Prompting-to-distill semantic knowledge for few-shot learning.

JI, H., GAO, Z., REN, J., WANG, X.-A., GAO, T., SUN, W. and MA, P.

2024

Prompting-to-Distill Semantic Knowledge for Few-shot Learning

Hong Ji, Zhi Gao, Jinchang Ren, Xing-ao Wang, Tianyi Gao, Wenbo Sun, Ping Ma

***Abstract*—Recognizing visual patterns in low-data regime necessitates deep neural networks to glean generalized representations from limited training samples. In this paper, we propose a novel few-shot classification method, namely ProDFSL, leveraging multi-modal knowledge and attention mechanism. We are inspired by recent advances of large language models and the great potential they have shown across a wide range of downstream tasks, and tailor it to benefit the remote sensing community. We utilize ChatGPT to produce class-specific textual inputs for enabling CLIP with rich semantic information. To promote the adaptation of CLIP in remote sensing domain, we introduce a Cross-modal Knowledge Generation Module, which dynamically generates a group of soft prompts conditioned on the few-shot visual samples and further uses a shallow Transformer to model the dependencies between language sequences. Fusing the semantic information with few-shot visual samples, we build representative class prototypes, which are conducive to both inductive and transductive inference. In extensive experiments on standard benchmarks, our ProDFSL consistently outperforms the state-of-the-art in few-shot learning.**

***Index Terms*—Few-shot learning, ChatGPT, CLIP, multi-modal knowledge, attention mechanism**

I. Introduction

OVER the past decade, the rapid expansion of Earth observation technologies has led to an exponential increase in the availability of remote sensing imagery, fueling the capabilities of deep learning models [1]. However, the insatiable hunger for vast amounts of high-quality, well-annotated data has become one of the major concerns that impedes deep models applying to real-world scenarios. Inspired by biological vision, few-shot learning (FSL) aims to generalize to novel tasks in the presence of extremely limited training samples, and it can be roughly categorized into transfer learning [2, 3], meta-learning [4], and metric-learning [5, 6]. Despite some successes, FSL still faces challenges such as model complexity and overfitting due to the scarcity of visual samples.

Foundation models, including large language models (LLMs) [7, 8] and vision language models (VLMs) [9, 10], are usually pre-trained on large-scale datasets and capture general patterns in various data modalities, such as text, images, or both. Current examples such as BERT [7], ChatGPT [11], and CLIP [9] have showcased powerful capabilities in tasks spanning natural language processing, computer vision, and beyond. One particular benefit of such models is the sample efficiency of adaptation, i.e., few-shot or zero-shot capabilities [9]. Most recently, some early attempts have explored the foundation models for various remote sensing tasks [12]. Although these initial endeavors have demonstrated the success of employing foundation models in remote sensing, it remains an emerging field fraught with unresolved challenges. Existing approaches in the remote sensing tasks may suffer from domain gaps due to the significant difference between natural scene images and remote sensing images, therefore demanding large-scale image-text pairs for fine-tuning.

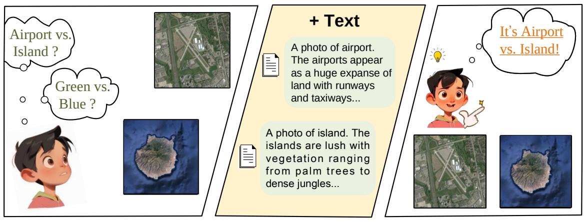
Airport vs. Island ?
Green vs. Blue ?
+ Text
A photo of airport. The airports appear as a huge expanse of land with runways and taxiways...
A photo of island. The islands are lush with vegetation ranging from palm trees to dense jungles...
It's Airport vs. Island!

Fig. 1. Multimodality reduces the ambiguity of few-shot learning. Classical few-shot leaning (left) is usually sub-optimal. Adding textual descriptions (middle) helps clarify the problem setup (right).

To address the above issues, this paper leverages cross-modal understanding of visual and language modalities to improve the few-shot adaptation. As shown in Fig. 1, reading about airport and island helps to build a better visual classifier to them. Specifically, we utilize ChatGPT [11] to provide prior knowledge. ChatGPT is based on GPT (Generative Pre-trained Transformer) architecture [8] and has acquired substantial language knowledge, allowing it to generate coherent and contextually relevant responses in a conversation. For each scene category, it takes a set of hand-written templates as input and produce human-like descriptions, as shown in Fig. 2 (a). To distill domain-specific semantic knowledge, we design a Cross-modal Knowledge Generation Module, which learns soft prompts conditioned on the remote sensing visual samples, and combine them with the class-specific descriptions to prompt CLIP textual encoder. By doing so, we take advantage of the differentiable nature of neural networks and reduces the negative impact of noise in the fixed language phrases, thus estimating representative semantic prototypes to assist the few-shot learner. In Fig. 2 (b), several independent semantic embeddings are extracted from CLIP

