Envisioning Class Entity Reasoning by Large Language Models for Few-shot Learning

Mushui Liu[1], Fangtai Wu[1], Bozheng Li[1], Ziqian Lu[2], Yunlong Yu[1]*, Xi Li[3]

[1]College of Information Science & Electronic Engineering, Zhejiang University
[2]School of Aeronautics and Astronautics, Zhejiang University
[3]College of Computer Science and Technology, Zhejiang University
{lms, wft, bozhengli, ziqianlu, yuyunlong, xilizju}@zju.edu.cn

Abstract

Few-shot learning (FSL) aims to recognize new concepts using a limited number of visual samples. Existing methods attempt to incorporate semantic information into the limited visual data for category understanding. However, these methods often enrich class-level feature representations with abstract category names, failing to capture nuanced features essential for effective generalization. To address this issue, we propose a novel framework for FSL, which incorporates both the abstract class semantics and the concrete class entities extracted from Large Language Models (LLMs), to enhance the representation of the class prototypes. Specifically, our framework composes a Semantic-guided Visual Pattern Extraction (SVPE) module and a Prototype-Calibration (PC) module, where the SVPE meticulously extracts semantic-aware visual patterns across diverse scales, while the PC module seamlessly integrates these patterns to refine the visual prototype, enhancing its representativeness. Extensive experiments on four few-shot classification benchmarks and the BSCD-FSL cross-domain benchmark showcase remarkable advancements over the current state-of-the-art methods. Notably, for the challenging one-shot setting, our approach, utilizing the ResNet-12 backbone, achieves an impressive average improvement of 1.95% over the second-best competitor.

Introduction

Deep learning has revolutionized numerous fields, notably computer vision (He et al. 2016; Tang et al. 2022b; Dan et al. 2023, 2024b) and natural language processing (Achiam et al. 2023; Liu et al. 2023; Tang et al. 2022a). These learning systems are inherently data-intensive, demanding substantial amounts of labeled data for model training. In some practical applications, the acquisition and annotation of data can be prohibitively expensive or unfeasible. Consequently, Few-Shot Learning (**FSL**) (Sun and Gao 2023; Zhang et al. 2024) has garnered significant attention as a promising solution, enabling learning from a scarce quantity of data, thereby mitigating the challenges associated with data scarcity.

Most FSL methods face significant challenges due to the scarcity of samples, especially when only one sample is available for each class (Baik et al. 2023). To mitigate this issue, several approaches (Yang, Wang, and Chen 2022; Zhang et al. 2024; Chen et al. 2023) have introduced semantic information to assist in constructing class understanding. However, semantic-incorporating methods enrich class representations with high-level information but often overlook the local structure of the visual environment. Many discriminative patterns in low-level descriptions are crucial for classifying samples, especially when only a few samples are available for each class.

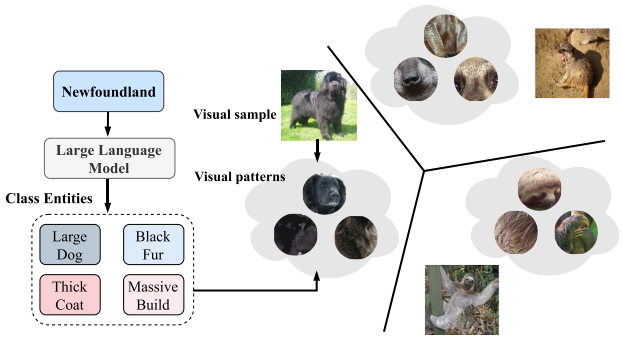


Figure 1: The key idea is to learn a classifier from data using a visual sample per category, extended with additional visual patterns derived from semantic entities generated by LLMs.

Humans have demonstrated the ability to recognize novel categories with minimal examples by combining abstract attributes and detailed category descriptions with visual signals. If machine vision systems could do such information abstraction and alignment, then the novel categories may be quickly learned with only a few samples. Unfortunately, directly extracting low-level descriptions of categories from visual samples can be challenging, as images often fail to comprehensively capture the full range of attributes and characteristics that define a class. The limitations of visual representation, *e.g.*, variations in lighting, angles, and occlusion, can hinder the models' capability to comprehensively describe a category based solely on the images at hand.

Considering the success of nature language processing (NLP) that the related semantic descriptions or attributes could be generated with the power of Large-Language Models (LLMs), we propose a novel paradigm that learns semantic-aware visual patterns from the semantic descriptions to enrich the class representations for FSL, as shown

*Corresponding Author.

in Fig. 1. However, such a paradigm still suffers from challenges: though producing some descriptions from the category names, most LLMs often encounter the hallucination problem that the generative descriptions fail to maintain a relevant correspondence with the categories. Further, the misalignment between the visual samples and the low-level descriptions makes it challenging to classify samples.

In this paper, we introduce a novel approach named **ECER-FSL**, designed to utilize concrete class entities generated by LLMs to enhance FSL. To address the hallucination issue in generating semantic entities for each class, we employ a filtering strategy that evaluates the similarity between the generated semantic descriptions and the corresponding class names, discarding entities with low similarity. Furthermore, to address the critical issue of misalignment, we propose a Progressive Visual-Semantic Aggregation (PVSA) framework that leverages both class names and class-related entities to generate semantic-aware visual patterns, progressively enriching the visual prototypes. The PVSA framework effectively integrates multiple stages of visual features with semantic information via the Semantic-guided Visual Pattern Extraction (SVPE) module, thereby comprehensively uncovering the potential visual characteristics of categories. By incorporating semantic-aware visual patterns into the visual prototypes through a well-designed Prototype Calibration (PC) module, discriminative classifiers are obtained under limited visual samples.

Overall, our contributions are as follows:

- We introduce a novel FSL paradigm that learns semantic-aware visual patterns from the semantic entities derived from the large-language models to enrich the visual prototype for each category.
- We propose a Progressive Visual-Semantic Aggregation (PVSA) framework that gradually captures the semantic-aware visual patterns at different network blocks guided by both class names and semantic entities.
- Extensive experiments on both FSL and cross-domain FSL tasks have demonstrated that our proposed method establishes a new benchmark for state-of-the-art performance, outperforming the second-best competitor by a substantial margin, specifically under the 1-shot scenario.

