifier, following the practice introduced in [1] and used in [3, 33], to ensure stable training.

4.3. Outer-Level Optimization

In conventional meta-learning, outer-level optimization updates prior knowledge θ by minimizing the average query loss across training tasks, as shown in Equation 2. However, when task distributions are diverse, the resulting gradi-

Algorithm 1 Data-Centric Meta-Learning (DCML)

Require: $p(\mathcal{T})$: distribution over tasks
Require: α, β: inner/outer learning rates
Require: T: number of inner-level steps
Initialize: prior knowledge $\theta = \{\theta^{\text{b}}, \theta^{\text{h}}\}$
Initialize: a meta-learnable visual prompt ϕ
while not converged **do**
  Sample batch of tasks $\{\mathcal{T}_i\} \sim p(\mathcal{T})$
  for each task $\mathcal{T}_i$ **do**
    Sample support set $S_i = \{(\mathbf{x}_i^s, y_i^s)\}_{s=1}^{K}$
    Sample query set $Q_i = \{(\mathbf{x}_i^q, y_i^q)\}_{q=1}^{M}$
    for $t = 0$ to $T - 1$ **do**
      Apply prompt to support inputs: $\mathbf{x}_i^s + \phi_{i,t}$
      Compute support loss $\mathcal{L}_{\mathcal{T}_i}^{\text{support}}$ using Eq. 3
      Update $\phi_{i,t}$ and $\theta_{i,t}^{\text{h}}$ using Eq. 4
    end for
    Apply updated prompt to query inputs: $\mathbf{x}_i^q + \phi_{i,T}$
    Compute query loss $\mathcal{L}_{\mathcal{T}_i}^{\text{query}}$ using Eq. 5
  end for
  Update θ and ϕ using Eq. 6
end while

ent discrepancies hinder the formation of a consistent meta-gradient, ultimately limiting generalization.

To address this issue, DCML applies the task-specific prompt $\phi_{i,T}$—updated during inner-level optimization to ensure better compatibility between each task's inputs and the shared prior knowledge—to the query set of each task, as illustrated in Figure 2b. This alignment reduces gradient discrepancies across tasks and promotes more consistent meta-gradient updates. Formally, the query loss for each task is defined as:

$$\mathcal{L}_{\mathcal{T}_i}^{\text{query}} := -\log p\left(y_i^q \mid f_{\theta_{i,T}^{\text{h}}}\left(f_{\theta^{\text{b}}}(\mathbf{x}_i^q + \phi_{i,T})\right)\right), \quad (5)$$

The outer-level optimization is then performed by computing meta-gradients of the aggregated query losses across tasks:

$$(\theta, \phi) \leftarrow (\theta, \phi) - \beta \nabla_{(\theta,\phi)} \mathbb{E}_{\mathcal{T}_i \sim p(\mathcal{T})}\left[\mathcal{L}_{\mathcal{T}_i}^{\text{query}}\right]. \quad (6)$$

where β denotes the outer-level learning rate. This joint optimization not only facilitates the acquisition of generalizable prior knowledge θ, but also enables the prompt ϕ to learn how to align diverse task input distributions with the shared knowledge.

By participating in both inner- and outer-level optimization, the prompt serves as a core mechanism that actively guides the meta-training process. The full bi-level optimization is summarized in algorithm 1. Upon completion of meta-training, we evaluate how the model $f_\theta(\cdot)$, equipped with the prior knowledge guided by the visual prompt, can rapidly and parameter-efficiently adapt to unseen tasks.

5. Experiments

We evaluate the effectiveness of DCML in two few-shot classification settings: *(i)* in-domain, where test tasks share the same distribution as training tasks; and *(ii)* cross-domain, where test tasks come from a different distribution. We also present additional analyses to investigate the design principles underlying DCML.

5.1. Experimental Setup

5.1.1. Datasets.

We evaluate DCML on five widely used few-shot classification benchmarks: miniImageNet [37], tieredImageNet [29], CIFAR-FS [6], FC100 [26], and CUB [40]. Each dataset is split into meta-training, validation, and test sets following standard few-shot learning protocols. Further details and class splits are provided in the supplementary material.

