Like Humans to Few-Shot Learning through Knowledge Permeation of Visual and Language

Yuyu Jia, Qing Zhou, Junyu Gao, *Member, IEEE,*
Qiang Li, *Member, IEEE,* and Qi Wang, *Senior Member, IEEE*

***Abstract*—Few-shot learning aims to generalize the recognizer from seen categories to an entirely novel scenario. With only a few support samples, several advanced methods initially introduce class names as prior knowledge for identifying novel classes. However, obstacles still impede achieving a comprehensive understanding of how to harness the mutual advantages of visual and textual knowledge. In this paper, we set out to fill this gap via a coherent Bidirectional Knowledge Permeation strategy called BiKop, which is grounded in human intuition: a class name description offers a more *general* representation, whereas an image captures the *specificity* of individuals. BiKop primarily establishes a hierarchical joint general-specific representation through bidirectional knowledge permeation. On the other hand, considering the bias of joint representation towards the base set, we disentangle base-class-relevant semantics during training, thereby alleviating the suppression of potential novel-class-relevant information. Experiments on four challenging benchmarks demonstrate the remarkable superiority of BiKop, particularly outperforming previous methods by a substantial margin in the 1-shot setting (improving the accuracy by 7.58% on *mini*ImageNet). The code is available at https://github.com/mrazhou/BiKop.**

***Index Terms*—Few-shot Learning, Knowledge Disparity, Class-relevant Information.**

I. Introduction

As a forefront area of research in computer vision, Few-Shot Learning (FSL) investigates how models can rapidly generalize existing knowledge to novel scenarios, mirroring human-like adaptability. It first acquires a base classifier through pre-training with ample samples, subsequently learning to identify unseen/novel classes with a few support samples [1], [2], [3], [4], [5]. In this setting, two significant challenges naturally arise: (1) the collapse of sparse feature representation induced by significant intra-class variations, (2) the bias toward the base set during training might compromise the model's fragile generalization. To tackle these issues, previous FSL methods primarily concentrate on the adaptive optimization of meta-learners [6], [7], [8], the effective mining of prior knowledge, [9], [10], [11], or the formulation of judicious metric functions [12], [13], [14].

Driven by advancements in Natural Language Processing (NLP) models [15], [16], [17], [18], class name embeddings as a direct form of prior knowledge are injected to enhance the few-shot recognizer [19], [20] as illustrated in Fig. 1(a). Not confined to this, recent studies leverage textual information to modulate the extraction of visual features, unlocking further potential (Fig. 1(b). For example, SP [21] utilizes textual information as prompts to fine-tune visual features. CMA [22] repurposes class names as additional one-shot training samples, implicitly intervening in visual features.

While these methods have exhibited considerable promise, there still exist crucial limitations. **(L1)** The effective utilization of mutual knowledge between vision and text has not been achieved. Typically, a class name offers a more general representation, while an image encapsulates the specificity of individuals. For instance, given a sentence like "*a photo of the dog*", we evoke the **general** concept of "dog" and its shared characteristics, while an image concretely characterizes the individual **specificity** of the category of "dog". In this spirit, we posit that explicitly constructing the **general-specific** hierarchical knowledge structure can effectively alleviate the collapse of sparse feature representation. **(L2)** Although textual knowledge brings gains, the inclusion of class names as base-class-relevant information during training could exacerbate the model bias towards the base set, which is an inherent challenge in FSL [23], [24], [25]. In other words, such gains may be significantly compromised when dealing with novel classes. This previously overlooked contradiction undermines the enhancements facilitated by textual knowledge.

In this paper, we propose BiKop, a novel approach comprising two modules, Bidirectional Knowledge Permeation (BKP) and Semantic Adversarial Disentanglement (SAD), to jointly mitigate the representation collapse (**L1**) and model bias (**L2**). BKP allows the bidirectional permeation of mutual knowledge derived from textual and visual modalities. It involves three steps. Firstly, inspired by the use of class names as prompts in [21], we creatively extend it to meta-class-specific prompts to equip textual knowledge with stronger meta-task adaptability. Secondly, a lightweight cross-attention block is introduced to facilitate the permeation of general knowledge, from textual prompts to visual features. In the opposite direction, individual specificity knowledge from vision features is permeated into textual prompts to enhance their diversity. Thirdly, the enhanced prompts and visual features are fed into a Transformer encoder, producing a robust joint feature representation for sparse support samples. Furthermore, introducing base-class-relevant knowledge (*i.e.,* class names) exacerbates the bias towards the base set. In other words, the model inevitably suppresses the potential semantics of novel categories during

This work was supported by the National Natural Science Foundation of China under Grant 62301385, 62471394, and U21B2041.

Yuyu Jia, Qing Zhou, Junyu Gao, Qiang Li, and Qi Wang are with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China.

E-mail: jyy2019@mail.nwpu.edu.cn, chautsing@gmail.com, gjy3035@gmail.com, liqmges@gmail.com, crabwq@gmail.com

Yuyu Jia and Qing Zhou contributed equally to this work. Qi Wang is the corresponding author.

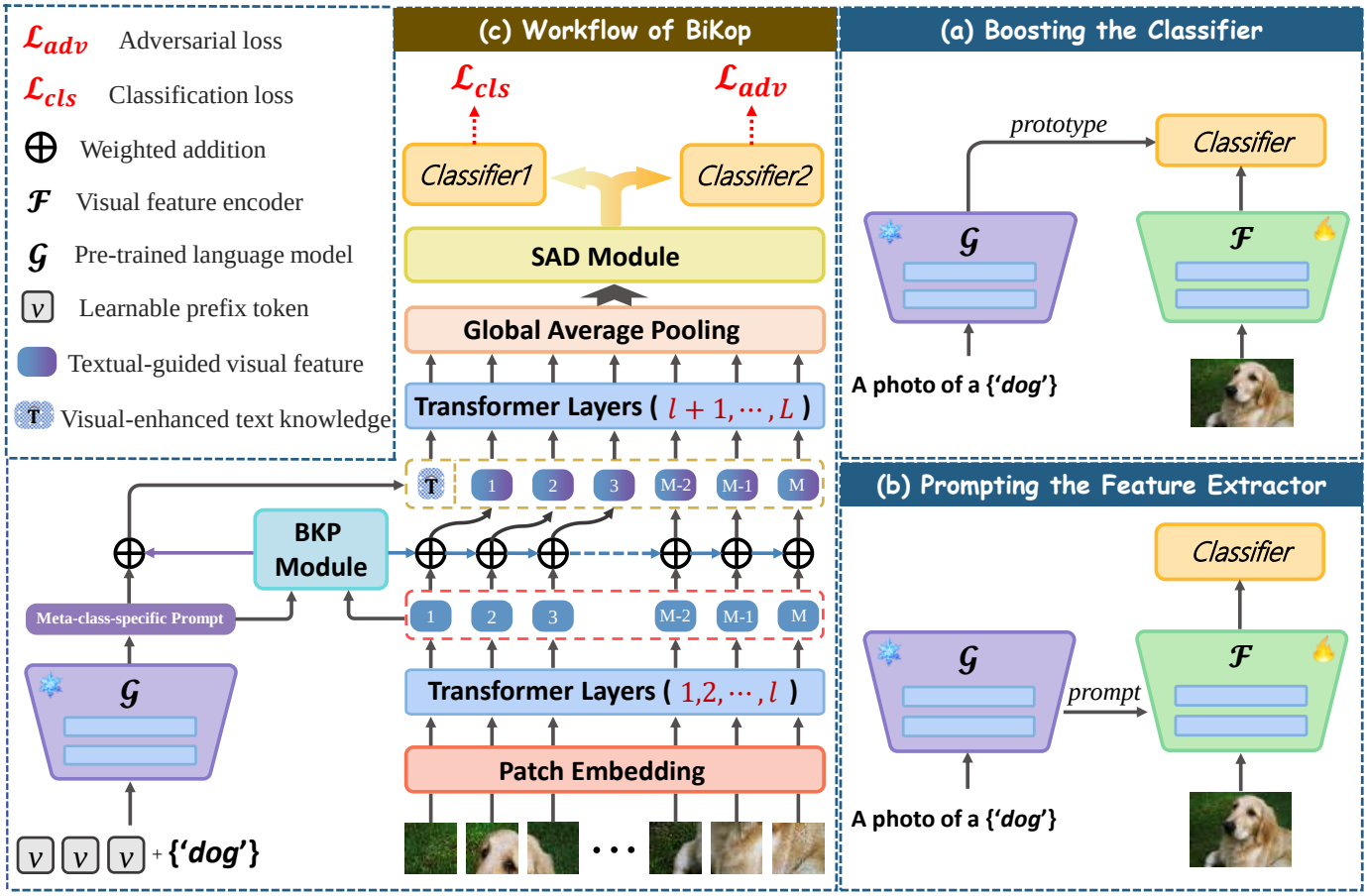

Fig. 1. Comparison of BiKop with studies of introducing textual knowledge. **(a)** They embed class names into text prototypes, directly employed to enhance classifiers. **(b)** Several recent methods utilize textual information to modulate the extraction of visual features. **(c)** Workflow of our BiKop. To alleviate the collapse of sparse feature representation, the BKP module harnesses the mutual advantages between textual and visual knowledge by the bidirectional permeation of both. Furthermore, the SAD module adversarially disentangles the base-class-relevant semantic to mitigate the base set bias and boost the model's generalization to novel categories.

training to achieve better convergence. In response, SAD adversarially disentangles base-class-relevant semantics from the joint feature representation, which balances the model's convergence on the base set and generalization to novel classes.

