PMCE: Probabilistic Multi-Granularity Semantics with Caption-Guided Enhancement for Few-Shot Learning

Jiaying Wu[1], **Can Gao**[1], **Jinglu Hu**[2], **Hui Li**[1], **Xiaofeng Cao**[3] and **Jingcai Guo**[4*]
[1]School of Computer Engineering, Jiangsu Ocean University, China
[2]Graduate School of Information, Production and Systems, Waseda University, Japan
[3]School of Computer Science and Technology, Tongji University, China
[4]Department of Computing & LSGI, The Hong Kong Polytechnic University, Hong Kong SAR
{jiaying, 2024220911}@jou.edu.cn, jinglu@waseda.jp, lih@jou.edu.cn, xiaofengcao@tongji.edu.cn, jc-jingcai.guo@polyu.edu.hk

Abstract

Few-shot learning aims to recognize novel categories from limited labeled samples, where prototypes estimated from 1–5 supports per class are often unreliable. Semantic-based approaches alleviate this by introducing class-level priors, but they often ignore instance-level cues and rarely optimize queries under the inductive protocol. We propose PMCE, a **P**robabilistic framework that leverages **M**ulti-granularity semantics with **C**aption-guided **E**nhancement for few-shot classification. On base classes, we build a knowledge bank with class-wise visual statistics and class-name embeddings. At test time, the semantic embedding of a novel class retrieves a few similar base classes whose visual priors are aggregated into a class-specific prior and combined with the support-based prototype via MAP estimation. Simultaneously, we train a lightweight enhancer on base classes that fuses frozen BLIP captions with visual features, and apply it to both supports and queries without using query-set statistics or novel labels. A simple caption-consistency regularizer further improves robustness to noisy captions. Experiments on four standard benchmarks with ResNet-12 and Swin-T backbones show that PMCE outperforms state-of-the-art few-shot baselines, achieving up to **7.71%** gains over the strongest competitor on MiniImageNet in the challenging 1-shot setting. Our code is available at https://github.com/channa419/PMCE.

1 Introduction

Few-shot learning studies the recognition of novel categories from only a handful of labeled examples. Although deep networks perform well with large-scale supervision, they often overfit when only 1–5 supports per class are available. As a result, the gains over simple baselines are often modest, especially in the 1-shot regime and on visually diverse benchmarks [Fei-Fei *et al.*, 2003; Chen *et al.*, 2019].

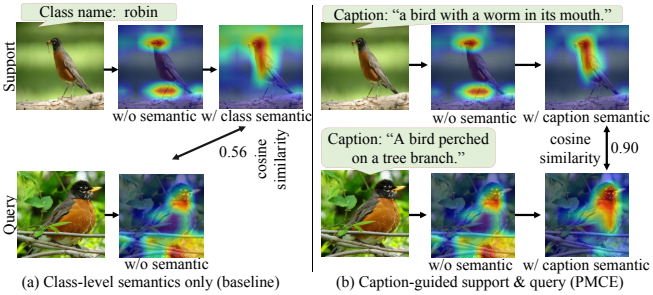

Figure 1: Semantic cues for support–query alignment visualized by Grad-CAM. **(a)** With class-name semantics applied only on the support side, the query remains purely visual and the alignment is weak (cosine similarity: 0.56). **(b)** PMCE uses a frozen BLIP captioner to provide label-free instance captions for both support and query, yielding more focused Grad-CAM heatmaps and substantially stronger alignment (cosine similarity: 0.90).

To alleviate data scarcity, metric-based and distribution-based methods learn transferable embedding spaces or reuse base-class statistics. Matching Networks [Vinyals *et al.*, 2016], Prototypical Networks [Snell *et al.*, 2017], MAML [Finn *et al.*, 2017] and later strong baselines [Chen *et al.*, 2019] improve the metric, while distribution calibration approaches [Yang *et al.*, 2022; Wu *et al.*, 2021; Wu and Hu, 2022; Xu *et al.*, 2026] transfer base statistics to novel classes by adjusting empirical feature distributions or sampling from Gaussian priors. These methods show the benefit of a probabilistic view, but their priors are defined and selected purely in the visual space, mainly according to feature similarity and hand-designed rules.

Semantic-based methods introduce attributes, word vectors or text embeddings to compensate for limited visual data. AM3 [Xing *et al.*, 2019] adapts visual prototypes with class-level semantics. With CLIP [Radford *et al.*, 2021], BMI [Li and Wang, 2023], SP-CLIP [Chen *et al.*, 2023b], SemFew [Zhang *et al.*, 2024], PRSN [Dong *et al.*, 2025] and related work encode class names with prompts and inject the resulting text features into metric learning pipelines or prototype resynthesis, leading to strong and robust baselines. However, most methods rely on coarse class-level text and mainly

*Corresponding author

modify support prototypes or classifier weights, while query features remain purely visual or are adjusted in a transductive manner. Instance-level semantics such as captions are rarely modeled under the standard inductive few-shot protocol.

In this work, we propose PMCE motivated by a simple yet critical observation, that is, in the inductive setting, improving the support prototype alone is often not sufficient when the query representation remains purely visual and deviates from the prototype manifold. We address this by organizing semantics at two complementary granularities. At the class level, text embeddings act as retrieval keys to identify relevant base-class statistics as priors for MAP prototype estimation. At the instance level, captions provide label-free descriptions that optimize both support and query representations into a more coherent space. This design is particularly helpful in 1-shot regime, where support evidence is scarce and support–query mismatch becomes a major source of errors.

In summary, for this paper, the main contributions are:

- We present PMCE, which connects semantic-based and distribution-based few-shot learning by using class-name embeddings to retrieve class-specific priors from a non-parametric knowledge bank and injecting them into support prototypes under a MAP view.
- We design a caption-guided dual-stream enhancer that integrates BLIP-generated captions with visual features for both supports and queries in an inductive setting, yielding a more coherent space for prototype matching.
- We introduce a lightweight caption consistency regularization and demonstrate consistent gains on four benchmarks with two backbones, with the largest improvements in the 1-shot regime.

2 Related Work

In this section, we briefly review three lines of work that are most related to PMCE and clarify how our framework fits into this landscape.

Metric- and distribution-based few-shot learning. Traditional few-shot methods often learn a transferable embedding space or a good initialization, such as Matching Networks [Vinyals *et al.*, 2016], Prototypical Networks [Snell *et al.*, 2017], and MAML [Finn *et al.*, 2017], as well as strong baselines that carefully revisit pre-training and episodic training [Chen *et al.*, 2019]. Recent work further improves the representations or matching rules, for example FGFL [Cheng *et al.*, 2023] and CPEA [Hao *et al.*, 2023]. Another line transfers base-class statistics to novel classes by modeling feature distributions, including distribution calibration [Yang *et al.*, 2022] and MAP-based prototype estimation [Wu *et al.*, 2021; Wu and Hu, 2022]. These methods are effective, but priors are usually selected only in the visual space based on feature similarity or simple heuristics.

