International Journal of Computer Vision
https://doi.org/10.1007/s11263-024-02086-8

HybridPrompt: Domain-Aware Prompting for Cross-Domain Few-Shot Learning

Jiamin Wu[1] · **Tianzhu Zhang**[1] · **Yongdong Zhang**[1]

Received: 5 December 2023 / Accepted: 17 April 2024

Abstract

Cross-Domain Few-Shot Learning (CD-FSL) aims at recognizing unseen classes from target domains that vastly differ from training classes from source domains, utilizing only a few labeled samples. However, the substantial domain disparities between target and source domains pose huge challenges to few-shot generalization. To resolve domain disparities, we propose HybridPrompt, a novel architecture for *Domain-Aware Prompting* that integrates a variety of cross-domain learned prompts as knowledge experts for CD-FSL. The proposed method enjoys several merits. First, to encode knowledge from diverse source domains, several *Domain Prompts* are introduced to capture domain-specific knowledge. Subsequently, to facilitate the cross-domain transfer of valuable knowledge, a *Transferred Prompt* is specifically tailored for each target task by retrieving highly relevant Domain Prompts based on domain properties. Finally, to complement insufficient transferred information, an *Adaptive Prompt* is learned to incorporate additional target characteristics for model adaptation. Consequently, the collaboration of these three types of prompts contributes to a hybridly prompted model that achieves domain-aware encoding, transfer, and adaptation, thereby enhancing adaptability on unseen domains. Extensive experimental results on the Meta-Dataset benchmark demonstrate that our method achieves superior performance against state-of-the-art methods. The source code is available at https://github.com/Jamine-W/HybridPrompt.

Keywords Few-shot learning · Cross-domain few-shot learning · Open-world recognition · Image classification

1 Introduction

Few-Shot Learning (FSL) (Fei-Fei et al., 2006; Finn et al., 2017; Vinyals et al., 2016; Snell et al., 2017; Oreshkin et al., 2018) has emerged as a prominent topic, with the objective of learning new **unseen** concepts from only a few labeled examples. Dissimilar to conventional supervised learning approaches that require a large amount of target data, a standard few-shot model is initially trained on *base* classes to learn a good feature representation, and subsequently applied or fine-tuned on *novel* classes with limited samples. Generally, the base and novel classes are presumed to stem from an identical domain, sharing similar data distribution. However, in a more realistic cross-domain scenario, the base and novel classes may exhibit considerable discrepancies, e.g., deploying an ImageNet-pretrained model to a medical imaging or remote sensing dataset. The substantial **domain disparities** between domains pose new challenges to generalization, as it requires the few-shot model to extrapolate source domain knowledge to tasks from new domains that were not encountered during training.

To facilitate few-shot generalization across different domains, recent FSL models have shifted towards *Cross-Domain Few-Shot Learning (CD-FSL)* (Guo et al., 2020; Triantafillou et al., 2019), where the model is initially trained on base classes drawn from N_s source domains, and is then transferred to novel classes from N_t target domains that exhibit extensive domain disparities, as shown in Fig. 2). To address this challenge, recent CD-FSL methods (Li et al., 2021, 2022; Hu et al., 2022) typically follow an **Encoding-Transfer-Adaptation (ETA)** paradigm (see Fig. 1a) including (1) **source domain encoding** to encode knowledge from

Communicated by Zhun Zhong.

✉ Tianzhu Zhang
tzzhang@ustc.edu.cn

Jiamin Wu
jiaminwu@mail.ustc.edu.cn

Yongdong Zhang
zhyd73@ustc.edu.cn

[1] School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China

Published online: 24 June 2024

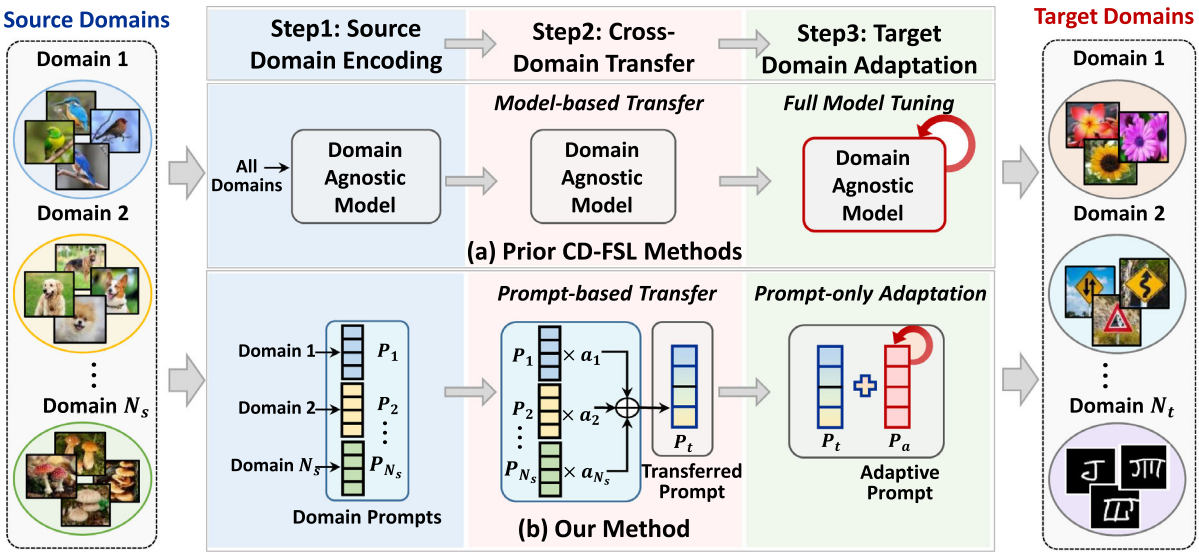

Fig. 1 Comparison between prior CD-FSL methods and ours under the ETA paradigm. **a** Prior works (Hu et al., 2022; Li et al., 2021, 2022) encode general domain knowledge by training a shared domain-agnostic model across various source domains. The learned model is directly transferred to unseen tasks from the target domain **without knowledge selection**. Then they fine-tune considerable model parameters from a tiny support set to adapt to the target task. **b** In contrast, HybridPrompt adopts a domain-aware prompting architecture by initially learning N_s **Domain Prompts** $P_1, P_2, \ldots, P_{N_s}$ to encode domain-specific information for N_s source domains, avoiding inter-domain interference. For a target task, our model can identify and aggregate highly relevant Domain Prompts using affinity scores $A = \{a_i\}_{i=1}^{N_s}$ to create the **Transferred Prompt** P_t. This selective approach allows us to transfer useful knowledge from source domains to the target task efficiently. Besides, we achieve efficient target domain adaptation by solely adapting the **Adaptive Prompt** P_a with a small number of parameters

N_s source domains; (2) **cross-domain transfer** to transfer the knowledge learned from source domains to N_t unseen target domains; (3) **target domain adaptation** to incorporate target characteristics for complementing insufficient transferred knowledge, particularly in complex target tasks. The ETA paradigm enables the model to maintain robust few-shot recognition capability when migrating across different domains.

Despite the success, the prior CD-FSL methods (Guo et al., 2020; Li et al., 2021, 2022; Hu et al., 2022) still struggle to adequately address the large domain shift issue. By diving into the essence of CD-FSL, we identify several limitations within the **ETA paradigm** (see Fig. 1a). (1) In the **encoding** step, preceding methods (Guo et al., 2020; Li et al., 2021, 2022; Hu et al., 2022) encode knowledge from diverse source domains by training a domain-agnostic feature extractor in a multi-domain learning manner. However, sharing the feature extractor across domains may inevitably introduce inter-domain interference, due to the fact that substantially varied source domains often prioritize different learning objectives. Therefore, it is crucial to incorporate ***domain-specific knowledge encoding*** for diverse source domains. (2) In the **transfer** step, a common strategy (Li et al., 2021; Hu et al., 2022; Li et al., 2022) is to apply the learned domain-agnostic model to vastly different target domains. However, directly reusing the model lacks interpretability, as it fails to identify which aspects of the source domains' knowledge are useful for the current task. Furthermore, transferring general knowledge without effective selection may lead to negative transfer, where irrelevant and even undesired knowledge hampers generalization. Thus, a ***domain-guided selection*** mechanism is essential to identify useful knowledge for proper transfer to target domains. (3) In the **adaptation** step, recent methods (Hu et al., 2022; Li et al., 2021, 2022; Requeima et al., 2019; Bateni et al., 2020) learn task-related parameters from a tiny support set by either fine-tuning the whole backbone (Hu et al., 2022) or tuning the last several network layers (Li et al., 2021). The former approach causes huge computation costs, while the latter is restricted in its ability for deep adaptation. Additionally, tuning a substantial number of parameters from only a few support samples may cause over-fitting. Therefore, an ***efficient adaptation*** mechanism tailored for data-scarce scenarios is imperative.

To overcome these limitations, we propose a domain-aware prompting architecture to ameliorate the domain disparity issue for CD-FSL. Prompting techniques (Lester et al., 2021; Brown et al., 2020; Liu et al., 2023; Shin et al., 2020) are originally introduced in the NLP field to condition the transformer by attaching a prompt to the input sequence. Different from simply learning additional prompts, we utilize prompts as **plug-and-play knowledge experts** to carry, transfer, and adapt domain knowledge,

with careful awareness of inter-domain disparities and relations. Specifically, we polish the ETA paradigm through three types of key prompts (see Fig. 1b): (1) a collection of **Domain Prompts (D-Prompts)** acting as specialized experts to encode domain-specific knowledge for source domains; (2) a **Transferred Prompt (T-Prompt)** that carries appropriate knowledge composed from highly relevant D-prompts for effectively transferring to the target domain, enforcing the transfer process concentrating on potent knowledge; (3) an **Adaptive Prompt (A-Prompt)** that captures adaptive task-related knowledge for complementing the insufficient transferred knowledge.