Overall, we posit that the BKP module is intricately linked with the SAD module, where SAD alleviates the derivative side effects caused by BKP. Comprehensive experiments conducted on four benchmarks consistently demonstrate that BiKop achieves performance improvements across different backbones. Particularly, with the popular backbones of Visformer-T [26] and ViT-S [27], the 1-shot classification accuracy on *mini*ImageNet has been improved by 5.76% and 7.58%, respectively. The contribution of this paper is threefold:

1) We take a close look at the complementary disparities within textual and visual knowledge, ingeniously combining the two to construct a novel few-shot image recognition method.
2) The proposed BKP module extends class names as meta-class-specific prompts and conducts bidirectional permeation of textual and visual knowledge, fully harnessing the mutual advantages of both to alleviate the collapse of sparse feature representation. Besides, the SAD module mitigates the base set bias exacerbated by base-class-relevant knowledge (*i.e.,* class names), further enhancing the model's generalization to novel categories.
3) Our proposed method is easy to implement and seamlessly compatible with different backbone architectures. Through extensive experiments, we validate that BiKop achieves state-of-the-art performance across different FSL datasets.

II. Related Works

A. Few-shot Classification

Over the past few years, a persistent fervor surrounds Few-Shot Learning (FSL) as scholars ardently seek to enhance the generalization of artificial intelligence models. Among the significant applications of FSL, few-shot classification can be broadly categorized into two groups: metric-based methods [12], [28], [29], [30], [31] and optimisation-based methods [6], [7], [32], [33], [34], [35]. Metric-based approaches concentrate on acquiring a feature space, in which a suitable distance function is employed for measuring similarity. Optimization-based methods conduct rapid adaption with a few training samples for novel categories through learning a good meta-optimizer. Additionally, recent self-supervised-based methods [36], [9],

[11], [37], [38], [39] have demonstrated their potential for exploring in few-shot settings. They fundamentally leverage pretext tasks on the base set to extract more prior knowledge, thereby benefiting the model's generalization.

B. Few-shot Learning with Textual Knowledge

Few-shot learning with textual knowledge is divided into two approaches: using VLMs, which involves fine-tuning foundation models for downstream tasks, and traditional methods, which utilize meta-learning strategies and feature extraction to enable small models to generalize to unseen classes. This study focuses on the traditional paradigm.

Instead of mining prior or generalizing knowledge from training on the base set, a series of advanced investigations [40], [41], [42], [19] directly incorporate prior information from textual modality, to facilitate the recognition of novel classes. For example, an adaptive fusion mechanism is introduced in [19] to merge a visual prototype with a semantic prototype derived from the class name embedding. An Image-guided Text Weighting module is introduced in [43] to adjust the influence of textual prompts based on their feature similarity with training images. In [42], the additional textual knowledge and visual information are effectively integrated to infer desired few-shot classifiers. SP [21] treats textual knowledge as prompts to modulate the extraction of visual features, extending its focus beyond just optimizing classifiers. Unlike studies using textual knowledge with classifiers or rigid prompts, we implement bidirectional knowledge permeation to leverage the mutual benefits of textual and visual modalities for sparse feature representation.

C. Prompt Learning

With the support of massive data, large-scale visual-language models [44], [45], [46], [47] have witnessed rapid development in recent years. However, in the face of diverse downstream tasks, efficiently adapting large models with limited data has become a new research focus. Prompt learning, introduced from natural language processing into visual tasks, adapts to various downstream tasks by optimizing prompts with few parameters instead of tuning deep models. For example, the manually crafted template "a photo of a [CLASS]" in CLIP [18] is employed to represent the textual embedding for zero-shot prediction. Considering the difficulty of learning task-specific knowledge with manually crafted template prompts, Context Optimization (CoOp) [48] introduces learnable tokens to construct soft prompts. Conditional Context Optimization (CoCoOp) [49] is proposed to learn instance-specific prompts for further generalization to unseen categories. DPT [50] incorporates class-aware visual prompt tuning, which dynamically generates visual prompts through cross-attention, effectively aligning image features with target concepts for improved performance. Comparatively, BiKop makes the first attempt to manage meta-class-specific prompts to adapt them to the meta-learning scheme, achieving meta-task-specific optimization for each category. It is worth noting that the aforementioned prompt learning methods focus on directly enhancing CLIP's zero-shot or few-shot classification performance through learnable prompt optimization. In contrast, our approach seeks to harness the textual modality knowledge embedded in CLIP and seamlessly integrate it with visual modality information to construct more robust and generalizable joint class representations.

III. Methodology

A. Preliminaries

Few-shot classification generalizes the knowledge learned on base classes $\mathcal{C}_{base}$ with abundant labeled data to sparse samples from novel classes $\mathcal{C}_{novel}$, where $\mathcal{C}_{base} \cap \mathcal{C}_{base} = \emptyset$. We adopt the popular episodic manner [12], [28] to define the N-way K-shot classification task, where N represents the number of classes, and K represents the number of labeled images per class. In each episode, we are given support set $\mathcal{S} = \{(x_i^s, y_i^s)\}_{i=1}^{NK}$ and query set $\mathcal{Q} = \{(x_i^q, y_i^q)\}_{i=1}^{NQ}$, where Q denotes the number of images to be classified for each category. We aim to acquire a model with good generalization to novel classes after meta-training on the base set.

B. Design Guideline of Network Structure

1) Preparation: Learning a generalized feature extractor has been proven to be beneficial for few-shot classification. Thus, Our approach involves two stages: the pre-training and fine-tuning. For the sake of ensuring a fair comparison, we followed consistent pre-training procedures with [9] and [21] for different feature extractors (*i.e.,* Visformer-T [26] and ViT-S [27]). Specific details can be referred to the aforementioned literature. We focus on utilizing BiKop for fine-tuning the feature extractor, addressing the challenges of sparse feature representation collapse and model bias, thereby improving the model's performance in scenarios with scarce data. Fig. 1(c) illustrates the pipeline of BiKop. The input image $\boldsymbol{x} \in \mathbb{R}^{H\times W\times C}$ (H, W, C are the height, width, and dimension) is divided into $M = H \cdot W/P^2$ image patches $\boldsymbol{p} = \left\{p^i\right\}_{i=1}^{M}$, with each patch $p^i \in \mathbb{R}^{P^2\times C}$. Then, all flattened patches of support and query images are fed into L-layers Transformer architecture $\mathcal{F}(\cdot)$ to extract visual features.

2) The collapse of sparse feature representation: The quality of the supporting representation determines the performance of the few-shot classifier. However, significant intra-class variations mean that a minimal number of support samples is challenging to organize a generalizable and discriminative representation. To mitigate this issue, we establish a hierarchical joint general-specific representation through bidirectional knowledge permeation, maximizing the mutual advantages of visual and textual modalities. Given a support image $\boldsymbol{x}^s$, the **B**idirectional **K**nowledge **P**ermeation (BKP) module (Sec.III-C) takes its visual features at the l^{th} layer $\boldsymbol{Z}_l^s$ and its class name y^{text} as inputs, producing textual-guided visual features $\hat{\boldsymbol{Z}}_l^s$ and visual-enhanced textual knowledge $\hat{\boldsymbol{T}}$:

$$\hat{\boldsymbol{Z}}_l^s, \hat{\boldsymbol{T}} = \mathcal{BKP}(\boldsymbol{Z}_l^s, y^{text}), \tag{1}$$

where $\hat{\boldsymbol{Z}}_l^s, \boldsymbol{Z}_l^s \in \mathbb{R}^{M\times C_d}$, $\hat{\boldsymbol{T}} \in \mathbb{R}^{C_d}$, and C_d is the number of output feature channels. We utilize $\hat{\boldsymbol{T}}$ as prompts to

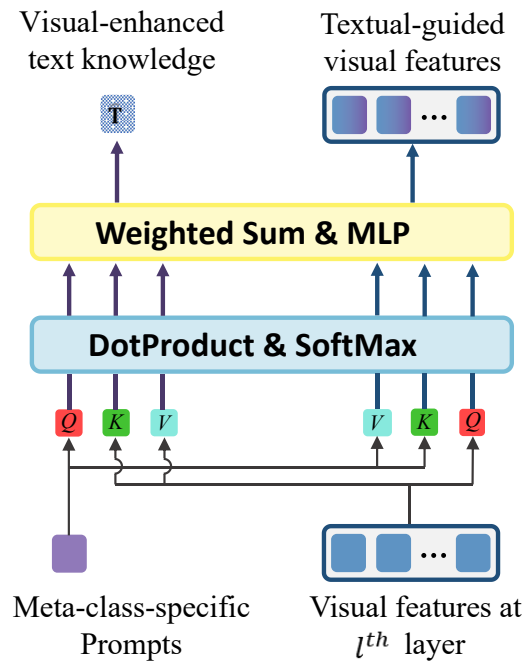

Fig. 2. Configuration diagram of BKP.