Semantic-based few-shot learning. Semantic-based methods introduce attributes or text embeddings to compensate for limited visual data [Guo and Guo, 2021; Guo *et al.*, 2023; Guo *et al.*, 2024]. AM3 [Xing *et al.*, 2019] adapts prototypes with class-level semantics, while CLIP-based [Radford *et al.*, 2021], approaches such as BMI [Li and Wang, 2023], SP-CLIP [Chen *et al.*, 2023b], SemFew [Zhang *et al.*, 2024], and PRSN [Dong *et al.*, 2025] encode class names with prompts and inject text features into metric learning or prototype construction. Most of them mainly modify support prototypes or classifier weights; query features are often left purely visual, or refined transductively using query-set statistics.

Vision–language models and caption-based semantics. Vision–language pre-training offers general-purpose text cues for images [Lu *et al.*, 2023]. CLIP learns aligned image–text embeddings [Radford *et al.*, 2021], and BLIP can generate high-quality captions [Li *et al.*, 2022]. In few-shot learning, these cues are commonly used at the class levels [Chen *et al.*, 2023b; Zhang *et al.*, 2024], whereas automatic captions are less explored under the inductive setting. PMCE keeps CLIP and BLIP frozen as off-the-shelf semantic extractors, using class names for stable prior retrieval and captions for query-side refinement.

3 Preliminaries

In this section, we explain the few-shot setting and review the MAP-based prototype estimation that our method builds on.

3.1 Problem Setting

In the standard few-shot setting, we have a base set $\mathcal{D}_b = \{(x_i, y_i)\}_{i=1}^{|\mathcal{D}_b|} \subset \mathcal{X} \times \mathcal{C}_b$ and a novel set $\mathcal{D}_n = \{(x_i, y_i)\}_{i=1}^{|\mathcal{D}_n|} \subset \mathcal{X} \times \mathcal{C}_n$, where $\mathcal{C}_b \cap \mathcal{C}_n = \varnothing$. $\mathcal{D}_b$ provides abundant annotated data needed to train the visual backbone and the enhancement module, while $\mathcal{D}_n$ is used only at meta-test time.

At meta-test time, we evaluate the performance on the N-way K-shot tasks. Each task contains a support set $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{N\times K}$ and a query set $\mathcal{Q} = \{(x_j, y_j)\}_{j=1}^{N\times M}$, where M is the number of query samples per class. Given support set $\mathcal{S}$ and a query image x_q, the goal is to predict its label by

$$\hat{y}_q = \arg\max_{c\in\mathcal{C}} \; p(c \,|\, x_q, \mathcal{S}, \mathcal{D}_b), \tag{1}$$

where $\mathcal{C}$ denotes the set of N classes sampled in the current episode.

3.2 MAP-Based Prototype Estimation

Following [Wu *et al.*, 2021; Wu and Hu, 2022], we construct a visual knowledge bank based on the priors from base set. Specifically, for each base class $j \in \mathcal{C}_b$, we store its mean feature

$$\mu_j = \frac{1}{|\mathcal{D}_b^j|} \sum_{(x_i,y_i)\in\mathcal{D}_b^j} f_\theta(x_i), \tag{2}$$

where $\mathcal{D}_b^j = \{(x_i, y_i) \in \mathcal{D}_b \mid y_i = j\}$.

For a novel class c in an episode, p^c_{init} is the empirical prototype calculated from the mean of its support features. Given a class-specific prior mean μ^c_{prior} obtained from the knowledge bank, we calibrate the prototype by a MAP estimator under the Gaussian assumptions, which admits a closed-form convex combination:

$$\tilde{p}^c = \alpha\, p^c_{\text{init}} + (1-\alpha)\, \mu^c_{\text{prior}}, \tag{3}$$

where $\alpha \in [0, 1]$ controls the strength of the prior.

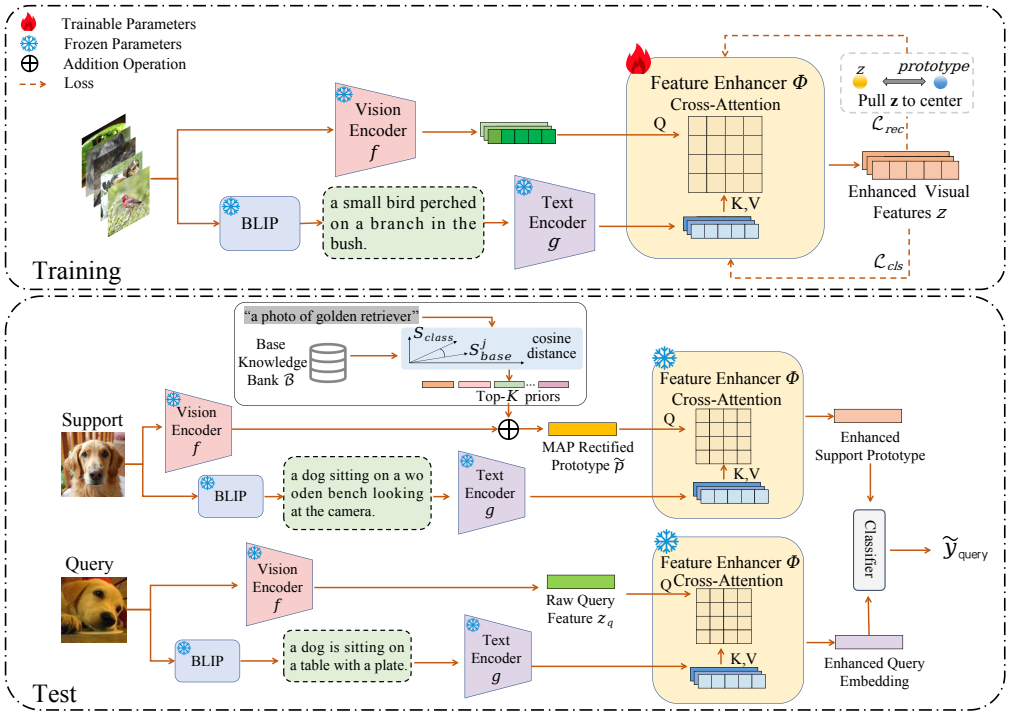

Figure 2: Overall framework of PMCE. **Training stage:** On base classes, we build a non-parametric knowledge bank that stores class-wise visual statistics and class-name embeddings, and train a caption-guided enhancer Φ by fusing visual features from f with BLIP-generated caption semantics from g. **Test stage:** (1) extract visual features and label-free captions for both supports and queries; (2) use class-name semantics to retrieve related base classes from the knowledge bank and obtain a class-specific prior, which is fused with the support prototype in a MAP-style update to get $\tilde{p}$; (3) apply Φ to enhance $\tilde{p}$ with aggregated support captions and to enhance query features with their own captions, and perform metric-based classification in the aligned feature space.