This work was supported in part by the National Natural Science Foundation of China under Grant 42192580 and Grant 42192583, in part by the Hubei Province Natural Science Foundation under Grant 2021CFA088, and in part by the Civilian Space Project under Grant D040103 and Grant D010206. The numerical calculations in this paper had been supported by the supercomputing system in the Supercomputing Center of Wuhan University. (*Corresponding author: Zhi Gao.*)

Hong Ji, Zhi Gao, Xing-ao Wang, Tianyi Gao and Wenbo Sun are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China. Zhi Gao is also with Hubei Luojia Laboratory, Wuhan 430079, China (e-mail:jihong@whu.edu.cn; gaozhinus@gmail.com; xingaowang@whu.edu.cn; tianyigao@whu.edu.cn; wenbosun@whu.edu.cn).

Jinchang Ren and Ping Ma are with the National Subsea Centre, Robert Gordon University, Aberdeen, AB21 0BH, U.K. (e-mail: jinchang.ren@ieee.org; p.ma2@rgu.ac.uk).

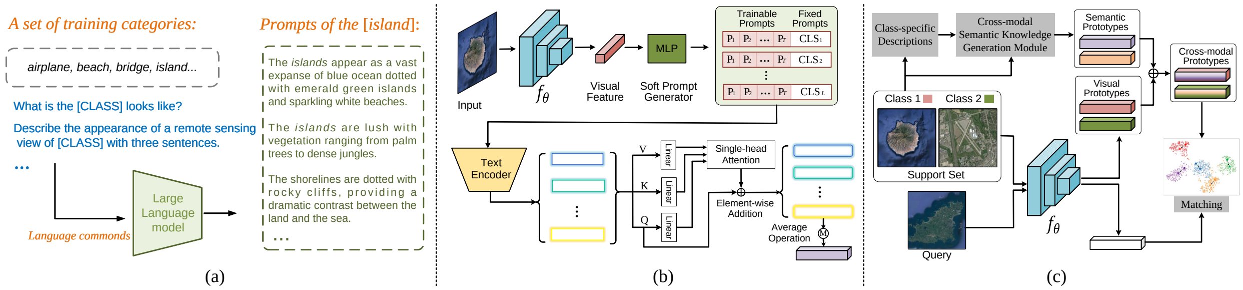
A set of training categories:
airplane, beach, bridge, island...
What is the [CLASS] looks like?
Describe the appearance of a remote sensing view of [CLASS] with three sentences.
...
Language commonds
Large Language model
Prompts of the [island]:
The islands appear as a vast expanse of blue ocean dotted with emerald green islands and sparkling white beaches.
The islands are lush with vegetation ranging from palm trees to dense jungles.
The shorelines are dotted with rocky cliffs, providing a dramatic contrast between the land and the sea.
...
(a)
Input
f_θ
Visual Feature
MLP
Soft Prompt Generator
Trainable Prompts
Fixed Prompts
P1 P2 ... PT CLS1
P1 P2 ... PT CLS2
P1 P2 ... PT CLSL
Text Encoder
V
K
Q
Linear
Linear
Linear
Single-head Attention
Element-wise Addition
Average Operation
M
(b)
Class-specific Descriptions
Cross-modal Semantic Knowledge Generation Module
Semantic Prototypes
Visual Prototypes
Cross-modal Prototypes
Class 1
Class 2
Support Set
Query
f_θ
Matching
(c)

Fig. 2. (a) Prompt with ChatGPT for each class. (b) Cross-modal Knowledge Generation Module. We utilize soft prompt generator to learn soft prompts conditioned on the input visual samples, which are combined with fixed prompts from ChatGPT. Then the CLIP textual encoder is adopted to extract semantic embeddings. Finally the Transformer is used to process the language sequences, thereby outputting a cross-modal semantic prototype. (c) The overview of our meta-learning framework. For each FSL task, cross-modal prototypes are generated to classify the query samples by Nearest Class Mean.

textual encoder via feeding the learned prompts into it. To produce abstract semantic information, we utilize attention architecture [13] to capture dependencies and finally output a semantic prototype for each category, which is then used to enhance the visual representations. As depicted Fig. 2 (c), we propose the meta-learning based cross-modal prompting to distill knowledge from both LLMs and VLMs. Furthermore, we utilize a transductive inference algorithm to alleviate the low-data problem. We have conducted thorough experiments on established benchmarks, and the quantitative outcomes affirm the efficacy of our method.