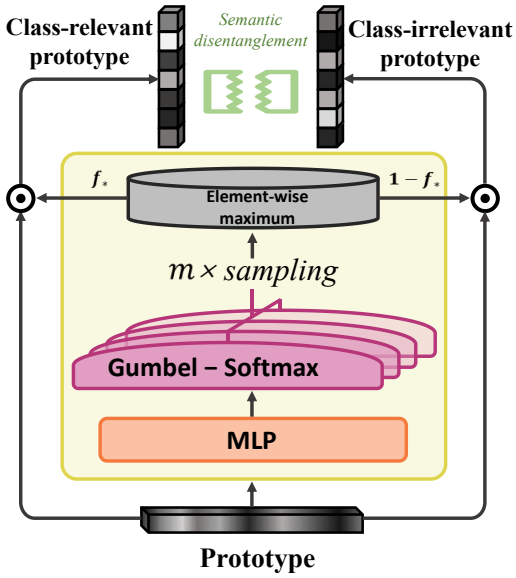

Fig. 3. Implementation of SAD.

modulate $\hat{\boldsymbol{Z}}_l^s$ through the remaining Transformer layers, in which the Multi-head Self-Attention (MSA) mechanisms allow the interaction between prompts and visual features. Then, we construct the robust joint feature representation for sparse support samples as:

$$\boldsymbol{F}^s = GAP(MSA([\hat{\boldsymbol{Z}}_l^s; \hat{\boldsymbol{T}}])), \tag{2}$$

where $\boldsymbol{F}^s \in \mathbb{R}^{C_d}$, $[\cdot;\cdot]$ denotes the concatenation operation, and $GAP(\cdot)$ stands for the Global Average Pooling function. Subsequently, the robust feature representation is averaged within each category to compute the class prototypes:

$$\boldsymbol{p}_c = \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{F}_k^{s|c}, \tag{3}$$

where $\boldsymbol{p}_c \in \mathbb{R}^{C_d}$ is the prototype, and $\boldsymbol{F}_k^{s|c}$ denotes the k^{th} support feature of the category c.

3) The model bias exacerbated by base-class-relevant knowledge: During training, the model is exposed solely to base class samples. The BKP module supplements visual features with textual descriptions of these classes, leading to a specialization in modeling the base class distribution. However, this also weakens the model's ability to generalize to unseen class distributions during the testing phase.

For a specific class c in the base set, despite its prototype $\boldsymbol{p}_c$ incorporating hierarchical knowledge of both general and specific aspects, introducing base-class-relevant knowledge from class names inevitably exacerbates the bias toward the base set. Specifically, we assume that $\boldsymbol{p}_c = Mix(\hat{\boldsymbol{p}}_c, \check{\boldsymbol{p}}_c)$ is constructed from a mix of class-relevant semantics $\hat{\boldsymbol{p}}_c$ and class-irrelevant semantics $\check{\boldsymbol{p}}_c$. During training, the model suppresses the learning of $\check{\boldsymbol{p}}_c$ to achieve better convergence to base classes. This undermines the model's representational capacity for the potential semantics of novel categories.

For a better trade-off between model convergence and model generalization, the **S**emantic **A**dversarial **D**isentanglement (SAD) module (Sec.III-D) adversarially disentangles base-class-relevant semantics $\hat{\boldsymbol{p}}_c$ and releases the suppression of base-class-irrelevant semantics $\check{\boldsymbol{p}}_c$:

$$\hat{\boldsymbol{p}}_c, \check{\boldsymbol{p}}_c = \mathcal{SAD}(\boldsymbol{p}_c). \tag{4}$$

During training, the pre-trained language model $\mathcal{G}(\cdot)$ is fixed and other parameters are optimized by *maximizing* the similarities between query features and $\hat{\boldsymbol{p}}_c$ with a cross-entropy loss:

$$\mathcal{L}_{cls} = -\mathbb{E}_{S,Q}\mathbb{E}_{\boldsymbol{x}^q} log \frac{\exp(\delta(\boldsymbol{F}(\boldsymbol{x}^q), \hat{\boldsymbol{p}}_{y^q})/\tau)}{\sum_{i=1}^{N}\exp(\delta(\boldsymbol{F}(\boldsymbol{x}^q), \hat{\boldsymbol{p}}_i)/\tau)}, \tag{5}$$

where $\hat{\boldsymbol{p}}_{y^q}$ is the base-class-relevant semantics of category y^q, τ is a temperature hyper-parameter, and $\delta(\cdot)$ represents the cosine similarity. Furthermore, we *minimize* the similarities between query features and their base-class-irrelevant semantics $\check{\boldsymbol{p}}_{y^q}$, thereby liberating the model for representing potential novel class semantics.

$$\mathcal{L}_{adv} = \mathbb{E}_{S,Q}\mathbb{E}_{\boldsymbol{x}^q} log \frac{\exp(\delta(\boldsymbol{F}(\boldsymbol{x}^q), \check{\boldsymbol{p}}_{y^q})/\tau)}{\sum_{i=1}^{N}\exp(\delta(\boldsymbol{F}(\boldsymbol{x}^q), \check{\boldsymbol{p}}_i)/\tau)}. \tag{6}$$

To be clear, the overall loss of BiKop is obtained by weighted summation:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \gamma \cdot \mathcal{L}_{adv}. \tag{7}$$

C. BKP Module

As shown in Fig. 2, given the meta-class-specific prompts and the l^{th}-layer visual features, the cross-attention block in the BKP module $\mathcal{BKP}(\cdot)$ goes beyond instance-level feature alignment. Instead, it fosters bidirectional permeation of category-level textual descriptive knowledge and instance-level visual representations, enabling effective adaptation to the few-shot learning task.

1) Meta-class-specific prompts: Unlike the prior FSL study [21] that directly employs class name embeddings as prompts, we extend it to meta-class-specific prompts for better adaptation to the meta-learning scheme. Experiments in Sec. IV-D5 prove the efficacy of this design. In an episode containing N categories, we use w learnable prefix tokens to create a meta-class-specific template for each category. Accordingly, the $\{class\ name\}$, *i.e.,* y^{text} is filled into the corresponding template:

$$\boldsymbol{E}_n = [v]_1^n[v]_2^n \dots [v]_w^n[class\ name]^n, \qquad n = 1, \dots N, \quad (8)$$

where $[v]_i^n, i \in \{1, 2, \dots, w\}$ denotes the learnable prefix tokens of n^{th} category, $[class\ name]^n$ represents the n^{th} class name. Each meta-class-specific prompt $\boldsymbol{E}_n$ is then fed into a pre-trained CLIP text encoder $\mathcal{G}(\cdot)$ to extract the corresponding prompt embedding $\boldsymbol{T}_n = \mathcal{H}(\mathcal{G}(\boldsymbol{E}_n))$ ($\mathcal{H}$ projects the prompt to the same dimension C_d of visual features).

2) Permeating textual knowledge into visual knowledge: It aims to supplement the lack of general knowledge in sparse visual features. Given the support visual features from the l^{th} Transformer layer $\boldsymbol{Z}_l^s \in \mathbb{R}^{M \times C_d}$ and prompt embedding $\boldsymbol{T} \in \mathbb{R}^{1 \times C_d}$ for its corresponding category, we get a visual-to-text similarity map $\boldsymbol{A}_{\boldsymbol{Z}} \in \mathbb{R}^{M \times 1}$:

$$\boldsymbol{A}_{\boldsymbol{Z}} = SoftMax((\boldsymbol{Z}_l^s W_q) \otimes (\boldsymbol{T} W_k)^\top / \sqrt{C_d}). \quad (9)$$

The textual-guided visual feature is then formulated as:

$$\hat{\boldsymbol{Z}}_l^s = \boldsymbol{Z}_l^s + \mu \cdot \mathcal{X}(\boldsymbol{A}_{\boldsymbol{Z}} \otimes \boldsymbol{T} W_v), \quad (10)$$

where μ is the weight coefficient, $\otimes$ denotes matrix multiplication, and $\mathcal{X}$ is a two-layer MLP [51] block. $W_q \in \mathbb{R}^{C_d \times C_d}$, $W_k \in \mathbb{R}^{C_d \times C_d}$, and $W_v \in \mathbb{R}^{C_d \times C_d}$ are linear transformation parameters.