5.1.2. Implementation Details.

We use two backbone architectures following prior works: a four-layer convolutional network (4-CONV) [16] and ResNet-12 [3, 33]. Following the protocol in [16], the inner-level optimization is performed for $T = 5$ steps with a learning rate of $\alpha = 0.01$, while the outer-level learning rate is set to $\beta = 0.0001$. All experiments are conducted under this setup on a single NVIDIA RTX 4090 GPU under this configuration. Further architectural details of the 4-CONV and ResNet-12 backbones, as well as other training settings, are provided in the supplementary material.

5.2. Main Results

5.2.1. Few-shot In-domain Classification

We evaluate the in-domain classification performance of DCML on four benchmark datasets—miniImageNet, tieredImageNet, CIFAR-FS, and FC100—under the 5-way 1-shot and 5-shot settings using the 4-CONV backbone. As reported in Tables 1 and 2, DCML consistently delivers strong performance across all datasets. For completeness, additional results using the ResNet-12 backbone are reported in the supplementary material, which further confirm the robustness of DCML. Importantly, the advantage of DCML becomes more pronounced under distribution shifts, as detailed in the following cross-domain evaluation.

5.2.2. Few-shot Cross-domain Classification

To evaluate generalization under distribution shifts, we conduct cross-domain classification experiments under the 5-way 5-shot setting. The model is meta-trained on miniImageNet and evaluated on two transfer scenarios: miniImageNet→CUB and miniImageNet→CIFAR-FS. As shown in Table 3, DCML achieves the highest accuracy across both transfer settings using the 4-CONV backbone. These results highlight the robustness of our approach in challenging cross-domain scenarios where ex-

Table 1. Few-shot in-domain classification accuracy on miniImageNet and tieredImageNet.

Method	miniImageNet		tieredImageNet	
	1-shot	5-shot	1-shot	5-shot
MAML [9]	48.70±1.75	63.11±0.92	49.06±0.50	67.48±0.47
ANIL [28]	47.92±0.16	63.04±0.42	49.35±0.26	65.82±0.12
BOIL [25]	49.61±0.16	66.45±0.37	48.58±0.27	69.37±0.12
L2F [4]	52.10±0.50	69.38±0.46	54.40±0.50	73.34±0.44
ALFA [3]	50.58±0.51	69.12±0.47	53.16±0.49	70.54±0.46
Sparse-MAML [38]	51.04±0.59	67.03±0.74	53.91±0.67	69.92±0.21
MeTAL [5]	52.63±0.37	70.52±0.29	54.34±0.31	70.40±0.21
CxGrad [18]	51.80±0.46	69.82±0.42	55.55±0.46	73.55±0.41
Meta-AdaM [33]	52.00±0.49	70.70±0.49	53.93±0.49	72.66±0.49
MetaProxNet+MAML [46]	53.70±1.40	70.08±0.69	54.56±1.44	71.80±0.73
ProtoNet+MGAug [41]	48.77±0.86	67.99±0.73	-	-
GAP [16]	54.86±0.85	71.55±0.61	57.60±0.93	74.90±0.68
DCML (Ours)	55.52±0.50	73.31±0.44	58.01±0.49	75.96±0.43

Table 2. Few-shot in-domain classification accuracy on CIFAR-FS and FC100.