4 Method

In this section, we describe semantics-guided prior retrieval for MAP prototype estimation, the caption-guided dual-stream enhancer, and the training objective. An overview is shown in Fig. 2.

4.1 Semantics-Guided Prior Selection

We demonstrate the prior selection with CLIP class-name semantics for MAP prototype estimation in Eq. (3).

Class-level semantics on base classes. Let n_y be the natural language name of class y. We follow the CLIP-style prompting strategy and encode

$$s^y_{\text{class}} = g_\phi(\text{“a photo of a } n_y\text{”}) \in \mathbb{R}^{d_t}, \tag{4}$$

which provides a coarse description of category y. For each base class j, we store its mean feature μ_j and class-name embedding $s^j_{\text{base}} = s^j_{\text{class}}$ in a non-parametric semantic knowledge bank

$$\mathcal{B} = \{(\mu_j, s^j_{\text{base}})\}_{j=1}^{|\mathcal{C}_b|}. \tag{5}$$

Prior selection for novel classes. At meta-test time, for a novel class c with class name n_c and embedding s^c_{class}, we measure its similarity to each base class via cosine similarity:

$$\text{score}_j = \frac{s^c_{\text{class}} \cdot s^j_{\text{base}}}{\|s^c_{\text{class}}\| \, \|s^j_{\text{base}}\|}. \tag{6}$$

We then select the indices of the top-K base classes to obtain a neighbor set $\mathcal{N}_{\text{top}}$ and aggregate their visual means into a class-specific prior. Instead of uniform averaging, we use a similarity-weighted aggregation:

$$w_k = \frac{\exp(\text{score}_k/\tau)}{\sum_{u\in\mathcal{N}_{\text{top}}} \exp(\text{score}_u/\tau)}, \qquad \mu^c_{\text{prior}} = \sum_{k\in\mathcal{N}_{\text{top}}} w_k \mu_k, \tag{7}$$

where τ is a temperature.

MAP view with spherical uncertainties. We interpret the support prototype p^c_{init} as a noisy observation of the true class mean μ^c under a spherical Gaussian likelihood $p^c_{\text{init}} \sim \mathcal{N}(\mu^c, \sigma^2_{\text{like}}\mathbf{I})$, while the retrieved prior mean provides a spherical Gaussian prior $p(\mu^c) = \mathcal{N}(\mu^c_{\text{prior}}, \sigma^2_{\text{prior}}\mathbf{I})$. Under Gaussian–Gaussian conjugacy, the MAP estimate has a closed form that reduces to a convex combination:

$$\tilde{p}^c = \alpha\, p^c_{\text{init}} + (1-\alpha)\, \mu^c_{\text{prior}}, \qquad \alpha = \frac{\sigma^2_{\text{prior}}}{\sigma^2_{\text{prior}} + \sigma^2_{\text{like}}}. \tag{8}$$

Thus, α is determined by the relative confidence between the support evidence and the retrieved semantic prior.

4.2 Caption-Guided Dual-Stream Enhancement

In this section, we introduce the enhancer in our framework. We optimize representations with instance-level captions by applying the same enhancer Φ to the support and

query streams. Given a visual feature and its caption-derived semantic representation, Φ outputs an enhanced visual embedding used for prototype construction and query matching.

Instance-level semantics. To provide finer-grained cues and cover unlabeled queries, we use image captions as instance-level semantics. For any image x, we first generate a caption with a frozen BLIP captioner,

$$c = \text{BLIP}(x), \tag{9}$$

and then encode it with the same text encoder $g_\phi(\cdot)$ to obtain a single sentence embedding:

$$s_{\text{inst}} = g_\phi(c) \in \mathbb{R}^{d_t}, \tag{10}$$

where d_t is the text feature dimension. s_{inst} is label-free and thus available for both supports and queries, serving as the semantic input to Φ on the two streams.

Enhancement module. We define a mapping

$$\mathbf{v}_{\text{out}} = \Phi(\mathbf{v}_{\text{in}}, s_{\text{in}}; \Theta), \tag{11}$$

where $\mathbf{v}_{\text{in}} \in \mathbb{R}^{1\times d_v}$ is a visual anchor (e.g., a prototype or a query feature), $s_{\text{in}} \in \mathbb{R}^{d_t}$ is a semantic context, and Θ are learnable parameters. Since visual and textual features live in different spaces, we first project semantics to the visual dimension:

$$\mathbf{s}_{\text{proj}} = \text{ReLU}(\text{LN}(s_{\text{in}} W_p + b_p)) \in \mathbb{R}^{d_v}, \tag{12}$$

with $W_p \in \mathbb{R}^{d_t\times d_v}$, $b_p \in \mathbb{R}^{d_v}$ and $\text{LN}(\cdot)$ layer normalization.

We then perform cross-attention by using the visual anchor as query and the projected semantics as key/value:

$$\mathbf{Q}_{att} = \mathbf{v}_{\text{in}} W_Q, \quad \mathbf{K}_{att} = \mathbf{s}_{\text{proj}} W_K, \quad \mathbf{V}_{att} = \mathbf{s}_{\text{proj}} W_V, \tag{13}$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d_v\times d_k}$ are learnable projections. The attention output is computed as

$$\text{Attn}(\mathbf{Q}_{att}, \mathbf{K}_{att}, \mathbf{V}_{att}) = \text{Softmax}\left(\frac{\mathbf{Q}_{att}\mathbf{K}_{att}^\top}{\sqrt{d_k}}\right)\mathbf{V}_{att}. \tag{14}$$

A residual connection with a learnable scale β (initialized to a small value) preserves the original visual structure:

$$\mathbf{v}_{\text{out}} = \mathbf{v}_{\text{in}} + \beta\,\text{Attn}(\mathbf{Q}_{att}, \mathbf{K}_{att}, \mathbf{V}_{att}). \tag{15}$$

Support and query streams. We apply $\Phi(\cdot)$ to supports and queries in the same form. For class c, let $\mathcal{S}_c$ be its support set and $s_{\text{inst}}^{(i)}$ be the caption embedding of support $x_i \in \mathcal{S}_c$. We average support caption embeddings to form a class-level semantic context:

$$s_{\text{proto}}^c = \frac{1}{|\mathcal{S}_c|}\sum_{x_i\in\mathcal{S}_c} s_{\text{inst}}^{(i)} \in \mathbb{R}^{d_t}. \tag{16}$$

We then use the MAP-estimated prototype $\tilde{p}^c$ as the visual anchor to obtain the enhanced prototype:

$$\mathbf{P}_{\text{final}}^c = \Phi(\tilde{p}^c, s_{\text{proto}}^c). \tag{17}$$

For a query image x_q with visual feature $\mathbf{z}_q = f_\theta(x_q)$ and caption embedding s_{inst}^q, we compute

$$\mathbf{Z}_{\text{final}}^q = \Phi(\mathbf{z}_q, s_{\text{inst}}^q). \tag{18}$$

Each query is processed independently, so the overall framework remains strictly inductive.