3) Permeating visual knowledge into textual knowledge: In this process, visual features enrich the individual specificity of textual knowledge, *i.e.,* prompt embedding. Similarly to Eqs. 9 and 10, using a shared structure and parameters, the visual-enhanced textual knowledge can be written as:

$$\boldsymbol{A}_{\boldsymbol{T}} = SoftMax((\boldsymbol{T} W_q) \otimes (\boldsymbol{Z}_l^s W_k)^\top / \sqrt{C_d}) \quad (11)$$

$$\hat{\boldsymbol{T}} = \boldsymbol{T} + \mu \cdot \mathcal{X}(\boldsymbol{A}_{\boldsymbol{T}} \otimes \boldsymbol{Z}_l^s W_v). \quad (12)$$

D. SAD Module

The proposed SAD module primarily serves as a patch for the BKP module, mitigating the side effects introduced by BKP. Inspired by [25], specific prototype channel distributions correspond to the semantic expressions of particular categories, and this distribution can be considered class-relevant, while the remaining channel distributions are regarded as class-irrelevant. We propose the SAD module $\mathcal{SAD}(\cdot)$ as shown in Fig. 3, to adversely disentangle base-class-relevant and base-class-irrelevant semantics from the perspective of channel representations, *i.e.,* prototypes. Then, they are used to release the suppression of potential novel class semantics through an adversarial optimization strategy, as in Eq. 6.

We design a differentially samplable filter $\boldsymbol{f}_* \in \mathbb{R}^{C_d}$ for each category to learn the contribution of each dimension of the prototype. In this way, the class-relevant and class-irrelevant prototypes can be written as:

$$\hat{\boldsymbol{p}}_* = \boldsymbol{p}_* \odot \boldsymbol{f}_*, \check{\boldsymbol{p}}_* = \boldsymbol{p}_* \odot (1 - \boldsymbol{f}_*), \quad (13)$$

where $\odot$ denotes element-wise multiplication. To approximate the optimal differentially samplable filter $\boldsymbol{f}_*$, we apply the Gumbel-Softmax trick [52], [53], [54] to weighted sample a C_d-dimensional random vector independently for m times:

$$O_j = \text{Gumbel} - \text{Softmax}(\mathcal{D}(\boldsymbol{p}_*)), \quad j = 1, 2, \dots m, \quad (14)$$

where $\mathcal{D}(\cdot)$ is an MLP block that maps the input to a C_d-dimensional vector. Then, $\boldsymbol{f}_*$ is accquired from the element-wise maximum of $O_1, O_2, \dots, O_m$:

$$\boldsymbol{f}_* = (f^1, f^2, \dots, f^{C_d}), \quad f^i = \max_j O_j^i, \quad (15)$$

where $f^i, i \in \{1, 2, \dots, C_d\}$represents the i^{th} dimension of $\boldsymbol{f}_*$, O_j^i stands for the i^{th} dimension of the vector obtained in the j^{th} sampling.

IV. Experiments

A. Datasets

Our method is evaluated on four widely used few-shot classification benchmarks: *mini*ImageNet [12], *tiered*ImageNet [55], CIFAR-FS [56], and FC100 [57]. *mini*ImageNet and *tiered*ImageNet are the subcollections of ImageNet [58], while CIFAR-FS and FC100 are derived from CIFAR100 [59]. We adhere to the common practice data splits, consistent with [21], [9], where the data is mutually disjointly divided into the meta-training, meta-validation, and meta-test sets.

B. Implementation Details

Our experiments are conducted under the 5-way 1-shot and 5-shot settings. To enable a broad comparison with peer algorithms, we employ the Visformer-T [26] and ViT-S [27] as backbone feature extractors, following the same data augmentation strategies and pre-training the model on the four datasets as mentioned above, as done in prior works [9], [21]. To derive comprehensive textual knowledge from class names, we employ the text encoder of CLIP [18], which has undergone pre-training on extensive corpora. We default to resizing the input image to 224×224. The number of learnable tokens in meta-class-specific prompts is 8. The weight factor μ used in bidirectional knowledge permeation is set to 0.2. The weight parameter γ in the loss $\mathcal{L}_{total}$ is empirically set to 0.5. Alternative selections for these parameters are validated in Sec. IV-D2.

BiKop framework is instantiated in PyTorch [60] and executes for 100 epochs utilizing an NVIDIA RTX3090 GPU equipped with 24GB of memory. During meta-training, the network is optimized using the AdamW optimizer [61] with a weight decay of 1e−5 and an initial learning rate of 2e−6. Additionally, the learning rate for the BKP module is increased by a factor of 10, and for the SAD module, it is increased by a factor of 50. During inference, a random sampling of 2,000 test episodes with 15 query images per class from the novel set is conducted, and the average accuracy along with a 95% confidence interval is reported.

TABLE I
COMPARISON WITH PRIOR APPROACHES FOR THE 5-WAY 1-SHOT AND 5-WAY 5-SHOT SETTINGS ON *mini*IMAGENET [12] AND *tiered*IMAGENET [55]. METHODS IN THE TOP PART DO NOT INVOLVE TEXTUAL INFORMATION, THE MIDDLE PART INTRODUCES TEXTUAL KNOWLEDGE FROM CLASS NAMES, AND THE BOTTOM PART ILLUSTRATES OUR METHODOLOGY.

Method	Backbone	Params	*mini*ImageNet 5-way 1-shot	*mini*ImageNet 5-way 5-shot	*tiered*ImageNet 5-way 1-shot	*tiered*ImageNet 5-way 5-shot
CC+rot [62]	WRN-28-10	36.5M	62.93±0.45	79.87±0.33	70.53±0.51	84.98±0.36
Align [63]	WRN-28-10	36.5M	65.92±0.60	82.85±0.55	74.40±0.68	86.61±0.59
MBSS [13]	WRN-28-10	36.5M	66.91±0.20	84.71±0.41	76.43±0.67	89.67±0.44
ProtoNet [28]	ResNet-12	12.5M	62.29±0.33	79.46±0.48	68.25±0.23	84.01±0.56
MetaOptNet [64]	ResNet-12	12.5M	62.64±0.61	78.63±0.46	65.99±0.72	81.56±0.53
Meta-Baseline [65]	ResNet-12	12.5M	63.17±0.23	79.26±0.17	68.62±0.27	83.74±0.18
DeepEMD [14]	ResNet-12	12.5M	65.91±0.82	82.41±0.56	71.16±0.87	86.03±0.58
Feat [66]	ResNet-12	12.5M	66.78±0.20	82.05±0.14	70.80±0.23	84.79±0.16
TPMN [67]	ResNet-12	12.5M	67.64±0.63	83.44±0.43	72.24±0.70	86.55±0.63
SetFeat [68]	ResNet-12	12.5M	68.32±0.62	82.71±0.46	73.63±0.88	87.59±0.57
FGFL [69]	ResNet-12	12.5M	69.14±0.80	86.01±0.62	73.21±0.88	87.21±0.61
SUN [70]	Visformer-S	12.4M	67.80±0.45	83.25±0.30	72.99±0.50	86.74±0.33
FewTURE [36]	ViT-S	22.0M	68.02±0.88	84.51±0.53	72.96±0.92	86.43±0.67
CPEA [9]	ViT-S	22.0M	**71.97±0.65**	**87.06±0.38**	**76.93±0.70**	**90.12±0.45**
KTN [42]	ResNet-12	12.5M	61.42±0.72	74.16±0.56	-	-
AM3 [19]	ResNet-12	12.5M	65.30±0.49	78.10±0.36	69.08±0.47	82.58±0.31
TRAML [41]	ResNet-12	12.5M	67.10±0.52	79.54±0.60	-	-
DeepEMD-BERT [20]	ResNet-12	12.5M	67.03±0.79	83.68±0.65	73.76±0.72	87.51±0.75
DHSRG [3]	WRN-28-10	36.5M	70.64±0.60	**84.71±0.41**	76.43±0.67	**89.67±0.44**
SP [21]	Visformer-T	10.0M	**72.31±0.40**	83.42±0.30	**78.03±0.46**	88.55±0.32
Pre-train (Ours)	Visformer-T	10.0M	67.36±0.46	81.25±0.40	72.55±0.47	86.59±0.29
BiKop (Ours)	Visformer-T	10.0M	78.07±0.36	83.53±0.29	80.29±0.43	88.26±0.32
BiKop (Ours)	ViT-S	22.0M	**79.55±0.50**	**88.07±0.37**	**82.98±0.63**	**90.33±0.44**

C. Comparison with State-of-the-arts (SOTAs)

In Tabel I and III, we comprehensively compare BiKop with previous algorithms under 1-shot and 5-shot settings, where methods KTN [42], AM3 [19], TRAML [41], DeepBERT [20], DHSRG [3], and SP [21] similarly leverage textual information from class names.