Related Works

Uni-Modal Few-Shot Learning

The uni-modal FSL approaches only use the visual support data to predict the query samples. The existing uni-modal FSL approaches could be roughly grouped into three categories, i.e., metric-based, optimization-based, and hallucination-based approaches. Metric-based approaches (Snell, Swersky, and Zemel 2017; Chen et al. 2019a) aim to learn a good metric space to evaluate the affinity similarities between the support-query pairs, where the query samples from the novel classes can be nicely categorized via the nearest neighbor classifier. To capture rich statistics, several works (Zhang et al. 2019; Xie et al. 2022) further exploit the second moment to enrich the image representations and perform the similarity metric with different metrics (Liu, Cao, and He 2023; Hu et al. 2023). Optimization-based approaches (Finn, Abbeel, and Levine 2017; Yu et al. 2022; Sun and Gao 2024) aim to learn how to train model parameters to produce good results on new tasks with a few optimization steps, or even a single optimization step. Hallucination-based approaches address FSL via augmenting the support data at either the image level (Zhang et al. 2018) or the feature level (Lazarou, Stathaki, and Avrithis 2022; Bär et al. 2024).

Multi-Modal Few-Shot Learning

In recent, multi-modal learning has been widely explored (Zhao et al. 2024b,c,a; Liu et al. 2025; He et al. 2025; Dan et al. 2024a). Multi-modal FSL approaches (Xing et al. 2019; Yang, Wang, and Chen 2022; Ji et al. 2022) learn novel categories from multiple modalities when support visual samples are scarce. These multi-modal methods can complement uni-modal FSL approaches. AM3 (Xing et al. 2019) enhances ProtoNet (Snell, Swersky, and Zemel 2017) by integrating class-level semantic and visual prototypes. SEGA (Yang, Wang, and Chen 2022) uses label embeddings to focus on semantic-related visual features. TRAML (Li et al. 2020) improves visual feature embeddings by aligning them with semantic similarities. SemFew (Zhang et al. 2024) proposes a simple two-layer network to transform semantic and visual features into robust category prototypes with rich discriminative features. Some studies use semantic features for data augmentation. ProtoComNet (Zhang et al. 2021) generates additional features using a Gaussian distribution based on hand-crafted attribute features. TriNet (Chen et al. 2019b) uses an encoder-decoder framework to align and augment visual features with semantic information. SIFT (Pan, Xin, and Shen 2024) uses class-specific semantic embeddings to generate high-quality features via an encoding-transformation-decoding process, enabling models to transfer features from base to novel categories.

Our approach also aligns with multi-modal methods, yet distinguishes itself by integrating expert knowledge from LLMs to enhance visual concept extraction across various stages of the visual encoder.

Methodology

Following the existing FSL literature (Chen et al. 2021a), our method encompasses a two-stage process: a pre-training phase that involves learning generalizable representations with the base set, followed by an episode-based fine-tuning stage that the model is further fine-tuned with various FSL tasks generated from the base dataset.

Multi-Modality Pre-training

As visual feature embedding plays a pivotal role in FSL (Tian et al. 2020), most existing approaches (Chen et al. 2021a) leverage the available base data to train the visual feature encoder with a proxy classification task. However, such a training paradigm overemphasizes the extraction of classification features related to base classes, which tends to overlook the inherent characteristics of the samples themselves when extracting samples from novel classes. Prior

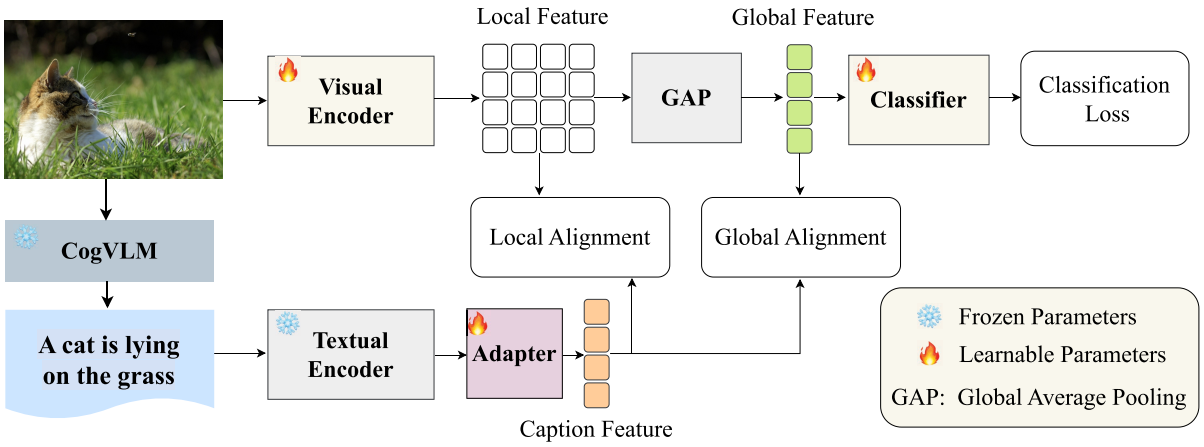


Figure 2: **Multi-Modality Pre-training Stage**. Our pre-training paradigm trains a visual encoder that captures more semantic-rich information under the guidance of the textual captions derived from an off-the-shelf model, which harnesses the power of natural language processing to enrich the visual representation, fostering a deeper understanding of the images and their underlying semantics.

work has attempted to mitigate this limitation by implicitly introducing feature constraints (Yang, Wang, and Zhu 2022) or instance enhancement (Fu and Zhu 2024).

Different from the pre-training paradigm in the existing FSL literature that trains the visual backbone in a uni-modality way, we introduce a multi-modality pre-training strategy that leverages both visual and textual data to enhance the representation learning capabilities of the visual backbone. As depicted in Fig. 2, our pre-training paradigm ingeniously harnesses semantic captions generated by the existing caption model CogVLM (Wang et al. 2023) from the input visual samples, where the captions serve as auxiliary information to guide and enhance the visual feature backbone's ability to extract semantic-rich feature representations. Specifically, we apply the off-the-shelf text encoder (Radford et al. 2021) to extract the feature embedding and introduce an adapter to handle the compatibility between the visual and text embeddings. To integrate textual information into the pre-training process, we introduce the multi-modality contrastive loss that operates at both local and global levels. The local level focuses on aligning specific regions of the image with corresponding segments of the text, while the global level ensures the overall semantic alignment between the entire image and the text description.