4.3 Training Objective With Consistency Regularization

In this section, we present the objective we trained in our framework. We freeze f_θ and g_ϕ and train only the enhancer Φ together with an auxiliary classifier on base classes. Let $v_{\text{out}}^{(i)}$ be the enhanced feature of a training sample x_i with label y_i.

Classification and prototype-preserving regularization. We use a standard cross-entropy loss

$$\mathcal{L}_{\text{cls}} = -\frac{1}{B}\sum_{i=1}^{B}\log P(y_i \mid v_{\text{out}}^{(i)}), \tag{19}$$

and a lightweight prototype-preserving term that discourages excessive drift from the frozen backbone structure:

$$\mathcal{L}_{\text{rec}} = \frac{1}{B}\sum_{i=1}^{B}\left\|v_{\text{out}}^{(i)} - \mu_{\text{proto}}^{y_i}\right\|_1. \tag{20}$$

We use a small λ_{rec} so that the enhancer can still introduce caption-guided adaptation while keeping base-class geometry stable.

Caption consistency via a contrastive loss. Caption quality may vary, so we add a consistency regularization on caption semantics in the projected space. For sample x_i, let $\mathbf{s}_{\text{proj}}^{(i)} \in \mathbb{R}^{d_v}$ be the projected caption embedding (Eq. (12)). We normalize it as

$$v_i^{\text{cap}} = \frac{\mathbf{s}_{\text{proj}}^{(i)}}{\|\mathbf{s}_{\text{proj}}^{(i)}\|} \in \mathbb{R}^{d_v}. \tag{21}$$

We encourage captions from the same class to be coherent while maintaining inter-class separation via a supervised contrastive loss [Khosla *et al.*, 2020]:

$$\mathcal{L}_{\text{con}} = -\frac{1}{B}\sum_{i=1}^{B}\frac{1}{|\mathcal{P}(i)|}\sum_{p\in\mathcal{P}(i)}\log\frac{\exp\left(s_{i,p}/\tau_c\right)}{\sum_{\substack{a=1\\a\neq i}}^{B}\exp\left(s_{i,a}/\tau_c\right)}. \tag{22}$$

where $s_{i,a} = \cos\left(v_i^{\text{cap}}, v_a^{\text{cap}}\right)$ denotes the cosine similarity between the normalized projected caption embeddings of samples i and a. Here, $\mathcal{P}(i) = \{p \neq i \mid y_p = y_i\}$ is the positive set of sample i in the mini-batch, and τ_c is the temperature.

Overall objective. The final objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}}, \tag{23}$$

with λ_{rec} and λ_{con} tuned on the validation set. During meta-test, f_θ, g_ϕ and Φ are all frozen, and classification is performed with logistic regression using enhanced support and query features.

5 Experiments

In this section, we present the datasets and implementation details in the experiments, followed by experimental results, ablation, and visualization analysis.

Methods	Backbone	Semantic	MiniImageNet		TieredImageNet	
			1-shot	5-shot	1-shot	5-shot
Metric-based						
MatchNet [Vinyals *et al.*, 2016]	ResNet-12	No	65.64 ± 0.20	78.72 ± 0.15	68.50 ± 0.92	80.60 ± 0.71
ProtoNet [Snell *et al.*, 2017]	ResNet-12	No	60.37 ± 0.83	78.02 ± 0.57	53.31 ± 0.89	72.69 ± 0.74
RelationNet [Sung *et al.*, 2018]	ConvNet64	No	52.19 ± 0.83	70.20 ± 0.66	54.48 ± 0.93	71.32 ± 0.78
FGFL [Cheng *et al.*, 2023]	ResNet-12	No	69.14 ± 0.80	86.01 ± 0.62	73.21 ± 0.88	87.21 ± 0.61
CPEA [Hao *et al.*, 2023]	ViT-S	No	71.97 ± 0.65	87.06 ± 0.38	76.93 ± 0.70	90.12 ± 0.45
MCNet [Chen *et al.*, 2024]	ResNet-18	No	72.13 ± 0.43	88.20 ± 0.23	73.14 ± 0.46	87.16 ± 0.32
Cads [Zhang *et al.*, 2025]	ResNet-12	No	66.56 ± 0.19	82.74 ± 0.13	72.04 ± 0.22	86.47 ± 0.15
SRE-ProtoNet* [Chen *et al.*, 2025b]	ResNet-12	No	70.86 ± 0.44	87.01 ± 0.27	73.02 ± 0.50	87.78 ± 0.31
Distribution-based						
ImprovedMAP [Wu and Hu, 2022]	WRN-28-10	Yes	71.56 ± 0.74	85.05 ± 0.49	79.26 ± 0.08	90.41 ± 0.72
TDO [Liu *et al.*, 2023]	WRN-28-10	No	74.10 ± 0.23	86.44 ± 0.13	82.68 ± 0.22	90.85 ± 0.13
PBML [Fu *et al.*, 2025]	NesT ViT	No	70.04 ± 0.67	85.21 ± 0.58	76.73 ± 0.88	90.11 ± 0.65
DDC [Chen *et al.*, 2025a]	WRN-28-10	No	69.17 ± 1.25	85.23 ± 0.97	–	–
Semantic-based						
KTN [Peng *et al.*, 2019]	ConvNet	Yes	64.42 ± 0.72	74.16 ± 0.56	63.43 ± 0.21	74.32 ± 0.58
AM3 [Xing *et al.*, 2019]	ResNet-12	Yes	65.30 ± 0.49	78.10 ± 0.36	69.08 ± 0.47	82.58 ± 0.31
BMI [Li and Wang, 2023]	ResNet-12	Yes	77.01 ± 0.34	84.85 ± 0.27	78.37 ± 0.44	86.30 ± 0.32
MAVSI [Zhao *et al.*, 2024]	ResNet-12	Yes	69.74 ± 0.21	82.23 ± 0.41	74.61 ± 0.18	87.45 ± 0.46
PRSN [Dong *et al.*, 2025]	ResNet-12	Yes	71.09 ± 0.90	86.82 ± 0.61	74.14 ± 0.86	88.54 ± 0.61
Vision–language based						
SP-CLIP [Chen *et al.*, 2023b]	ViT-T	Yes	72.31 ± 0.40	83.42 ± 0.30	78.03 ± 0.46	88.55 ± 0.32
SP-SBERT [Chen *et al.*, 2023b]	ViT-T	Yes	70.70 ± 0.42	83.55 ± 0.30	73.31 ± 0.50	88.56 ± 0.32
SP-Glove [Chen *et al.*, 2023b]	ViT-T	Yes	70.81 ± 0.42	83.31 ± 0.30	74.68 ± 0.50	88.64 ± 0.31
SemFew-Trans [Zhang *et al.*, 2024]	Swin-T	Yes	78.94 ± 0.66	86.49 ± 0.50	82.37 ± 0.77	89.89 ± 0.52
SYNTRANS [Tang *et al.*, 2025]	ResNet-12	Yes	76.20 ± 0.69	86.12 ± 0.54	79.69 ± 0.81	87.78 ± 0.60
PMCE(Ours)	ResNet-12	Yes	82.90 ± 0.78	92.29 ± 0.46	**83.50 ± 0.32**	**95.03 ± 0.34**
	Swin-T	Yes	**85.03 ± 0.67**	**92.77 ± 0.37**	83.13 ± 0.33	93.23 ± 0.43