1) miniImageNet: BiKop demonstrates a significant performance improvement compared to existing methods. With the Visformer-T backbone, BiKop achieves a classification accuracy of 5.76% higher than SP [21], which also incorporates textual information as prior knowledge. When employing the meticulously pre-trained ViT-S backbone, we attain a performance improvement of 7.58% compared to CPEA [9]. Under the 5-shot setting, our approach achieves an obvious lead, *i.e.,* 1.01% with the ViT-S backbone. Moreover, it is still on par with previous SOTAs when using the Visformer-T backbone.

2) tieredImageNet: By adopting the same backbone ViT-S, BiKop outperforms the previous best method CPEA [9] with advantages of 6.05% and 0.21% in 1-shot and 5-shot settings, respectively. When the backbone of Visformer-T is utilized, BiKop obtains an improvement of 2.26% under the 1-shot setting.

3) CIFAR-FS: Compared to the competitive SP [21], BiKop achieves 2.65% and 0.65% higher accuracies in the 1-shot and 5-shot settings, respectively. With the backbone of ViT-S, the proposed BiKop outperforms the preeminent CPEA [9] with a performance improvement of 6.37% in 1-shot and 0.72% in 5-shot.

TABLE II
ABLATION STUDY UNDER THE 1-SHOT SETTING. I→T AND I←T REPRESENT THE UNIDIRECTIONAL PERMEATION FROM VISUAL TO TEXTUAL MODALITIES ONLY AND THE OPPOSITE DIRECTION.

Backbone	M-Prompt	BKP i2t	BKP t2i	BKP SAD	*mini*	*tiered*	CIFAR	FC100
Visformer-T (10M)	✗	✗	✗	✗	71.88	76.51	81.34	47.71
	✔	✗	✗	✗	76.36	78.76	82.50	52.12
	✔	✗	✔	✗	77.77	80.89	85.23	51.16
	✔	✔	✗	✗	75.22	78.11	82.07	51.43
	✔	✔	✔	✗	77.89	80.21	84.43	50.94
	✗	✔	✔	✔	76.45	77.21	84.04	49.68
	✔	✔	✔	✔	78.07	80.29	**84.83**	52.14
ViT-S (22M)	✔	✔	✔	✔	**79.55**	**82.98**	84.19	53.25
Swin-T (28M)	GPT Prompt [71]				78.94	82.37	84.34	**54.27**

4) FC100: For the more challenging dataset FC100, BiKop again sets a new SOTA performance across various settings. Based on the backbone of ViT-S, BiKop obtains a significant improvement of 6.01% in the 1-shot scenario and gets a similar result to CPEA [9] in the 5-shot setting. Meanwhile, with the backbone of Visformer-T, there is an increase of 3.61% (1-shot) and 0.86% (5-shot).

In summary, BiKop demonstrates a substantial performance lead in the extremely limited samples, *i.e.,* 1-shot setting, affirming that our strategy takes mutual advantage of the textual and visual knowledge, establishing a more robust sparse feature representation. In the 5-shot scenario, we observe a slight performance improvement, as the richer visual features of the support samples take precedence over the complementarity of textual knowledge. For subsequent experiments, we adopt Visformer as the default backbone.

TABLE III
COMPARISON WITH PEERS FOR THE 5-WAY 1-SHOT AND 5-WAY 5-SHOT SETTINGS ON CIFAR-FS [56] AND FC100 [57].

Method	Backbone	Params	CIFAR-FS 5-way		FC100 5-way	
			1-shot	5-shot	1-shot	5-shot
PN+rot [62]	WRN-28-10	36.5M	69.55±0.34	82.34±0.24	-	-
Align [63]	WRN-28-10	36.5M	-	-	45.83±0.48	59.74±0.56
Meta-QDA [74]	WRN-28-10	36.5M	75.83±0.88	88.79±0.75	-	-
ProtoNet [28]	ResNet-12	12.5M	72.20±0.70	83.50±0.50	37.50±0.60	52.50±0.60
MetaOptNet [64]	ResNet-12	12.5M	72.60±0.70	84.30±0.50	41.10±0.60	55.50±0.60
Distill [75]	ResNet-12	12.5M	73.90±0.80	86.90±0.50	44.60±0.70	60.90±0.60
BML [76]	ResNet-12	12.5M	73.45±0.47	88.04±0.33	-	-
CG [77]	ResNet-12	12.5M	73.00±0.70	85.80±0.50	-	-
TPMN [67]	ResNet-12	12.5M	75.50±0.90	87.20±0.60	46.93±0.71	63.26±0.74
MixFSL [78]	ResNet-12	12.5M	-	-	44.89±0.63	60.70±0.60
infoPatch [79]	ResNet-12	12.5M	-	-	43.80±0.40	58.00±0.40
SUN [70]	Visformer-S	12.4M	78.37±0.46	88.84±0.32	-	-
SP [21]	Visformer-T	10.0M	**82.18±0.40**	88.24±0.32	**48.53±0.38**	61.55±0.41
FewTURE [36]	ViT-S	22.0M	76.10±0.88	86.14±0.64	46.20±0.79	63.14±0.73
CPEA [9]	ViT-S	22.0M	77.82±0.66	**88.98±0.45**	47.24±0.58	**65.02±0.60**
Pre-train (Ours)	Visformer-T	10.0M	72.24±0.49	85.76±0.36	43.56±0.39	59.47±0.39
BiKop (Ours)	Visformer-T	10.0M	**84.83±0.35**	88.89±0.31	52.14±0.39	62.41±0.40
BiKop (Ours)	ViT-S	22.0M	84.19±0.58	**89.70±0.47**	**53.25±0.53**	**65.07±0.56**

TABLE IV
COMPARISON WITH PEERS FOR THE 5-WAY 1-SHOT SETTING ON THE META-DATASET [72].

Method	Testing Set							
	Mini-test	CU-birds	Fungi	Omniglot	Traffic-Sign	Quickdraw	Flower	DTD
SimpleShot [80]	67.18	49.68	43.79	78.19	54.04	54.50	71.68	51.19
ZN [81]	67.05	48.15	43.24	78.80	53.92	52.86	72.01	52.20
TCPR [82]	69.52	53.83	46.28	80.88	56.65	57.31	75.37	54.38
AFR [73]	72.98	54.45	47.93	81.84	60.12	58.20	76.11	57.47
BiKop (Visformer-T)	**78.07**	**58.29**	**48.12**	**83.74**	**61.40**	**59.92**	**78.28**	**57.51**

We further evaluate the performance of the proposed BiKop on the larger-scale Meta-Dataset [72] with four advanced techniques following the AFR [73] setting. Specifically, training is conducted on the base set of *mini*ImageNet, while the test phase incorporates the other eight datasets, including CU-birds, Fungi, Omniglot, Traffic-Sign, Quickdraw, Flower, and DTD. As shown in Table IV, the experimental results demonstrate that BiKop achieves superior performance, showcasing its exceptional adaptability to diverse domain data.

D. Ablation Studies

1) Efficacy of different components in BiKop: The ablation study results on key components of BiKop are shown in Table II. First, we use the original class name embedding and remove the proposed BKP and SAD modules, considering this as our baseline. Then, incrementally introducing the meta-class-specific prompt improves accuracy significantly (2.39% on average over four datasets). To validate the effectiveness of BKP, we transform it into two unidirectional knowledge permeations, *i.e.,* from text to vision and vice versa. Observing the results, bidirectional knowledge permeation significantly outperforms unidirectional permeation, indicating that enhancing the generalization of visual knowledge and the individual diversity of textual knowledge is beneficial in establishing a robust joint feature representation. Finally, by incorporating the SAD module, the performance is further enhanced.

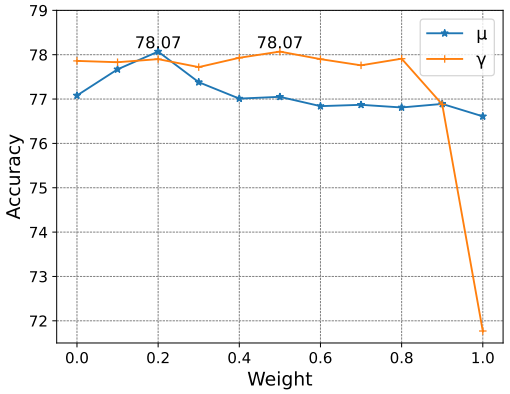

Fig. 4. Effect of weight coefficients μ in the BKP module and γ in the overall loss on *mini*ImageNet under 1-shot setting with the Visformer-T backbone.