Method	CIFAR-FS		FC100	
	1-shot	5-shot	1-shot	5-shot
MAML [9]	56.55 ± 0.45	70.10 ± 0.29	35.99 ± 0.48	47.58 ± 0.30
ANIL [28]	57.13 ± 0.47	69.87 ± 0.39	36.37 ± 0.33	45.65 ± 0.44
BOIL [25]	58.03 ± 0.43	73.61 ± 0.32	38.93 ± 0.45	51.66 ± 0.32
L2F [4]	60.35 ± 0.48	76.76 ± 0.42	38.96 ± 0.49	53.23 ± 0.48
ALFA [3]	60.56 ± 0.49	75.43 ± 0.43	38.20 ± 0.49	52.98 ± 0.50
MeTAL [5]	59.19 ± 0.56	74.62 ± 0.43	37.46 ± 0.39	51.34 ± 0.25
CxGrad [18]†	61.29 ± 0.49	77.56 ± 0.42	40.31 ± 0.49	53.89 ± 0.50
GAP [16]†	62.13 ± 0.49	78.53 ± 0.41	41.36 ± 0.49	55.53 ± 0.50
DCML (Ours)	64.42 ± 0.47	80.32 ± 0.40	42.08 ± 0.49	56.36 ± 0.50

† Reproduced results.

Table 3. Few-shot cross-domain classification accuracy under the 5-way 5-shot setting.

Method	miniImagenet	
	→ CUB	→ CIFAR-FS
MAML [9]	52.70±0.35	55.82±0.50
ANIL [28]	55.82±0.21	61.45±0.49
BOIL [25]	60.92±0.11	63.28±0.45
L2F [4]	60.89±0.22	63.73±0.48
ALFA [3]	58.35±0.25	59.76±0.49
MeTAL [5]	58.20±0.24	63.31±0.47
CxGrad [18]	63.92±0.44	64.85±0.44
GAP [16]	64.88±0.72	65.27±0.47
DCML (Ours)	66.14±0.47	67.33±0.47

isting optimization-based methods typically underperform. For completeness, experiments using the ResNet-12 backbone are presented in the supplementary material, which further confirm the consistent cross-domain generalization capability of DCML.

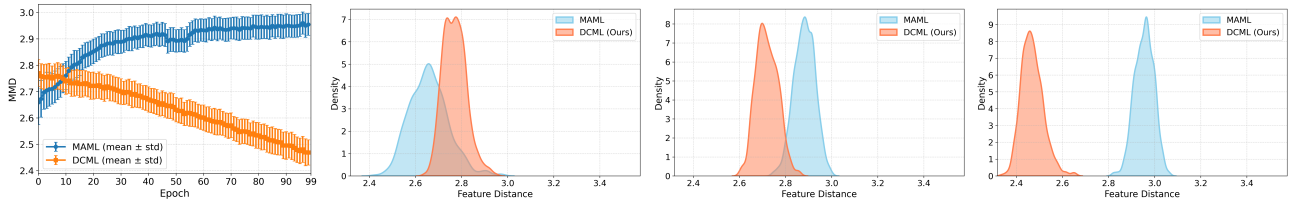
Table 4. Effect of prompt usage during meta-training and meta-test on miniImageNet under 5-way 1-shot and 5-shot settings.

Prompt Usage Configuration	1-shot	5-shot
(1) Test-time only	46.40±0.50	65.10±0.47
(2) Training-time only	50.44±0.50	72.97±0.44
(3) Training & Test-time (DCML)	55.52±0.50	73.31±0.44

5.3. Additional Analysis

5.3.1. Effect of Integrating the Visual Prompt into the Entire Meta-Learning Process

To examine whether the visual prompt ϕ contributes beyond test-time adaptation, we compare three settings designed to isolate its role during training and testing: (1) *Test-time only*, where the model is meta-trained without any prompt, following the standard MAML procedure, and the prompt is randomly initialized and adapted only at test time, evaluating the effect of using the prompt solely for adaptation; (2) *Training-time only*, where the prompt is used



(a) Maximum Mean Discrepancy (b) Epoch 1 (Early) (c) Epoch 25 (Mid-stage) (d) Epoch 99 (Late)

Figure 3. Analysis of data-space alignment across training tasks using pairwise Maximum Mean Discrepancy (MMD). (a) MMD trend averaged over tasks at each epoch, showing that MAML exhibits increasing feature distance across tasks, whereas DCML consistently reduces it throughout meta-training. (b–d) Kernel density estimates (KDEs) of MMD values at early (epoch 1), mid-stage (epoch 25), and late (epoch 99) stages. DCML progressively shifts the MMD distributions leftward while maintaining non-zero variance, indicating that tasks become more aligned with the shared prior knowledge without losing variation across tasks.

during meta-training to guide learning but removed at test time, assessing whether the prompt contributes to learning generalizable prior knowledge; and (3) *Training & Test-time (DCML)*, where the prompt is used throughout both training and testing phases. As shown in Table 4, incorporating the prompt during training notably improves performance over test-time adaptation alone, while full integration achieves the best results. This demonstrates that the prompt in DCML serves not merely as a test-time auxiliary but as a core component for learning generalizable knowledge and facilitating effective adaptation.