Table 1: Comparison with state-of-the-art few-shot learning methods under the 5-way 1-shot and 5-way 5-shot settings on MiniImageNet and TieredImageNet. All results are reported as mean accuracy (%) with 95% confidence intervals. The best performance in each column is highlighted in bold, and "–" indicates that the result is not reported.

5.1 Datasets

To evaluate the performance of proposed method, we conduct experiments on MiniImageNet [Vinyals *et al.*, 2016] , TieredImageNet [Ren *et al.*, 2018], CIFAR-FS [Bertinetto *et al.*, 2019], and FC100 [Oreshkin *et al.*, 2018]. They cover both ImageNet-style datasets with richer visual details and CIFAR-style datasets with lower resolution.

MiniImageNet contains 100 classes with 600 images per class. We follow the standard split of 64, 16, and 20 classes for training, validation and test, respectively.

TieredImageNet is a larger subset of ImageNet with 608 classes grouped into 34 higher-level categories. We use the standard split with 351, 97, and 160 classes for training, validation, and test, respectively.

CIFAR-FS is created by randomly splitting 100 classes of CIFAR-100 into 64 training, 16 validation and 20 test classes with image size of 32 ×32.

FC100 is another CIFAR-100 derived benchmark with 100 classes grouped into 20 super-classes. We follow the standard split of 60, 20, and 20 classes for training, validation, and test, respectively.

5.2 Implementation Details

Following [Zhao *et al.*, 2024], we use pre-trained ResNet-12 and Swin-T as visual backbones. We generate captions with a frozen BLIP [Li *et al.*, 2022] captioner and encode both class names and captions using the frozen ViT-B/16 CLIP text encoder [Radford *et al.*, 2021], obtaining 512-d sentence embeddings. We intentionally adopt BLIP to keep the captioner modest in size, while stronger captioners (e.g., BLIP-2) can be used as drop-in replacements. The enhancer Φ is a projection block (Linear+LayerNorm+ReLU) followed by multi-head cross-attention, with a residual scale β initialized to 0.1. We use $h = 4$ heads and dropout 0; the key dimension is $d_k = 160$ for ResNet-12 ($d_v = 640$) and $d_k = 192$ for Swin-T ($d_v = 768$). All backbones and BLIP/CLIP modules remain frozen, and inference is strictly inductive.

We optimize only Φ and a shallow auxiliary classifier for 50 epochs using Adam (lr 10^{-4}, batch size 128) with StepLR (step 30, gamma 0.1). At test time, we evaluate 5-way 1-shot and 5-way 5-shot tasks over 600 episodes and report mean accuracy with 95% confidence intervals; logistic regression is used by default, and nearest-prototype yields similar trends. Unless otherwise specified, we set $\alpha = 0.33$ (1-shot) and $\alpha = 0.7$ (5-shot) in Eq. (3), use top-k retrieval with $k = 7$ and temperature $\tau = 1.0$, and set $\tau_c = 0.1$, $\lambda_{\text{rec}} = 1.0$, and $\lambda_{\text{con}} = 1.0$ for training. We use class-name embeddings for retrieval since they are stable and class-consistent, while captions can be noisier on low-resolution images. We also experimented with token-level caption features, but they lead to significantly larger intermediate files and storage overhead, so we report sentence-embedding results by default.

Methods	Backbone	Semantic	CIFAR-FS		FC100	
			1-shot	5-shot	1-shot	5-shot
Metric-based						
ProtoNet [Snell *et al.*, 2017]	ResNet-12	No	55.50 ± 0.70	72.00 ± 0.60	41.54 ± 0.76	57.08 ± 0.76
MetaOptNet [Lee *et al.*, 2019]	ResNet-12	No	72.80 ± 0.70	84.30 ± 0.50	47.20 ± 0.60	55.50 ± 0.60
CORL [He *et al.*, 2023]	ResNet-12	No	74.13 ± 0.71	87.54 ± 0.51	44.82 ± 0.73	61.31 ± 0.54
QSFormer [Wang *et al.*, 2023]	ResNet-12	No	46.51 ± 0.26	61.58 ± 0.25	–	–
SSFormers [Chen *et al.*, 2023a]	ResNet-12	No	74.50 ± 0.21	86.61 ± 0.23	43.72 ± 0.21	58.92 ± 0.18
Cads [Zhang *et al.*, 2025]	ResNet-12	No	73.23 ± 0.21	87.67 ± 0.14	–	–
Distribution-based						
PBML [Fu *et al.*, 2025]	ResNet-12	No	73.07 ± 0.59	85.51 ± 0.41	47.92 ± 0.49	62.96 ± 0.51
DDC [Chen *et al.*, 2025a]	WRN-28-10	No	76.47 ± 1.86	87.24 ± 0.89	–	–
Semantic-based						
PRSN [Dong *et al.*, 2025]	ResNet-12	Yes	78.72 ± 0.96	**90.22 ± 0.60**	47.91 ± 0.57	62.95 ± 0.52
Vision–language based						
LPE [Yang *et al.*, 2023]	ResNet-12	Yes	74.88 ± 0.45	85.30 ± 0.35	–	–
SP-Pretrain [Chen *et al.*, 2023b]	ViT-S	Yes	71.99 ± 0.47	85.98 ± 0.34	43.77 ± 0.39	59.48 ± 0.39
PMCE(Ours)	ResNet-12	Yes	76.97 ± 0.81	84.76 ± 0.61	49.81 ± 0.82	63.67 ± 0.73
	Swin-T	Yes	**80.92 ± 0.72**	89.02 ± 0.58	**52.46 ± 0.82**	**67.00 ± 0.75**

Table 2: Comparison with state-of-the-art few-shot learning methods under the 5-way 1-shot and 5-way 5-shot settings on CIFAR-FS and FC100. All results are reported as mean accuracy (%) with 95% confidence intervals. The best performance in each column is highlighted in bold, and "–" indicates that the result is not reported.