Motivated by the above discussion, we propose a novel architecture for **Domain-Aware Prompting**, dubbed as **HybridPrompt**, to integrate a range of cross-domain learned prompts for CD-FSL, as shown in Fig. 1b. Initially, we introduce N_s **D-Prompts** to augment a domain-agnostic Vision Transformer (ViT) for N_s source domains. Each D-Prompt ***encodes domain-specific knowledge*** by attentively interacting with instances from its corresponding source domain. This approach enables our method could learn specialized domain patterns without inter-domain interference, while still maintaining general knowledge shared within the domain-agnostic model. To fully consolidate domain knowledge, a self-supervised prompt regularization is proposed to enforce the compatibility between D-Prompts and domain instances. When encountering a target task $\mathcal{T}$, a **T-Prompt** containing effective transferable knowledge is generated by composing highly relevant D-Prompts. Specifically, we utilize the task embedding aggregated from the support features in $\mathcal{T}$ to ***selectively retrieve*** D-prompts based on inter-domain affinity scores $A = \{a_i\}_{i=1}^{N_s}$. T-Prompt can be seamlessly integrated into ViT layers to inject valuable source domain knowledge, contributing to positive transfer and dynamic task adaptation. Finally, a small-sized **A-Prompt** is optimized on the low-shot data while keeping the remaining network frozen, thus ***efficiently*** rendering an adaptive representation for challenging target tasks. The collaborative learning of D-Prompt, T-Prompt, and A-Prompt contributes to a hybridly prompted model that incorporates domain-aware encoding, transfer, and adaptation to enhance the model's adaptability and generalizability. Consequently, our method effectively boosts out-of-domain generalization capability, particularly in scenarios with substantial domain disparities, as depicted in Fig. 2.

The contributions of HybridPrompt could be summarized as follows: (1) We propose a Domain-Aware Prompting architecture that integrates a variety of cross-domain learned prompts to address significant domain disparities for CD-FSL. To the best of our knowledge, this is the first work. (2) Our HybridPrompt facilitates effective knowledge delivery from source to target domains by incorporating D-Prompts for domain-specific knowledge encoding, a T-Prompt for selective cross-domain knowledge transfer, and an A-Prompt for efficient target domain adaptation. (3) Extensive experimental results on the Meta-Dataset benchmark demonstrate that our method achieves superior performance against state-of-the-art CD-FSL methods.

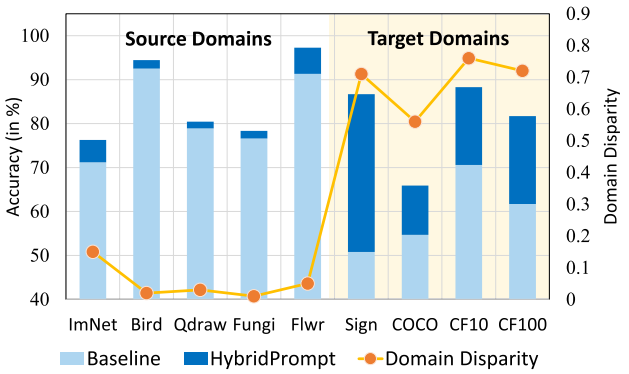

Fig. 2 (1) We illustrate the Domain Disparity of each domain, including both source and target domains. This metric is computed as the earth mover's distance (Rubner et al., 1998) between the test split of a given domain and the training data of all source domains, following (Cui et al., 2018). Notably, target domains exhibit significantly larger domain disparities from the training data of source domains. (2) We also present the accuracy improvements achieved by our proposed HybridPrompt over the baseline (non-prompted ViT, see Sect. 4.3). Our method substantially boosts the performance of target domains, particularly in scenarios with large domain disparities

2 Related Work

In this section, we introduce several lines of research in few-shot learning, cross-domain few-shot learning, and prompt learning.

2.1 Few-Shot Learning (FSL)

The literature on FSL can be mainly divided into two main streams, optimization-based methods and metric-based methods. **Optimization-based** methods (Finn et al., 2017; Antoniou et al., 2018; Ravi & Larochelle, 2017; Sun et al., 2019; Lee et al., 2019) use a meta-learner as the optimizer to update the network parameters for new tasks with limited samples available. The most representative one is MAML (Finn et al., 2017), which attempts to learn a generalizable model initialization that allows the few-shot learner to rapidly adapt to new tasks with only a few gradient descent steps. **Metric-based** methods (Vinyals et al., 2016; Snell et al., 2017; Sung et al., 2018; Tian et al., 2020; Liu et al., 2020; Simon et al., 2020; Ye et al., 2020; Wu et al., 2021) aim at learning a discriminative embedding space for the selected similarity metrics. Prototypical Network (ProtoNet) (Snell et al., 2017) computes Euclidean distances

between query samples and class prototypes obtained by averaging the embedded support samples. The few-shot prediction is performed by a nearest neighbour classifier. Several methods (Hou et al., 2019; Wu et al., 2021; Zhang et al., 2020; Wu et al., 2022) explore local parts for each sample to construct fine-grained similarity measurements, thereby deriving discriminative and transferable representations. Few-shot learning has also been actively studied for the tasks of semantic segmentation (Cheng et al., 2022; Lang et al., 2023a; 2023b; 2023c), object detection (Bulat et al., 2023; Sun et al., 2021; Zhu et al., 2021), and action recognition (Wu et al., 2022; Perrett et al., 2021; Kumar Dwivedi et al., 2019). Our method belongs to the metric-based category. However, directly transferring the above approaches into CD-FSL may be suboptimal, due to substantial inter-domain disparities. In this paper, we propose a novel approach that jointly utilizes multiple cross-domain learned prompts to build a generalizable few-shot classifier.

2.2 Cross-Domain Few-Shot Learning (CD-FSL)

A plethora of recent FSL work (Triantafillou et al., 2019; Guo et al., 2020; Hu et al., 2022; Li et al., 2021) has been dedicated to the cross-domain setting where source and target domains are significantly dissimilar. To leverage the potentially transferable knowledge from source domains, most of the previous methods (Guo et al., 2020; Li et al., 2021, 2022; Hu et al., 2022) directly apply the shared domain-agnostic model trained on source domains to target domains. However, these approaches lack the ability to discriminate useful knowledge for the current task. Other methods (Dvornik et al., 2020; Liu et al., 2021a) make improvements by designing feature aggregation mechanisms to blend the weighted features from multiple domain-specific networks. However, training a domain-specific network for each source domain is time-consuming and cumbersome. Another line of CD-FSL methods (Requeima et al., 2019; Li et al., 2022, 2021; Liu et al., 2021a; Bateni et al., 2020) adapt the model from the given limited support samples via test-time tuning. A straightforward solution is to directly fine-tune the whole network (Hu et al., 2022) or the last network layers (Li et al., 2021). However, the former incurs substantial memory and computational costs, while the latter may lead to insufficient adaptation power. Some other methods (Requeima et al., 2019; Bateni et al., 2020; Liu et al., 2021a) turn to meta-learn auxiliary networks to estimate task-related parameters from the support samples. CNAPs (Requeima et al., 2019) and SCNAPs (Bateni et al., 2020) estimate task-specific FiLM layers (including the scale and shift parameters) to obtain adapted features. However, as the auxiliary network is solely learned on source domains, it may fail to generalize to the target domain. Differently, our method enhances adaptability for target domains by constructing a domain-aware prompting architecture that achieves effective source domain encoding, selective cross-domain transfer, and efficient target domain adaptation, thereby bridging the domain gap for CD-FSL.

2.3 Prompt Learning

To employ large-scale pre-trained language models to downstream tasks, prompt learning (Liu et al., 2023; Li & Liang , 2021; Lester et al., 2021; Shin et al., 2020; Brown et al., 2020) has become topical in the natural language processing field. Initially, prompt learning prepends text instructions to the input text to enable the language model to understand the downstream task. To avoid heuristic manual prompt design, recent works (Lester et al., 2021; Li & Liang, 2021; Jia et al., 2022; Zhou et al., 2022a, 2022b) propose to learn soft prompts, *i.e.,* continuous vectors that are tailored for each downstream task, to capture high-level task characteristics for instructing the frozen models. The idea of soft prompt learning has been extended to cross-modality tasks (Zhou et al., 2022a, 2022b; Zhang et al., 2023). CoOp (Zhou et al., 2022a) and CoCoOp (Zhou et al., 2022b) build upon the vision-language model CLIP (Radford et al., 2021) by introducing soft prompt tuning, attempting to adapt the pre-trained model to various cross-modality tasks. Recently, prompt learning has also been explored in vision tasks. VPT (Jia et al., 2022) attaches randomly initialized prompt tokens to the Vision Transformer and optimizes them to incorporate instructive information. However, it is non-trivial to directly use prompt learning to adapt the cross-domain few-shot model due to limited supervised samples and non-negligible domain shift. To unveil the important values of prompt learning in CD-FSL, our method designs a collection of cross-domain learned prompts for domain-aware encoding, transfer, and adaptation, significantly enhancing the generalization capability of the model against substantial domain disparities.

3 Our Approach

In this section, we first introduce some preliminaries including the problem setting and backbone network. Then we provide a detailed introduction of the proposed HybridPrompt, which consists of three modules (as shown in Fig. 3): (1) the **Source Domain Prompt Learning** module (Sect. 3.2) learns a collection of Domain Prompts to independently encode domain-specific knowledge for source domains; (2) the **Cross-domain Prompt Transfer** module (Sect. 3.3) generates the Transferred Prompt for arbitrary target tasks by retrieving relevant Domain Prompts; (3) the **Target Domain Prompt Adaptation** module (Sect. 3.4) complements the