Furthermore, we compare the proposed BiKop with the GPT-based prompt approach used in SemFew [71]. While both methods achieve comparable performance, BiKop offers two distinct advantages: (i) The proposed meta-prompt introduces only a minimal number of additional parameters (e.g., 4096 parameters with 8 learnable tokens), whereas GPT involves a more complex computational process, with inference durations extending to several hundred milliseconds (Table V presents the inference speed of BiKop). (ii) BiKop utilizes significantly

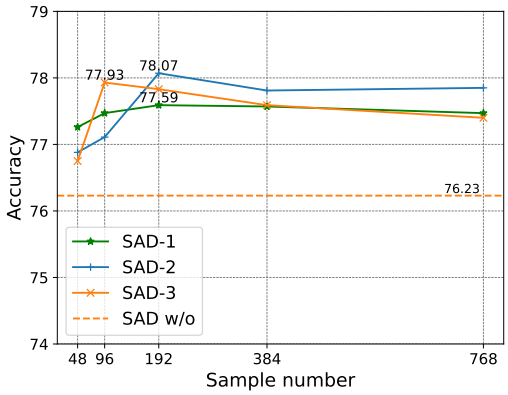

Fig. 5. Effect of sampling times m and layer number of MLP block $\mathcal{D}(\cdot)$ in the SAD module on *mini*ImageNet under 1-shot setting with the Visformer-T backbone.

TABLE V
IMPACT OF SAMPLING NUMBERS ON MODEL EFFICIENCY WITH THE VISFORMER-T BACKBONE.

Sampling numbers	48	96	192	384	768
Training time (s/epoch)	99.41	109.91	129.32	175.01	265.17
Inference time (ms/episode)	22.73	22.58	22.63	22.79	22.75

TABLE VI
IMPACT OF SUPPORT NUMBERS ON THE MODEL'S PERFORMANCE ACROSS VARIOUS DATASETS WITH THE VISFORMER-T BACKBONE.

Shot	1	3	5	7	10
mini	78.07	81.80	83.53	83.93	84.68
tiered	80.29	85.44	88.26	88.92	89.57
CIFAR-FS	84.83	86.91	88.89	88.35	89.67
FC100	52.14	55.33	62.41	62.47	62.67

TABLE VII
IMPACT OF DIFFERENT IMPLEMENTATIONS OF BKP UNDER THE 1-SHOT SETTING WITH THE VISFORMER-T BACKBONE.

	mini	*tiered*	CIFAR-FS	FC100
Dot product	59.91	67.63	64.38	38.02
Addition	75.91	79.44	81.83	47.58
Concatenation	74.36	78.24	82.76	49.02
Cross-attention (ours)	**78.07**	**80.29**	**84.83**	**52.14**

TABLE VIII
THE IMPACT OF DIFFERENT TEXT ENCODERS ON MODEL PERFORMANCE UNDER THE 1-SHOT SETTING WITH THE VISFORMER-T BACKBONE.

Text encoder	1-shot			
	mini	*tiered*	CIFAR-FS	FC100
SBERT [83]	77.58	79.26	84.01	51.22
GloVe [17]	77.14	79.31	83.44	50.76
CLIP [18]	**78.07**	**80.29**	**84.83**	**52.14**

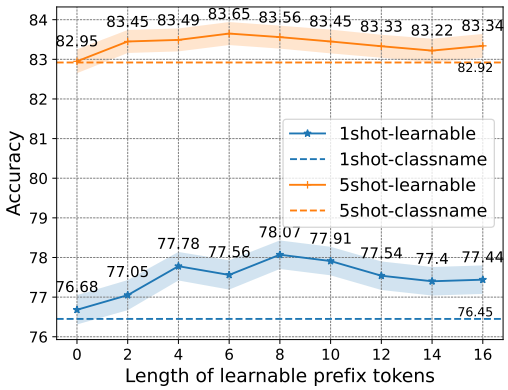

Fig. 6. Designs for the meta-class-specific prompt on *mini*ImageNet with the Visformer-T backbone.

smaller backbone networks, such as Visformer-T and ViT-S, resulting in a more efficient parameter footprint.

2) Influence of hyperparameters: From Fig. 4, as μ increases, the performance improves initially but then experiences a significant decrease. We posit that an excessively large μ leads to the suppression of intrinsic modality knowledge, which might be detrimental to establishing a robust joint feature representation. Meanwhile, a sharp decline occurred in performance when γ becomes too large since an overly high proportion of adversarial loss might bring catastrophic optimization of the model.

For the choice of sampling numbers m and the layer number of the MLP block $\mathcal{D}(\cdot)$ in the SAD module, we compare the corresponding performance in Fig. 5. Additionally, experimental results in V show that training times increase significantly with the number of samples. However, since the SAD module does not participate in inference, the sampling frequency has little to no impact on inference time. Considering the trade-off between time overhead and performance gain, we set the number of samples to 192 and employ a two-layer MLP block.

3) Expand more support samples: In the Sec. IV-C experiments, we observe that the advantages of algorithms utilizing textual knowledge are generally diminished in the 5-shot setting. Therefore, we explore BiKop's performance trends when more support samples are provided. The performance change in Table VI corroborates our perspective: The BiKop method establishes robust joint class representations by permeating textual knowledge and visual features, demonstrating superior performance in extremely low-shot scenarios. However, as the number of visual samples increases, visual information progressively dominates the joint representation, thereby diminishing the gains contributed by textual knowledge.

4) Implementation of BKP: We investigate the impact of different implementation strategies for BKP. Let $GAP(\cdot)$ stands for the Global Average Pooling function, $\boldsymbol{Z}_l^s \in \mathbb{R}^{M\times C_d}$ and $\boldsymbol{T} \in \mathbb{R}^{1\times C_d}$ represent the visual feature and textual knowledge, respectively. "Dot product" can be formulated as:

$$\hat{\boldsymbol{Z}}_l^s = \boldsymbol{Z}_l^s \cdot \boldsymbol{T}, \quad \hat{\boldsymbol{T}} = \boldsymbol{T} \cdot GAP(\boldsymbol{Z}_l^s). \tag{16}$$

However, it can only model the relationships of local information and overlooks the interaction of global information across modalities. "Addition" can be denoted by:

$$\hat{\boldsymbol{Z}}_l^s = \boldsymbol{Z}_l^s + \boldsymbol{T}, \quad \hat{\boldsymbol{T}} = \boldsymbol{T} + GAP(\boldsymbol{Z}_l^s). \tag{17}$$

While this approach maintains the independence of the two modalities, it does not fully utilize their complementarity because of insufficient weight allocation. "Concatenation" concatenates $\boldsymbol{Z}_l^s$ and $\boldsymbol{T}$ along the spatial dimensions, similar

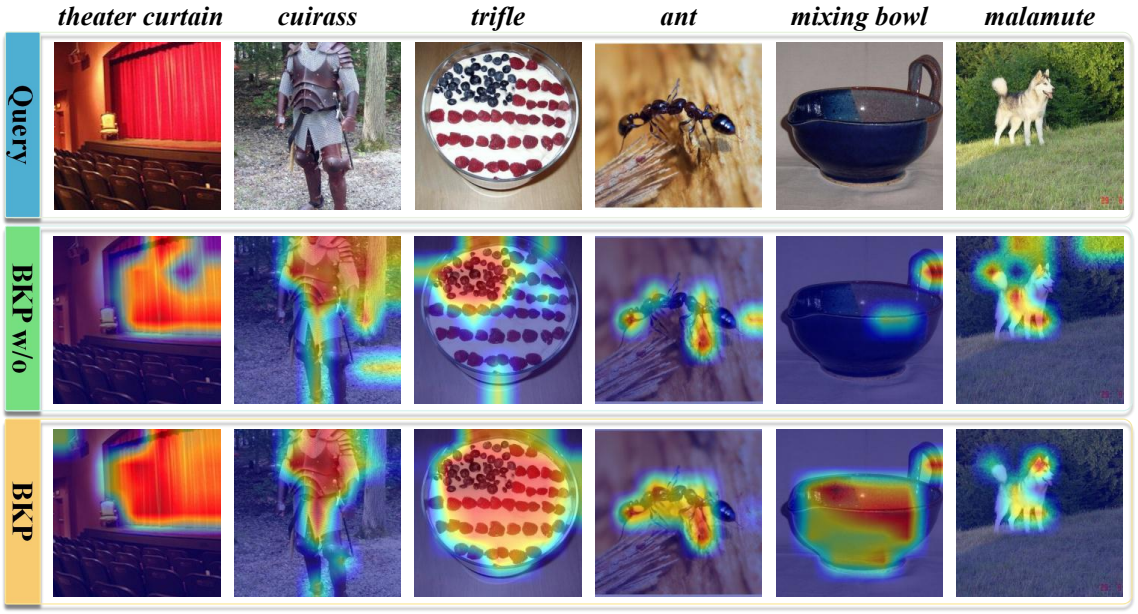

Fig. 7. Quantitative comparison of the segmentation effects of the HSE and baseline methods on iSAID-5^i dataset under the 1-shot setting.

to the "spatial interaction" in SP [21], preserving the integrity of the original features. However, this approach increases the computational burden of the model and lacks the flexibility to model global information effectively. Compared to other methods, cross-attention dynamically allocates attention weights during the modality fusion process, capturing the semantic relationships and contextual information between modalities more accurately, as shown in Table VII.