MAP	Enh.(S)	Enh.(Q)	MiniImageNet	FC100
×	×	×	71.02 ± 0.79	45.27 ± 0.75
×	✓	×	72.18 ± 0.78	46.03 ± 0.74
×	×	✓	72.49 ± 0.81	45.47 ± 0.77
✓	×	×	74.93 ± 0.76	45.36 ± 0.75
✓	✓	×	74.99 ± 0.74	45.15 ± 0.77
✓	×	✓	77.17 ± 0.76	45.27 ± 0.78
×	✓	✓	79.42 ± 0.73	48.79 ± 0.80
✓	✓	✓	$\mathbf{85.03 \pm 0.67}$	$\mathbf{52.46 \pm 0.82}$

Table 3: Component ablation on 5-way 1-shot MiniImageNet and FC100 with Swin-T. MAP is semantics-guided estimation; Enh.(S)/Enh.(Q) are caption-guided enhancement on support/query. Results are mean accuracy (%) ± 95% CIs.

5.3 Comparison With State-of-the-Art Methods

To evaluate the performance of our method, we make a comprehensive comparison with the state-of-the-art methods. These include metric-based methods MatchNet [Vinyals *et al.*, 2016], ProtoNet [Snell *et al.*, 2017], RelationNet [Sung *et al.*, 2018], FGFL [Cheng *et al.*, 2023], CPEA [Hao *et al.*, 2023], MCNet [Chen *et al.*, 2024], Cads [Zhang *et al.*, 2025], and SRE-ProtoNet* [Chen *et al.*, 2025b]; distribution-based methods ImprovedMAP [Wu and Hu, 2022], TDO [Liu *et al.*, 2023], PBML [Fu *et al.*, 2025], and DDC [Chen *et al.*, 2025a]; semantic-based methods KTN [Peng *et al.*, 2019], AM3 [Xing *et al.*, 2019], BMI [Li and Wang, 2023], MAVSI [Zhao *et al.*, 2024], and PRSN [Dong *et al.*, 2025]; and Vision–language based methods SP-CLIP [Chen *et al.*, 2023b], SP-SBERT [Chen *et al.*, 2023b], SP-Glove [Chen *et al.*, 2023b], SemFew-Trans [Zhang *et al.*, 2024], and SYN-TRANS [Tang *et al.*, 2025]. Results on MiniImageNet and TieredImageNet are reported in Tab. 1, and results on CIFAR-FS and FC100 in Tab. 2. The results of these compared methods are quoted from their original work or relevant references.

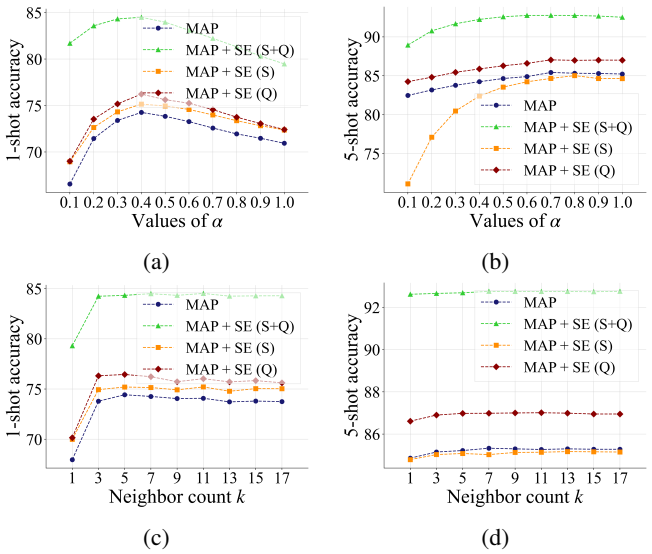

Figure 3: Sensitivity of PMCE to the prior weight α and the number of retrieved base classes k on MiniImageNet. Performance is stable over a wide range of values, and our default settings lie in the high-accuracy region.

Tab. 1 and Tab. 2 show the main comparison on four benchmarks with different lines of state-of-the-art methods. On MiniImageNet, PMCE achieves the best performance under both backbones. With Swin-T, PMCE reaches 85.03% and 92.77%, improving over SemFew-Trans by 7.71% and 7.26%, respectively. On TieredImageNet, PMCE also gives clear gains on the 1-shot setting and is consistently stronger in 5-shot. On CIFAR-style benchmarks, PMCE remains strong, it achieves the best 1-shot accuracy, which is 2.79 higher than PRSN, while it is slightly lower in 5-shot. On FC100, PMCE sets the best results with 52.46% and 67.00%, improving over

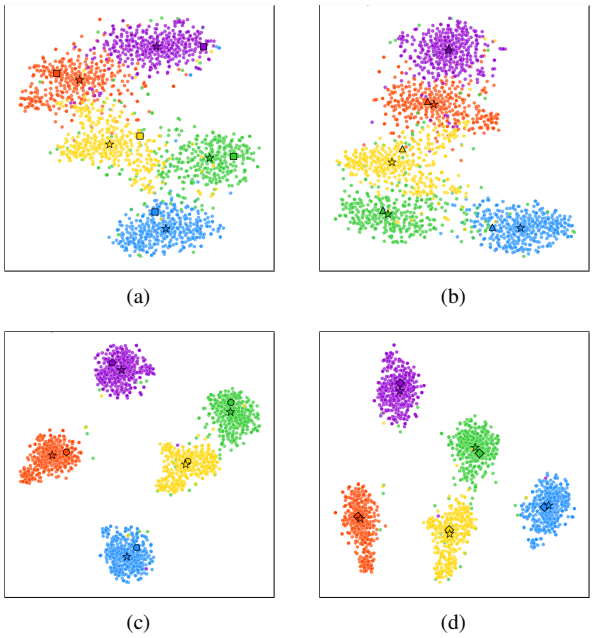

Figure 4: t-SNE on a 5-way 1-shot MiniImageNet episode. Colors denote classes; ⋆ are class centers, ■ are empirical prototypes, △ are MAP rectified prototypes, ● are caption-enhanced prototypes, and ⋄ are PMCE final prototypes. (a) Empirical, (b) +MAP, (c) +prototype enhancement, (d) +prototype and query enhancement.

PRSN by 9.5% and 6.43%, respectively.

We believe the improvements mainly come from two design choices. First, semantics-guided prior retrieval provides a class-specific prior that stabilizes prototype estimation when supports are extremely limited, which is especially important in 1-shot. Second, PMCE optimizes the query side with label-free captions under the inductive protocol, so support and query features are better aligned before matching.