Given an image I, the visual model extracts the feature map $\mathbf{F}_i \in \mathbb{R}^{HW \times C}$ and global representation $\mathbf{v}_i \in \mathbb{R}^{1 \times C}$. The caption representation $\mathbf{t}$ of the image I is obtained with the off-the-shelf text encoder followed by an adapter module. The image-text contrastive loss is defined as:

$$\mathcal{L}_{IT} = -\sum_{\mathbf{I}_i, \mathbf{t}_i \in \mathcal{B}} \log \frac{\exp(\text{sim}(\mathbf{I}_i, \mathbf{t}_i)/\tau)}{\sum_{\mathbf{t}_j \in \mathcal{B}} \mathbb{I}_{j \neq i} \exp(\text{sim}(\mathbf{I}_i, \mathbf{t}_j)/\tau)} + \log \frac{\exp(\text{sim}(\mathbf{t}_i, \mathbf{I}_i)/\tau)}{\sum_{\mathbf{I}_j \in \mathcal{B}} \mathbb{I}_{j \neq i} \exp(\text{sim}(\mathbf{t}_i, \mathbf{I}_j)/\tau)}, \quad (1)$$

where $\mathbb{I}$ is the indicator function, τ represents the temperature parameter, and $\mathcal{B}$ signifies the batch size.

For the local-level visual-semantic alignment, the similarity between the visual and semantic modality is obtained with:

$$\text{sim}(\mathbf{I}_i, \mathbf{t}_i) = \frac{1}{HW} \sum_{k=1}^{HW} \cos(\mathbf{f}_i^k, \mathbf{t}_i), \quad (2)$$

where $\mathbf{f}_i^k$ is the k-th local feature representation of feature map $\mathbf{F}_i$. For the global-level visual-semantic alignment, the visual-semantic similarity is:

$$\text{sim}(\mathbf{I}_i, \mathbf{t}_i) = \cos(\mathbf{v}_i, \mathbf{t}_i), \quad (3)$$

where cos is the cosine similarity. Then the overall loss in a batch size during the pre-training stage is:

$$\mathcal{L}_{\text{pre-trained}} = \mathcal{L}_{CE} + \lambda \mathcal{L}_{IT}^{global} + \eta \mathcal{L}_{IT}^{local}, \quad (4)$$

where $\mathcal{L}_{CE}$ denotes the cross-entropy loss, λ and η are hyper-parameters that balance losses. By leveraging this semantic guidance, our approach fosters a deeper understanding of the visual content and strengthens the alignment between visual and semantic modalities, facilitating the model to learn more meaningful and contextually relevant features that better capture the characteristics of the image.

Entities-Assisted Episode-based Fine-Tuning

Given an FSL episode task, which comprises query samples $\mathcal{Q}$ and a support set $\mathcal{S}$ with a limited number of visual samples and corresponding class names for each class, the fine-tuning stage aims to learn discriminative feature prototypes for each class and effective feature representations for the query samples. We utilize the visual encoder pre-trained in the first stage as the initial weight for this stage, while keeping the text encoder and text adapter frozen during the second stage to obtain the text features without any further training. To extract a representative feature prototype for each support class, we incorporate some prior semantic information related to the key attributes of the categories into

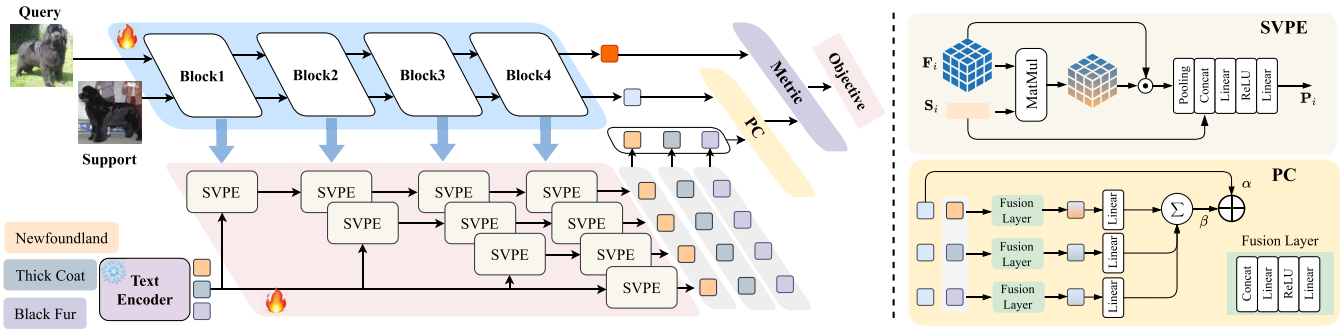


Figure 3: The framework of our progressive visual-semantic aggregation (PVSA) leverages both the class name ("Newfoundland") and class-related entities ("Thick Coat", "Black Fur") to enrich the visual prototypes. PVSA consists of a semantic-guided visual pattern extraction (SVPE) module that extracts visual patterns that are related to the class semantics and a prototype calibration (PC) module that enriches the visual prototype by incorporating the semantic-aware visual features.

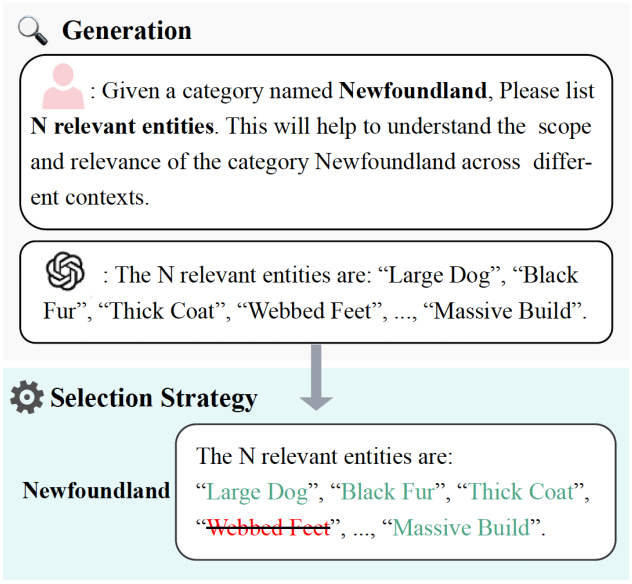


Figure 4: The entities produced with LLMs prompt and filtered with the proposed selection strategy.

the fine-tuning phase. Unlike the abstract conceptualization process that human beings naturally undertake when interpreting visual samples, our approach leverages the derivative and reasoning capabilities of large language models (LLMs) to envision the class entities from the class names, as shown in Fig. 4.

