Leveraging Bilateral Correlations for Multi-Label Few-Shot Learning

Yuexuan An, Hui Xue*, *Member, IEEE*, Xingyu Zhao, Ning Xu, Pengfei Fang
and Xin Geng, *Senior Member, IEEE*

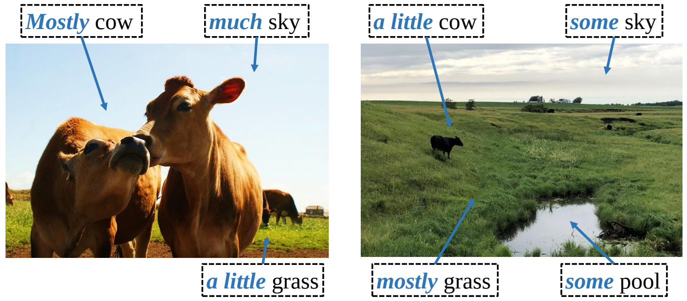
***Abstract*—Multi-label few-shot learning (ML-FSL) refers to the task of tagging previously unseen images with a set of relevant labels, giving a small number of training examples. Modeling the correlations between instances and labels, formulated in the existing methods, allows us to extract more available knowledge from limited examples. However, they simply explore the instance and label correlations with a uniform importance assumption without considering the discrepancy of importance in different instances or labels, making the utilization of instance and label correlations a bottleneck for ML-FSL. To tackle the issue, we propose a unified framework named Bilateral Correlation Reconstruction (BCR) to enable the network to effectively mine underlying instance and label correlations with varying importance information from both instance-to-label and label-to-instance perspectives. Specifically, from the instance-to-label perspective, we refine prototypes per category by reweighting each image with its specific instance-importance degree extracted from the similarity between the instance and the corresponding category. From the label-to-instance perspective, we smooth labels for each image by recovering latent label-importance with considering the integrated topology of all samples in a task. Experimental results on multiple benchmarks validate that BCR could outperform existing ML-FSL methods by large margins.**

***Index Terms*—Multi-Label Learning, Few-Shot Learning, Multi-Label Few-Shot Learning, Label Correlation**

I. Introduction

Recent advances in data-driven deep learning have flourished in several tasks related to artificial intelligence, including object detection [1], [2], action recognition [3], [4] and robotics [5]. However, most deep learning approaches rely heavily on large quantities of labeled data and struggle with problems when labeled data is scarce [6]. In contrast, humans exhibit a remarkable ability to rapidly acquire knowledge about new classes from limited instances. This significant gap between human learning and machine learning provides fertile ground for the development of deep learning [7].

For this reason, recent works of few-shot learning aim to obtain the human-like meta learner that can learn novel concepts with very few samples [8], [9]. However, most existing few-shot learning approaches concentrate solely on the scenario where each image is annotated with a single label, ignoring a more realistic scenario where each image may be interpreted by different concepts [10]–[12]. In such cases, the meta learner should possess more intelligence to simultaneously recognize multiple novel unseen classes with limited examples. Unfortunately, these approaches struggle to handle the challenge posed by multiple labels.



(a) *cow≻sky≻grass* (b) *grass≻sky≻cow*

Fig. 1. An illustration of an image and its corresponding labels. For each image, label-importance (a): *cow≻sky≻grass* and (b): *grass≻sky≻cow*. For each label, instance-importance *cow*: (a) ≻ (b) and *grass*: (b) ≻ (a).

Recently, a few studies have begun to focus on the challenging topic of multi-label few-shot learning (ML-FSL). To address this problem, they attempt to capture instance and label correlations by modeling label dependencies to extract more available knowledge from limited examples. LaSO [13] redeploys the label dependency by generating synthesized feature vectors with a series of label-set operations. KGGR [14] and CMW [15] exploit the statistical label co-occurrences to capture the label dependencies of different labels. Simon et al. [16] proposes a label count module containing context information to tackle the problem. However, the previous methods leverage correlations between instance and label to mitigate the risk of limited examples, with the uniform assumption of instance-importance and label-importance. In practice, the instance-importance corresponding to each instance for a possible label and label-importance corresponding to each relevant label for an instance are often diverse, as illustrated in Figure 1. As a result, these methods may not perform as well as expected due to the unrealistic assumption of uniform latent bilateral correlation.

As a matter of fact, instance-importance and label-importance naturally exist in real-world applications, representing how important an instance is in describing a label and how a particular label is in describing an instance. On the one hand, for each image, the label-importance corresponding to relevant labels is different. For instance, in Figure 1(a), the label of *cow* corresponds to strong label importance while

Y. An, H. Xue, X. Zhao, N. Xu, P. Fang and X. Geng are with the School of Computer Science and Engineering, and the Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications, Ministry of Education, Southeast University, Nanjing 211189, China.
E-mail: {yx_an, hxue, xyzhao, xning, fangpengfei, xgeng}@seu.edu.cn

* Corresponding author

grass corresponds to weak importance. On the other hand, the significance of the label-specific feature for a particular label can vary from image to image, due to the diversity of the label-importance across instances. As a result, the instance-importance for different instances to describe a label is different. For instance, in Figures 1(a) and 1(b), even though two images pose the same labels, the instance-importance for particular labels are quite different. If the diverse instance-importance and label-importance could be fully extracted and leveraged, the performance of the model would be further improved and strengthened.

In light of the above observations, we propose a unified framework called *Bilateral Correlation Reconstruction* (BCR) to effectively leverage the underlying label correlations, along with the varying bilateral importance from both instance-to-label and label-to-instance perspectives. Specifically, to construct instance-to-label correlation, we reweight images according to the similarity-based instance-importance of a certain label in images to refine the representative prototype. To construct label-to-instance correlations, we smooth the ground-truth logical labels into numerical labels with label-importance by aggregating the relation of paired examples in a joint feature and label embedding space. To ensure that the model generates reasonably soft label vectors, we impose a constraint on the soft label to prevent it from deviating too far from the original ground-truth label and lead it to consistently approaching the ground truth.

Our contributions can be summarized as follows.

- We analyze the problem of the uniform label-importance and instance-importance assumption and emphasize the significance of utilization of the latent varying importance information for multi-label few-shot learning (ML-FSL).
- We propose a novel framework, Bilateral Correlation Reconstruction (BCR) for ML-FSL. BCR effectively leverages the underlying label correlations, along with the varying importance from both instance-to-label and label-to-instance perspectives.
- We conduct extensive experiments to validate that our BCR can effectively extract the underlying label correlations without any auxiliary information and outperforms existing ML-FSL methods by large margins.

Other sections of the paper are organized as follows. Firstly, Section II provides a brief review and discussion of related research. Secondly, Section III presents the technical details of BCR. After that, Section IV reports detailed experimental results. Finally, Section V concludes the paper and discusses prospects for future research.

II. Related Work

A. Few-Shot Learning

Few-shot classification aims to acquire profound visual representation by learning to recognize unseen novel classes from a few samples with abundant training on base classes [17], [18]. Many efforts have been dedicated to handling the challenge of data efficiency [19]–[23]. The most related work to our method is the metric-based model, which leverages similarity information in samples to identify novel classes with few examples [24]–[28]. DeepEMD [29] divides images into different local regions and adopts the Earth Mover's Distance to measure the similarity of two images. Prototypical Networks [26] measures the Euclidean distance between a query image and the class centroids of support images. Since the method was proposed, many approaches [30]–[33] have been developed to generate appropriate prototypes to better depict the representation of each class. ProtoComNet [30] utilizes word embedding with WordNet to assist in modifying prototypes. ReProto [31] and RDC [32] leverage the Mixup strategy to generate new prototypes combining original prototypes with context information. CLIP-Adapter [34] imprints the language information with CLIP into the classifier weighting. This method adopts two additional linear layers to residual style connect with original visual or language features to improve the flexibility of few-shot tasks. CLIP-FSAR [35] leverages the textual concepts from CLIP to refine visual prototypes. Different from the above studies, our method aims to adaptively generate appropriate prototypes according to the instance-importance of each instance to the particular labels without any auxiliary prior knowledge.