5.3.2. Effect of Aligning Data Spaces with the Shared Prior Knowledge

To evaluate how well the data spaces across tasks are aligned with the shared prior knowledge, we compute the pairwise Maximum Mean Discrepancy (MMD) [14] between task-specific feature distributions extracted using the shared prior knowledge. MMD serves as a statistical measure of the distance between task-specific feature distributions in the representation space. As shown in Figure 3 (a), MAML exhibits increasing MMD over training epochs, indicating that the feature distributions of tasks become progressively more distant and less aligned with the shared prior knowledge. In contrast, DCML steadily reduces MMD, demonstrating that the feature distributions of tasks become more closely aligned in the representation space. Figures 3 (b–d) further illustrate this trend through Kernel Density Estimates (KDEs) [35] of MMD distributions at early, mid-stage, and late stages of training, showing that the distributions of DCML gradually shift leftward, indicating a progressive reduction in MMD. Importantly, while DCML reduces the mean MMD, the variance of MMD values remains non-zero, indicating that the resulting alignment does not make task distributions indistinguishable. In other words, these results demonstrate that DCML encourages task data space to align toward a feature space shaped by the shared prior knowledge, while maintaining sufficient variation across tasks.

Table 5. DCML performance under different numbers of inner steps on the miniImageNet in 5-way 1-shot and 5-shot settings.

Inner Steps	1-shot	5-shot
1	54.31±0.50	68.40±0.46
2	55.48±0.50	70.32±0.46
3	56.31±0.50	71.70±0.45
4	55.21±0.50	73.08±0.44
5	55.52±0.50	73.31±0.44

5.3.3. Effect of Data Space Alignment on Outer-level Optimization

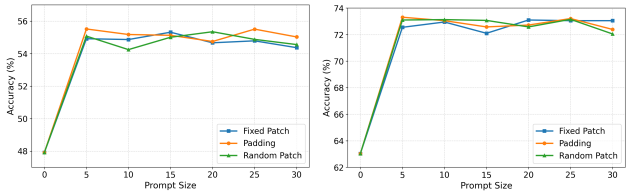
To evaluate how data space alignment influences the meta-optimization process, we revisit Figure 1, which compares DCML and MAML in terms of gradient discrepancy and optimization behavior. The top row shows cosine similarity and ℓ_2 distance between task-specific gradients and the meta-gradient, while the bottom row reports the meta-gradient norm and parameter shift over epochs. Compared to MAML, DCML aligns each task's data space with the shared prior knowledge, resulting in higher cosine similarity and lower ℓ_2 distance throughout meta-training. This alignment reduces conflicting update directions and yields more consistent meta-gradient updates. Although DCML exhibits smaller meta-gradient norms, these updates accumulate more effectively over time, producing a larger parameter shift. Intuitively, DCML takes shorter but steadier steps in a consistent direction, whereas MAML takes larger yet misaligned steps that often cancel each other out, leading to less consistent optimization toward generalizable prior knowledge.

5.3.4. Evaluation of Rapid Adaptation under Different Numbers of Inner-Level Steps

To evaluate how rapidly the model adapts to unseen tasks during adaptation, we vary the number of inner-level steps $T \in \{1, 2, 3, 4, 5\}$. As reported in Table 5, DCML maintains strong performance even with a small number of inner steps, demonstrating its ability to adapt rapidly with min-

Table 6. Learnable components during meta-test time adaptation.