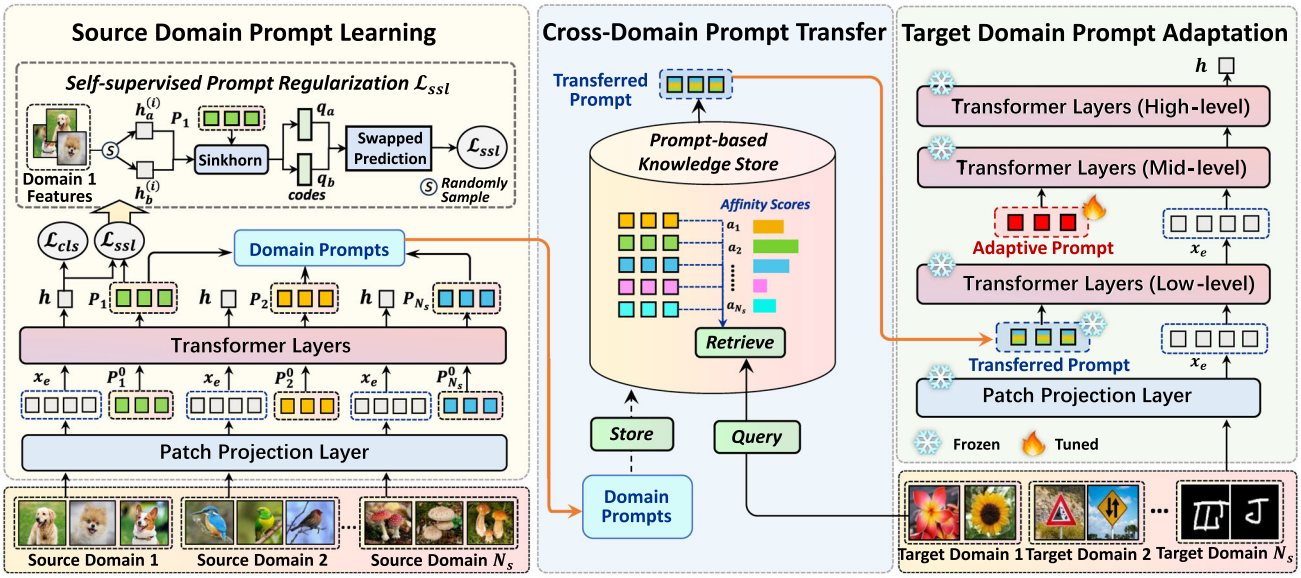

Fig. 3 The architecture of HybridPrompt: (1) The **source domain prompt learning** module initially learns N_s ***Domain Prompts (D-Prompts)*** $P_1, \ldots, P_{N_s}$ to encode domain-specific knowledge for various source domains. A self-supervised prompt regularization $\mathcal{L}_{ssl}$ is proposed to ensure alignment between the D-Prompts and corresponding domain instances. (2) The **cross-domain prompt transfer** module establishes a prompt-based knowledge store K_{store} to store D-Prompts, serving as a cross-domain bridge between the source and target domains. For each target task, a ***Transferred Prompt (T-Prompt)*** is generated by retrieving and composing highly-relevant D-Prompts stored in K_{store}, with affinity scores $\{a_1, \ldots, a_{N_s}\}$ as weights. (3) The **target domain prompt adaptation** module introduces an ***Adaptive Prompt (A-Prompt)*** and optimizes it to incorporate task characteristics from low-shot target data while keeping the remaining network frozen. Ultimately, the T- and A-Prompts are inserted into different ViT layers to simultaneously activate the transferred and adaptive knowledge

insufficient transferred information by learning an Adaptive Prompt.

3.1 Preliminaries

Problem Setting. In the general few-shot learning protocol, a large-scale meta-training set $\mathbb{D}_{train}$ is utilized to train a generalizable model capable of classifying novel classes from the meta-test set $\mathbb{D}_{test}$. Notably, $\mathbb{D}_{train}$ and $\mathbb{D}_{test}$ contain mutually exclusive classes, *i.e.,* the classes in the two sets have no overlap. In traditional few-shot learning, $\mathbb{D}_{train}$ and $\mathbb{D}_{test}$ are split from the same dataset. However, in the challenging cross-domain setting, $\mathbb{D}_{train}$ and $\mathbb{D}_{test}$ are defined over a union of diverse datasets. Thus the new classes in $\mathbb{D}_{test}$ originate from different datasets from $\mathbb{D}_{train}$, resulting in a large domain gap. Here, we denote the N_s datasets in $\mathbb{D}_{train}$ as source domains, and the N_t datasets in $\mathbb{D}_{test}$ as target domains. Classification is performed over a series of tasks (or episodes) $\mathcal{T}$ to evaluate the model's few-shot learning ability. Each few-shot task $\mathcal{T}$ is composed of a support set $\mathcal{S}$ and a query set $\mathcal{Q}$ drawn from the same dataset (or domain). Specifically, an N-way K-shot task $\mathcal{T}$ contains N classes in the support set $\mathcal{S}$ with K samples per class, i.e., $\mathcal{S} = \{(\boldsymbol{x}_i^s, \boldsymbol{y}_i^s)\}_{i=1}^{NK}$, where $\boldsymbol{x}_i^s$ and $y_i^s \in \{1, 2, \ldots, N\}$ denote the images and labels, respectively. The query set $\mathcal{Q} = \{(\boldsymbol{x}_i^q, y_i^q)\}_{i=1}^{|\mathcal{Q}|}$ consists of $|\mathcal{Q}|$ query samples. The meta-training process is also conducted on tasks to resemble the meta-test process. Each iteration randomly samples an N-way K-shot task as a mini-batch. To elaborate, we initially randomly select N classes from the training class set. Subsequently, for each class, we sample K support images and $|\mathcal{Q}|$ query images. Consequently, each mini-batch comprises $N \times (K + |\mathcal{Q}|)$ samples. During the meta-test, the ultimate goal of each task is to classify a query sample $\boldsymbol{x}_i^q \in \mathcal{Q}$ into one of the N support classes, given only a few labeled samples from $\mathcal{S}$.

Vision Transformer. We introduce the details of our backbone network, the Vision Transformer (Dosovitskiy et al., 2021), which is a pure Transformer architecture for the vision domain. It first reshapes and divides the input image $\boldsymbol{x} \in \mathbb{R}^{H \times W \times C}$ into a sequence of patches $\boldsymbol{x}_p \in \mathbb{R}^{L \times (b^2 \times C)}$, where L is the number of patches and b is the patch size. A patch projection layer embeds the image patches into patch tokens $\boldsymbol{x}_e \in \mathbb{R}^{L \times D}$, where D is the embedding dimension. To simplify notation, we assume the first token of $\boldsymbol{x}_e$ is the classification token. The patch token sequence $\boldsymbol{x}_e$ is fed into the Transformer model $\mathcal{H}_t$ that consists of a series of Transformer layers containing attention and MLP sub-layers. During the classification task, the first token of $\boldsymbol{h} = \mathcal{H}_t([\boldsymbol{x}_e])[0, :]$ is utilized for prediction.

3.2 Source Domain Prompt Learning

Different from previous methods (Guo et al., 2020; Li et al., 2021, 2022; Hu et al., 2022) that learn a shared domain-agnostic feature extractor, our method designs a flexible feature backbone integrated with Domain Prompts **(D-Prompts)** to decouple the encoding of domain-specific and domain-shared knowledge for different source domains.

Domain Prompt Learning. During the meta-training stage, we introduce a collection of D-Prompts $\boldsymbol{P} = [P_1, P_2, \ldots, P_{N_s}] \in \mathbb{R}^{N_s \times m \times D}$ as plug-and-play knowledge experts to encode domain-specific knowledge for N_s source domains, where m denotes the prompt length. For each source domain, we learn a unique domain prompt. Following the common practice in prior prompting methods (Lester et al., 2021; Zhou et al., 2022a), the prompt tokens are prepended to the patch token sequence before being fed into the domain-agnostic ViT encoder $\mathcal{H}_t$. The domain-specific representation $\boldsymbol{h}$ for the sample $\boldsymbol{x}$ from the i-th source domain is extracted as:

$$\boldsymbol{h} = \mathcal{H}_t([P_i; \boldsymbol{x}_e])[0, :], \tag{1}$$

where P_i denotes the D-Prompt associated with the i-th source domain. During meta-training, $\boldsymbol{P}$ is optimized via attention interaction with a multitude of image tokens from the same domain, effectively encoding frequently occurring specialized domain patterns into semantic meta-knowledge across diverse tasks. By decoupling domain-specific knowledge from the domain-agnostic backbone, our method incorporates domain-dependent inductive biases without compromising the general representation.

Self-supervised Prompt Regularization. The ideal D-Prompts are expected to be class-agnostic yet domain-aware, i.e., capturing essential domain patterns shared across classes. To steer D-Prompts to comprehensively grasp the domain knowledge, we propose a self-supervised prompt regularization motivated by the SwAV framework (Caron et al., 2020), a contrastive self-supervised learning approach. The prompt tokens are regarded as prototypes in SwAV (Caron et al., 2020) to compute codes for images. Specifically, given a pair of extracted features $\boldsymbol{h}_a^{(i)}, \boldsymbol{h}_b^{(i)} \in \mathbb{R}^D$ randomly sampled from the i-th source domain in the current batch, the objective for regulating the D-Prompt $P_i \in \mathbb{R}^{m \times D}$ is formulated as:

$$\mathcal{L}_{ssl} = l(\boldsymbol{h}_b^{(i)}, \boldsymbol{q}_a) + l(\boldsymbol{h}_a^{(i)}, \boldsymbol{q}_b), \tag{2}$$

where the function $l(\boldsymbol{h}, \boldsymbol{q})$ measures the fit between the feature $\boldsymbol{h}$ and a code $\boldsymbol{q}$:

$$l(\boldsymbol{h}_b^{(i)}, q_a) = -\sum_{k=1}^{m} \boldsymbol{q}_a^k \log(\boldsymbol{p}_b^k), \tag{3}$$

where $\boldsymbol{p}_b^k = \frac{exp(\boldsymbol{h}_b^{(i)} \cdot P_i^k)}{\sum_{k'} exp(\boldsymbol{h}_b^{(i)} \cdot P_i^{k'})}$, and P_i^k is the k-th token (*prototype*) of the i-th D-Prompt. $\boldsymbol{q}_a \in \mathbb{R}^m$ denotes a code calculated by the Sinkhorn algorithm (Cuturi, 2013), where $\boldsymbol{h}_a^{(i)}$ is mapped to the prompt tokens $\{P_i^k\}_{k=1}^m$, and $\boldsymbol{q}_a^k$ denotes the k-th dimension. Please refer to SwAV (Caron et al., 2020) for more details. By enforcing consistency between prompt-based predictions of instances from the same domain, the compatibility between D-Prompts and instances from the same domain can be enhanced.