Entity Generation via LLMs. As depicted in Fig. 4, the class-specific concepts or entities are first generated by off-the-shelf LLMs and then filtered the irrelevant entities with a selection strategy. Given a class name, *e.g.*, Newfoundland, we elaborate the specific prompts for LLMs to generate N class-related entities, i.e., $E = \{e_n\}_{n=1}^{N}$. However, the generated entities may exhibit hallucinations, leading to the production of irrelevant or unreliable entities. To overcome this limitation, we introduce a selection strategy that assesses the textual similarity between the class name and the generated entities, enabling the identification and selection of only the most relevant entities for inclusion. Given a category name t, we modify it using the prompt "a photo of a $[t]$" and then calculate the similarity between each entity e_i and the category name t using the text encoder θ_T, with the following equation:

$$\text{sim}(e_n, t) = \cos(\theta_T(e_n),\ \theta_T(t)), \tag{5}$$

where cos represents the cosine similarity metric. Then, we rank the entities based on their similarity scores and select the top-K entities as the class-specific entities.

Progressive Visual-Semantic Aggregation

As illustrated in Fig. 3, we introduce a progressive visual-semantic aggregation (PVSA) framework that leverages both the class name and class-related entities generated by the LLMs to enrich the visual prototypes. This focuses on concentrating class-specific visual features at each stage of information extraction, allowing for more refined visual patterns through semantic guidance. Specifically, PVSA mainly consists of a semantic-guided visual pattern extraction module and a prototype calibration module.

Semantic-guided Visual Pattern Extraction (SVPE). SVPE is the key module to extract visual patterns that are related to the class names or class-wise entities. Given the feature map $\mathbf{F}_i \in \mathbb{R}^{HW \times D}$ from i-th block, the semantic-aware visual feature $\mathbf{P}_i$ is obtained with a SVPE module, which is formulated with:

$$\mathbf{P}_i = \text{SVPE}(\mathbf{S}_i, \mathbf{F}_i). \tag{6}$$

For the first block, $\mathbf{S}_i$ is the feature embedding from the semantic feature set that consists of the semantic entities $\mathbf{T}_k (k \leq K)$ and class names $\mathbf{T}_{cls}$. For the other block, $\mathbf{S}_i$ is the feature embedding from the joint set that consists of both semantic features and the semantic-aware visual feature $\mathbf{P}_{i-1}$ from the previous block. Specifically, the SVPE module is formulated as:

$$\begin{aligned} \mathbf{F}'_i &= \text{Pooling}(\text{MatMul}(\mathbf{S}_i, \mathbf{F}_i) \cdot \mathbf{F}_i), \\ \mathbf{P}_i &= \text{FusionLayer}(\mathbf{F}'_i,\ \mathbf{S}_i), \end{aligned} \tag{7}$$

where $\mathrm{FusionLayer}$ integrates input pairs $\mathbf{F}'_i$, $\mathbf{S}_i$ by first concatenating and then processing them through a two-layer network with LeakyReLU activation.

Once the semantic-aware visual features about the specific semantic entity or class name from the final block of the visual backbone have been obtained, they are then concatenated with the visual support prototypes for further refinement. For k-th semantic entity, we obtain:

$$\mathbf{P}^k_{entity} = \mathrm{FusionLayer}(\{\mathbf{P}^n_{entity\text{-}k}\}^4_{n=1}), \tag{8}$$

where $\mathrm{FusionLayer}$ is a concatenation operation followed by a two-layer network. Thus, we obtain the corresponding class-level semantic prototype as $\mathbf{P}_{cls}$ and the entity-level semantic prototypes as $\mathbf{P}_{entity} = \{\mathbf{P}^k_{entity} \mid k = 1, \ldots, K\}$, K is the entity number.

Prototype Calibration. The prototype calibration (PC) module incorporates diverse semantic-aware visual features into the feature embedding of the support samples, thereby generating a refined, calibrated visual prototype that captures a more comprehensive and semantically enriched representation. The process is formulated as:

$$\begin{aligned}
&\mathbf{C}_{cls} = \mathrm{FusionLayer}(\mathbf{f}^s, \mathbf{P}_{cls}),\\
&\mathbf{C}^k_{entity} = \mathrm{FusionLayer}(\mathbf{f}^s, \mathbf{P}^k_{entity}),\\
&\mathbf{C} = \alpha \cdot \mathbf{f}^s + \beta \cdot (\mathrm{FC}(\mathbf{C}_{cls}) + \sum_{k=1}^{K} \mathrm{FC}(\mathbf{C}^k_{entity})),
\end{aligned} \tag{9}$$

where $\mathbf{f}^s = \mathrm{AveragePooling}(\mathbf{F}^N)$ represents the average-pooled feature from the final block, $\mathbf{P}_{cls}$ and $\mathbf{P}_{entity}$ respectively denote the class fusion prototype and entities fusion prototype obtained with Eq. (8). FC denotes a fully connected layer for feature dimensionality mapping. α and β are two hyper-parameters.

To this end, the model culminates in the generation of a comprehensive representative prototype, denoted as $\mathbf{C}$, which seamlessly integrates both semantic and visual features. The multi-layered fusion strategy, where semantic and visual signals are harmoniously combined at each stage, empowers our model to leverage the full potential of both information streams.

Objective Function. The probability of the query sample q to the i-th class can be obtained with:

$$p_i = \frac{\exp\left(\cos\left(f\left(q_i\right), \mathbf{C}_i\right)/\tau\right)}{\sum_{j=1}^{N} \exp\left(\cos\left(f\left(q_i\right), \mathbf{C}_j\right)/\tau\right)}, \tag{10}$$

where $f(q_i)$ is the feature embedding of query sample q, $\cos$ denotes the cosine similarity and τ is the temperature ratio. The objective function is defined with a cross-entropy loss:

$$\mathcal{L} = \sum_{i=1}^{M} \mathrm{CrossEntropy}(p_i, y_i), \tag{11}$$

where y_i donates the corresponding ground-truth label of the query image q_i, and M is the number of query samples in each episode.

Experiments

Experimental Details

Datasets. We evaluate the proposed method across two primary tasks: the traditional FSL and the cross-domain FSL (CD-FSL). The traditional FSL is evaluated on four datasets, namely **MiniImageNet** (Vinyals et al. 2016), **TieredImageNet** (Ren et al. 2018), **CIFAR-FS** (Lee et al. 2019), and **FC100** (Oreshkin, Rodríguez López, and Lacoste 2018). Following (Guo et al. 2020), we evaluate the CD-FSL on BSCD-FSL benchmark, which involves training on MiniImageNet and testing on four unrelated datasets: **ChestX** (Wang et al. 2017), **ISIC** (Tschandl, Rosendahl, and Kittler 2018), **EuroSAT** (Helber et al. 2019), and **CropDisease** (Mohanty, Hughes, and Salathé 2016).