B. Multi-Label Learning

Recently, there has been a growing interest in tacking with label ambiguity, particularly in the context of multi-label learning (MLL) [36]–[46]. Traditional MLL approaches can be broadly categorized into three types based on the order of label correlations [47]: the first-order approaches decompose MLL into a series of binary classification tasks, which may overlook the potential information from one label that could benefit the learning of others [48]; the second-order approaches consider the correlations between pairs of labels, but they focus primarily on distinguishing relevant from irrelevant labels [49]; the high-order approaches take into account the correlations among label subsets or all class labels [50]. However, these approaches often assume equal label importance and overlook latent discrepancies [51]–[54]. Some techniques have been proposed to mitigate these challenges. For instance, distribution-balanced loss [55] aims to rebalance weights and alleviate the undue suppression of negative labels. ASL [56] introduces an asymmetric loss that employs distinct γ values to weigh positive and negative samples in the focal loss [57]. Another study [58] combines negative-tolerant regularization [55] with class-balanced focal loss [59]. The balanced softmax method [60] transforms the multi-label learning loss into a comparison of scores between relevant and irrelevant labels. Nevertheless, these methods intricately design correlation rules extracted from a vast number of samples and often lack tailored strategies for few-shot learning scenarios. In our approach, we utilize prototypes to represent label-specific features and adjust these prototypes based on varying importance information in labels. This allows us to effectively harness underlying label correlations, enhancing the performance of the few-shot model.

C. Multi-Label Few-Shot Learning

Recently, a few works have begun to focus on the profound but difficult ML-FSL. LaSO [13] employs a data augmentation

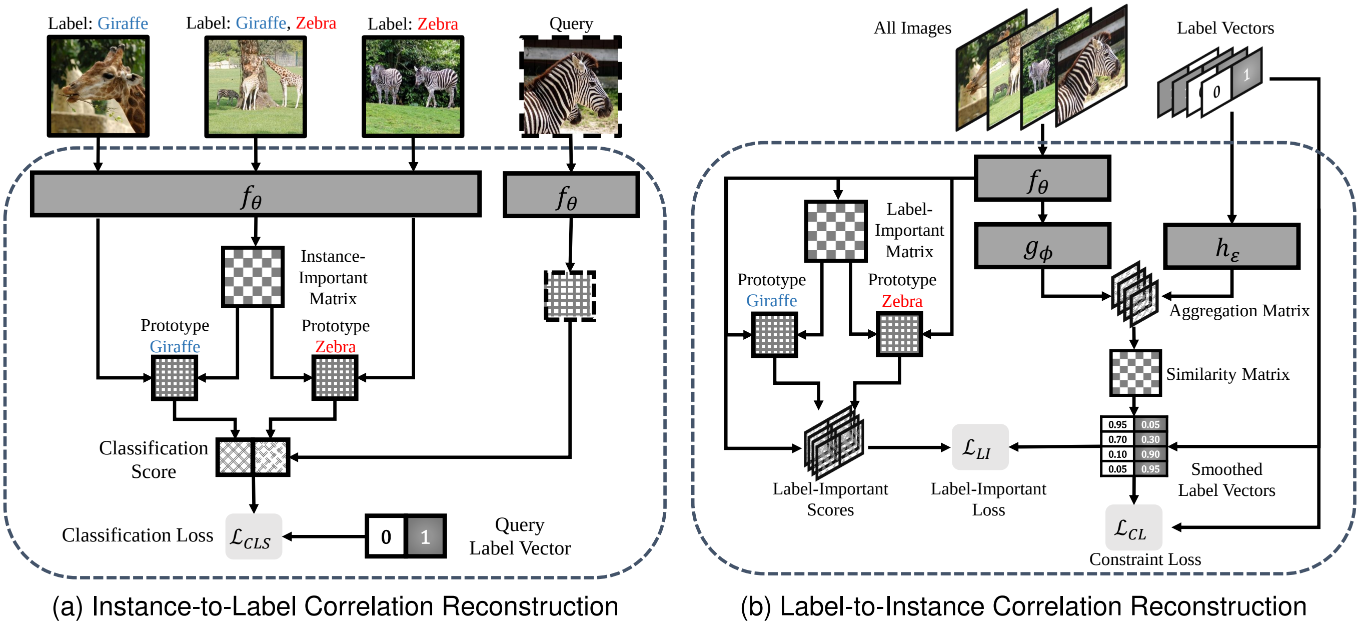


Fig. 2. An overview of Bilateral Correlation Reconstruction. The prototypes in both subfigures are refined prototypes according to the instance-importance of samples.

strategy that generates synthetic feature vectors through label-set operations. KGGR [14] adopts a GCN where labels are modeled as nodes and statistical label co-occurrences are modeled as edges to exploit the label dependencies. To address the statistical bias of limited examples. CMW [15] enhances the label co-occurrence by utilizing word embeddings as auxiliary prior knowledge about label meanings and aggregates the local feature maps of the support images to generate prototypes. [16] designs three baselines by modifying the single-label few-shot learning methods, i.e., ProtoNets [26], RelationNets [61] and Label Propagation Networks with label count module for the multi-label regime. However, the above methods have uniform assumptions about the instance-importance of images to a particular label and the label-importance of relevant labels to each instance. They indiscriminately utilize each label and instance. In the context of few-shot scenarios with limited examples, relying on such an assumption becomes unreasonable. This constraint hinders the effective utilization of underlying label correlations, consequently impeding further enhancements in model performance. In this paper, we resort to diverse instance-importance and label-importance and propose a unified framework for ML-FSL.

III. Bilateral Correlation Reconstruction

A. Problem Setting

In this paper, we consider the following ML-FSL setting. We adopt the episode training process and each episode samples a task. The whole process contains two stages: the meta-training stage and the meta-testing stage. In the meta-training stage, base class dataset $\mathcal{D}_{base}$ with a set of base labels $\mathcal{C}_{base}$ is adopted. In the meta-testing stage, novel class dataset $\mathcal{D}_{novel}$ with a set of novel labels $\mathcal{C}_{novel}$, where $\mathcal{C}_{novel} \bigcap C_{base}=\emptyset$ is used. The meta-training stage can be regarded as a rehearsal to mimic the learning process of the meta-testing stage to assist in model training and generalize the knowledge to meta-testing. Specifically, in the meta-training stage, a series of tasks are sampled from $\mathcal{D}_{base}$ and each task $\mathcal{T} = \{\mathcal{S}, \mathcal{Q}\} = \{(x_t, y_t)\}_{t=1}^{T}$ is composed of a support set $\mathcal{S}$ and a query set $\mathcal{Q}$. The query set $\mathcal{Q}$ is composed of samples that are unseen in $\mathcal{S}$. The goal of each task is to estimate the label vectors of query samples in $\mathcal{Q}$ with limited support samples in $\mathcal{S}$. For a N-way K-shot task, N labels are randomly selected from $\mathcal{C}_{base}$ and K samples are randomly selected for each label to compose support set $\mathcal{S}$. It should be noted that, in multi-label regimes, since each sample can be annotated with multiple labels, each label may correspond to more than K samples. In the meta-testing stage, a series of ML-FSL tasks are sampled for testing where the labels are from $\mathcal{C}_{novel}$. In this stage, query samples are used for the final prediction with limited support samples in $\mathcal{D}_{novel}$.