5) Design of the meta-class-specific prompt.: We explore the impact of different prompt design strategies, including the original class name embedding (dashed lines) and various sizes of the meta-class-specific prompt, *i.e.,* the number of learnable tokens. From Fig. 6, as the size of the prompt increases, the performance gradually improves, followed by a slight decline. The underlying cause can be deduced to be two-fold. Increasing the size of the prompt enhances the discrimination of textual knowledge between the meta-classes. An oversized prompt might potentially result in the model being overfitted. Therefore, the number of learnable tokens is set to 8 to achieve optimal performance.

6) Text encoder selection: To evaluate the impact of different text encoders on model performance, Table VIII reports results under the 1-shot setting using SBERT [83], GloVe [17], and CLIP [18] for textual knowledge extraction. Leveraging superior visual-text alignment, CLIP-based textual knowledge demonstrates the highest robustness. Notably, all three encoders outperform alternative algorithms, underscoring BiKop's remarkable adaptability. This further highlights that BiKop's core strength lies in the BKP module's ability to construct robust joint class representations rather than relying solely on CLIP's inherent cross-modal alignment capabilities.

E. Visualization and Analysis

Luo *et al.* [25] indicated that channel importance is closely related to model bias, impacting the generalization of FSL. Mean Magnitude of Channels (MMC) can reveal channels' responses, high-performing few-shot methods may exhibit a more uniform MMC curve on the test set. To this end, we depict the MMC over the test set of *mini*ImageNet. In Fig. 8, the channel magnitude becomes more uniform after applying the SAD module, which validates that the proposed SAD module effectively alleviates the suppression on potential novel class semantics, *i.e.,* the issue of model bias. Furthermore, we visualize the spatial attention map of the output feature in Fig. 7, where the red regions indicate higher attention values. We notice that the BKP module assists the model in adapting attention to the objects responsible for classification, utilizing a classifier built with only a few support images. This underscores its efficacy in establishing the robust joint feature representation.

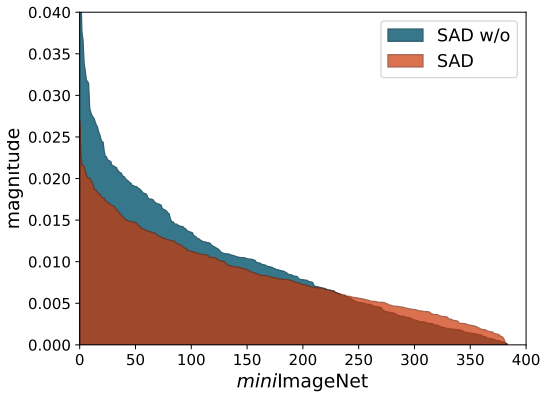

Fig. 8. Visualization results on Mean Magnitude features of Channels (MMC).

V. Conclusion

In this paper, we proposed a novel BiKop framework for FSL. BiKop extends class names to meta-class-specific prompts and incorporates a bidirectional knowledge permeation module, fully leveraging the mutual advantages of textual

and visual knowledge. This effectively mitigates the sparse feature representation collapse. In addition, considering the model bias exacerbated by base-class-relevant information, the Semantic Adversarial Disentanglement module loosens the suppression of potential novel-class semantics during training, further enhancing the model's generalization. Experimental results on four datasets show that BKP performs much better than previous methods. More in-depth ablation studies validate that BiKop responds to the advocated motivation.

REFERENCES

[1] Y. Hu, J. Gao, and C. Xu, "Learning dual-pooling graph neural networks for few-shot video classification," *IEEE Transactions on Multimedia*, vol. 23, pp. 4285–4296, 2021.

[2] R. Zhang, J. Tan, Z. Cao, L. Xu, Y. Liu, L. Si, and F. Sun, "Part-aware correlation networks for few-shot learning," *IEEE Transactions on Multimedia*, vol. 26, pp. 9527–9538, 2024.

[3] H. Wu, G. Ye, Z. Zhou, L. Tian, Q. Wang, and L. Lin, "Dual-view data hallucination with semantic relation guidance for few-shot image recognition," *arXiv preprint arXiv:2401.07061*, 2024.

[4] Y. Jia, J. Gao, W. Huang, Y. Yuan, and Q. Wang, "Exploring hard samples in multiview for few-shot remote sensing scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023.

[5] X. Yang, M. Han, Y. Luo, H. Hu, and Y. Wen, "Two-stream prototype learning network for few-shot face recognition under occlusions," *IEEE Transactions on Multimedia*, vol. 25, pp. 1555–1563, 2023.

[6] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 70. PMLR, 06–11 Aug 2017, pp. 1126–1135.

[7] A. Nichol, "On first-order meta-learning algorithms," *arXiv preprint arXiv:1803.02999*, 2018.

[8] A. A. Rusu, D. Rao, J. Sygnowski, O. Vinyals, R. Pascanu, S. Osindero, and R. Hadsell, "Meta-learning with latent embedding optimization," *arXiv preprint arXiv:1807.05960*, 2018.

[9] F. Hao, F. He, L. Liu, F. Wu, D. Tao, and J. Cheng, "Class-aware patch embedding adaptation for few-shot image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 18 905–18 915.

[10] C. Doersch, A. Gupta, and A. Zisserman, "Crosstransformers: spatially-aware few-shot transfer," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 21 981–21 993.

[11] P. Mangla, N. Kumari, A. Sinha, M. Singh, B. Krishnamurthy, and V. N. Balasubramanian, "Charting the right manifold: Manifold mixup for few-shot learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

[12] O. Vinyals, C. Blundell, T. Lillicrap, k. kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[13] J. Cheng, F. Hao, F. He, L. Liu, and Q. Zhang, "Mixer-based semantic spread for few-shot learning," *IEEE Transactions on Multimedia*, vol. 25, pp. 191–202, 2023.

[14] C. Zhang, Y. Cai, G. Lin, and C. Shen, "Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[15] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[16] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training."

[17] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 139. PMLR, 18–24 Jul 2021, pp. 8748–8763.

[19] C. Xing, N. Rostamzadeh, B. Oreshkin, and P. O. O. Pinheiro, "Adaptive cross-modal few-shot learning," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[20] K. Yan, Z. Bouraoui, P. Wang, S. Jameel, and S. Schockaert, "Aligning visual prototypes with bert embeddings for few-shot learning," in *Proceedings of the 2021 International Conference on Multimedia Retrieval*, ser. ICMR '21, 2021, p. 367–375.

[21] W. Chen, C. Si, Z. Zhang, L. Wang, Z. Wang, and T. Tan, "Semantic prompt for few-shot image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 581–23 591.

[22] Z. Lin, S. Yu, Z. Kuang, D. Pathak, and D. Ramanan, "Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 325–19 337.

[23] Q. Fan, C.-K. Tang, and Y.-W. Tai, "Few-shot object detection with model calibration," in *Computer Vision – ECCV 2022*, pp. 720–739.

[24] T. Ma, Y. Sun, Z. Yang, and Y. Yang, "Prod: Prompting-to-disentangle domain knowledge for cross-domain few-shot image classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 19 754–19 763.

[25] X. Luo, J. Xu, and Z. Xu, "Channel importance matters in few-shot image classification," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 162. PMLR, 17–23 Jul 2022, pp. 14 542–14 559.

[26] Z. Chen, L. Xie, J. Niu, X. Liu, L. Wei, and Q. Tian, "Visformer: The vision-friendly transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 589–598.

[27] D. Alexey, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv: 2010.11929*, 2020.

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

[29] S. Fort, "Gaussian prototypical networks for few-shot learning on omniglot," *arXiv preprint arXiv:1708.02735*, 2017.

[30] R. Ma, P. Fang, G. Avraham, Y. Zuo, T. Zhu, T. Drummond, and M. Harandi, "Learning instance and task-aware dynamic kernels for few-shot learning," in *Computer Vision – ECCV 2022*, pp. 257–274.

[31] Z. Zhao, Z. Cao, H. Xin, R. Wang, D. Wu, Z. Wang, and F. Nie, "Enhancing clustering performance with tensorized high-order bipartite graphs: A structured graph learning approach," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2024.

[32] J. Oh, H. Yoo, C. Kim, and S.-Y. Yun, "Boil: Towards representation change for few-shot learning," *arXiv preprint arXiv:2008.08882*, 2020.