II. Our Method

Notation. $\mathcal{D}_{base}$ constitutes an extensive set of labeled data featuring a collection of base classes $\mathcal{C}_{base}$. $\mathcal{D}_{novel}$ comprises instances from a distinct set of novel classes $\mathcal{C}_{novel}$ that are mutually exclusive with the base classes. In FSL, a task $\mathcal{T}$ involves K classes. Within this framework, the support data is represented as $\mathcal{S} = (\mathbf{x}_i, y_i)_{i=1}^{NK}$, where each class is characterized by N examples, and the query data is denoted as $\mathcal{Q} = (\mathbf{x}_i, y_i)_{i=1}^{MK}$. Here, $\mathbf{x}_i$ and $y_i \in \{1, 2, ..., K\}$ correspond to the input data and its categorical label, respectively. This is referred to as K-way N-shot setting. Our goal is to train a few-shot learner on $\mathcal{D}_{base}$ and perform evaluation on $\mathcal{D}_{novel}$.

A. Overview

Our framework overview is illustrated in Fig. 2, which aims to estimate representative prototypes by the aid of language modality. Before elaborating each part of the framework, we introduce the meta-learning method ProtoNet [5], which is a few-shot learner designed to classify novel classes effectively using prototype representations derived from a limited support set. Formally, given an FSL task $\mathcal{T} = \{\mathcal{S}, \mathcal{Q}\}$, $\mathcal{S}_k$=$\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ is the support set of k-th category, its corresponding prototype is computed as $\boldsymbol{c}_k^v = \frac{1}{|\mathcal{S}_k|}\sum_{\mathbf{x}_i \in \mathcal{S}_k} f_\theta(\mathbf{x}_i)$, where f_θ is the feature extractor.

For a query sample $\mathbf{x}_i \in \mathcal{Q}$, the probability distribution is determined through a Softmax operation applied to the distances between the instance and each class prototype. i.e., $p_\theta(\hat{y}_i = k|\mathbf{x}_i) = \frac{\exp(\langle f_\theta(\mathbf{x}_i), \boldsymbol{c}_k^v \rangle)}{\sum_k \exp(\langle f_\theta(\mathbf{x}_i), \boldsymbol{c}_k^v \rangle)}$, where $\langle \cdot, \cdot \rangle$ represents the similarity matching between two feature embeddings. $\hat{y}_i$ is the prediction of query image $\mathbf{x}_i$. Then the training loss can be calculated as a standard cross-entropy loss $\mathcal{L}_b = \sum_{i=1}^{MK} \mathcal{L}_{ce}(p_\theta(\hat{y}_i = k|\mathbf{x}_i), y_i)$.

B. Prompt with ChatGPT

Neuroscience suggests that cognitive representations are inherently multi-modal. In FSL, the scarcity of visual modality data makes it difficult to learn effective representations. Therefore textual knowledge can supplement a more abstract and comprehensive semantic understanding to make up for the lack of visual information and improve the adaptation ability.

Instead of hand-crafted image-text pairs as in prior research [12], we leverage existing LLMs to extract text knowledge. Given a K-way N-shot task, we have N training samples for each K categories. As shown in Fig. 2 (a), for each category, we define detailed text templates to guide the ChatGPT to generate precise text descriptions, e.g., "Describe the appearance of a remote sensing view of [CLASS] with three sentences.". We denote the output prompts as:

$$\mathbf{t} = \{\text{CLS}_l\}_{l=1}^{L} = \text{ChatGPT}(\text{Commands}) \tag{1}$$

where CLS_l is the l-th sentence and there are totally L sentences for each category. In this work, we set L as 4. The text descriptions are mainly centered around general attributes of the target categories. For example, given a category [island], we get the prompts as in Fig. 2 (a). To minimize the negative impact of noise brought by LLMs, we set the first sentence of $\mathbf{t}$ as "This is a remote sensing photo of [CLASS].".

C. Cross-modal semantic knowledge generation module

Due to the inherent domain gaps between nature scene images and remote sensing images (i.e., the pre-training data for CLIP model mostly consists of natural scene images, thus the semantic features obtained from encoding simple class names might not be helpful for remote sensing image classification task), even a subtle alteration in wording can yield a significant impact on performance. To overcome this problem, we leverage automatic prompt learning to learn task-specific prompts. As shown in Fig. 2 (b), we introduce the concept of learning a limited set of new parameters to model the prompt's text words as trainable vectors while maintaining the frozen state of the CLIP textual encoder. Consider a sentence composed of a sequence of words (tokens), such as "This is a remote sensing photo of an island.", CLIP initially converts each token into a 512-dimensional token embedding (capped at a maximum of 77 tokens for efficient minibatch processing). Then the token embeddings (with dimensionality of 77× 512) are fed into the Transformer architecture to generate a 512-dimensional semantic vector for the given sentence.