B. Overview

As illustrated in Figure 2, our framework consists of two main components, i.e., instance-to-label correlation reconstruction (I2L-CR) and label-to-instance correlation reconstruction (L2I-CR). I2L-CR in Figure 2(a) aims to exploit the instance-importance of a certain label in all samples to generate refined prototypes. I2L-CR first adopts similarity scores between samples and the prototype corresponding to a label to profile the instance-importance of the particular label in samples. Then, the prototypes are reconstructed with similarity-based instance-importance scores to achieve the classification loss. L2I-CR in Figure 2(b) aims to recover and leverage the label-importance of relevant labels for each instance. It first concatenates both feature embedding and label embedding for feature information aggregation. Then, the similarity matrix

Algorithm 1 BCR Framework

```
Require: Base class dataset $\mathcal{D}_{base} = \{\mathcal{T}_i\}_{i=1}^{I}$ and meta-testing tasks $\mathcal{T}^* = \{\mathcal{S}^*, \mathcal{Q}^*\}$
Require: Feature extractor $f_\theta(\cdot)$, feature encoder $g_\phi(\cdot)$ and label encoder $h_\epsilon(\cdot)$

procedure TRAIN($\mathcal{D}_{base}$, $f_\theta$, $g_\phi$, $h_\epsilon$)
    while not done do
        Sample task $\mathcal{T} = \{\mathcal{S}, \mathcal{Q}\}$
        Calculate the classification loss $\mathcal{L}_{CLS}$ according to Eq.(6)
        Calculate the label-importance loss $\mathcal{L}_{LI}$ according to Eq.(15)
        Calculate the constraint loss $\mathcal{L}_{CL}$ according to Eq.(16)
        Update $f_\theta$, $g_\phi$, $h_\epsilon$ based on Eq.(17) with $\mathcal{L}_{CLS}$, $\mathcal{L}_{LI}$ and $\mathcal{L}_{CL}$
    end while
end procedure

procedure TEST($\mathcal{T}^*$, $f_\theta$, $g_\phi$, $h_\epsilon$)
    Obtain the refined prototype $\hat{p}^{(c)}$ according to Eq.(3)
    Obtain the final prediction of the classification module $\hat{r}_{(c)}$ according to Eq.(5)
end procedure
```

containing the similarity scores between aggregated features is established in the joint space to smooth the labels. In the meantime, the label-importance scores are estimated with refined prototypes and are leveraged to align smoothed labels. Therefore, the label-importance loss is established. Moreover, a constraint loss is established to constrain smoothed labels from deviating too far from the ground-truth labels.

It should be noted that the importance of instance-to-label and label-to-instance perspectives are different. In the former perspective, the instance-importance denotes the importance degree of any sample to a certain label, which is used for reconstructing prototypes. In the latter perspective, the label-importance denotes the importance degree of all relative labels to each sample, which is leveraged for smoothing label vectors.

C. Instance-to-Label Correlation Reconstruction

For the instance-to-label correlation reconstruction, we explore the correlation of different samples annotated with the same label and adopt prototypes as feature representatives for few-shot scenarios. The original idea of prototypes is the average embeddings of all samples with the same label. Considering the discrepancy of instance-importance inherent in different images, we refine the prototype of the label by reweighting samples with the corresponding instance-importance to the label.

Given a specific task $\mathcal{T} = \{\mathcal{S}, \mathcal{Q}\}$ in an episode, we first use the feature extractor $f_\theta(\cdot)$ to transform the original images from the input space to the embedding space and obtain the original prototype:

$$p^{(c)} = \frac{1}{\left|\mathcal{S}^{(c)}\right|} \sum_{x_i \in \mathcal{S}^{(c)}} f_\theta(x_i), \tag{1}$$

where $\mathcal{S}^{(c)}$ is a subset of $\mathcal{S}$ with the label c and x_i is from $\mathcal{S}^{(c)}$.

Original prototypes are with the assumption that different samples hold uniform instance-importance for a label. According to previous analyses, the assumption is irrational and the learned feature representations might be biased.

To eliminate the problem, for each label, we further refine its prototype according to the instance-importance of the support samples for specific labels, as shown in Figure 2(a). In this paper, for a specific label, we adopt the similarity between a sample and the prototype for the label to indicate the instance-importance of the sample to the label. Therefore, we calculate the normalized similarity score between the sample and average prototype $p^{(c)}$:

$$s_i^{(c)} = \frac{\left[\cos\left(f_\theta(x_i), p^{(c)}\right)\right]_+}{\sum\limits_{x_j \in \mathcal{S}^{(c)}} \left[\cos\left(f_\theta(x_j), p^{(c)}\right)\right]_+}, \tag{2}$$

where $\cos(\cdot,\cdot)$ denotes cosine similarity. We adopt cosine similarity since it is not only insensitive to the absolute values of different numerical values, but can normalize these values to a unified order of magnitude to calculate the similarity between two vectors. Moreover, since the range of cosine similarity is [-1, 1], we leverage the "$[*]_+$" function (if $*$ is less than 0, then $[*]_+$ is 0, else it is $*$) to guarantee non-zero denominators. That is, if the cosine similarity between the specific sample and the average prototype is negative, we assume that it does not contribute to the class prototype.

After that, the normalized similarity score $s_i^{(c)}$ can be used to reweight the samples and the prototype is refined as:

$$\hat{p}^{(c)} = \sum_{x_i \in \mathcal{S}^{(c)}} s_i^{(c)} \cdot f_\theta(x_i). \tag{3}$$

For a query sample x_q in $\mathcal{Q}$, the classification score corresponding to label c is calculated as negative Euclidean distance between its feature and the refined prototype of label c:

$$r_q^{(c)} = -\left\|\hat{p}^{(c)} - f_\theta(x_q)\right\|_2^2 / \tau. \tag{4}$$

where τ is the temperature scalar to scale τ to the appropriate scope. Since ML-FSL focuses more on the classification of a specific label rather than the overall distribution of all labels, the Euclidean distance with a wider range is better than cosine similarity as the measure used in classification.

Finally, the ouput of the classification module $\hat{r}_{(c)}$ which is scaled into $(0,1)$ can be obtained by

$$\hat{r}_q^{(c)} = 2 \cdot \texttt{sigmoid}\left(r_q^{(c)}\right), \tag{5}$$

where $\texttt{sigmoid}(\cdot)$ is the sigmoid function.

Classification Loss. In the meta-training stage, the binary cross-entropy loss function is adopted to update the parameters of the classification module:

$$\begin{aligned}\mathcal{L}_{CLS} = \sum_{q=1}^{|\mathcal{Q}|}\sum_{c=1}^{N} & \Big[y_q^{(c)} \log\left(\hat{r}_q^{(c)}\right) \\ & + \left(1 - y_q^{(c)}\right)\log\left(1 - \hat{r}_q^{(c)}\right)\Big],\end{aligned} \tag{6}$$

where $y^{(c)}$ is the ground-truth label indicator for label c.

D. Label-to-Instance Correlation Reconstruction

For label-to-instance correlation reconstruction, we explore the correlation of different labels in a sample. The core idea is to aggregate the topology of all samples to recover and leverage the latent distribution of label-importance in images.

1) Label Smoothing: The goal of our label smoothing is to utilize existing images and their corresponding label vectors as much as possible to recover the label-to-instance correlation in the given image. As shown in Figure 2(b), given a specific task $\mathcal{T} = \{\mathcal{S}, \mathcal{Q}\}$ in an episode, we merge support set $\mathcal{S}$ with query set $\mathcal{Q}$, and then divide the whole task dataset into the image part $\mathcal{X} \in \mathbb{R}^{(NK+Q)\times C\times H\times W}$ and the label matrix $Y \in \mathbb{R}^{(NK+Q)\times N}$, where Q denote the number of query samples, C, H and W denote the channel, height, and width of the feature maps of the images, respectively. We extract features with $f_\theta(\cdot)$ from $\mathcal{X}$ to generate feature matrix $X \in \mathbb{R}^{(NK+Q)\times D}$, where D denote the dimension of the output space of feature extractor $f_\theta(\cdot)$. Then, feature encoder $g_\phi(\cdot)$ and label encoder $h_\epsilon(\cdot)$ are used to encode X and Y respectively. After that, the embeddings are concatenated in the joint embedding space to construct the aggregation matrix Z from both feature and label perspectives:

$$Z = \texttt{Concat}\left[g_\phi(X), h_\epsilon(Y)\right], \tag{7}$$

where $\texttt{Concat}$ performs concatenation operation. Then, the similarity information of samples in Z is used to enhance the label matrix Y and recover the latent label-importance in labels. Concretely, the similarity matrix M are first constructed with each element in Z:

$$M_{i,j} = \cos\left(Z_i, Z_j\right), \tag{8}$$

where Z_i and Z_j are the i-th and j-th element vector.