[33] L. Zintgraf, K. Shiarli, V. Kurin, K. Hofmann, and S. Whiteson, "Fast context adaptation via meta-learning," in *Proceedings of the 36th International Conference on Machine Learning*, vol. 97. PMLR, 2019, pp. 7693–7702.

[34] A. Jelley, A. Storkey, A. Antoniou, and S. Devlin, "Contrastive meta-learning for partially observable few-shot learning," *arXiv preprint arXiv:2301.13136*, 2023.

[35] Z. Zhao, F. Nie, R. Wang, Z. Wang, and X. Li, "An balanced, and scalable graph-based multiview clustering method," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 12, pp. 7643–7656, 2024.

[36] M. Hiller, R. Ma, M. Harandi, and T. Drummond, "Rethinking generalization in few-shot classification," *Advances in Neural Information Processing Systems*, vol. 35, pp. 3582–3595, 2022.

[37] Y. Lu, L. Wen, J. Liu, Y. Liu, and X. Tian, "Self-supervision can be a good few-shot learner," in *Computer Vision – ECCV 2022*, pp. 740–758.

[38] Z. Yang, J. Wang, and Y. Zhu, "Few-shot classification with contrastive learning," in *Computer Vision – ECCV 2022*, pp. 293–309.

[39] Z. Zhao, R. Wang, Z. Wang, F. Nie, and X. Li, "Graph joint representation clustering via penalized graph contrastive learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 12, pp. 17 650–17 661, 2024.

[40] A. F. Akyürek, E. Akyürek, D. T. Wijaya, and J. Andreas, "Subspace regularizers for few-shot class incremental learning," *arXiv preprint arXiv:2110.07059*, 2021.

[41] A. Li, W. Huang, X. Lan, J. Feng, Z. Li, and L. Wang, "Boosting few-shot learning with adaptive margin loss," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[42] Z. Peng, Z. Li, J. Zhang, Y. Li, G.-J. Qi, and J. Tang, "Few-shot image recognition with knowledge transfer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[43] T. Xun, W. Chen, Y. He, D. Wu, Y. Gao, J. Zhu, and W. Zheng, "Distinguishing textual prompt importance: Image-guided text weighting for clip-based few-shot learning," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*, 2024, pp. 1–6.

[44] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. van den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[45] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[46] K. Xu, "Show, attend and tell: Neural image caption generation with visual attention," *arXiv preprint arXiv:1502.03044*, 2015.

[47] T. Qiao, J. Zhang, D. Xu, and D. Tao, "Mirrorgan: Learning text-to-image generation by redescription," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[48] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, p. 2337–2348, Sep 2022. [Online]. Available: http://dx.doi.org/10.1007/s11263-022-01653-1

[49] ——, "Conditional prompt learning for vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 16 816–16 825.

[50] Y. Xing, Q. Wu, D. Cheng, S. Zhang, G. Liang, P. Wang, and Y. Zhang, "Dual modality prompt tuning for vision-language pre-trained model," *IEEE Transactions on Multimedia*, vol. 26, pp. 2056–2068, 2024.

[51] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, M. Lucic, and A. Dosovitskiy, "Mlp-mixer: An all-mlp architecture for vision," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 24 261–24 272.

[52] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," *arXiv preprint arXiv:1611.01144*, 2016.

[53] F. Lv, J. Liang, S. Li, B. Zang, C. H. Liu, Z. Wang, and D. Liu, "Causality inspired representation learning for domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 8046–8056.

[54] J. Chen, L. Song, M. Wainwright, and M. Jordan, "Learning to explain: An information-theoretic perspective on model interpretation," in *Proceedings of the 35th International Conference on Machine Learning*, vol. 80. PMLR, 10–15 Jul 2018, pp. 883–892.

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

[56] L. Bertinetto, J. F. Henriques, P. H. Torr, and A. Vedaldi, "Meta-learning with differentiable closed-form solvers," *arXiv preprint arXiv:1805.08136*, 2018.

[57] B. Oreshkin, P. Rodríguez López, and A. Lacoste, "Tadam: Task dependent adaptive metric for improved few-shot learning," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[58] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[59] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[60] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[61] I. Loshchilov, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[62] S. Gidaris, A. Bursuc, N. Komodakis, P. Perez, and M. Cord, "Boosting few-shot visual learning with self-supervision," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[63] A. Afrasiyabi, J.-F. Lalonde, and C. Gagné, "Associative alignment for few-shot image classification," in *Computer Vision – ECCV 2020*, pp. 18–35.

[64] K. Lee, S. Maji, A. Ravichandran, and S. Soatto, "Meta-learning with differentiable convex optimization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[65] Y. Chen, Z. Liu, H. Xu, T. Darrell, and X. Wang, "Meta-baseline: Exploring simple meta-learning for few-shot learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 9062–9071.

[66] H.-J. Ye, H. Hu, D.-C. Zhan, and F. Sha, "Few-shot learning via embedding adaptation with set-to-set functions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[67] J. Wu, T. Zhang, Y. Zhang, and F. Wu, "Task-aware part mining network for few-shot learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 8433–8442.

[68] A. Afrasiyabi, H. Larochelle, J.-F. Lalonde, and C. Gagné, "Matching feature sets for few-shot image classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 9014–9024.

[69] H. Cheng, S. Yang, J. T. Zhou, L. Guo, and B. Wen, "Frequency guidance matters in few-shot learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 11 814–11 824.

[70] B. Dong, P. Zhou, S. Yan, and W. Zuo, "Self-promoted supervision for few-shot transformer," in *Computer Vision – ECCV 2022*, pp. 329–347.

[71] H. Zhang, J. Xu, S. Jiang, and Z. He, "Simple semantic-aided few-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 28 588–28 597.

[72] E. Triantafillou, T. Zhu, V. Dumoulin, P. Lamblin, U. Evci, K. Xu, R. Goroshin, C. Gelada, K. Swersky, P.-A. Manzagol *et al.*, "Meta-dataset: A dataset of datasets for learning to learn from few examples," *arXiv preprint arXiv:1903.03096*, 2019.

[73] X. Zhu, S. Wang, J. Lu, Y. Hao, H. Liu, and X. He, "Boosting few-shot learning via attentive feature regularization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, 2024, pp. 7793–7801.

[74] X. Zhang, D. Meng, H. Gouk, and T. M. Hospedales, "Shallow bayesian meta learning for real-world few-shot recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 651–660.

[75] Y. Tian, Y. Wang, D. Krishnan, J. B. Tenenbaum, and P. Isola, "Rethinking few-shot image classification: A good embedding is all you need?" in *Computer Vision – ECCV 2020*, pp. 266–282.

[76] Z. Zhou, X. Qiu, J. Xie, J. Wu, and C. Zhang, "Binocular mutual learning for improving few-shot classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 8402–8411.

[77] Z. Gao, Y. Wu, Y. Jia, and M. Harandi, "Curvature generation in curved spaces for few-shot learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 8691–8700.

[78] A. Afrasiyabi, J.-F. Lalonde, and C. Gagné, "Mixture-based feature space learning for few-shot image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 9041–9051.

[79] C. Liu, Y. Fu, C. Xu, S. Yang, J. Li, C. Wang, and L. Zhang, "Learning a few-shot embedding model with contrastive learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 10, pp. 8635–8643, May 2021.

[80] Y. Wang, W.-L. Chao, K. Q. Weinberger, and L. Van Der Maaten, "Simpleshot: Revisiting nearest-neighbor classification for few-shot learning," *arXiv preprint arXiv:1911.04623*, 2019.

[81] N. Fei, Y. Gao, Z. Lu, and T. Xiang, "Z-score normalization, hubness, and few-shot learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 142–151.

[82] J. Xu, X. Luo, X. Pan, Y. Li, W. Pei, and Z. Xu, "Alleviating the sample selection bias in few-shot learning by removing projection to the centroid," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 21 073–21 086.

[83] N. Reimers, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

Yuyu Jia received the B.E. degree and the M.S. degree in control theory and engineering from the Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China, in 2019 and 2022, respectively. He is currently pursuing the Ph.D. degree at the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include few-shot learning, deep learning, and remote sensing.

Qing Zhou is currently pursuing the Ph.D degree in computer science and technology with the school of Artificial Intelligence, Optics and Electronics (iOPEN). His research interests include computer vision and pattern recognition.

Junyu Gao received the B.E. degree and the Ph.D. degree in computer science and technology from the Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China, in 2015 and 2021, respectively. He is currently an associate professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.

Qiang Li is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University. His research interests include remote sensing image processing, particularly for image quality enhancement, object/change detection.

Qi Wang (Senior Member, IEEE) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, machine learning, pattern recognition, and remote sensing. For more information, visit the link (https://crabwq.github.io/).