3.3 Cross-Domain Prompt Transfer

In contrast to previous methods (Li et al., 2022, 2021; Hu et al., 2022) that rely on implicit knowledge transfer through indiscriminate feature reuse, our approach establishes explicit cross-domain knowledge transfer by the construction of a prompt-based knowledge store K_{store} as an inter-domain bridge. In K_{store}, the learned N_s D-Prompts $\mathbf{P}$ are stored as values and keys. The acquired knowledge from K_{store} can be flexibly combined and extended to an arbitrary task $\mathcal{T}$ with domain awareness. To this end, a **domain-guided prompt retrieval** mechanism is proposed. Specifically, for the task $\mathcal{T}$, we aggregate its support features $\boldsymbol{h}_{supp}^j$ into a task descriptor $\boldsymbol{t}_e$:

$$\boldsymbol{t}_e = \frac{1}{NK} \sum_{j=1}^{NK} \boldsymbol{h}_{supp}^j, \tag{4}$$

where $\boldsymbol{h}_{supp}^j = \tilde{\mathcal{H}}_t(\boldsymbol{x}_{supp}^j)$, and $\tilde{\mathcal{H}}_t$ denotes the frozen ViT. The task descriptor $\boldsymbol{t}_e$ is treated as the retrieval query to match with the prompt keys. During knowledge retrieval, we first compute inter-domain affinities $\boldsymbol{A} = \{a_i\}_{i=1}^{N_s}$ between $\boldsymbol{t}_e$ and D-Prompts $\boldsymbol{P}$:

$$a_i = \frac{\Phi(\boldsymbol{t}_e, P_i)}{\sum_{j=1}^{N_s} \Phi(\boldsymbol{t}_e, P_j)}, \tag{5}$$

where Φ denotes the cosine similarity metric. Subsequently, a **Transferred-Prompt (T-Prompt)** $P_t \in \mathbb{R}^{m \times D}$ for $\mathcal{T}$ can be generated as the combination of selected D-Prompts with $\boldsymbol{A}$ acting as gating:

$$P_t = \sum_{i=1}^{N_s} a_i P_i. \tag{6}$$

The T-Prompt can be seamlessly inserted into the network by input appending to provide task-level instructions: $\mathcal{H}_t([P_t; \boldsymbol{x}_e])$. The domain-guided prompt composition and retrieval enable flexible knowledge reuse and fine-grained knowledge transfer for diverse domains, thus dynamically

assisting in producing task-adaptive representations for novel tasks.

3.4 Target Domain Prompt Adaptation

When newly emerged target domains exhibit substantial disparities from source domains, it may be insufficient to solely rely on the transferred knowledge. Therefore, we develop a target domain adaptation mechanism with a learnable **Adaptive-Prompt (A-Prompt)** to incorporate additional task-related characteristics during the meta-test stage. The A-Prompt $P_a \in \mathbb{R}^{m\times D}$ is inserted into the ViT in the same manner as the T-Prompt: $\mathcal{H}_t([P_a; \boldsymbol{x}_e])$. P_a is updated by minimizing a classification loss computed over a few labeled support samples. Notably, only A-Prompt is tunable, while the other components, including the ViT model and the T-Prompt P_t, remain fixed. The entire adaptation process can be accomplished by several gradient descent steps. In contrast to full fine-tuning, A-Prompt adaptation is parameter-efficient and has favorable test-time computational complexity, leading to less risk of over-fitting to the support set. By cooperating with T- and A-Prompts, our model can simultaneously activate transferred and adaptive knowledge for arbitrary target domains.

3.5 Training Objectives

The T-Prompt and A-Prompt inject source-transferred knowledge and task-specific knowledge into the Transformer model, respectively. In Sects. 3.3 and 3.4 we introduced how to insert them into the input Transformer layer, a strategy commonly used in prior prompting methods (Lester et al. 2021; Zhou et al. 2022a, b). However, it's worth noting that different backbone layers of representations exhibit varying levels of abstraction (Wang et al., 2022; Raghu et al., 2021; Zeiler & Fergus, 2014), thus giving distinct activations to different knowledge encoded in T- and A-prompts.

Therefore, we propose a **multi-layered prompting strategy** to explore the optimal configuration for placing T- and A-Prompts in separate layers. The multi-layered prompts are denoted as $\boldsymbol{P}_t = \{P_t^{(l)} | l \in \boldsymbol{\pi}_t\}$, $\boldsymbol{P}_a = P_a^{(l)} | l \in \boldsymbol{\pi}_a\}$, where $\boldsymbol{\pi}_t$ and $\boldsymbol{\pi}_a$ denote the sets of contiguous layer indices selected for P_t and P_a, respectively. For the l-th layer, the prompting function is formulated as:

$$[\_, \boldsymbol{x}_e^l] = \mathcal{H}_t^l([P_x^{(l-1)}; \boldsymbol{x}_e^{l-1}]), \tag{7}$$

where $\boldsymbol{P}_x^{(l-1)} \in \{\boldsymbol{P}_t^{(l-1)}, \boldsymbol{P}_a^{(l-1)}\}$ and $l \in \{\boldsymbol{\pi}_t, \boldsymbol{\pi}_a\}$. We search for the optimal $\boldsymbol{\pi}_t$ and $\boldsymbol{\pi}_a$ by grid search (see Sect. 4.5 for more details), which is empirically proved to demonstrate consistently strong performance across different tasks and domains.

The final prediction is performed by a prototype-based classifier (Snell et al., 2017). Specifically, given a task $\mathcal{T}$, we first compute the prototype for each class in $\mathcal{T}$: $\hat{\boldsymbol{h}}_c = \frac{1}{K}\sum_{i=1}^{K} \boldsymbol{h}_i^s$, where $\hat{\boldsymbol{h}}_c$ denotes the prototype for the c-th class, and $\boldsymbol{h}_i^s$ denotes the support feature extracted by $\boldsymbol{H}_t$ with its label $y_i^s = c$. Then, the class probability over class $c \in \{1, 2, \ldots, N\}$ for each query point $\boldsymbol{h}^q \in Q$ is calculated as

$$p(y = c|x^q) = \frac{exp(\Phi(\boldsymbol{h}^q, \hat{\boldsymbol{h}}_c))}{\sum_{c'=1}^{N} exp(\Phi(\boldsymbol{h}^q, \hat{\boldsymbol{h}}_{c'}))}, \tag{8}$$

where Φ denotes the cosine similarity metric. The classification loss is formulated as negative log-probability:

$$\mathcal{L}_{cls} = -\frac{1}{|Q|}\sum_{(\boldsymbol{h}^q, y^q)\in\mathcal{Q}} \log p(y = y^q|\boldsymbol{h}^q). \tag{9}$$

As such, the final training objective for HybridPrompt is as follows: $\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda_{ssl}\mathcal{L}_{ssl}$, where λ_{ssl} is the weight coefficient of $\mathcal{L}_{ssl}$.

4 Experiment

In this section, we conduct extensive experiments to evaluate the performance of our proposed algorithm on the large-scale CD-FSL benchmark Meta-Dataset (Triantafillou et al., 2019). We first introduce the used datasets and implementation details, and then present extensive experimental results to verify the efficacy of our method.

4.1 Experimental Setup

Datasets. We evaluate our HybridPrompt on Meta-Dataset (Triantafillou et al., 2019), a comprehensive cross-domain few-shot learning benchmark collecting ten public image datasets from a diverse range of domains: ILSVRC-2012 (ImNet) (Deng et al., 2009), Omniglot (Omni) (Lake et al., 2015), FGVC-Aircraft (Acraft) (Maji et al., 2013), CUB-200-2011 (Bird) (Wah et al., 2011), Describable Textures (DTD) (Cimpoi et al., 2014), QuickDraw (QDraw) (Jongejan et al., 2016), FGVCx Fungi (Fungi) (Schroeder & Cui, 2018), VGG Flower (Flwr) (Nilsback & Zisserman, 2008), Traffic Signs (Sign) (Houben et al., 2013) and MSCOCO (COCO) (Lin et al., 2014). We adopt the standard evaluation protocol (Triantafillou et al., 2019) to use the first eight datasets as *source domains* for meta-training, where each dataset is further divided into train/val/test splits with disjoint classes. The remaining two datasets, i.e., Sign (Houben et al., 2013) and COCO (Lin et al., 2014), are reserved as unseen *target domains* for meta-test to measure the ***out-of-domain*** generalization performance. The test splits of source

domain datasets are used to evaluate the performance of ***in-domain*** generalization during the meta-test. Besides, we follow the practice in CNAPS (Requeima et al., 2019) to incorporate three additional datasets to enrich the unseen domains, i.e., MNIST (MNS) (LeCun & Cortes, 2010), CIFAR-10 (CF10) (Krizhevsky et al., 2009), and CIFAR-100 (CF100) (Krizhevsky et al., 2009).

Implementation Details. We adopt the Vision Transformer (ViT-Small) (Dosovitskiy et al., 2021) pre-trained by the self-supervised learning algorithm DINO (Caron et al., 2021) as our backbone model, following the practice in PMF (Hu et al., 2022). The prompt length m is set as 8 for D-, T- and A-Prompts. For multi-layered prompting, we insert T-Prompts into shallow ViT layers ($\boldsymbol{\pi}_t = [0, 1, 2, 3]$), and A-Prompts into middle layers ($\boldsymbol{\pi}_a = [4, 5, 6, 7, 8]$). λ_{ssl} is set as 0.1. During meta-training, we adopt an episodic training protocol to train the model on source datasets. We use an SGD optimizer with a cosine annealing learning rate schedule plus a warm-up strategy, with the learning rate starting from 10^{-6}, increasing to 5×10^{-5} in 5 warm-up epochs, and gradually decreasing to 10^{-6}. In the meta-test stage, we report the average classification accuracy of 600 sampled tasks/episodes from the meta-test split of each dataset. We conduct experiments under the varied-way varied-shot setting, where the number of ways N, shots K, and query images $|Q|$ are randomly sampled with respect to the dataset specifications. Given that the Meta-dataset has 8 source domains, the number of domain prompts is accordingly set as 8. In the target domain prompt adaptation module, the A-Prompt is tuned for 30 iterations by using Adadelta optimizer (Zeiler , 2012), with the learning rate set as 0.1 for in-domain datasets and 1 for out-of-domain datasets.

4.2 Comparison to State-of-the-Art Methods

We compare our HybridPrompt with several state-of-the-art methods on Meta-Dataset benchmark (Triantafillou et al., 2019). The results of the Overall Average Accuracy (*Overall Avg.*), In-domain Accuracy (*ID Avg.*) and Out-of-domain Accuracy (*OD Avg.*) are reported in Table 1.