After that, in order to profile the label-importance of relevant labels in a sample, we reconstruct the label matrix with smoothed numerical values by aggregating the similarity information of paired examples in the joint embedding space.

$$\hat{Y} = MY, \tag{9}$$

where $\hat{Y}$ is the enhanced label matrix. Subsequently, the smoothed label vector d_i with varying label-importance corresponding to a sample x_i can be obtained by the softmax operation on each element of $\hat{Y}$:

$$d_i = \texttt{softmax}\left(\hat{Y}_i\right). \tag{10}$$

where $\hat{Y}_i$ is the label vector corresponding to x_i.

2) Label-Importance Estimation: The goal of label-importance estimation is to estimate label-important scores and guide the training of the classification model, which illustrates the label-to-instance correlation of an image. To effectively leverage the obtained label-importance information, samples in the task are further used to construct a label-importance estimation. Similar to I2L-CR, for a specific label c, we first attain the original prototype $p^{*(c)}$:

$$p^{*(c)} = \frac{1}{\left|\mathcal{X}^{(c)}\right|}\sum_{x_i \in \mathcal{X}^{(c)}} f_\theta(x_i), \tag{11}$$

where $\mathcal{X}^{(c)}$ is the subset of $\mathcal{X}$ with the label c. Note that $p^{*(c)}$ is different from $p^{(c)}$ in Eq.(1), where the former only uses the support samples to generate prototypes while the latter use all samples in a task.

Then, the similarity-based instance-importance is leveraged to refine prototypes and eliminate this representation bias. We calculate the normalized similarity score between the sample x_i and prototype $p^{*(c)}$:

$$s_i^{*(c)} = \frac{\left[\cos\left(f_\theta(x_i), p^{*(c)}\right)\right]_+}{\sum\limits_{x_j \in \mathcal{X}^{(c)}} \left[\cos\left(f_\theta(x_j), p^{*(c)}\right)\right]_+}. \tag{12}$$

The prototype of label c is refined as:

$$\hat{p}^{*(c)} = \sum_{x_i \in \mathcal{X}^{(c)}} s_i^{*(c)} \cdot f_\theta(x_i). \tag{13}$$

For each sample $x_i \in \mathcal{X}$, the label-importance score corresponding to label c is calculated as follows:

$$r_i^{*(c)} = \frac{e^{-\left\|\hat{p}^{*(c)} - f_\theta(x_i)\right\|_2^2/\tau}}{\sum_{n=1}^{N} e^{-\left\|\hat{p}^{*(n)} - f_\theta(x_i)\right\|_2^2/\tau}}. \tag{14}$$

Label-Importance Loss. We leverage the label-importance loss to guide the label-importance score to approximate the recovered label smoothing vector, so that it can more accurately reflect the relative relationship between different labels to the given image, thereby achieving a more accurate prediction. We adopt the Kullback-Leibler divergence between the smoothed label vector d and the label-importance score vector r^* as the loss function:

$$\mathcal{L}_{LI} = \sum_{i=1}^{|\mathcal{X}|}\sum_{c=1}^{N} d_i^{(c)} \log\frac{d_i^{(c)}}{r_i^{*(c)}}, \tag{15}$$

where i denotes the sample index in $\mathcal{X}$.

Constraint Loss. Additionally, in order to make feature encoder $g_\phi(\cdot)$ and label encoder $h_\epsilon(\cdot)$ recover reasonably label-importance information, we constrain the smoothed numerical label matrix $\hat{Y}$ so that it does not deviate too far from the

TABLE I

Comparison in mAP(%) with existing SOTA methods in 16-way 1-shot and 5-shot scenarios on COCO dataset. The results of compared methods with presented backbones have been reported in the original paper.

Method	Backbone	1-shot	5-shot
LaSO [13]	GoogleNet-v3	45.3	58.1
KGGR [14]	GoogleNet-v3	49.4	61.0
KGGR [14]	ResNet-101	52.3	63.5
CMW [15]	GoogleNet-v3	53.4	65.1
CMW [15]	ResNet-101	55.7	68.1
ProtoNet [16]	Conv-4-64	48.7	59.9
RelationNet [16]	Conv-4-64	49.5	58.5
LPN [16]	Conv-4-64	56.1	63.4
LPN + NLC [16]	Conv-4-64	56.8	64.8
BCR(Ours)	Conv-4-64	58.6	66.5
BCR(Ours)	GoogleNet-v3	61.5	70.1
BCR(Ours)	ResNet-101	**63.7**	**72.2**

ground-truth label. Concretely, we adopt the constraint with the following loss function:

$$\mathcal{L}_{CL} = \sum_{i=1}^{|\mathcal{X}|} \sum_{c=1}^{N} \Big[y_i^{(c)} \log \Big(\tilde{Y}_i^{(c)} \Big) + \Big(1 - y_i^{(c)} \Big) \log \Big(1 - \tilde{Y}_i^{(c)} \Big) \Big], \tag{16}$$

where $\tilde{Y}_i^{(c)} = \texttt{sigmoid}(\hat{Y}_i^{(c)})$ and $y^{(c)}$ is the ground-truth label indicator for label c.

E. Model Training and Prediction

In the meta-training stage, we leverage the following loss function to optimize the model in each episode:

$$\mathcal{L} = \mathcal{L}_{CLS} + \lambda \mathcal{L}_{LI} + \eta \mathcal{L}_{CL}, \tag{17}$$

where λ and η are hyper-parameters to balance the classification loss, label-importance loss and constraint loss.

In the meta-testing stage, the output of the classification score is the final prediction in Eq.(5). The pseudocode for the training and testing processes of the BCR framework is given in Algorithm 1.

IV. Experiments

In this section, the effectiveness of BCR is verified through a series of experiments. All methods are implemented using the PyTorch framework. The computations are performed on a GPU server with NVIDIA Tesla V100 GPU, Intel Xeon Gold 6240 CPU 2.60 GHz processor and 32 GB GPU memory.

A. Datasets and Preprocessing

To evaluate the performance of BCR, we conduct experiments on several real-world datasets, including MS-COCO [62], CUB-200-2011 [63], NUS-WIDE [64] and Visual Genome [65]. For each dataset, we split three disjoint sets of labels for training, validating and testing respectively, and the images with annotated labels in one of these sets do not appear in the others. Following [16], we remove images that have less than two labels.

- MS-COCO (COCO) is a benchmark dataset for object detection and recognition tasks, comprising 80K training images annotated with 80 different labels. Following [13], we split the training set, validation set and test set into 48, 16 and 16 labels, respectively.
- CUB-200-2011 (CUB) is a prominent image benchmark dataset for fine-grained classification tasks. It encompasses a total of 11,788 images annotated with 200 labels. For the CUB dataset, we narrow our focus to the top 100 most prevalent categories, forming a relevant subset. Subsequently, we partition these labels into distinct sets: a training set containing 60 labels, a validation set containing 20 labels, and a test set containing 20 labels.
- NUS-WIDE (NUS) is a public multi-label image classification dataset comprising 260K images labeled with 81 visual concepts. To facilitate our experimentation, we distribute these labels across distinct sets: a training set incorporating 41 labels, a validation set encompassing 20 labels, and a test set encompassing 20 labels.
- The Visual Genome (VG) dataset consists of an extensive collection of 108,077 images, each meticulously annotated with object labels spanning a multitude of categories. In our exploration of the VG dataset, we narrow our focus to the top 100 most frequently occurring categories, thereby establishing a subset for our analysis. Consequently, we partition these selected labels into distinct subsets: a training set incorporating 60 labels, a validation set encompassing 20 labels, and a test set containing 20 labels.