Component	MAML	ANIL	BOIL	DCML (Ours)
Backbone (θ^{b})	✓	–	✓	–
Classifier (θ^{h})	✓	✓	–	✓
Visual Prompt (ϕ)	–	–	–	✓
# Parameters	4.6234×10^5	0.1601×10^5	4.4634×10^5	0.2075×10^5



(a) 5-way 1-shot (b) 5-way 5-shot

Figure 4. Impact of the template and size of the visual prompt

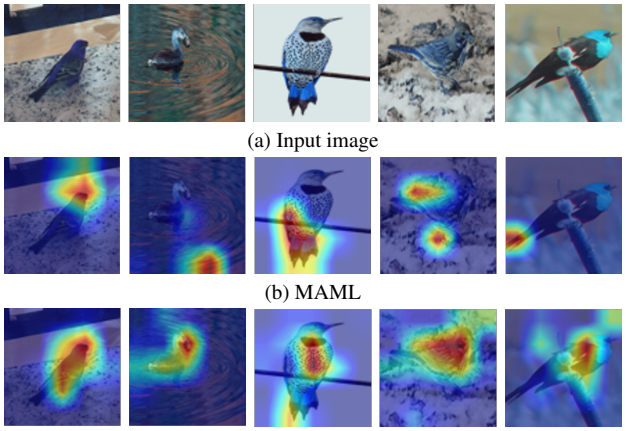
imal computation. This efficiency stems from the meta-learnable visual prompt, which is optimized during training to reduce gradient discrepancies and facilitate the acquisition of generalizable prior knowledge.

5.3.5. Evaluation of Parameter-Efficient Test-Time Adaptation

Table 6 summarizes the learnable parameters used during test-time adaptation for various meta-learning methods with a 4-CONV backbone. MAML [9] updates all parameters θ, including the backbone θ^{b} and classifier θ^{h}. BOIL [25] adapts only the backbone θ^{b}, while ANIL [28] improves efficiency by updating only the classifier θ^{h}. DCML follows a similar strategy to ANIL, but additionally updates a lightweight visual prompt ϕ alongside θ^{h}, while keeping θ^{b} frozen. Although introducing only minimal overhead, DCML outperforms all baselines in both in-domain and cross-domain settings (Tables 1 and 3). These results demonstrate that DCML achieves parameter-efficient test-time adaptation.

5.3.6. Impact of Meta-learnable Visual Prompt Design

We analyze how the template and size of the meta-learnable visual prompt ϕ affect performance, following the design strategy of [2]. The templates include padding (added to image boundaries), a fixed patch (added at a fixed location), and a random patch (added at random locations), while the prompt size p determines its spatial extent. The number of learnable parameters is computed as Cp^2 for patch-based templates and $2Cp(H+W-2p)$ for the padding template, where C, H, and W denote the image channels, height, and width. As shown in Figure 4, accuracy remains stable across different prompt templates and sizes in both 5-way 1-shot and 5-shot settings on miniImageNet, indicating that ϕ is robust to variations in prompt design. We therefore adopt the padding template with $p = 5$ as the default configuration for simplicity and efficiency.



(a) Input image (b) MAML (c) DCML

Figure 5. Grad-CAM visualizations on 5-way 5-shot miniImageNet→CUB, showing that DCML attends to more class-relevant regions and remains robust under domain shift.

5.3.7. Visualizing Adaptation Effects with Grad-CAM

To qualitatively assess cross-domain adaptation, we visualize class-discriminative regions using Grad-CAM [31], which highlights the areas the model focuses on when making predictions. Figure 5 compares activation maps from MAML and DCML on miniImageNet→CUB. MAML often focuses on irrelevant background areas or incomplete object parts, likely due to mismatches between training and test distributions. In contrast, DCML consistently focuses on class-relevant regions, even under significant domain shifts. This suggests that DCML effectively transforms the input distribution to better match the prior knowledge acquired during meta-training.