Training Details. We employ two vision backbones for comparison: ResNet12 (He et al. 2016) and Visformer-T (Chen et al. 2021b). For text encoding, we use ViT-B/32 CLIP (Radford et al. 2021) and we utilize GPT-4-o (OpenAI 2023) to generate related entities. The pre-training stage is set to 200 epochs for all datasets, while the meta-training stage is set to 50 epochs. The α and β in Eq. (9) are consistently assigned values of 0.2 and 0.8, respectively, across all datasets. During the pre-training phase, we set the batch size to 128, leveraging the Adam optimizer (Kingma and Ba 2014) with a learning rate of 1e-4 for optimization of the model parameters.

Evaluation Protocol. For the evaluation, we uniformly sampled 600 classification tasks from a novel set that comprises classes that are disjoint from those in the base set. In each task, there are 15 query samples for each class. The mean and 95% confidence interval of the accuracy are reported.

Performance Comparison

Few-Shot Classification. Table 1 provides a comprehensive comparison of the performance of our method against both visual-only-based and semantic-incorporated-based competitors under the 5-way 1-shot task across four benchmarks. From the results, we observe that our method achieves the best with both CNN and Transformer backbones on four datasets, outperforming the second-best competitor SemFew (Zhang et al. 2024) by significant margins. We speculate that the superior performance is due to our method capturing semantic-rich feature representations for each support class. Besides, we observe that the methods incorporating semantic information exhibit significantly superior performance compared to those solely reliant on visual cues, suggesting that prior semantic knowledge can serve as valuable hints, enhancing the overall effectiveness and accuracy of the tasks at hand. Moreover, the results obtained with the Visformer-T backbone perform better than those obtained with the ResNet backbone.

Cross-Domain Classification. Table 2 shows the comparison results of our method and six competitors on the BSCD-FSL benchmark. Our approach demonstrates a substantial advantage over the other competitors, achieving a notable improvement ranging from 1.39% to 3.71% over the second-best performing method across four diverse datasets.

	Method	Backbone	MiniImageNet	TieredImageNet	CIFAR-FS	FC100	Average
Visual	SUN (Dong et al. 2022)	ViT-S	67.80 ± 0.45	72.99 ± 0.50	78.37 ± 0.46	-	-
	FewTURE (Hiller et al. 2022)	Swin-T	72.40 ± 0.78	76.32 ± 0.87	77.76 ± 0.81	47.68 ± 0.78	68.54
	FGFL (Cheng et al. 2023)	ResNet-12	69.14 ± 0.80	73.21 ± 0.88	-	-	-
	CPEA (Hao et al. 2023)	ViT-S	71.97 ± 0.65	76.93 ± 0.70	77.82 ± 0.66	47.24 ± 0.58	68.49
	Meta-AdaM (Sun and Gao 2023)	ResNet-12	59.89 ± 0.49	65.31 ± 0.48	-	41.12 ± 0.49	-
	ALFA (Baik et al. 2023)	ResNet-12	66.61 ± 0.28	70.29 ± 0.40	76.32 ± 0.43	44.54 ± 0.50	64.44
	LastShot (Ye et al. 2022)	ResNet-12	67.35 ± 0.20	72.43 ± 0.23	76.76 ± 0.21	44.08 ± 0.18	65.16
Semantic	SVAE-Proto (Xu and Le 2022)	ResNet-12	74.84 ± 0.23	76.98 ± 0.65	-	-	-
	SP-CLIP (Chen et al. 2023)	Visformer-T	72.31 ± 0.40	78.03 ± 0.46	82.18 ± 0.40	48.53 ± 0.38	70.26
	SIFT (Pan, Xin, and Shen 2024)	WRN-28-10	77.31 ± 0.67	77.86 ± 0.77	81.35 ± 0.75	-	-
	Sem-Few (Zhang et al. 2024)	ResNet-12	77.63 ± 0.63	78.96 ± 0.80	83.65 ± 0.70	54.36 ± 0.71	73.65
	ECER-FSL (Ours)	ResNet-12	80.34 ± 0.21	80.79 ± 0.57	85.13 ± 0.61	56.12 ± 0.41	75.60
	ECER-FSL (Ours)	Visformer-T	**81.14 ± 0.15**	**81.81 ± 0.51**	**86.01 ± 0.35**	**57.34 ± 0.31**	**76.58**

Table 1: Results (%) on four datasets on **5-way 1-shot** tasks. The ± shows 95% confidence intervals. "Visual" and "Semantic" indicate the visual-only-based and semantic-incorporated-based methods. The results for the competitors are directly from the published literature. The best and second-best results are shown in **bold** and underline, respectively.

Method	ChestX	ISIC	EuroSAT	CropDisease
GNN (Satorras and Estrach 2018)	22.00	32.02	63.69	64.48
ATA (Wang and Deng 2021)	22.10	33.21	61.35	67.47
AFA (Hu and Ma 2022)	22.92	33.21	63.12	67.61
StyleAdv (Fu et al. 2023)	22.64	33.96	70.94	74.13
LDP-net (Zhou et al. 2023)	23.01	33.97	65.11	69.64
Dara (Zhao et al. 2023)	22.92	36.42	67.42	80.74
ECER-FSL (Ours)	**25.12**	**40.13**	**74.13**	**82.13**
△	2.11	3.71	3.19	1.39

Table 2: Average results (%) on BSCD-FSL benchmarks. △ denotes ECER-FSL's gain over the second-best competitors.

$\mathcal{L}_{CE}$	$\mathcal{L}_{global}$	$\mathcal{L}_{local}$	*Adapter*	MiniImageNet	TieredImageNet
✓	✗	✗	✗	78.30 ± 0.38	79.10 ± 0.47
✓	✓	✗	✗	79.21 ± 0.29	79.35 ± 0.62
✓	✓	✓	✗	79.78 ± 0.31	80.32 ± 0.27
✓	✓	✓	✓	80.34 ± 0.21	80.79 ± 0.57

Table 3: Ablation results (%) of during pre-training stage.

This underscores the significant value of incorporating rich semantic information in cross-domain scenarios.

Ablation Study

To comprehensively evaluate the effectiveness of our method, we conduct experiments on both MiniImageNet and TieredImageNet datasets with ResNet12 as the backbone.