B. Methods and Hyper-Parameters

To validate the efficacy of BCR, several competitive and state-of-the-art ML-FSL methods, including LaSO [13], KGGR [14], CMW [15], ProtoNet [16], RelationNet [16], LPN [16] and LPN+NLC [16] are used as baselines. According to the original manuscripts of compared methods, the former two methods adopt the traditional training process while the latter five methods are with the episode training process. Detailed descriptions of these methods are given as follows:

- LaSO [13] introduces a data augmentation technique that involves generating synthesized feature vectors through label-set operations. This innovative approach enables the generation of novel multi-label samples by combining elements from other samples, thereby expanding the scope of data augmentation.
- KGGR [14] presents a novel approach that integrates statistical label correlations with deep neural networks. By incorporating prior knowledge, it guides adaptive information propagation across various categories within a graph. This strategy enhances multi-label analysis and reduces reliance on training samples.
- CMW [15] leverages word embeddings as a source of prior knowledge concerning label meanings. By aggregating local feature maps from the support images, it derives visual prototypes that incorporate this prior information.

TABLE II

COMPARISON IN MAP(%) WITH DIFFERENT DATASETS ON 10-WAY 1-SHOT AND 5-SHOT SCENARIOS. ALL RESULTS ARE AVERAGED OVER 1000 TEST EPISODES WITH 95% CONFIDENCE. ALL METHODS ADOPT CONV-4-64 [25] AS THE BACKBONE AND THE FEATURE EMBEDDING DIMENSIONS ARE SET TO 1600. ● / ○ INDICATES WHETHER BCR IS STATISTICALLY SUPERIOR/INFERIOR TO THE COMPARING ALGORITHMS.

Method	COCO		CUB		NUS		VG	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
LaSO [13]	53.17±0.75●	53.55±0.74●	59.85±0.56●	60.37±0.57●	49.27±0.80●	49.61±0.82●	52.91±0.67●	53.31±0.67●
KGGR [14]	53.87±0.70●	54.07±0.73●	59.88±0.58●	59.90±0.57●	49.93±0.80●	50.81±0.81●	54.40±0.71●	55.23±0.69●
CMW [15]	53.41±0.48●	53.68±0.48●	60.00±0.57●	61.04±0.59●	49.10±0.40●	49.82±0.40●	52.74±0.49●	53.05±0.50●
ProtoNet [16]	66.58±0.76●	73.35±0.78●	60.23±0.59●	62.65±0.61●	53.59±0.85●	64.41±0.93●	56.57±0.73●	59.73±0.73●
RelationNet [16]	65.90±0.77●	73.97±0.73●	61.84±0.61	62.97±0.58●	52.89±0.85●	61.22±0.89●	55.11±0.69●	58.77±0.68●
LPN [16]	61.04±0.81●	63.57±0.81●	61.37±0.60●	63.13±0.62●	54.89±0.90●	63.63±0.96●	56.04±0.72●	60.28±0.71●
LPN + NLC [16]	61.97±0.79●	66.32±0.83●	61.53±0.59●	63.36±0.59	55.11±0.92●	63.92±0.95●	56.29±0.72●	60.62±0.72●
BCR(Ours)	**71.63±0.69**	**77.23±0.67**	**62.57±0.58**	**64.09±0.61**	**61.82±0.90**	**70.28±0.86**	**58.81±0.69**	**63.48±0.68**

TABLE III

COMPARISON IN MAP(%) WITH DIFFERENT CASES OF BCR ON 10-WAY 1-SHOT AND 5-SHOT SCENARIOS. ALL RESULTS ARE AVERAGED OVER 1000 TEST EPISODES WITH 95% CONFIDENCE. ● / ○ INDICATES WHETHER BCR IS STATISTICALLY SUPERIOR/INFERIOR TO THE COMPARING CASES.

Case	COCO		CUB		NUS		VG	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
ProtoNet	66.58±0.76●	73.35±0.78●	60.23±0.59●	62.65±0.61●	53.59±0.85●	64.41±0.93●	56.57±0.73●	59.73±0.73●
Pre-Trained Directly	62.11±0.68●	62.99±0.65●	62.14±0.59●	62.80±0.58●	56.65±0.71●	59.01±0.69●	56.05±0.58●	56.41±0.58●
Average prototypes	69.95±0.71●	75.73±0.68●	61.13±0.60●	62.77±0.58●	59.99±0.89●	67.89±0.85●	57.29±0.69●	61.44±0.68●
EUC+COS	57.03±0.77●	58.10±0.74●	59.79±0.60●	60.06±0.59●	50.81±0.69●	52.88±0.72●	54.66±0.65●	55.01±0.63●
COS+COS	57.89±0.72●	59.06±0.69●	60.23±0.59●	60.44±0.59●	52.04±0.74●	54.29±0.78●	55.27±0.70●	55.67±0.65●
EUC+EUC	69.91±0.73●	76.40±0.70●	61.14±0.59●	62.61±0.58●	59.28±0.90●	68.80±0.89●	57.73±0.68●	61.94±0.71●
$\mathcal{L}_{CLS}$	70.14±0.74●	75.93±0.68●	61.14±0.59●	62.79±0.58●	59.54±0.86●	68.27±0.85●	57.16±0.67●	61.23±0.68●
$\mathcal{L}_{CLS}$ + $\mathcal{L}_{CL}$	70.33±0.75●	76.01±0.71●	61.34±0.59●	62.93±0.58●	59.74±0.86●	68.36±0.85●	57.44±0.70●	61.26±0.70●
$\mathcal{L}_{CLS}$ + $\mathcal{L}_{LI}$	70.30±0.72●	76.02±0.69●	61.63±0.59●	62.95±0.58●	59.98±0.89●	68.41±0.86●	57.46±0.69●	61.48±0.69●
BCR(Ours)	**71.63±0.69**	**77.23±0.67**	**62.57±0.58**	**64.09±0.61**	**61.82±0.90**	**70.28±0.86**	**58.81±0.69**	**63.48±0.68**

- ProtoNet [16] extends the concept of prototypical networks to operate within a multi-label context. It accomplishes this by utilizing a softmax function to perform the multi-label classification.
- RelationNet [16] introduces modifications to the relation module, transforming it into a non-linear metric. Additionally, it employs a log-loss function to seamlessly transition RelationNet from single-label to multi-label scenarios.
- LPN [16] formulates few-shot learning within a graph-based framework. It harnesses label propagation techniques for query sample classification. By capitalizing on the smoothness property, LPN predicts concept scores, attributing similar concept scores to akin instances.
- LPN + NLC [16] synergizes the LPN approach with neural label count module (NLC). This integration enhances the ability of the method to estimate the number of labels associated with a given input, thereby contributing to overall performance improvement.

For fair comparisons, in each experiment, the maximum number of training episodes is set to 50000. All algorithms leverage the Adam optimizer [66] for parameter optimization, uniformly employing a learning rate of 10^{-3}. In BCR, the hyper-parameters λ and η is chosen from the set $\{0.01, 0.05, 0.1, 0.5, 1.0\}$ on the validation set using grid search. The performances of different methods are evaluated using the extensively adopted mean average precision (mAP) [13]–[16].