6. Conclusion

We introduced Data-Centric Meta-Learning (DCML), a framework that enhances few-shot generalization by mitigating gradient discrepancies through data-space alignment. By integrating a meta-learnable visual prompt into the meta-learning process, DCML aligns task-specific input distributions with shared prior knowledge, leading to more consistent meta-gradient updates. This enables robust prior knowledge that generalizes well across diverse tasks while allowing parameter-efficient test-time adaptation by updating only the lightweight prompt and classifier. Overall, DCML establishes a data-centric perspective for optimization-based meta-learning, paving the way toward robust few-shot generalization. Although effective, DCML currently employs a shared prompt design across tasks. Future work will explore task-adaptive prompts conditioned on task embeddings to better mitigate gradient discrepancies across diverse tasks.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No.RS-2024-00459618, Research on improving intelligent command and control capabilities based on generative AI and real-time 3D digital twin construction) and partly supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (RS-2019-II190421, AI Graduate School Support Program(Sungkyunkwan University)).

References

[1] Antreas Antoniou, Harrison Edwards, and Amos Storkey. How to train your maml. *arXiv preprint arXiv:1810.09502*, 2018. 4

[2] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 2022. 3, 8

[3] Sungyong Baik, Myungsub Choi, Janghoon Choi, Heewon Kim, and Kyoung Mu Lee. Meta-learning with adaptive hyperparameters. *Advances in neural information processing systems*, 33:20755–20765, 2020. 4, 5, 6

[4] Sungyong Baik, Seokil Hong, and Kyoung Mu Lee. Learning to forget for meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2379–2387, 2020. 3, 6

[5] Sungyong Baik, Janghoon Choi, Heewon Kim, Dohee Cho, Jaesik Min, and Kyoung Mu Lee. Meta-learning with task-adaptive loss function for few-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9465–9474, 2021. 3, 6

[6] Luca Bertinetto, Joao F Henriques, Philip HS Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. *arXiv preprint arXiv:1805.08136*, 2018. 5

[7] Wentao Chen, Chenyang Si, Zhang Zhang, Liang Wang, Zilei Wang, and Tieniu Tan. Semantic prompt for few-shot image recognition. *arXiv preprint arXiv:2303.14123*, 2023. 3

[8] Yixiong Chen, Li Liu, Jingxian Li, Hua Jiang, Chris Ding, and Zongwei Zhou. Metalr: meta-tuning of learning rates for transfer learning in medical imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 706–716. Springer, 2023. 2

[9] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017. 1, 2, 6, 8

[10] Chelsea Finn, Kelvin Xu, and Sergey Levine. Probabilistic model-agnostic meta-learning. *Advances in neural information processing systems*, 31, 2018. 2

[11] Sebastian Flennerhag, Andrei A Rusu, Razvan Pascanu, Francesco Visin, Hujun Yin, and Raia Hadsell. Meta-learning with warped gradient descent. *arXiv preprint arXiv:1909.00025*, 2019. 2

[12] Yulu Gan, Yan Bai, Yihang Lou, Xianzheng Ma, Renrui Zhang, Nian Shi, and Lin Luo. Decorate the newcomers: Visual domain prompt for continual test time adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 7595–7603, 2023. 3

[13] Hassan Gharoun, Fereshteh Momenifar, Fang Chen, and Amir H Gandomi. Meta-learning approaches for few-shot learning: A survey of recent advances. *ACM Computing Surveys*, 56(12):1–41, 2024. 1

[14] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773, 2012. 7

[15] Qidong Huang, Xiaoyi Dong, Dongdong Chen, Weiming Zhang, Feifei Wang, Gang Hua, and Nenghai Yu. Diversity-aware meta visual prompting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10878–10887, 2023. 3

[16] Suhyun Kang, Duhun Hwang, Moonjung Eo, Taesup Kim, and Wonjong Rhee. Meta-learning with a geometry-adaptive preconditioner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16080–16090, 2023. 1, 2, 5, 6

[17] Suhyun Kang, Jungwon Park, Wonseok Lee, and Wonjong Rhee. Task-specific preconditioner for cross-domain few-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 17760–17769, 2025. 3