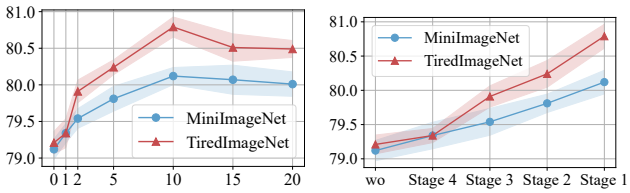
Ablation Study during Pre-training. Table 3 shows the ablation results during the pre-training stage. From the results, we observe that our multi-modality pre-training method performs a substantial improvement over the baseline model trained solely with CE loss. Specifically, combined $\mathcal{L}_{CE}$ with $\mathcal{L}_{global}$ improves the accuracy by an average of 0.91% and 0.25% on MiniImageNet and TieredImageNet, respectively. Meanwhile, the addition of $\mathcal{L}_{local}$ further enhances the performance, demonstrating that semantic alignment of local information helps feature representation learning. Notably, we observe an even greater enhancement in performance when employing the adapter to map textual features to visual representations, as opposed to the reverse direction.

SVPE	PC	Entity	MiniImageNet	TieredImageNet
✗	✗	✗	65.39 ± 0.29	68.32 ± 0.78
✓	✗	✗	78.30 ± 0.38	79.10 ± 0.47
✓	✓	✗	79.34 ± 0.33	79.51 ± 0.49
✓	✗	✓	79.92 ± 0.27	80.21 ± 0.29
✓	✓	✓	80.34 ± 0.21	80.79 ± 0.57

Table 4: Ablation results (%) during fine-tuning stage.

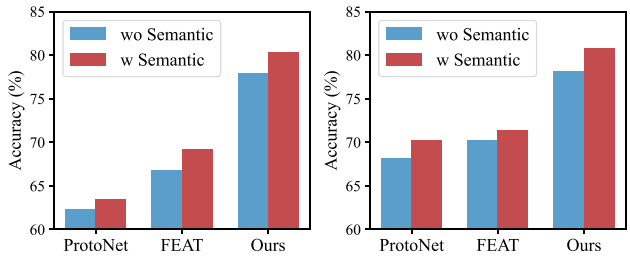
Ablation Study during Fine-tuning. Table 4 demonstrates that each module in our proposed PVSA framework during the fine-tuning stage. By incorporating the SVPE during the fine-tuning phase, we achieve substantial performance gains, with notable improvements of 12.91% and 10.78% on minImageNet and TieredImageNet, respectively. This suggests that progressively fusing semantic features into visual signals facilitates a better understanding of related class concepts. Moreover, the inclusion of the PC module significantly amplifies performance. Besides, the integration of entity concepts within this framework yields an additional leap in performance.

Impacts of Entity Numbers and SVPE Inserted Stages. In this experiment, we evaluate how the number of both entities and the SVPE modules inserted in the framework affected the performance of the method for the FSL task. To do so, we varied the number of entities from 0 to 20 and the number of SVPE modules. According to Fig. 5, we can observe that the performance is enhanced with an increased number of entities and plateaus when the number of entities is above 10, possibly because there is no further additional class information. Thus, the optimal number of entities is therefore set to 10. Furthermore, as the number of stages where semantic information is infused into the model increases from 1 to 4 blocks, we observe a consistent and incremental improvement in performance. This suggests that incorporating semantic information at multiple stages within



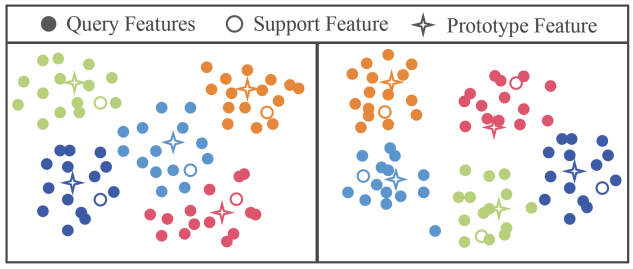
(a) Entity Numbers (b) SVPE Stages

Figure 5: Effects (%) of entity number and SVPE stages.



(a) MiniImageNet (b) TieredImageNet

Figure 6: Experimental results of combination FSL method with our pre-training model.



(a) MiniImageNet (b) TieredImageNet

Figure 7: Visualization of a random 5-way 1-shot task, where 15 query images and 1 support image for each test class from both MiniImageNet and TieredImageNet datasets.

the model architecture contributes positively to enhancing its overall capabilities.

Further Analysis

Combination FSL Competitors with Our Pre-training Model. In this experiment, we evaluate the performance when combining our pre-trained visual backbone with the existing FSL methods ProtoNet (Snell, Swersky, and Zemel 2017) and FEAT (Ye et al. 2020). As depicted in Fig. 6, it is evident that the existing methods, when coupled with our pre-training strategy, exhibit significantly superior performance compared to those utilizing a pre-training approach that solely focuses on a visual classification task. This advantage is consistent across both the miniImageNet and TieredImageNet datasets, underscoring the effectiveness of our pre-training strategy.

t-SNE Visualization. To further evaluate the impacts of our method in the feature embedding space, we visualize a random 5-way 1-shot test task sampled from both MiniImageNet and TieredImageNet datasets in Fig. 7. Each class contains 15 query samples, 1 support sample, and 1 class prototype. The features of query and support samples are derived from the visual backbone, and the class prototype is after the semantic fusion of PSVF and PC. The visualization results indicate that the refined class prototypes can represent class concepts more effectively than those obtained with the original limited visual samples.

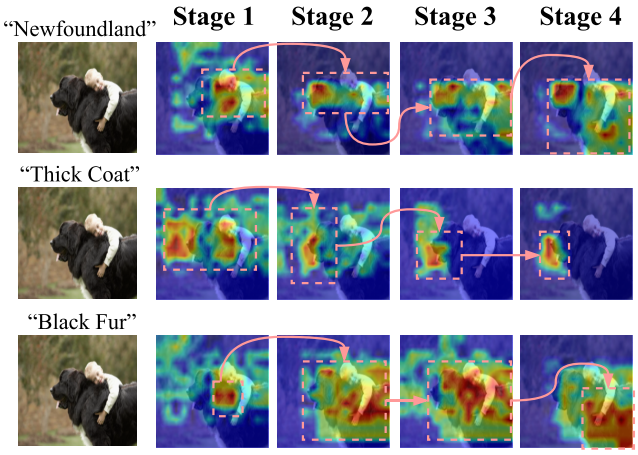


Figure 8: The feature map is computed based on the similarity between feature maps at different stages and the corresponding semantic features. The three rows correspond to "Newfoundland", "Thick Coat", and "Black Fur" respectively.

Feature Map Visualization. Fig. 8 shows the feature map. In this experiment, we further evaluate the visualization results at different network blocks under different semantic entities and the class name ("Newfoundland"). The provided visual sample may contain extraneous elements, such as people, grass, and other irrelevant details. From the result, our method effectively extracts information across a hierarchy of features, from low-level to high-level, allowing class-related entities to distill pertinent visual concepts and maintain a sharper focus. This is evident in distinct regions, with respective feature maps undergoing targeted optimization at different stages.