To bring about task-specific knowledge, we use the visual modality to guide the learning of prompt's context words, dubbed conditional prompt learning. Compared to discrete language phrases and randomly initialized learnable vectors, such a paradigm can optimize token embeddings to characterize each class, being more robust to class shifts, rather than merely serving certain specific classes. We denote the visual feature of a training sample in k-th class as $\mathbf{z}_k^v = f_\theta(\mathbf{x}_k)(k \in \{1, ...K\})$. Then the generated soft prompt can be written as:

$$\mathbf{P}^k = s(\mathbf{z}_k^v) \in \mathbb{R}^{L \times T \times 512} \tag{2}$$

where $s(\cdot)$ is the soft prompt generator. L denotes the number of sentences, and T represents the number of learnable prompts. We then concatenate the $\mathbf{P}^k$ with the embeddings of existing fixed prompts. The created token embeddings $\hat{\mathbf{t}}$ of the k-th class can be written as:

$$\hat{\mathbf{t}} = \mathrm{CAT}(\mathbf{P}^k, \mathbf{t}) = [\mathbf{p}]_1[\mathbf{p}]_2...[\mathbf{p}]_T[\delta(\mathbf{t})] \tag{3}$$

where $[\mathbf{p}]_t$ $(t \in \{1, ..., T\})$ denotes learned token embeddings with dimensionality of $L{\times}512$. $\delta(\cdot)$ indicates the CLIP token embedding encoder. Accessing the final prompt $\hat{\mathbf{t}}$, we utilize the CLIP textual encoder $g(\cdot)$ to extract the cross-modal semantic embeddings for the k-th class:

$$\mathbf{a} = \{\mathbf{a}_l\}_{l=1}^L = \{g(\hat{\mathbf{t}}_l)\}_{l=1}^L \in \mathbb{R}^{L \times 512} \tag{4}$$

where $\mathbf{a}_l$ denotes the semantic embedding for the l-th sentence.

The individual semantic embedding in $\mathbf{a}$ are independent of each other, without taking into account the dependencies between them. To fully exploit the semantic knowledge, we utilize a light-weight Transformer architecture [13] to model the dependencies between the embeddings and capture the contextual relationships. Concretely, the semantic embeddings $\{\mathbf{a}_l\}_{l=1}^L$ are served as query($\mathcal{Q}$), key($\mathcal{K}$) and value($\mathcal{V}$). The interaction between the semantic embeddings of the L sentences can be achieved by:

$$\hat{\mathbf{a}} = \{\hat{\mathbf{a}}_l\}_{l=1}^L = LN(softmax(\frac{\mathcal{Q}\mathcal{K}^T}{\sqrt{d}})\mathcal{V} + \mathcal{Q}) \in \mathbb{R}^{L \times 512} \tag{5}$$

where $\hat{\mathbf{a}}$ is the result of the self-attention module. $LN(\cdot)$ represents Layer Normalization. d is the embedding dimension.

Once $\hat{\mathbf{a}}$ is obtained, we adopt a straightforward averaging operation to achieve a cross-modal semantic embedding $\hat{\mathbf{a}}^m$, and fuse it with the visual sample $\boldsymbol{z}_k^v$ by:

$$\hat{\mathbf{a}}^m = \frac{1}{L}\sum_{l=1}^{L} \hat{\mathbf{a}}_l \qquad \boldsymbol{z}_k = \boldsymbol{z}_k^v + \alpha \hat{\mathbf{a}}^m \tag{6}$$

where α is a balancing hyper-parameter. Then we use $\boldsymbol{z}_k$ to compute final cross-modal prototype $\boldsymbol{c}_k$ for the k-th class.

Moreover, to avoid the prediction ambiguity after multimodal fusion of the prototypes, we further enforce the semantic prototypes from different classes to be distinctive. It is implemented by a cross-entropy loss $\mathcal{L}_m$ encouraging the semantic prototypes to fall in the class each image belongs to. Thus the final training loss is $\mathcal{L} = \mathcal{L}_b + \beta\mathcal{L}_m$, where β controls the contribution of $\mathcal{L}_m$.

D. Inference

Once the training is done, we sample a collection of FSL tasks from $\mathcal{D}_{novel}$ to perform evaluation. For each query $\mathbf{x}_{B+i} \in \mathcal{Q}$ (note $\mathbf{x}_1, ...\mathbf{x}_B$ where $B = NK$ are support samples), we predict its label by finding its nearest cross-modal prototype, namely inductive inference.

TABLE I
DESCRIPTION OF DATASETS COMPOSITION. # DENOTES THE NUMBER OF TOTAL IMAGES OF EACH DATASET, AND OTHER COLUMNS SHOW THE CLASS SPLITS FOR TRAINING, VALIDATION, AND TESTING.