C. Experimental Results

1) Comparison on COCO Dataset: We first evaluate the performances of different algorithms on the COCO dataset under the setting of 16-way 1-shot and 5-shot, using the class split proposed in [13]. For BCR, we test its performance when using Conv-4-64, GoogleNet-v3, and ResNet-101 as backbones. Since GoogleNet-v3 and ResNet-101 require more training resources compared with Conv-4, the model is first pre-trained on the training dataset with binary cross-entropy loss function for 50 epochs. The performance comparisons are presented in Table I. For each compared method, we use the reported result from the original paper. We highlight the best result among all methods in bold. From Table I, it can be observed that our proposed BCR exhibits significant performance advantages. Using a simple and shallow Conv-4-64 backbone, BCR can outperform existing methods with deeper architecture by a considerable margin on the 1-shot setting. Moreover, our proposed BCR method is better than existing methods on the GoogleNet-v3 and ResNet-101 backbones. Especially, BCR with ResNet-101 outperforms all existing leading methods by around 7% in 1-shot setting and 4% in

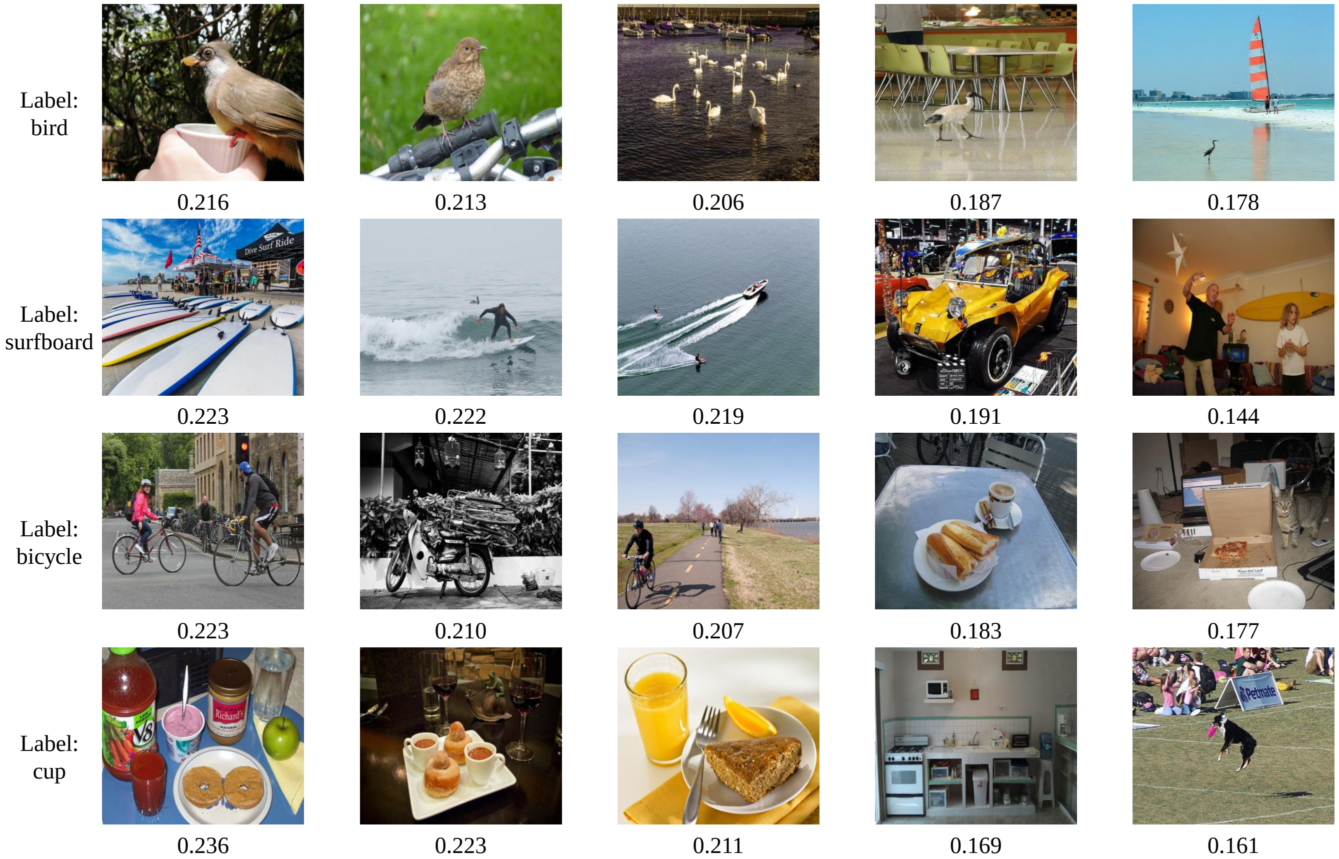


Fig. 3. Quantitative analysis of instance-importance in an exemplar ML-FSL task.

5-shot setting. These results effectively illustrate the efficacy and flexibility of our proposed BCR.

2) Comparison on Different Datasets: To further substantiate the superior performance of BCR and ensure a fair comparison, we conduct experiments across all methods under the same settings. We adopt Conv-4-64 [25] as the backbone and set the feature embedding dimensions as 1600 for all methods. The evaluation metric employed is the average mAP calculated over 1000 test episodes, accompanied by 95% confidence intervals. These evaluations are conducted on the COCO, CUB, NUS, and VG datasets, encompassing both 10-way 5-shot and 1-shot scenarios. Table II presents the experimental results. For each comparison, the best result is highlighted in bold. Moreover, we perform two-tailed t-tests at a significance level of 0.05 to determine whether BCR exhibits statistically superior or inferior performance compared to comparing algorithms.

From Table II, we can observe that BCR achieves the highest accuracies and outperforms the previous methods with significant margins in all cases on all four datasets. Especially, since the classes in these two datasets are more explicit and have more significant differences, BCR outperforms current leading methods by around 6% on COCO and NUS datasets in all cases. In CUB, the fine-grained dataset with significant class overlap and VG the visual question-answering dataset with poor-quality annotations and ambiguous object names [67], BCR still improved the performances consistently. Moreover, BCR establishes statistical superiority over the compared algorithms in almost all cases. Experimental results demonstrate that BCR has superior performance in ML-FSL tasks.

D. Quantitative Analysis of Instance-Importance and Label-Importance

We conduct quantitative analysis to illustrate the effectiveness of instance-to-label correlation reconstruction and label-to-instance correlation reconstruction. Concretely, we analyze the instance-importance scores for different images to a particular label and label-importance score distribution for different labels to an instance in exemplar ML-FSL tasks.

Figure 3 gives a quantitative analysis of instance-to-label correlation reconstruction in an exemplar ML-FSL task. From Figure 3, it can be observed that BCR can generate reasonable instance-importance score for different images to a particular label. For each label, images that can describe the label more effectively have higher instance-importance scores, while images with less instance-to-label correlation have lower scores.

Figure 4 illustrates a quantitative analysis of label-to-instance correlation reconstruction in an exemplar ML-FSL task. From Figure 3, it can be observed that BCR can generate suitable label-importance distribution for different labels to an instance. For each image, the label that can describe the image more accurately has higher label-importance scores, while the label with less label-to-instance correlation has lower scores.