5.4 Ablation Study

To test whether our performance is solely brought by MAP or Enhancer, we evaluate their importance separately. Tab. 3 shows the results of our method when trained with or without each component. Starting from the visual baseline, adding only MAP estimation already brings a 5.51% gain on MiniImageNet, showing that semantics-guided priors effectively correct prototype bias. Caption enhancement on either the support side (Enh.(S)) or the query side (Enh.(Q)) yields further improvements, and enabling both sides jointly brings clear gains on both benchmarks. With all components activated, PMCE improves the baseline by up to 19.73% on MiniImageNet and 15.88% on FC100. Therefore, both MAP estimation and Enhancer for both supports and queries play important roles in our method.

5.5 Impact of Hyperparameters

We study the sensitivity to α in Eq. (3) and the neighbor count k. Fig. 3a and Fig. 3b show that the optimal α lies around 0.3–0.4 in the 1-shot case and around 0.7 in the 5-shot case, which matches the intuition that the prior should be stronger when supports are scarce. Fig. 3c and Fig. 3d indicate that the accuracy is stable for a wide range of k, suggesting that PMCE is robust to the exact number of retrieved base classes.

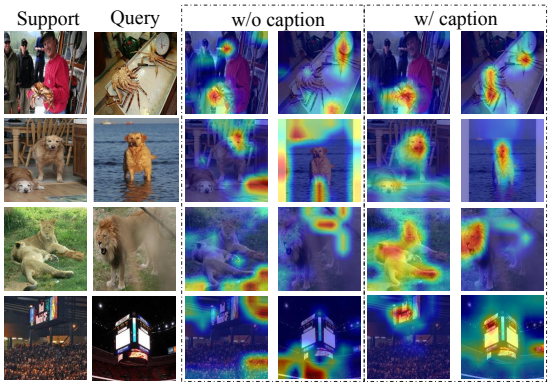

Figure 5: Grad-CAM visualizations on the MiniImageNet test set. Compared with **w/o caption**, **w/ caption** yields more focused activations on query images, suppressing background responses and highlighting discriminative object regions.

5.6 Visualization Analysis

In this section, we present visualization results of both t-SNE and Grad-CAM on MiniImageNet.

Prototype evolution. Fig. 4 illustrates how prototypes and features evolve on a 5-way 1-shot episode from MiniImageNet. Initial prototypes lie near cluster boundaries and are far from the ground-truth centers, indicating strong bias under very few supports. After MAP estimation they move closer to the centers, and caption-guided enhancement further tightens and separates the clusters; the final prototypes almost coincide with the ground-truth centers. Overall, the visualization suggests that PMCE turns scarce supports into stable, semantically anchored prototypes while improving the separability of the surrounding query features.

Effect of caption guidance. We further visualize Grad-CAM activation maps for several MiniImageNet episodes. The comparison between w/o caption and w/ caption (Fig. 5) shows that caption-guided enhancement suppresses background responses and concentrates attention on the target object for both supports and queries. This happens because instance captions act as semantic cues, highlighting key visual parts that global features often miss.

6 Conclusion

In this paper, we propose PMCE, a probabilistic few-shot framework that couples semantics-guided prior retrieval with caption-guided dual-stream enhancement. By calibrating prototypes with class-specific priors and enhancing both supports and queries using label-free captions under the inductive protocol, PMCE effectively incorporates multi-granularity semantic knowledge into the adaptation process and consistently improves few-shot classification. Experiments on four benchmarks with ResNet-12 and Swin-T validate the effectiveness of PMCE, with the largest gains observed in 1-shot settings.

Acknowledgements

This work was supported by the Hong Kong RGC General Research Fund under Grant Nos. 15221123, 15216424, and 15211525, and the Hong Kong PolyU Internal Research Fund under Grant Nos. P0058468 and P0056171.

References

[Bertinetto *et al.*, 2019] Luca Bertinetto, João F. Henriques, Philip H. S. Torr, and Andrea Vedaldi. Meta-Learning with Differentiable Closed-Form Solvers. In *Proceedings of the International Conference on Learning Representations (ICLR 2019) (New Orleans)*, 2019.

[Chen *et al.*, 2019] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A Closer Look at Few-shot Classification. In *Proceedings of the International Conference on Learning Representations (ICLR 2019) (New Orleans)*, 2019.

[Chen *et al.*, 2023a] Haoxing Chen, Huaxiong Li, Yaohui Li, and Chunlin Chen. Sparse Spatial Transformers for Few-Shot Learning. *Science China Information Sciences*, 66(11), 2023.

[Chen *et al.*, 2023b] Wentao Chen, Chenyang Si, Zhang Zhang, Liang Wang, Zilei Wang, and Tieniu Tan. Semantic Prompt for Few-Shot Image Recognition. *arXiv preprint arXiv:2303.14123*, 2023.

[Chen *et al.*, 2024] Derong Chen, Feiyu Chen, Deqiang Ouyang, and Jie Shao. Mutual Correlation Network for Few-Shot Learning. *Neural Networks*, 175:106289, 2024.

[Chen *et al.*, 2025a] Lingxing Chen, Yang Gu, Yi Guo, Fan Dong, Dongmei Jiang, and Yiqiang Chen. DDC: Dynamic Distribution Calibration for Few-Shot Learning Under Multi-Scale Representation. *Knowledge-Based Systems*, 311:113030, 2025.

[Chen *et al.*, 2025b] Xingye Chen, Wenxiao Wu, Li Ma, Xinge You, Changxin Gao, Nong Sang, and Yuanjie Shao. Exploring Sample Relationship for Few-Shot Classification. *Pattern Recognition*, 159:111089, 2025.

[Cheng *et al.*, 2023] Hao Cheng, Siyuan Yang, Joey Tianyi Zhou, Lanqing Guo, and Bihan Wen. Frequency Guidance Matters in Few-Shot Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2023) (Paris)*, pages 11780–11790, 2023.

[Dong *et al.*, 2025] Mengping Dong, Fei Li, Zhenbo Li, and Xue Liu. PRSN: Prototype Resynthesis Network with Cross-Image Semantic Alignment for Few-Shot Image Classification. *Pattern Recognition*, 159:111122, 2025.

[Fei-Fei *et al.*, 2003] Li Fei-Fei, Robert Fergus, and Pietro Perona. A Bayesian Approach to Unsupervised One-Shot Learning of Object Categories. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV 2003) (Nice)*, pages 1134–1141, 2003.

[Finn *et al.*, 2017] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proceedings of the International Conference on Machine Learning (ICML 2017) (Sydney)*, pages 1126–1135, 2017.

[Fu *et al.*, 2025] Meijun Fu, Xiaomin Wang, Jun Wang, and Zhang Yi. Prototype Bayesian Meta-Learning for Few-Shot Image Classification. *IEEE Transactions on Neural Networks and Learning Systems*, 36(4):7010–7024, 2025.