Name	#	Training	Validation	Testing
NWPU-RESISC45	31500	25	10	10
AID	10000	16	7	7
miniImageNet	60000	64	16	20
tieredImageNet	779165	351	97	160

TABLE IV
ABLATION STUDY ON THE PROMPTING METHODS.

	Prompt	*Fusing*	**NWPU-RESISC45**	
			1-shot	5-shot
(a)	-	-	72.16±0.53	86.16±0.28
(b)	[Class_name]	Addition	72.75±0.46	86.11±0.25
(c)	"This is a photo of [Class_name]"	Addition	73.21±0.53	86.18±0.23
(d)	ChatGPT w/o soft prompting	Addition	73.04±0.48	86.20±0.29
(e)	ChatGPT w/o soft prompting	Our ProDFSL	73.45±0.51	86.28±0.31
(f)	ChatGPT w/ soft prompting	Our ProDFSL	**76.94±0.41**	**86.34±0.22**

The above prediction is performed independently on each query sample, ignoring the distribution of the queries. With this in mind, we utilize the query samples to update the class prototypes, called transductive inference. The objective is:

$$\min_{\tilde{\mathbf{Y}}, \tilde{\mathbf{C}}} \sum_{i,k} \tilde{\mathbf{Y}}_{ik} \|\mathbf{z}_i - \tilde{\mathbf{c}}_k\|_2^2 \quad \textbf{s.t. } \tilde{\mathbf{Y}}\mathbf{1}_K = \mathbf{1} \tag{7}$$

where $\tilde{\mathbf{Y}} \in \mathbb{R}^{U \times K}$ $(U = MK)$ is a mapping matrix, and $\tilde{\mathbf{Y}}_{ik}$ denotes the allocated portion of i-th query sample to k-th class. $\tilde{\mathbf{C}} = \{\tilde{\mathbf{c}}_1, ..., \tilde{\mathbf{c}}_K\}$ is the updated class centers that are calculated by original cross-modal prototypes and query samples, which is initialized by the cross-modal prototypes.

Inspired by [14], we leverage the prior of uniform distribution to normalize $\tilde{\mathbf{Y}}$. The normalization itself is a projection of $\tilde{\mathbf{Y}}$ on to the set $\mathbb{S}_{\mathbf{r},\mathbf{d}}$ of non-negative matrices with row-wise sum $\mathbf{r}$ and column-wise sum $\mathbf{d}$:

$$\mathbb{S}_{\mathbf{r},\mathbf{d}} := \{\tilde{\mathbf{Y}} \in \mathbb{R}^{U \times K} | \tilde{\mathbf{Y}}\mathbf{1}_K = \mathbf{r}, \tilde{\mathbf{Y}}^\top\mathbf{1}_U = \mathbf{d}\} \tag{8}$$

We adopt the Sinkhorn-Knopp algorithm [15] for the projection, by alternating between rescaling the rows of $\tilde{\mathbf{Y}}$ to sum to $\mathbf{r}$ and its columns to sum to $\mathbf{d}$:

$$\begin{aligned} \tilde{\mathbf{Y}} &\leftarrow \mathrm{diag}(\mathbf{r})\mathrm{diag}(\tilde{\mathbf{Y}}\mathbf{1}_\mathrm{K})^{-1}\tilde{\mathbf{Y}} \\ \tilde{\mathbf{Y}} &\leftarrow \tilde{\mathbf{Y}}\mathrm{diag}(\tilde{\mathbf{Y}}^\top\mathbf{1}_\mathrm{U})^{-1}\mathrm{diag}(\mathbf{d}) \end{aligned} \tag{9}$$

After convergence, we use $\tilde{\mathbf{Y}}$ to update the class centers:

$$\tilde{\mathbf{H}} = \frac{\tilde{\mathbf{C}} + \tilde{\mathbf{Y}}^\top\mathbf{Z}}{B + U} \quad \tilde{\mathbf{C}} = \tilde{\mathbf{C}} + \gamma(\tilde{\mathbf{H}} - \tilde{\mathbf{C}}) \tag{10}$$

where γ is a hyper-parameter that controls the update speed.

III. EXPERIMENTS AND ANALYSIS

A. Implementation Details

Our methods undergo evaluation on commonly used benchmarks, encompassing both remote sensing datasets and natural scene datasets. We follow prior works to determine class splits (Table I), and implement a deep backbone ResNet12 [1] and a shallow backbone Conv-256 for feature extraction. The soft prompt generator is a single fully-connected layer.