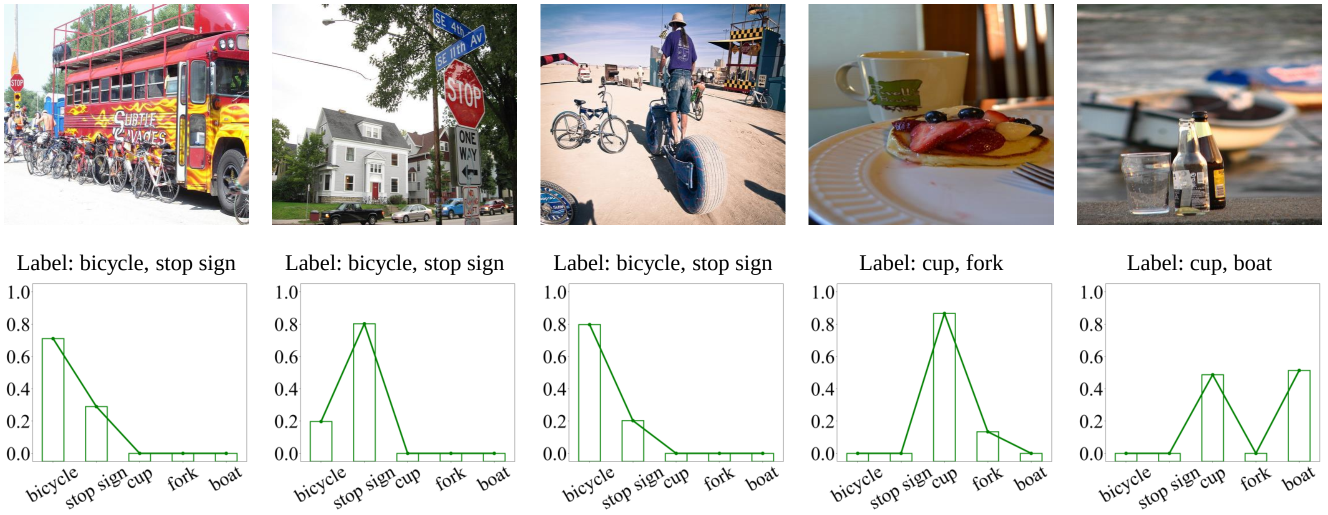
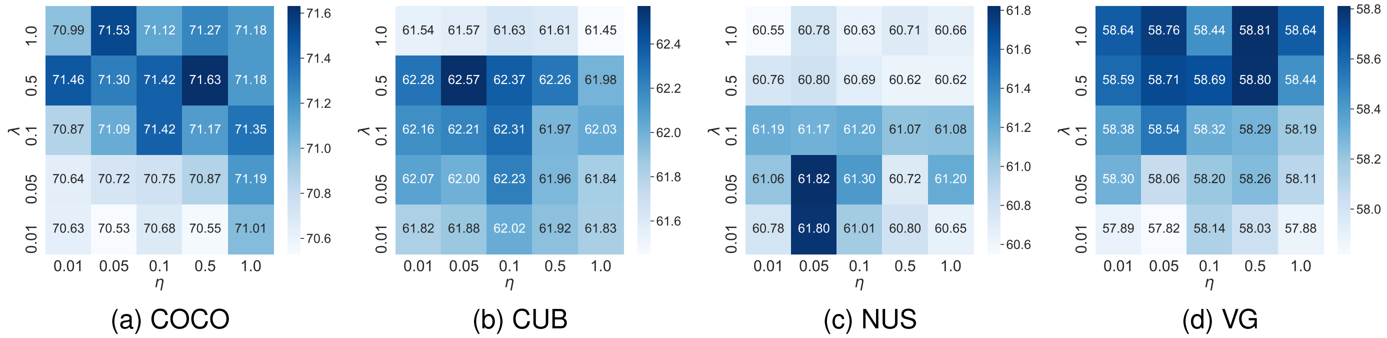


Fig. 4. Quantitative analysis of label-importance in an exemplar ML-FSL task.



(a) COCO (b) CUB (c) NUS (d) VG

Fig. 5. Performance comparison with between varying values of λ and η on different datasets.

E. Ablation Study

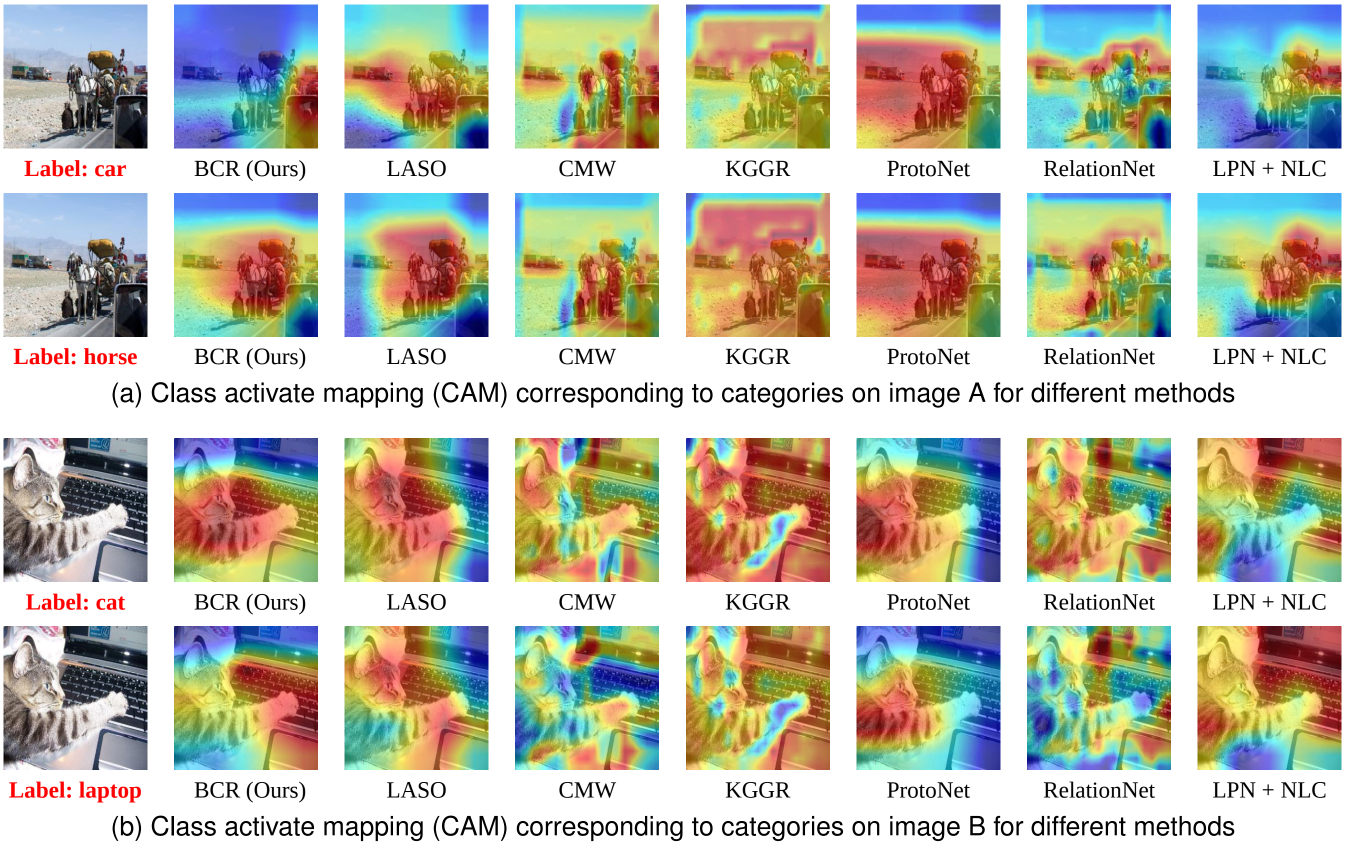
1) Efficiency of Different Components: To ensure a comprehensive grasp of our model, we devise four distinct scenarios to assess the individual impact of each component within the proposed BCR framework.