Compared Methods. We compare HybridPrompt against several state-of-the-art (SOTA) CD-FSL methods including ProtoNet (Snell et al., 2017), CNAPs (Requeima et al., 2019), SCNAPs (Bateni et al., 2020), URT (Liu et al., 2021a), tri-M (Liu et al., 2021b), FLUTE (Triantafillou et al., 2021), URL (Li et al., 2021), TSA (Li et al., 2022), and PMF (Hu et al., 2022). Most of these SOTA methods, except for PMF (Hu et al., 2022), utilize ResNet-18 as the backbone, which is somewhat out-of-date compared to the leading work (Dosovitskiy et al., 2021; Carion et al., 2020; Xie et al., 2021) in other computer vision tasks. Therefore, we follow PMF (Hu et al., 2022) to build our model upon a pre-trained ViT backbone, using the same pre-training strategy (Caron et al., 2021) and model scale. For a fair comparison, we re-implement several state-of-the-art CD-FSL methods (*i.e.,* URL (Li et al., 2021) and TSA (Li et al., 2022)) by using the same ViT backbone as HybridPrompt and strictly following their released implementation,[1] represented as URL* and TSA* in Table 1. Specifically, We follow TSA (Li et al., 2022) to insert linear adapters after each ViT layer in a residual manner. As for URL (Li et al., 2021), we insert linear layers after the ViT output embedding layer. Besides, we complement the results of PMF on MNS, CF10, and CF100 by using their officially released code.[2]

Result Analysis. As observed in Table 1, HybridPrompt achieves the highest performance in both in-domain and out-of-domain accuracies, establishing a new state-of-the-art record (+ 2.2%) on Meta-Dataset. Specifically, our method outperforms previous methods on **9** out of 13 datasets. HybridPrompt notably surpasses the best-performing method, PMF (Hu et al., 2022), particularly on Fungi (+ 4.1%), COCO (+ 3.3%), and CF100 (+ 4.1%). Based on these findings, we have several observations. **(1) In-domain Results.** On in-domain datasets, HybridPrompt achieves superior generalization performance compared to previous methods, particularly on fine-grained datasets such as *e.g.,* Fungi (+ 4.1%). These outcomes demonstrate the effectiveness of HybridPrompt in adapting to in-domain unseen classes, indicating that D-Prompts adeptly capture domain-specific characteristics for the source domains. **(2) Out-of-domain Results.** Our method demonstrates remarkable performance and high robustness in the out-of-domain setting against substantial domain disparities. Possible reasons can be attributed to the collaboration of three types of cross-domain learned prompts, facilitating effective knowledge encoding, transfer, and adaptation across domains. **(3)** State-of-the-art Comparison. Compared to test-time tuning-based methods (URL (Li et al., 2021), TSA (Li et al., 2022), and PMF (Hu et al., 2022)) with the *SAME* ViT backbone, HybridPrompt consistently leads in performance in both in-domain and out-of-domain generalization. Moreover, our method only adapts a tiny amount of prompt parameters, making it more efficient than the best-performing method, PMF, which fine-tunes the entire network for every task. Detailed efficiency comparison results are provided in Sect. 4.6. Compared with methods like CNAPS (Requeima et al., 2019) and SCNAPs (Bateni et al., 2020) that rely on generation networks to estimate task-related parameters, our method also achieves superior performance. Generation networks trained on source domains may not generalize effectively to target domains. In contrast, our method exhibits higher task adaptability by dynamically incorporating the

[1] https://github.com/VICO-UoE/URL.

[2] https://github.com/hushell/pmf_cvpr22.

Table 1 Comparison of HybridPrompt to state-of-the-art methods on Meta-Dataset

Method	In-domain								Out-of-domain					ID Avg.	OD Avg.	Overall Avg.
	ImNet	Omni	Acraft	Bird	DTD	QDraw	Fungi	Flwr	Sign	COCO	MNS	CF10	CF100			
ProtoNet	44.5	79.6	71.1	67.0	65.2	64.9	40.3	86.9	46.5	39.9	NA	NA	NA	64.9	43.2	60.6
CNAPs	52.3	88.4	80.5	72.2	58.3	72.5	47.4	86.0	56.5	42.6	92.7	61.5	50.1	69.7	60.7	66.2
SCNAPs	58.6	91.7	82.4	74.9	67.8	77.7	46.9	90.7	73.5	46.2	93.9	74.3	60.5	73.8	69.7	72.2
URT	55.7	94.4	85.8	76.3	71.8	**82.5**	63.5	88.2	51.1	52.2	94.8	67.3	56.9	77.3	64.5	72.3
tri-M	58.6	92.0	82.8	75.3	71.2	77.3	48.5	90.5	78.0	52.8	96.2	75.4	62.0	74.5	72.9	73.9
FLUTE	51.8	93.2	87.2	79.2	68.8	79.5	58.1	91.6	58.4	50.0	95.6	78.6	67.1	76.2	69.9	73.8
URL	58.8	94.5	89.4	80.7	77.2	**82.5**	68.1	92.0	63.3	57.3	94.7	74.2	63.5	80.4	70.6	76.6
TSA	59.5	94.9	89.9	81.1	77.5	81.7	66.3	92.2	82.8	57.6	96.7	82.9	70.4	80.4	78.1	79.5
URL*	74.2	90.1	86.9	93.3	86.9	74.0	70.1	96.1	54.8	62.6	93.3	88.4	78.3	84.0	75.3	80.7
TSA*	72.2	**95.2**	86.4	92.5	87.7	74.7	67.9	96.0	**89.7**	60.2	96.4	**89.2**	76.9	84.1	82.5	83.5
PMF	74.6	91.8	88.3	91.0	86.6	79.2	74.2	94.1	88.9	62.6	95.6	85.4	77.6	85.0	82.0	83.8
HybridPrompt	**76.3**	92.8	**91.5**	**94.4**	**87.8**	80.5	**78.3**	**97.2**	86.7	**65.9**	**96.8**	88.4	**81.7**	**87.3**	**83.9**	**86.0**

The in-domain (source) datasets (i.e., the first 8 datasets) are seen during meta-training, while the out-of-domain (target) datasets (i.e., the last 5 datasets) are unseen. We report the In-Domain Average Accuracy (**ID Avg.**), Out-of-Domain Average Accuracy (**OD Avg.**), and the overall results (**Overall Avg.**). The best results in each column are shown in bold font. *Denotes the reimplemented version with the SAME ViT backbone as HybridPrompt

Table 2 Ablation results on the Meta-Dataset

M	Method	ID Avg.	OD Avg.	Overall Avg.
1	Baseline	83.21	66.31	76.71
2	T-Prompt	85.41	74.10	81.06
3	T-Prompt + $\mathcal{L}_{ssl}$	86.99	75.38	82.52
4	A-Prompt	84.05	79.53	82.31
5	G-Prompt	83.89	67.14	77.45
6	G-Prompt + A-Prompt	84.57	77.82	81.97
7	T-Prompt + $\mathcal{L}_{ssl}$ + A-Prompt (HybridPrompt)	**87.34**	**83.90**	**86.02**

The bold font numbers denote the best results

transferred experience and task-specific knowledge via various prompts.

4.3 Ablation Studies

In this section, we conduct comprehensive ablation studies to illustrate the effectiveness of each model design.

Baseline. In establishing the baseline model, we utilize the non-prompted ViT model as the backbone for feature extraction and adopt the same training strategy as HybridPrompt. During inference, the meta-trained baseline model is directly applied to the target domain for prototype-based classification.

The Effectiveness of T-Prompt and A-Prompt. As shown in Table 2, the complete version of HybridPrompt significantly surpasses the classical baseline by **17.5%** in OD Avg. and **9.3%** in Overall Avg. (M1 vs. M7). From Table 2, we have several observations. **(1)** The improvement of different prompts on OD Avg. is generally higher than ID Avg. This aligns with our design rationale, which prioritized addressing the issue of domain disparities and enhancing out-of-domain generalization. **(2) The utilization of T-Prompt** with $\mathcal{L}_{ssl}$ (see M3) yields conspicuous improvements in OD Avg. (+ 9%) compared to the baseline, demonstrating the crucial role of knowledge selection in cross-domain transfer for resolving domain discrepancies. Besides, T-Prompt also enhances in-domain generalization (+ 3.8%), which can be attributed to the effective transfer of well-encoded domain-specific knowledge from source domains. The decoupled learning of domain-shared and domain-specific knowledge eliminates cross-domain interference and contributes to both in-domain and cross-domain perception. **(3) The introduction of A-Prompt** (see M4) brings significant improvement in the out-of-domain setting (+ **13**% on average), verifying the strong adaptation capability of A-Prompt on target domains. Furthermore, as Table 2 indicates, the utilization of A-prompt yields only marginal improvements in in-domain performance, in contrast to its impact on out-of-domain scenarios. The inconsistent improvement arises from the instability of prompt tuning with a fixed learning rate on in-domain datasets. In scenarios where domain disparities are minimal, the model trained on source domains alone suffices, rendering prompt tuning potentially distorting the well-established representation.

The Effectiveness of Self-Supervised Prompt Regularization $\mathcal{L}_{ssl}$. $\mathcal{L}_{ssl}$ is used to guide the learning of D-Prompts for improving source domain knowledge learning. Here, we investigate the effect of $\mathcal{L}_{ssl}$ and present the results in Table 2. Compared to the complete T-Prompt (M3), the removal of $\mathcal{L}_{ssl}$ (M2) causes a severe performance decline, particularly on in-domain datasets (-1.6%), justifying the effectiveness of $\mathcal{L}_{ssl}$ in steering the D-Prompt to consolidate domain-specific knowledge. By encouraging the D-Prompt to align closer to the centroid of its associated domain, $\mathcal{L}_{ssl}$ contributes to the generation of more reliable and informative D-Prompts with discriminative domain patterns captured.

Comparison with the General Prompt. To further verify the efficacy of the cross-domain prompt transfer module, we substitute the D-Prompt and T-Prompt with a General Prompt (denoted as G-Prompt) shared across all source domains, maintaining the same number of additional parameters. After training on source domains, the G-Prompt is directly transferred to the target domain. As shown in Table 2, introducing the G-Prompt (M5) yields only marginal improvements (+ 0.7% in Overall Avg.) compared to the baseline, which is substantially inferior to the T-Prompt that carries selected source domain knowledge. Moreover, even when further learning the A-Prompt alongside the G-Prompt (M6), its performance still significantly lags behind our HybridPrompt (- 4.0%). In fact, learning a G-Prompt is essentially sharing the feature backbone across domains. Contrary to D-Prompts, the G-Prompt only captures domain-general information, failing to handle domain disparities when simultaneously training on all source domains. Additionally, transferring G-Prompt to the target domain overlooks inter-domain relations and the identification of useful knowledge suited for transfer, which are key aspects addressed by our method.

Comparison with Other Prompt Learning Methods We compare our approach with other prompt learning methods, including CoOp (Zhou et al., 2022a), CoCoOp (Zhou et al., 2022b), and VPT (Jia et al., 2022). We re-implement these methods using the same backbone as HybridPrompt.