The training process involves two stages: pre-training and meta-training. The backbone is trained on base classes with a classifier and standard cross-entropy loss. The meta-training uses pre-trained weights as initialization. The models are trained using an SGD optimizer with a momentum of 0.9,

TABLE II
THE 5-WAY CLASSIFICATION ACCURACY (%) ON THE REMOTE SENSING DATASETS. THE **BEST** AND ***SECOND BEST*** RESULTS ARE HIGHLIGHTED.

FSL method	Backbone	NWPU-RESISC45		AID	
		1-shot	5-shot	1-shot	5-shot
MAML[4]	Conv-4-64	58.99±0.45	72.67±0.38	60.11±0.50	70.28±0.41
Meta-SGD[16]	ResNet12	60.63±0.90	75.75±0.65	53.14±1.46	66.94±1.20
MatchingNet[17]	ResNet12	61.57±0.49	76.02±0.34	64.30±0.46	74.49±0.35
ProtoNet[5]	ResNet12	64.52±0.48	81.95±0.30	67.08±0.47	82.44±0.29
RelationNet[18]	ResNet12	65.52±0.85	78.38±0.31	68.56±0.49	79.21±0.35
RS-MetaNet[19]	ResNet50	52.78±0.09	71.49±0.81	53.37±0.56	72.59±0.73
SCL-MLNet[20]	Conv-256	62.21±1.12	80.86±0.76	59.49±0.96	76.31±0.68
SPNet[21]	ResNet18	67.84±0.87	83.94±0.50	-	-
DLA-MatchNet[22]	ResNet12	71.56±0.30	83.77±0.64	-	-
IDLN [23]	ResNet12	75.25±0.75	84.67±0.23	-	-
DCN [24]	ResNet12	74.40±0.83	89.22±0.41	-	-
Ji et.al [3]	ResNet12	76.70±0.44	89.87±0.21	72.67±0.43	***87.33±0.23***
Wang et.al [25]	ResNet12	77.96±0.87	91.17±0.51	-	-
ProDFSL (*In.*)	Conv-256	70.86±0.59	83.19±0.40	67.45±0.41	79.13±0.36
	ResNet12	76.94±0.54	86.34±0.26	72.98±0.43	81.64±0.25
SIB [26]	WRN-28-10	67.34±0.81	78.28±0.48	60.72±0.78	71.86±0.49
CAN+T [27]	ResNet12	69.89±0.58	81.04±0.33	63.82±0.56	73.77±0.41
MTL-trans [28]	ResNet12	80.58±0.48	89.20±0.21	***76.42±0.47***	87.31±0.23
MES^{2}L-Net [29]	ResNet12	***86.55±0.18***	91.06±0.11	-	-
ProDFSL (*Trans.*)	Conv-256	81.78±0.60	***91.27±0.29***	73.59±0.63	87.00±0.33
	ResNet12	**89.34±0.45**	**94.12±0.25**	**80.21±0.51**	**91.04±0.29**

TABLE III
CROSS-DOMAIN PERFORMANCE OF INDUCTIVE (I) AND TRANSDUCTIVE (T) INFERENCE WITH RESNET12.

Source		NWPU-RESISC45		AID	
		1-shot	5-shot	1-shot	5-shot
miniImageNet	I	56.44±0.75	70.24±0.39	65.89±0.72	78.85±0.43
	T	68.64±0.62	86.24±0.23	80.12±0.56	90.32±0.31
tieredImageNet	I	**61.88±0.71**	**74.68±0.33**	**70.14±0.57**	**82.25±0.41**
	T	**75.53±0.64**	**88.58±0.36**	**85.27±0.54**	**92.94±0.35**

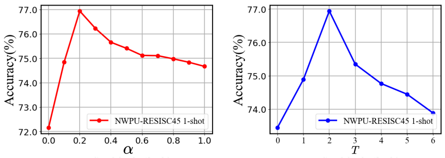

Fig. 3. Selection of hyper-parameters α and T (ResNet12).

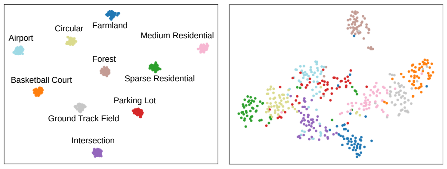

Fig. 4. The t-SNE visualization of prototype vectors w/wo semantic information (ResNet12).

TABLE V
ABLATION STUDY ON THE INFERENCE ALGORITHMS WITH RESNET12.