Conclusion

In this paper, we have revisited the pivotal role of semantics in the domain of few-shot learning (FSL). Our proposed model, ECER-FSL, has been designed to harness the full spectrum of semantic information. By leveraging the powerful expert knowledge in LLMs, we develop a strategy to generate and filter the class-related entities for constructing the comprehensive class concept. Additionally, by leveraging the progressive semantic-visual aggregation, ECER-FSL has demonstrated its capability to generate representative class prototypes, which are instrumental in enhancing classification accuracy, particularly in one-shot learning tasks. Experimental results on both the FSL task and cross-domain FSL task have shown the effectiveness of our ECER-FSL.

Acknowledgments

This research was supported in part by Zhejiang Provincial Natural Science Foundation of China under Grant LD24F020016, the Key R&D Program of Zhejiang Province, China 2023C01043, 2024C01G1752215, Science and Technology Innovation 2025 Major Project of Ningbo (2023Z236), National Science Foundation for Distinguished Young Scholars under Grant 62225605, Project 12326608 supported by NSFC.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Baik, S.; Choi, M.; Choi, J.; Kim, H.; and Lee, K. M. 2023. Learning to learn task-adaptive hyperparameters for few-shot learning. *IEEE TPAMI*, 46(3): 1441–1454.

Bär, A.; Houlsby, N.; Dehghani, M.; and Kumar, M. 2024. Frozen Feature Augmentation for Few-Shot Image Classification. In *CVPR*, 16046–16057.

Chen, W.; Liu, Y.; Kira, Z.; Wang, Y. F.; and Huang, J. 2019a. A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232*.

Chen, W.; Si, C.; Zhang, Z.; Wang, L.; Wang, Z.; and Tan, T. 2023. Semantic prompt for few-shot image recognition. In *CVPR*, 23581–23591.

Chen, Y.; Liu, Z.; Xu, H.; Darrell, T.; and Wang, X. 2021a. Meta-Baseline: Exploring Simple Meta-Learning for Few-Shot Learning. In *ICCV*, 9062–9071.

Chen, Z.; Fu, Y.; Zhang, Y.; Jiang, Y.; Xue, X.; and Sigal, L. 2019b. Multi-level semantic feature augmentation for one-shot learning. *IEEE TIP*, 28(9): 4594–4605.

Chen, Z.; Xie, L.; Niu, J.; Liu, X.; Wei, L.; and Tian, Q. 2021b. Visformer: The vision-friendly transformer. In *ICCV*, 589–598.

Cheng, H.; Yang, S.; Zhou, J. T.; Guo, L.; and Wen, B. 2023. Frequency Guidance Matters in Few-Shot Learning. In *ICCV*, 11814–11824.

Dan, J.; Liu, W.; Liu, M.; Xie, C.; Dong, S.; Ma, G.; Tan, Y.; and Xing, J. 2024a. HOGDA: Boosting Semi-supervised Graph Domain Adaptation via High-Order Structure-Guided Adaptive Feature Alignment. In *ACM MM*, 11109–11118.

Dan, J.; Liu, Y.; Deng, J.; Xie, H.; Li, S.; Sun, B.; and Luo, S. 2024b. Topofr: A closer look at topology alignment on face recognition. *arXiv preprint arXiv:2410.10587*.

Dan, J.; Liu, Y.; Xie, H.; Deng, J.; Xie, H.; Xie, X.; and Sun, B. 2023. Transface: Calibrating transformer training for face recognition from a data-centric perspective. In *ICCV*, 20642–20653.

Dong, B.; Zhou, P.; Yan, S.; and Zuo, W. 2022. Self-promoted supervision for few-shot transformer. In *ECCV*, 329–347.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Int. Conf. Mach. Learn.*, 1126–1135. PMLR.

Fu, M.; and Zhu, K. 2024. Instance-based Max-margin for Practical Few-shot Recognition. In *CVPR*, 28674–28683.

Fu, Y.; Xie, Y.; Fu, Y.; and Jiang, Y.-G. 2023. StyleAdv: Meta Style Adversarial Training for Cross-Domain Few-Shot Learning. In *CVPR*, 24575–24584.

Guo, Y.; Codella, N. C.; Karlinsky, L.; Codella, J. V.; Smith, J. R.; Saenko, K.; Rosing, T.; and Feris, R. 2020. A broader study of cross-domain few-shot learning. In *ECCV*, 124–141.

Hao, F.; He, F.; Liu, L.; Wu, F.; Tao, D.; and Cheng, J. 2023. Class-Aware Patch Embedding Adaptation for Few-Shot Image Classification. In *ICCV*, 18905–18915.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

He, W.; Fu, S.; Liu, M.; Wang, X.; Xiao, W.; Shu, F.; Wang, Y.; Zhang, L.; Yu, Z.; Li, H.; et al. 2025. Mars: Mixture of auto-regressive models for fine-grained text-to-image synthesis. In *AAAI*.

Helber, P.; Bischke, B.; Dengel, A.; and Borth, D. 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*

Hiller, M.; Ma, R.; Harandi, M.; and Drummond, T. 2022. Rethinking generalization in few-shot classification. In *NeurIPS*, 3582–3595.

Hu, M.; Chang, H.; Guo, Z.; Ma, B.; Shan, S.; and Chen, X. 2023. Understanding few-shot learning: Measuring task relatedness and adaptation difficulty via attributes. In *NeurIPS*.

Hu, Y.; and Ma, A. J. 2022. Adversarial feature augmentation for cross-domain few-shot classification. In *ECCV*, 20–37.

Ji, Z.; Hou, Z.; Liu, X.; Pang, Y.; and Han, J. 2022. Information Symmetry Matters: A Modal-Alternating Propagation Network for Few-Shot Learning. *IEEE TIP*, 31: 1520–1531.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lazarou, M.; Stathaki, T.; and Avrithis, Y. 2022. Tensor feature hallucination for few-shot learning. In *WACV*, 3500–3510.

Lee, K.; Maji, S.; Ravichandran, A.; and Soatto, S. 2019. Meta-learning with differentiable convex optimization. In *CVPR*, 10657–10665.

Li, A.; Huang, W.; Lan, X.; Feng, J.; Li, Z.; and Wang, L. 2020. Boosting few-shot learning with adaptive margin loss. In *CVPR*, 12576–12584.

Liu, M.; Ma, Y.; Zhen, Y.; Dan, J.; Yu, Y.; Zhao, Z.; Hu, Z.; Liu, B.; and Fan, C. 2025. Llm4gen: Leveraging semantic representation of llms for text-to-image generation. In *AAAI*.