[Guo and Guo, 2021] Jingcai Guo and Song Guo. A novel perspective to zero-shot learning: Towards an alignment of manifold structures via semantic feature expansion. *IEEE Transactions on Multimedia*, 23:524–537, 2021.

[Guo *et al.*, 2023] Jingcai Guo, Song Guo, Qihua Zhou, Ziming Liu, Xiaocheng Lu, and Fushuo Huo. Graph knows unknowns: Reformulate zero-shot learning as sample-level graph recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2023) (Washington)*, pages 7775–7783, 2023.

[Guo *et al.*, 2024] Jingcai Guo, Qihua Zhou, Xiaocheng Lu, Ruibin Li, Ziming Liu, Jie Zhang, Bo Han, Junyang Chen, Xin Xie, and Song Guo. Parsnets: A parsimonious composition of orthogonal and low-rank linear networks for zero-shot learning. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI 2024) (Jeju)*, pages 4062–4070, 2024.

[Hao *et al.*, 2023] Fusheng Hao, Fengxiang He, Liu Liu, Fuxiang Wu, Dacheng Tao, and Jun Cheng. Class-Aware Patch Embedding Adaptation for Few-Shot Image Classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2023) (Paris)*, pages 18859–18869, 2023.

[He *et al.*, 2023] Ju He, Adam Kortylewski, and Alan L. Yuille. CORL: Compositional Representation Learning for Few-Shot Classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2023) (Waikoloa)*, pages 3879–3888, 2023.

[Khosla *et al.*, 2020] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised Contrastive Learning. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS 2020) (Virtual)*, volume 33, pages 18661–18673, 2020.

[Lee *et al.*, 2019] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-Learning with Differentiable Convex Optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019) (Long Beach)*, pages 10657–10665, 2019.

[Li and Wang, 2023] Zhuoling Li and Yong Wang. Better Integrating Vision and Semantics for Improving Few-shot Classification. In *Proceedings of the 31st ACM International Conference on Multimedia (MM 2023) (Ottawa)*, pages 4737–4746, 2023.

[Li *et al.*, 2022] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and

Generation. In *Proceedings of the International Conference on Machine Learning (ICML 2022) (Baltimore)*, pages 12888–12900, 2022.

[Liu *et al.*, 2023] Xinyue Liu, Ligang Liu, Han Liu, and Xiaotong Zhang. Capturing the Few-Shot Class Distribution: Transductive Distribution Optimization. *Pattern Recognition*, 138:109371, 2023.

[Lu *et al.*, 2023] Xiaocheng Lu, Song Guo, Ziming Liu, and Jingcai Guo. Decomposed soft prompt guided fusion enhancing for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2023) (Vancouver)*, pages 23560–23569, 2023.

[Oreshkin *et al.*, 2018] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. TADAM: Task Dependent Adaptive Metric for Improved Few-Shot Learning. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS 2018) (Montréal)*, pages 719–729, 2018.

[Peng *et al.*, 2019] Zhimao Peng, Zechao Li, Junge Zhang, Yan Li, Guo-Jun Qi, and Jinhui Tang. Few-Shot Image Recognition With Knowledge Transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2019) (Seoul)*, pages 441–449, 2019.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, et al. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the International Conference on Machine Learning (ICML 2021) (Virtual Event)*, pages 8748–8763, 2021.

[Ren *et al.*, 2018] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B. Tenenbaum, Hugo Larochelle, and Richard S. Zemel. Meta-Learning for Semi-Supervised Few-Shot Classification. In *Proceedings of the International Conference on Learning Representations (ICLR 2018) (Vancouver)*, 2018.

[Snell *et al.*, 2017] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical Networks for Few-Shot Learning. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS 2017) (Long Beach)*, pages 4077–4087, 2017.

[Sung *et al.*, 2018] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to Compare: Relation Network for Few-Shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018) (Salt Lake City)*, pages 1199–1208, 2018.

[Tang *et al.*, 2025] Hao Tang, Shengfeng He, and Jing Qin. Connecting Giants: Synergistic Knowledge Transfer of Large Multimodal Models for Few-Shot Learning. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI 2025) (Montréal)*, pages 6227–6235, 2025.

[Vinyals *et al.*, 2016] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching Networks for One Shot Learning. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS 2016) (Barcelona)*, pages 3630–3638, 2016.

[Wang *et al.*, 2023] Xixi Wang, Xiao Wang, Bo Jiang, and Bin Luo. Few-Shot Learning Meets Transformer: Unified Query-Support Transformers for Few-Shot Classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(12):7789–7802, 2023.

[Wu and Hu, 2022] Jiaying Wu and Jinglu Hu. Improved Prior Selection Using Semantics in Maximum A Posteriori for Few-Shot Learning. *Knowledge-Based Systems*, 237:107688, 2022.

[Wu *et al.*, 2021] Jiaying Wu, Ning Dong, Fan Liu, Sai Yang, and Jinglu Hu. Feature Hallucination via Maximum A Posteriori for Few-Shot Learning. *Knowledge-Based Systems*, 225:107129, 2021.

[Xing *et al.*, 2019] Chen Xing, Negar Rostamzadeh, Boris N. Oreshkin, and Pedro O. Pinheiro. Adaptive Cross-Modal Few-shot Learning. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS 2019) (Vancouver)*, pages 4848–4858, 2019.

[Xu *et al.*, 2026] Mingwei Xu, Xiaofeng Cao, Ivor W Tsang, James T Kwok, and Heng Tao Shen. Generalized distribution aggregation protocol for federated statistical heterogeneity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

[Yang *et al.*, 2022] Shuo Yang, Songhua Wu, Tongliang Liu, and Min Xu. Bridging the Gap Between Few-Shot and Many-Shot Learning via Distribution Calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9830–9843, 2022.

[Yang *et al.*, 2023] Fengyuan Yang, Ruiping Wang, and Xilin Chen. Semantic Guided Latent Parts Embedding for Few-Shot Learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2023) (Waikoloa)*, pages 5436–5446, 2023.

[Zhang *et al.*, 2024] Hai Zhang, Junzhe Xu, Shanlin Jiang, and Zhenan He. Simple Semantic-Aided Few-Shot Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024) (Seattle)*, pages 28588–28597, 2024.

[Zhang *et al.*, 2025] Yourun Zhang, Maoguo Gong, Jianzhao Li, Kaiyuan Feng, and Mingyang Zhang. Few-Shot Learning With Enhancements to Data Augmentation and Feature Extraction. *IEEE Transactions on Neural Networks and Learning Systems*, 36(4):6655–6668, 2025.

[Zhao *et al.*, 2024] Peng Zhao, Yin Wang, Wei Wang, Jie Mu, Huiting Liu, Cong Wang, and Xiaochun Cao. Multi-Attention Based Visual-Semantic Interaction for Few-Shot Learning. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI 2024) (Jeju)*, pages 1753–1761, 2024.