	Feature Extractor	*Inference Algorithm*	NWPU-RESISC45	
			1-shot	5-shot
(a)	Baseline	Nearest Class Mean (*In.*)	68.77±0.46	84.12±0.25
(b)	Baseline	Sinkhorn-Knopp (*Trans.*)	83.45±0.51	89.97±0.27
(c)	ProNet	Nearest Class Mean (*In.*)	72.16±0.53	86.16±0.28
(d)	ProNet	Sinkhorn-Knopp (*Trans.*)	85.67±0.44	93.67±0.30
(e)	ProDFSL	Nearest Class Mean (*In.*)	76.94±0.49	86.34±0.25
(f)	ProDFSL	Label Propagation (*Trans.*)	79.44±0.48	87.22±0.22
(g)	ProDFSL	Prototype Rectification (*Trans.*)	82.67±0.39	87.99±0.23
(h)	ProDFSL	Sinkhorn-Knopp (*Trans.*)	**89.34±0.39**	**94.12±0.23**

an initial learning rate of 0.001, and a weight decay of 5e-4. Training spans 100 epochs, with a decay factor of 0.5 applied every 10 epochs. β and γ are set as 0.01 and 0.2, respectively. We assess performance through a random selection of 1000 tasks designed for 5-way classification, with each task comprising 15 queries. The evaluation metric is the average accuracy, accompanied by 95% confidence intervals.

B. Comparison to the state-of-the-art methods

We compare our ProDFSL with other methods, including transfer learning, meta-learning, metric-learning, and those transductive based. The results are shown in Table II, from which we find that without transductive inference, our ProDFSL achieves fairly competitive performance with both deep and shallow backbones in 1-shot tasks. Our method uses semantic knowledge to assist FSL, while other works [24, 3, 25] introduce various pretext tasks and sophisticated networks to improve their generalization ability. This demonstrates the effectiveness of the multi-modal information. Nevertheless, we can see that our method underperforms the above three methods in 5-shot setting. This is because the advantage of the semantic modality lies in compensating for the lack of visual samples. Therefore, with the increase in visual samples, the gain from semantic information becomes less pronounced. Regarding those methods using unlabeled data, our results exceed the second best results by a large margin, e.g., 89.34% is higher than 86.55%. Therein, MTL-trans [28] uses label propagation to make predictions based on a multi-task learning network, however there is still a significant gap with our results, which proves the superiority of our transductive inference algorithm as well as the learned representations.

In Table III, we show the cross-domain FSL performance. Training on natural scene images, our method achieves good results on remote sensing images, underlining the generalization ability. The results of the second block consistently outperforms the first block. This is because tieredImageNet has 351 training classes, while miniImageNet only 64 training classes. Extensive training data assimilates more profound knowledge and thereby achieving better transfer results.

C. Ablation study

1) Selection of hyper-parameters: We devise several experiments to select suitable hyper-parameter α in Eq. (6), and the number of learnable prompts T in Eq. (3). As shown in Fig. 3 left, when we set α as 0, the classification results under 1-shot setting is only 72.16%. This indicates the representation capability of the visual prototypes is insufficient. With increasing α, the performance improves rapidly. The accuracy finally peaks at 0.2. Fixing α, we show in Fig. 3 right, that the performance improves when increasing the number of learnable tokens from 0 to 2. However, the performance is saturated and the improvements diminish if further increasing the context length. Thus we set T as 2 by default.

2) Influence of our cross-modal semantic knowledge generation module: We compare different prompting methods in Table IV. Overall, we find that prompt learning can improve few-shot classification, e.g., 72.75% is higher than 72.16%. Setting (b) shows that only using the class names as prompt is sub-optimal. As such a method does not introduce any domain-specific guidance. Setting (c) is widely-used in many works [11], which brings about obvious improvements compared to setting (b). Besides, from setting (d), we see that, without semantic knowledge interaction, the results are still not satisfactory even with comprehensive descriptions from ChatGPT, e.g., 73.04% is lower than 73.45%. Besides, distinction between the results of setting (e) and (f) clarifies the importance of soft prompting. Our proposed cross-modal semantic knowledge generation module leverages the thorough information from ChatGPT, and dynamically generates soft prompts conditioned on remote sensing visual samples. Moreover, we use a shallow Transformer architecture to integrate the cross-modal semantic features. Therefore, our method learns domain-specific knowledge under few-shot scenarios and achieves the best performance. In Fig. 4 left, we visualize the cross-modal prototypes used for testing by t-SNE [30]. It can be seen that the support vectors are prone to cluster together when they belong to the same class and repulse those from the other classes. We also compare the same visualization of visual prototypes in the right, where the vectors are more diverse and the inter-class variance is very large.

3) Influence of the transductive inference: In Table V, we discuss the effect of the transductive inference. It is evident that the Sinkhorn algorithm brings large gains over the inductive inference, e.g., more than 10.0%. This indicates the effectiveness of using query samples to assist the classification. Comparing (b), (d) and (h), we find that better initial prototypes can yield better results. Besides, our method also outperforms other transductive methods, including label propagation [31] and prototype rectification [28]. This is because that the Sinkhorn algorithm balances the class distribution.