[18] Sanghyuk Lee, Seunghyun Lee, and Byung Cheol Song. Contextual gradient scaling for few-shot learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 834–843, 2022. 6

[19] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 3045–3059, 2021. 3

[20] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, 2021. 3

[21] Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Meta-sgd: Learning to learn quickly for few-shot learning. *arXiv preprint arXiv:1707.09835*, 2017. 1

[22] Zhuang Liu, Zhiqiu Xu, Joseph Jin, Zhiqiang Shen, and Trevor Darrell. Dropout reduces underfitting. In *International Conference on Machine Learning*, pages 22233–22248. PMLR, 2023. 2

[23] Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Locoop: Few-shot out-of-distribution detection via prompt learning. *Advances in Neural Information Processing Systems*, 36:76298–76310, 2023. 3

[24] Tsendsuren Munkhdalai and Hong Yu. Meta networks. In *International conference on machine learning*, pages 2554–2563. PMLR, 2017. 2

[25] Jaehoon Oh, Hyungjun Yoo, ChangHwan Kim, and Se-Young Yun. Boil: Towards representation change for few-

shot learning. *arXiv preprint arXiv:2008.08882*, 2020. 2, 6, 8

[26] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. *Advances in neural information processing systems*, 31, 2018. 5

[27] Eunbyung Park and Alexander C Berg. Meta-tracker: Fast and robust online adaptation for visual object trackers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 569–585, 2018. 2

[28] Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of maml. *arXiv preprint arXiv:1909.09157*, 2019. 2, 6, 8

[29] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*, 2018. 5

[30] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pages 1842–1850. PMLR, 2016. 2

[31] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 8

[32] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017. 2

[33] Siyuan Sun and Hongyang Gao. Meta-adam: An meta-learned adaptive optimizer with momentum for few-shot learning. *Advances in Neural Information Processing Systems*, 36:65441–65455, 2023. 1, 2, 4, 5, 6

[34] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018. 2

[35] George R Terrell and David W Scott. Variable kernel density estimation. *The Annals of Statistics*, pages 1236–1265, 1992. 7

[36] Hung-Yu Tseng, Shalini De Mello, Jonathan Tremblay, Sifei Liu, Stan Birchfield, Ming-Hsuan Yang, and Jan Kautz. Few-shot viewpoint estimation. *arXiv preprint arXiv:1905.04957*, 2019. 2

[37] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016. 2, 5

[38] Johannes Von Oswald, Dominic Zhao, Seijin Kobayashi, Simon Schug, Massimo Caccia, Nicolas Zucchet, and João Sacramento. Learning where to learn: Gradient sparsity in meta and continual learning. *Advances in Neural Information Processing Systems*, 34:5250–5263, 2021. 6

[39] Risto Vuorio, Shao-Hua Sun, Hexiang Hu, and Joseph J Lim. Multimodal model-agnostic meta-learning via task-aware modulation. *Advances in neural information processing systems*, 32, 2019. 3

[40] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 5

[41] Ren Wang, Haoliang Sun, Yuxiu Lin, Xinxin Zhang, and Yilong Yin. Improving generalization in meta-learning via meta-gradient augmentation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 6388–6396. International Joint Conferences on Artificial Intelligence Organization, 2025. Main Track. 6

[42] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 139–149, 2022. 3

[43] Junyang Wu, Xianhang Li, Chen Wei, Huiyu Wang, Alan Yuille, Yuyin Zhou, and Cihang Xie. Unleashing the power of visual prompting at the pixel level. *arXiv preprint arXiv:2212.10556*, 2022. 3

[44] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6757–6767, 2023. 3

[45] Jaesik Yoon, Taesup Kim, Ousmane Dia, Sungwoong Kim, Yoshua Bengio, and Sungjin Ahn. Bayesian model-agnostic meta-learning. *Advances in neural information processing systems*, 31, 2018. 2

[46] Yilang Zhang and Georgios B Giannakis. Meta-learning priors using unrolled proximal networks. In *The Twelfth International Conference on Learning Representations*, 2024. 6

[47] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 3