Table 3 Comparison with other prompt learning methods

M	Method	ID Avg.	OD Avg.	Overall Avg.
1	Baseline	83.21	66.31	76.71
2	VPT*/Visual-CoOp*	84.13	79.26	82.26
3	Visual-CoCoOp*	83.78	71.52	79.06
4	HybridPrompt	**87.34**	**83.90**	**86.02**

The bold font numbers denote the best results

For VPT, prompts are inserted at every ViT layer, following the original VPT-deep methodology. While CoOp and CoCoOp are built upon the vision-language model, our method employs a pure vision backbone. For fair comparison, we re-implement the visual version by integrating prompts into the ViT backbone, yielding Visual-CoOp and Visual-CoCoOp. For Visual-CoOp, a prompt is learned for each task shared by categories. It is evident that the Visual-CoOp shares essential similarities with VPT, hence we represent their results together. Regarding Visual-CoCoOp, we employ a linear meta-net to generate image-specific prompts. The outcomes are depicted in Table 3. Notably, VPT/Visual-CoOp and Visual-CoCoOp significantly behind behind our method, providing strong evidence for the efficacy of our domain-aware prompting architecture. These methods are not explicitly tailored for CD-FSL, thus struggling in out-of-domain scenarios. Conversely, our method harnesses the prompt-based cross-domain transfer and adaptation to effectively mitigate domain gap issues.

Length of Prompt. The length of the prompt is related to the number of introduced additional parameters. We examine the impact of prompt length m in {1, 4, 6, 8, 12, 16} and fix other settings as default. The results for D-Prompt and A-Prompt are shown in Fig. 4a, b. The optimal length for both prompts is 8. For D-Prompt, accuracy exhibits relatively smooth changes for different prompt lengths. Notably, even using a single prompt token yields competitive performance compared to prior approaches, validating the effectiveness of prompt-based knowledge transfer and task adaptation. When increasing the prompt length to 8, overall accuracies demonstrate growth, given that longer prompts are more expressive for conveying more domain knowledge. However, excessively increasing the prompt size leads to saturated performance, as redundant information and noises are introduced. Overall, the model performance is robust against different prompt length choices.

Number of Tuning Iterations. To determine the optimal optimization configuration for A-Prompt adaptation, we vary the number of tuning iterations in {5, 10, 20, 30, 40, 60} and keep other settings as default. Here we only use the A-Prompt without employing the T-Prompt. As depicted in Fig. 4b, the model starts converging after 20 iterations and achieves peak performance when tuned for 30 iterations. However, performance decreases with more tuning iterations, implying that prompt adaptation suffers from the over-fitting issue. These findings verify the efficiency of prompt adaptation which requires only a few dozens of optimization steps and limited adaptation time for updating to a new task.

4.4 Different Ways of Retrieving Domain Prompts

During the cross-domain prompt transfer, we retrieve relevant D-Prompts $\boldsymbol{P}$ by weighted interpolation (denoted as *Weighted Sum*) as the Transferred Prompt P_t. Here we compare several alternative retrieval strategies (conducted experiments without P_a), including selecting with the maximum affinity score (*Max*), concatenating top 3 or all D-Prompts (*Concat-Top3* and *Concat-All*), and averaging (*Mean*). As shown in Table 4, *Weighted Sum* outperforms *Mean* in out-of-domain accuracy by a considerable margin, indicating that carefully selecting knowledge according to domain relevance is essential for effective knowledge transfer. Indiscriminate transfer of prompts may introduce irrelevant knowledge noise, leading to negative transfer. Concatenating all prompts as the T-Prompt yields the worst results due to the introduction of considerable noises and redundancies. The *Concat-Top3* strategy eliminates the influence of less relevant prompts but still underperforms, implying that prompt aggre-

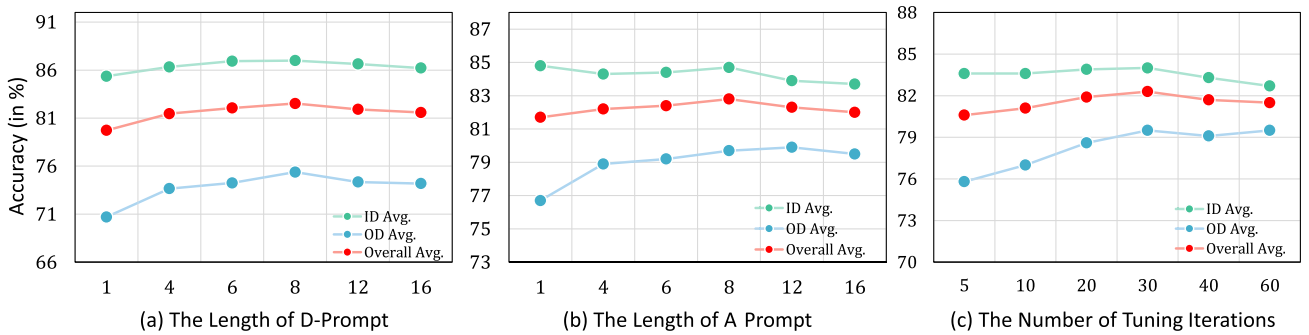

Fig. 4 The impact of **a** the length of the D-prompt, **b** the length of the A-prompt, and **c** the number of tuning iterations for optimizing the A-Prompt

Table 4 Comparison of different strategies of retrieving D-Prompts, including selecting with the maximum affinity (*Max*), concatenating top 3 or all D-Prompts (*Concat-Top3* and *Concat-All*), averaging (*Mean*), and interpolation with affinities (*Weighted Sum*)

Method	ID Avg.	OD Avg.	Overall Avg.
Max	86.14	73.36	81.22
Concat-Top 3	86.41	74.15	81.69
Concat-All	86.06	72.34	80.78
Mean	86.40	74.44	81.80
Weighted Sum	**86.99**	**75.38**	**82.52**

The bold font numbers denote the best results

gation via concatenation is inferior to weighted summation. The *Max* strategy only considers the most relevant prompt, neglecting the composite power of related knowledge from other prompts that could also be crucial for the current task. Based on these observations, we select the *Weighted Sum* as the retrieval strategy for achieving selective cross-domain transfer.

4.5 The Selection of Layers to Insert the Prompt

We have demonstrated the best-performing model equipped with a multi-layered prompt learning strategy in Sect. 4.2. In this section, we explore the optimal configuration for how to insert P_t and P_a in different network layers.

Comparison of Inserting the Prompt to Different Single Layers. We employ a heuristic search strategy to determine the optimal single-layered prompting strategy for the T-Prompt P_t and A-Prompt P_a. As shown in Fig. 5, inserting the T-Prompt at the 3rd layer yields the best performance. The performance gradually drops for higher layers. As for the A-Prompt, the 4th layer emerges as the optimal choice. Attaching A-Prompt to the 4th to 8th layers achieves significantly higher results compared other layers. Notably, the accuracy declines when prompts are inserted into deeper layers (i.e., 9th to 11th layers). This observation may be attributed to the fact that inserting prompts into deep layers affect fewer subsequent layers compared to shallower layers. Another noteworthy observation is that the optimal layers for P_t and P_a do not overlap. P_t proves more effective in shallow layers, while P_a performs better in middle layers. This finding aligns with the intuition that different network layers of representation capture varying types of knowledge (Wang et al., 2022; Raghu et al., 2021; Zeiler & Fergus, 2014). Shallow layers tend to learn low-level generic features (such as edges or colors) shared across different domains, while higher layers capture specific semantic patterns for a particular domain. Thus, T- and A-Prompts are indispensable, and their collaboration provides complementary knowledge for bridging large domain gaps.

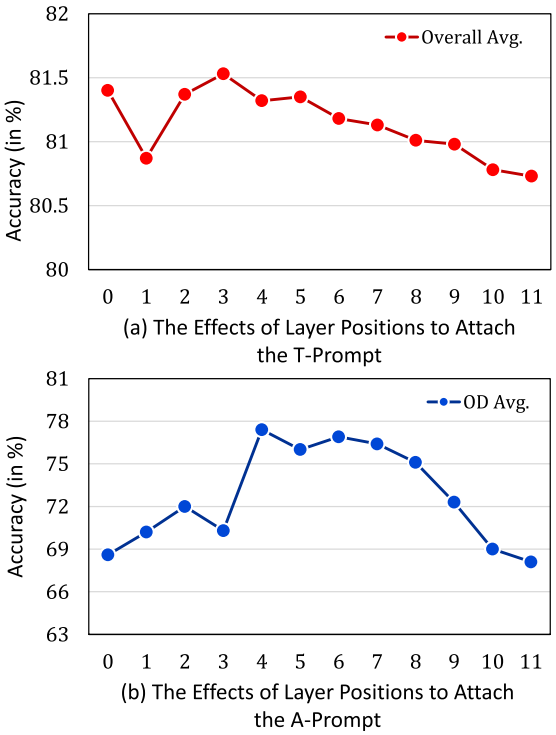

Fig. 5 The effect of attaching. **a** the T-Prompt and **b** the A-Prompt to different single network layers

Comparison of Single-layered and Multi-layered Prompting. Based on the results from single-layered experiments, we conduct further experiments to search the optimal configuration for multi-layered prompting: $\boldsymbol{\pi}_t = [0, 1, 2, 3]$ for the T-Prompt and $\boldsymbol{\pi}_a = [4, 5, 6, 7, 8]$ for the A-Prompt. We compare the single-layered prompting (i.e., attaching the P_t and P_a to the 3rd and 4th layers, respectively) with the multi-layered version in Table 5. The latter consistently outperforms the former for both T- and A-Prompts. Notably, the multi-layered T- and A-Prompts claim a significant lead in out-of-domain accuracy (+ 1.5% and + 2.1%), verifying the critical role of prompting depth in cross-domain generalization. Compared to single-layered prompting, multi-layered prompting attains more comprehensive adaptation by dynamically activating different layers of knowledge.