IV. Conclusion

In this work, we have presented a novel method, ProDFSL, to tackle few-shot classification problem, leveraging multi-modal knowledge. Motivated by the large language models and the multi-modal nature of cognitive representations, we propose to meta-learn cross-modal understanding for few-shot classification. Specifically, we use ChatGPT to provide textual descriptions for each classes. Then we introduce a cross-modal semantic knowledge generation module to learn soft prompts conditioned on the few visual samples, which enables the CLIP model to distill domain-specific knowledge. By fusing the cross-modal semantic information with visual samples, we build better prototypes, which serves as a good initialization for the transductive inference. Extensive experiments demonstrates the effectiveness of our ProDFSL. We hope to design interactive systems that allow users to provide feedback to improve the text descriptions, making cross-modal learning become a tool for future research on multi-modal adaptation.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[2] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. Wang, and J.-B. Huang, "A closer look at few-shot classification," in *International Conference on Learning Representations*, 2019.

[3] H. Ji, Z. Gao, Y. Zhang, Y. Wan, C. Li, and T. Mei, "Few-shot scene classification of optical remote sensing images leveraging calibrated pretext tasks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[4] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1126–1135.

[5] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 4080–4090.

[6] L. Li, X. Yao, G. Cheng, and J. Han, "Aifs-dataset for few-shot aerial image scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[7] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, vol. 1, 2019, p. 2.

[8] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding with unsupervised learning," 2018.

[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[10] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International conference on machine learning*. PMLR, 2021, pp. 4904–4916.

[11] OpenAI, "Introducing chatgpt," https://openai.com/blog/chatgpt, 2022.

[12] Z. Shi and Z. Zou, "Can a machine generate humanlike language descriptions for a remote sensing image?" *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 6, pp. 3623–3634, 2017.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[14] Y. Hu, V. Gripon, and S. Pateux, "Leveraging the feature distribution in transfer-based few-shot learning," in *International Conference on Artificial Neural Networks*. Springer, 2021, pp. 487–499.

[15] C. Villani, *Optimal transport: old and new*. Springer, 2009, vol. 338.

[16] Z. Li, F. Zhou, F. Chen, and H. Li, "Meta-sgd: Learning to learn quickly for few-shot learning," *arXiv preprint arXiv:1707.09835*, 2017.

[17] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," *Advances in Neural Information Processing Systems*, vol. 29, pp. 3630–3638, 2016.

[18] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1199–1208.

[19] H. Li, Z. Cui, Z. Zhu, L. Chen, J. Zhu, H. Huang, and C. Tao, "Rs-metanet: Deep metametric learning for few-shot remote sensing scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 8, pp. 6983–6994, 2021.

[20] X. Li, D. Shi, X. Diao, and H. Xu, "Scl-mlnet: Boosting few-shot remote sensing scene classification via self-supervised contrastive learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022.

[21] G. Cheng, L. Cai, C. Lang, X. Yao, J. Chen, L. Guo, and J. Han, "Spnet: Siamese-prototype network for few-shot remote sensing image scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2021.

[22] L. Li, J. Han, X. Yao, G. Cheng, and L. Guo, "Dla-matchnet for few-shot remote sensing image scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 9, pp. 7844–7853, 2021.

[23] Q. Zeng, J. Geng, W. Jiang, K. Huang, and Z. Wang, "Idln: Iterative distribution learning network for few-shot remote sensing image scene classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[24] Z. Ji, L. Hou, X. Wang, G. Wang, and Y. Pang, "Dual contrastive network for few-shot remote sensing image scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–12, 2023.

[25] L. Wang, L. Zhuo, and J. Li, "Few-shot remote sensing scene classification with spatial affinity attention and class surrogate-based supervised contrastive learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023.

[26] S. X. Hu, P. G. Moreno, Y. Xiao, X. Shen, G. Obozinski, N. D. Lawrence, and A. C. Damianou, "Empirical bayes transductive meta-learning with synthetic gradients," in *ICLR*, 2020.

[27] R. Hou, H. Chang, B. Ma, S. Shan, and X. Chen, "Cross attention network for few-shot classification," in *NeurIPS*, 2019.

[28] H. Ji, H. Yang, Z. Gao, C. Li, Y. Wan, and J. Cui, "Few-shot scene classification using auxiliary objectives and transductive inference," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[29] J. Li, M. Gong, H. Liu, Y. Zhang, M. Zhang, and Y. Wu, "Multiform ensemble self-supervised learning for few-shot remote sensing scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.

[30] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[31] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Advances in neural information processing systems*, 2004, pp. 321–328.