Liu, Q.; Cao, W.; and He, Z. 2023. Cycle optimization metric learning for few-shot classification. *PR*, 109468.

Liu, Y.; Zhang, J.; Peng, D.; Huang, M.; Wang, X.; Tang, J.; Huang, C.; Lin, D.; Shen, C.; Bai, X.; et al. 2023. Spts v2: single-point scene text spotting. *T-PAMI*.

Mohanty, S. P.; Hughes, D. P.; and Salathé, M. 2016. Using deep learning for image-based plant disease detection. *Frontiers in plant science*.

OpenAI. 2023. ChatGPT. https://chat.openai.com.

Oreshkin, B.; Rodríguez López, P.; and Lacoste, A. 2018. Tadam: Task dependent adaptive metric for improved few-shot learning. In *NeurIPS*.

Pan, M.-H.; Xin, H.-Y.; and Shen, H.-B. 2024. Semantic-Based Implicit Feature Transform for Few-Shot Classification. *IJCV*, 1–16.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763.

Ren, M.; Triantafillou, E.; Ravi, S.; Snell, J.; Swersky, K.; Tenenbaum, J. B.; Larochelle, H.; and Zemel, R. S. 2018. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*.

Satorras, V. G.; and Estrach, J. B. 2018. Few-shot learning with graph neural networks. In *ICLR*.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. In *NeurIPS*.

Sun, S.; and Gao, H. 2023. Meta-AdaM: A Meta-Learned Adaptive Optimizer with Momentum for Few-Shot Learning. In *NeurIPS*.

Sun, S.; and Gao, H. 2024. Meta-AdaM: An meta-learned adaptive optimizer with momentum for few-shot learning. In *NeurIPS*.

Tang, J.; Qian, W.; Song, L.; Dong, X.; Li, L.; and Bai, X. 2022a. Optimal boxes: boosting end-to-end scene text recognition by adjusting annotated bounding boxes via reinforcement learning. In *ECCV*, 233–248.

Tang, J.; Zhang, W.; Liu, H.; Yang, M.; Jiang, B.; Hu, G.; and Bai, X. 2022b. Few could be better than all: Feature sampling and grouping for scene text detection. In *CVPR*, 4563–4572.

Tian, Y.; Wang, Y.; Krishnan, D.; Tenenbaum, J. B.; and Isola, P. 2020. Rethinking few-shot image classification: a good embedding is all you need? In *ECCV*, 266–282.

Tschandl, P.; Rosendahl, C.; and Kittler, H. 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D.; et al. 2016. Matching networks for one shot learning. In *NeurIPS*.

Wang, H.; and Deng, Z.-H. 2021. Cross-Domain Few-Shot Classification via Adversarial Task Augmentation. In *IJCAI*, 1075–1081.

Wang, W.; Lv, Q.; Yu, W.; Hong, W.; Qi, J.; Wang, Y.; Ji, J.; Yang, Z.; Zhao, L.; Song, X.; et al. 2023. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; and Summers, R. M. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *CVPR*, 2097–2106.

Xie, J.; Long, F.; Lv, J.; Wang, Q.; and Li, P. 2022. Joint Distribution Matters: Deep Brownian Distance Covariance for Few-Shot Classification. In *CVPR*.

Xing, C.; Rostamzadeh, N.; Oreshkin, B.; and O Pinheiro, P. O. 2019. Adaptive cross-modal few-shot learning. In *NeurIPS*.

Xu, J.; and Le, H. 2022. Generating Representative Samples for Few-Shot Classification. In *CVPR*, 9003–9013.

Yang, F.; Wang, R.; and Chen, X. 2022. SEGA: semantic guided attention on visual prototype for few-shot learning. In *WACV*, 1056–1066.

Yang, Z.; Wang, J.; and Zhu, Y. 2022. Few-shot classification with contrastive learning. In *ECCV*, 293–309.

Ye, H.-J.; Hu, H.; Zhan, D.-C.; and Sha, F. 2020. Few-shot learning via embedding adaptation with set-to-set functions. In *CVPR*, 8808–8817.

Ye, H.-J.; Ming, L.; Zhan, D.-C.; and Chao, W.-L. 2022. Few-shot learning with a strong teacher. *IEEE TPAMI*, 46(3): 1425–1440.

Yu, Y.; Zhang, D.; Li, Y.; and Zhang, Z. 2022. Multi-Proxy Learning from an Entropy Optimization Perspective. In *IJCAI*, 1594–1600.

Zhang, B.; Li, X.; Ye, Y.; Huang, Z.; and Zhang, L. 2021. Prototype completion with primitive knowledge for few-shot learning. In *CVPR*, 3754–3762.

Zhang, H.; Xu, J.; Jiang, S.; and He, Z. 2024. Simple Semantic-Aided Few-Shot Learning. In *CVPR*, 28588–28597.

Zhang, J.; Zhao, C.; Ni, B.; Xu, M.; and Yang, X. 2019. Variational few-shot learning. In *CVPR*, 1685–1694.

Zhang, R.; Che, T.; Ghahramani, Z.; Bengio, Y.; and Song, Y. 2018. Metagan: An adversarial approach to few-shot learning. In *NeurIPS*.

Zhao, W.; Feng, H.; Liu, Q.; Tang, J.; Wei, S.; Wu, B.; Liao, L.; Ye, Y.; Liu, H.; Zhou, W.; Li, H.; and Huang, C. 2024a. TabPedia: Towards Comprehensive Visual Table Understanding with Concept Synergy. In *NeurIPS*.

Zhao, Y.; Zhang, T.; Li, J.; and Tian, Y. 2023. Dual Adaptive Representation Alignment for Cross-domain Few-shot Learning. In *IEEE TPAMI*.

Zhao, Z.; Tang, J.; Lin, C.; Wu, B.; Huang, C.; Liu, H.; Tan, X.; Zhang, Z.; and Xie, Y. 2024b. Multi-modal In-Context Learning Makes an Ego-evolving Scene Text Recognizer. In *CVPR*, 15567–15576.

Zhao, Z.; Tang, J.; Wu, B.; Lin, C.; Wei, S.; Liu, H.; Tan, X.; Zhang, Z.; Huang, C.; and Xie, Y. 2024c. Harmonizing Visual Text Comprehension and Generation. In *NeurIPS*.

Zhou, F.; Wang, P.; Zhang, L.; Wei, W.; and Zhang, Y. 2023. Revisiting Prototypical Network for Cross Domain Few-Shot Learning. In *CVPR*, 20061–20070.