4.6 Efficiency Comparison

We evaluate the efficiency of HybridPrompt by comparing it to several test-time tuning-based methods (Li et al., 2021, 2022; Hu et al., 2022) in terms of the number of tuned parameters and tuning time per task, as shown in Table 6. URL (Li et al., 2021) tunes a pre-classifier linear mapping to adapt learned features to the current task, while TSA (Li et al., 2022) tunes task-specific adapters inserted in every network module. PMF (Hu et al., 2022) fine-tunes the entire

Table 5 Comparison between single-layered prompting strategy and multi-layered prompting for the T- and A-Prompts

Method type		ID Avg.	OD Avg.	Overall Avg.
T-Prompt	Single-layered	86.32	73.87	81.53
	Multi-layered	**86.99**	**75.38**	**82.52**
A-Prompt	Single-layered	83.92	77.44	81.43
	Multi-layered	**84.05**	**79.53**	**82.31**

The bold font numbers denote the best results

Table 6 Efficiency comparison with other test-time tuning-methods with regard to the number of tuned parameters (Tuned Params.) and the adaptation time per task (Time/Task)

Method	ID Avg.	OD Avg.	Overall Avg.	Tuned Params.	Time per task
URL*	84.0	75.3	80.7	262 k	**2.93 s**
TSA*	84.1	82.5	83.5	1.48 m	19.72 s
PMF	85.0	82.0	83.8	21.7 m	27.17 s
HybridPrompt	**87.3**	**83.9**	**86.0**	**15.5 k**	9.75 s

The bold font numbers denote the best results
*Denotes the re-implementation with the same ViT backbone

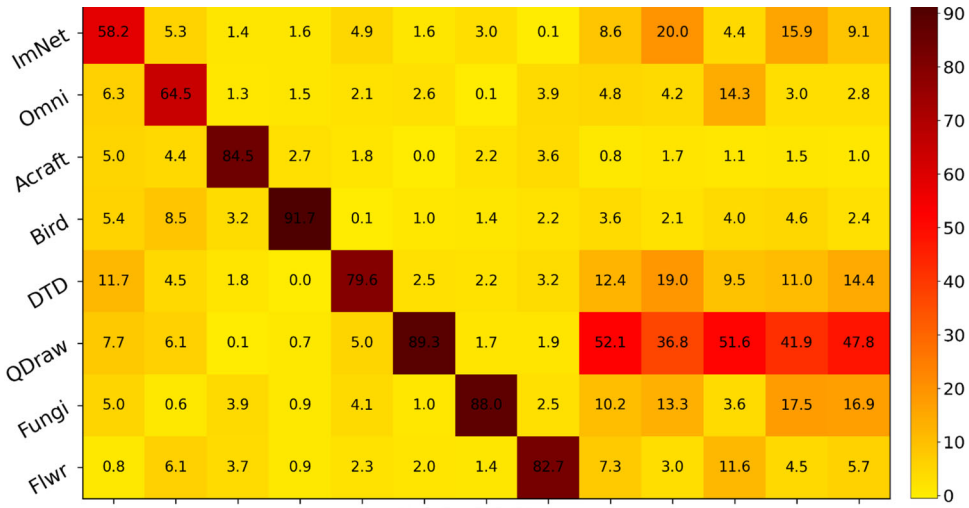

Fig. 6 The heatmap of affinity scores between test domains (columns, each with 600 tasks sampled) and D-Prompts learned on source domains (row)s

backbone network for each task by using the support set. Compared to PMF (Hu et al., 2022) based on full fine-tuning, HybridPrompt simultaneously achieves higher efficiency and adaptation performance simultaneously, requiring significantly less tuning time (9.75 s vs. 27.17 s) and fewer tuned parameter amount (15.5 K vs. 21.7 M). Our method updates only a small number of prompt parameters but demonstrates stronger task adaptation power and lower risk of over-fitting. Compared with URL, HybridPrompt requires slightly longer tuning time, but notably surpasses it in accuracy while reducing the number of tuned parameters by 90%. Overall, these results verify that our method strikes an optimal balance between generalization performance and parameter efficiency.

4.7 Interpretability of Prompt Selection

To illustrate the details of cross-domain prompt transfer process, we visualize the distribution of affinity scores between each dataset (600 tasks sampled) and the learned D-Prompts $\boldsymbol{P}$ during prompt selection. The visualization results are depicted in Fig. 6.

For novel categories from seen domains (the first 8 datasets), the highest affinity scores are consistently assigned to the D-prompt of the same domain, validating that the learned D-Prompts are effectively aligned with their corresponding source domains to encode domain-specific knowledge. Furthermore, for some in-domain datasets, D-Prompts from other domains also exhibit a certain level of contribution (e.g., InNet, Omni). Unseen target domains exhibit varied inter-domain relations. For instance, QDraw shows relevance to several target domains, as it comprises graffiti drawings with generalizable visual patterns (such as shapes and edges). Some domains have high affinities as they contain similar classes. For example, MNS has a certain probability of selecting Omni, as both datasets comprise digit images. Target domains COCO, CF10, and CF100 are relevant to ImNet, as these domains contain similar common objects.

Table 7 The results of the varying-way five-shot setting and the five-way one-shot setting

Accuracy	SCNAPS	SUR	URT	URL*	TSA*	PMF	Ours
Varying-way five-shot setting							
ID Avg.	69.0	71.2	73.8	80.4	79.8	80.1	**84.3**
OD Avg.	62.6	56.0	59.6	60.3	72.3	66.6	**76.6**
Overall Avg.	66.5	65.4	68.3	72.6	76.9	74.9	**81.3**
Five-way one-shot setting							
ID Avg.	65.0	64.0	70.6	74.8	73.6	74.9	**80.9**
OD Avg.	57.7	49.6	57.5	59.2	58.4	57.2	**66.8**
Overall Avg.	62.2	58.5	65.5	68.8	67.7	68.1	**75.5**

The bold font numbers denote the best results

These observations indicate that the selection of the domain knowledge during knowledge transfer has a non-negligible impact on the model generalization. Our method can identify the most transferable knowledge from the large-scale source spectrum, thereby providing interpretable and useful prompts automatically.

4.8 Further Results

In the previous experiments, we evaluated the standard varying-way varying-shot tasks in the multi-domain setting. Next, we conduct further experiments in more CD-FSL settings to show the robustness of our method.

The Varying-way Five-shot Setting. We follow (Li et al., 2021, 2022) to experiment with the varying-way five-shot setting, where the number of classes is varying and each class is sampled with five support samples. As presented in Table 7, the restrictions on the number of support samples make the setting more challenging, resulting in consistently decreased accuracies compared to the standard setting. Notably, among methods using the ViT backbone (i.e., URL*, TSA* and PMF), PMF exhibits significantly inferior performance in out-of-domain scenarios, as full fine-tuning causes severe over-fitting on low-shot data. Nevertheless, even in scenarios with reduced available samples, our method maintains superior performance, with approximately 5% higher overall accuracy compared to PMF. This highlights the effectiveness of HybridPrompt in generalizing to the extreme few-shot setting.

The Five-way One-shot Setting. We further evaluate the more challenging five-way one-shot setting, where each task is sampled with five classes, with each class only having one support sample. As shown in Table 7, the overall performance declines significantly compared to the varying-way five-shot setting. Owing to its parameter efficiency and cross-domain transferability, our method consistently surpasses previous state-of-the-art methods in all cases. These findings validate that HybridPrompt can effectively sustain high generalization performance in the extreme few-shot case.

Table 8 The results of the single-domain setting, which considers only ImageNet-1k's train-split for meta-training, and other datasets as out-of-domain test set

Accuracy	Baseline	PMF	URL*	TSA*	Ours
ID Avg.	74.8	74.7	73.4	74.2	76.3
OD Avg.	65.7	77.8	69.5	76.5	79.2
Overall Avg.	66.6	77.5	70.1	76.3	78.9

Single-Domain Setting. In the single-domain setting, the model is trained exclusively on ImageNet's train-split and tested on both test split of ImageNet and other 9 domains of Meta-dataset. Our method is applicable to various cross-domain scenarios. In the single domain setting, a large-scale complex source domain S can be seen as a blend of several smaller domains, allowing us to incorporate domain knowledge from diverse perspectives by learning several D-Prompts for S. When transferred to target domains, we carefully select relevant knowledge in a fine-grained manner by aggregating the learned D-Prompts. As shown in Table 8, our method significantly outperforms other state-of-the-art methods in both in-domain and out-of-domain conditions, which demonstrates the effectiveness of our approach in single-domain setting.

5 Discussions

In this section, we compare our method with relevant works and discuss the differences. Prompt learning methods (Zhou et al. 2022a, b; Lester et al. 2021) learn soft prompts with a small proportion of additional parameters, which can be attached to the Transformer model to adapt the model in a parameter-efficient manner. Our method significantly differs from previous prompt learning methods in two aspects. (1) Previous methods simply learn prompts for new tasks but fail to properly utilize the already learned experience of source domains related to target tasks. Differently, we regard

prompts as plug-and-play knowledge experts to store domain knowledge, which can be discarded, or selected, combined, and transferred according to their relevance with the target domain. (2) In contrast to the common practice of attaching a single prompt to the input layer, our method inserts two kinds of multi-layered prompts for the target task, simultaneously activating the transferred and adaptive knowledge in different layers, respectively.

Recently, Pro-D (Ma et al., 2023) has introduced a prompt learning strategy aimed at enhancing generalization by acquiring a shared prompt to encapsulate domain-agnostic knowledge. However, this approach primarily focuses on sharing feature backbones across domains without actively identifying relevant knowledge for transfer, while our method achieves selective knowledge transfer.

6 Conclusion

In this paper, we propose a Domain-aware Prompting architecture to mitigate substantial domain shift for cross-domain few-shot learning. Our method could achieve effective knowledge delivery from source to target domains, by including D-Prompts to encode domain-specific knowledge of source domains, a T-Prompt to achieve selective cross-domain knowledge transfer, and an A-Prompt to accomplish efficient target domain adaptation. The collaborative learning of these three types of prompts contributes to a hybridly prompted model with enhanced adaptability and generalizability. Extensive experimental results demonstrate that our method achieves superior performance against state-of-the-art methods.

7 Future Work and Impact

While our approach presents an advancement over the current state-of-the-art in comprehending rare out-of-domain objects, one limitation is that it sometimes may fail to clearly distinguish different categories that possess similar visual appearances. In the future, we will attempt to strengthen the fine-grained semantic understanding capability by combining pre-trained visual-language models and local part exploration. Another limitation is that the A-Prompt tuning is sensitive to learning rates (lr), posing challenges in selecting an appropriate lr for unknown tasks. Thus, automated tuning techniques may hold significance for future work.

Unlike closed-world classification approaches that specify a fixed set of classes during training, the proposed prompting framework empowers the model to identify arbitrary novel objects that are **unseen** in the training set, leveraging limited amounts of new data in the wild. Additionally, closed-world methods presume test classes belong to the known domains in the training set, while our method exhibits adaptable flexibility to **unknown domains** in the **open-world** context through domain-aware prompt learning, providing new perspectives for open-world visual recognition. While our current algorithm only addresses classification, many other works in computer vision, such as object detection and segmentation, that based on the Transformer architecture can be made more robust to domain variances by seamlessly integrating our proposed domain-aware prompts in a plug-and-play fashion.

Acknowledgements This work was supported by the Excellent Young Scientists Fund (Grant 62022078).

Data Availability The datasets generated during and/or analyzed during the current study are available in the original references, *i.e.,* Meta-Dataset (Triantafillou et al., 2019) https://github.com/google-research/meta-dataset.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Antoniou, A., Edwards, H., & Storkey, A. (2018). How to train your MAML. In *International conference on learning representations*

Bateni, P., Goyal, R., Masrani, V., Wood, F., & Sigal, L. (2020) Improved few-shot visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 14493–14502).

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

Bulat, A., Guerrero, R., Martinez, B., & Tzimiropoulos, G. (2023) FS-DETR: Few-shot detection transformer with prompting and without re-training. In *Proceedings of the IEEE/CVF international conference on computer vision*. (pp. 11793–11802).

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020) End-to-end object detection with transformers. In *European conference on computer vision* (pp. 213–229).

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems, 33*, 9912–9924.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE international conference on computer vision* (pp. 9650–9660).

Cheng, G., Lang, C., & Han, J. (2022). Holistic prototype activation for few-shot segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(4), 4650–4666.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., & Vedaldi, A. (2014). Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3606–3613).

Cui, Y., Song, Y., Sun, C., Howard, A., & Belongie, S. (2018). Large scale fine-grained categorization and domain-specific transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4109–4118).

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems, 26*, 2292–2300.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE international conference on computer vision* (pp. 248–255). IEEE.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16 x 16 words: Transformers for image recognition at scale. In *International conference on learning representations*.

Dvornik, N., Schmid, C., & Mairal, J. (2020). Selecting relevant features from a multi-domain representation for few-shot classification. In *European conference on computer vision* (pp. 769–786).

Fei-Fei, L., Fergus, R., & Perona, P. (2006). One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28*(4), 594–611.

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

Guo, Y., Codella, N. C., Karlinsky, L., Codella, J. V., Smith, J. R., Saenko, K., Rosing, T., & Feris, R. (2020). A broader study of cross-domain few-shot learning. In *European conference on computer vision* (pp. 124–141).

Hou, R., Chang, H., Ma, B., Shan, S., & Chen, X. (2019). Cross attention network for few-shot classification. In *Advances in neural information processing systems* (pp. 4003–4014).

Houben, S., Stallkamp, J., Salmen, J., Schlipsing, M., & Igel, C. (2013). Detection of traffic signs in real-world images: The German traffic sign detection benchmark. In *International joint conference on neural networks* (pp. 1–8). IEEE

Hu, S. X., Li, D., Stühmer, J., Kim, M., & Hospedales, T. M. (2022). Pushing the limits of simple pipelines for few-shot learning: External data and fine-tuning make a difference. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 9068–9077).

Jia, M., Tang, L., Chen, B. C., Cardie, C., Belongie, S., Hariharan, B., & Lim, S. N. (2022). Visual prompt tuning. In *European conference on computer vision* (pp. 709–727).

Jongejan, J., Rowley, H., Kawashima, T., Kim, J., & Fox-Gieg, N. (2016). The quick, draw!-ai experiment. http://quickdraw.withgoogle.com

Krizhevsky, A., & Hinton, G. (2009). *Learning multiple layers of features from tiny images*. Technical report, Citeseer.

Kumar Dwivedi, S., Gupta, V., Mitra, R., Ahmed, S., & Jain, A. (2019). Protogan: Towards few shot learning for action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision workshops* (pp. 0–0).

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science, 350*(6266), 1332–1338.

Lang, C., Cheng, G., Tu, B., & Han, J. (2023a). Few-shot segmentation via divide-and-conquer proxies. *International Journal of Computer Vision, 132*, 1–23.

Lang, C., Cheng, G., Tu, B., Li, C., & Han, J. (2023b). Base and meta: A new perspective on few-shot segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2023.3265865

Lang, C., Cheng, G., Tu, B., Li, C., & Han, J. (2023c). Retain and recover: Delving into information loss for few-shot segmentation. *IEEE Transactions on Image Processing*. https://doi.org/10.1109/TIP.2023.3315555

LeCun, Y., & Cortes, C. (2010). *MNIST handwritten digit database*. http://yann.lecun.com/exdb/mnist

Lee, K., Maji, S., Ravichandran, A., & Soatto, S. (2019). Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 10657–10665).

Lester, B., Al-Rfou, R., & Constant, N. (2021). The power of scale for parameter-efficient prompt tuning. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 3045–3059).

Li, W., Liu, X., & Bilen, H. (2022). Cross-domain few-shot learning with task-specific adapters. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7161–7170).

Li, W. H., Liu, X., & Bilen, H. (2021). Universal representation learning from multiple domains for few-shot classification. In *Proceedings of the IEEE international conference on computer vision* (pp. 9526–9535).

Li, X. L., & Liang, P. (2021). 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* (pp. 4582–4597).

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740–755). Springer.

Liu, B., Cao, Y., Lin, Y., Li, Q., Zhang, Z., Long, M., & Hu, H. (2020). Negative margin matters: Understanding margin in few-shot classification. In *European conference on computer vision* (pp. 438–455).

Liu, L., Hamilton, W., Long, G., Jiang, J., & Larochelle, H. (2021a). A universal representation transformer layer for few-shot image classification. In *International conference on learning representations*.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys, 55*(9), 1–35.

Liu, Y., Lee, J., Zhu, L., Chen, L., Shi, H., & Yang, Y. (2021b). A multi-mode modulator for multi-domain few-shot classification. In *Proceedings of the IEEE international conference on computer vision* (pp. 8453–8462).

Ma, T., Sun, Y., Yang, Z., & Yang, Y. (2023). Prod: Prompting-to-disentangle domain knowledge for cross-domain few-shot image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 19754–19763).

Maji, S., Rahtu, E., Kannala, J., Blaschko, M., & Vedaldi, A. (2013). Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151

Nilsback, M. E., & Zisserman, A. (2008). Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision* (pp. 722–729). IEEE: Graphics & Image Processing.

Oreshkin, B., Rodríguez López, P., & Lacoste, A. (2018). TADAM: Task dependent adaptive metric for improved few-shot learning. In *Advances in neural information processing systems* (pp. 721–731).

Perrett, T., Masullo, A., Burghardt, T., Mirmehdi, M., & Damen, D. (2021). Temporal-relational crosstransformers for few-shot action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 475–484).

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from nat-

ural language supervision. In *International conference on machine learning* (pp. 8748–8763). PMLR.
Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., & Dosovitskiy, A. (2021). Do vision transformers see like convolutional neural networks?. In *Advances in neural information processing systems* (pp. 12116–12128).
Ravi, S., & Larochelle, H. (2017). Optimization as a model for few-shot learning. In *International conference on learning representations*.
Requeima, J., Gordon, J., Bronskill, J., Nowozin, S., & Turner, R. E. (2019). Fast and flexible multi-task classification using conditional neural adaptive processes. In *Advances in neural information processing systems* (pp. 7959–7970).
Rubner, Y., Tomasi, C., & Guibas, L. J. (1998). A metric for distributions with applications to image databases. In *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)* (pp. 59–66). IEEE.
Schroeder, B., & Cui, Y. (2018). FGVCx fungi classification challenge 2018. https://github.com/visipedia/fgvcx_fungi_comp
Shin, T., Razeghi, Y., Logan IV, R. L., Wallace, E., & Singh, S. (2020). Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *EMNLP* (pp. 4222–4235).
Simon, C., Koniusz, P., Nock, R., & Harandi, M. (2020). Adaptive subspaces for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4136–4145).
Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. In *Advances in neural information processing systems* (pp. 4077–4087).
Sun, B., Li, B., Cai, S., Yuan, Y., & Zhang, C. (2021). FSCE: Few-shot object detection via contrastive proposal encoding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7352–7362).
Sun, Q., Liu, Y., Chua, T. S., & Schiele, B. (2019). Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 403–412).
Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H., & 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*.
Tian, Y., Wang, Y., Krishnan, D., Tenenbaum, J. B., & Isola, P. (2020). Rethinking few-shot image classification: A good embedding is all you need? In *European conference on computer vision* (pp. 266–282).
Triantafillou, E., Zhu, T., Dumoulin, V., Lamblin, P., Evci, U., Xu, K., Goroshin, R., Gelada, C., Swersky, K., Manzagol, P. A., & Larochelle, H. (2019). Meta-dataset: A dataset of datasets for learning to learn from few examples. arXiv preprint arXiv:1903.03096
Triantafillou, E., Larochelle, H., Zemel, R., & Dumoulin, V. (2021). Learning a universal template for few-shot dataset generalization. In *International conference on machine learning* (pp. 10424–10433).
Vinyals, O., Blundell, C., Lillicrap, T., & Wierstra, D. (2016). Matching networks for one shot learning. In *Advances in neural information processing systems* (pp. 3630–3638).
Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. (2011). *The Caltech-UCSD birds-200-2011 dataset*. Technical report.
Wang, Z., Zhang, Z., Ebrahimi, S., Sun, R., Zhang, H., Lee, C. Y., Ren, X., Su, G., Perot, V., Dy, J. & Pfister, T. (2022). Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European conference on computer vision* (pp. 631–648).
Wu, J., Zhang, T., Zhang, Y., & Wu, F. (2021). Task-aware part mining network for few-shot learning. In *Proceedings of the IEEE international conference on computer vision* (pp. 8433–8442).
Wu, J., Zhang, T., Zhang, Z., Wu, F., & Zhang, Y. (2022). Motion-modulated temporal fragment alignment network for few-shot action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 9151–9160).
Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems, 34*, 12077–12090.
Ye, H. J., Hu, H., Zhan, D. C., & Sha, F. (2020). Few-shot learning via embedding adaptation with set-to-set functions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8808–8817).
Zeiler, M. D. (2012). Adadelta: An adaptive learning rate method. arXiv preprint arXiv:1212.5701
Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision* (pp. 818–833). Springer.
Zhang, C., Cai, Y., Lin, G., & Shen, C. (2020). DeepEMD: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 12203–12213).
Zhang, R., Hu, X., Li, B., Huang, S., Deng, H., Qiao, Y., Gao, P., & Li, H. (2023) Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 15211–15222).
Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022a). Learning to prompt for vision-language models. *International Journal of Computer Vision, 130*(9), 2337–2348.
Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022b) Conditional prompt learning for vision-language models. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 16816–16825).
Zhu, C., Chen, F., Ahmed, U., Shen, Z., & Savvides, M. (2021). Semantic relation reasoning for shot-stable few-shot object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8782–8791).