- Case 1 (Pre-trained directly): Directly using the encoder pre-trained on all base classes using binary cross-entropy loss [68], [69] for the final prediction with Eq. (5) without meta-training.
- Case 2 (Average prototypes): Constructing prototypes without reweighting each sample according to instance-importance.
- Case 3 (EUC+COS): Use Euclidean distance to replace cosine similarity in Eqs.(2) and (12), and use cosine similarity to replace Euclidean distance in Eqs.(4) and (14).
- Case 4 (COS+COS): Use cosine similarity to replace Euclidean distance in Eqs.(4) and (14).
- Case 5 (EUC+EUC): Use Euclidean distance to replace cosine similarity in Eqs.(2) and (12).
- Case 6 ($\mathcal{L}_{CLS}$): Training the model without $\mathcal{L}_{LI}$ and $\mathcal{L}_{CL}$ in Eq. (17).
- Case 7 ($\mathcal{L}_{CLS}$ +$\mathcal{L}_{CL}$): Training the model without $\mathcal{L}_{LI}$ in Eq. (17).
- Case 8 ($\mathcal{L}_{CLS}$ +$\mathcal{L}_{LI}$): Training the model without $\mathcal{L}_{CL}$ in Eq. (17).

The experimental results of the various cases on the four datasets are consolidated in Table III. From Table III, we can observe that each part of our proposed BCR is essential. Specifically, meta-training process of BCR is necessary (case 1). The refined prototype can promote the classification ability of the model (case 2). The usages of cosine similarity in Eqs.(2) and (12) and Euclidean distance in Eqs.(4) and (14) are essential (Case 3-5). These observations illustrate that leveraging cosine similarity to calibrate prototypes and Euclidean distance for classification is beneficial for the whole learning system. Both $\mathcal{L}_{LI}$ and $\mathcal{L}_{CL}$ contribute positively to performance enhancement (Case 6-8). In the meantime, these observations further prove the rationality of latent instance-importance and label-importance extraction and the validity of each component in our proposed BCR.

2) Sensitivity Analysis of λ and η: In this part, we delve into the effectiveness of the hyperparameters λ and η. We undertake a comparative analysis of BCR across various values of λ and η under the 10-way 1-shot setting. The experimental results achieved by BCR with varying values of λ and η are illustrated in Figure 5.

From Figure 5, we can find the below.



(a) Class activate mapping (CAM) corresponding to categories on image A for different methods

(b) Class activate mapping (CAM) corresponding to categories on image B for different methods

Fig. 6. Comparison of the feature maps generated by different methods.

1) Overall, BCR has a stable performance with a wide range of hyper-parameter values on all four datasets.
2) Model performance variations remain within a minimal threshold of 1% across different parameter values.
3) Suitable value of η contribute to enhanced model performance. In particular, on the COCO dataset, the model achieves the best performance when both η and λ are set to 0.5. On the CUB dataset, optimal performance is reached with η set to 0.05 and λ set to 0.5. This happens on the NUS dataset when the values of η and λ are equal to 0.05. Lastly, on the VG dataset, BCR demonstrates superior performance with η at 0.5 and λ at 1. These findings further verify the robustness of the proposed BCR in practical application.

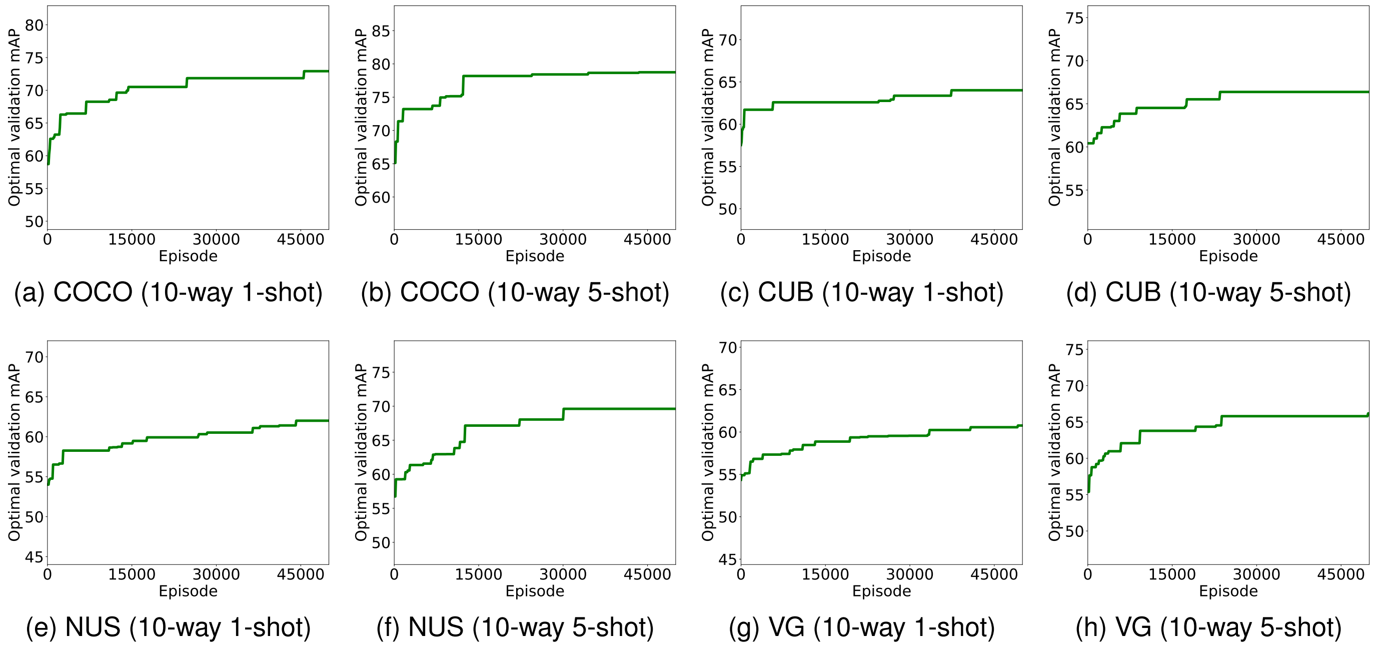
F. Visualization Analysis

To further demonstrate that our BCR can effectively mitigate the deviation of the learning for different categories, we utilize Class Activation Mapping (CAM) [70] to exhibit the visualization results of the compared methods and BCR in Figure 6. Concretely, we illustrate the responses of feature maps generated by different methods to different categories. The first column of Figure 6 showcases original images and their corresponding labels, drawn from the COCO dataset. The second column illustrates CAMs generated by BCR and the subsequent columns display CAMs generated by the comparing methods.

From Figures 6, we can find that the category representations extracted by the compared methods without distinguishing instance-importance and label-importance have serious deviations. For example, the category representations of *car* and *horse* are seriously affected by each other in Figure 6(a), and the category representation of *laptop* is also confused by *cat* in Figure 6(b). On the contrary, from the CAMs generated by BCR, we can find that BCR considering the various instance-importance and label-importance can effectively mitigate the representations bias, which can accurately respond to different categories and generate reasonable category representations. These observations further demonstrate that our BCR can effectively utilize underlying label correlations when handling ML-FSL tasks.

G. Convergence Analysis

We further conduct experiments to test the convergence of the optimal validation mAP for the BCR method with changes in the number of training episodes. The experimental results are illustrated in Figure 7. From Figure 7, we can find that BCR can achieve high precision in early training, while the overall training process can further improve the performance of the model. Therefore, BCR can achieve effective representation and reweighting in the early stages of training through instance-to-label correlation reconstruction, thereby ensuring the effectiveness of the training of the feature extractor.



(a) COCO (10-way 1-shot) (b) COCO (10-way 5-shot) (c) CUB (10-way 1-shot) (d) CUB (10-way 5-shot)

(e) NUS (10-way 1-shot) (f) NUS (10-way 5-shot) (g) VG (10-way 1-shot) (h) VG (10-way 5-shot)

Fig. 7. The convergence of the optimal validation mAP for the BCR method with changes in the number of training episodes.

V. Conclusion

In this paper, we delve into multi-label few-shot learning (ML-FSL), which is a profound and practical topic. We analyze and reveal the problem in existing ML-FSL methods that they model the label correlations with an irrational assumption of the uniform instance-importance and label-importance, which might affect the improvement of model performance. To address this issue, we develop a novel framework named Bilateral Correlation Reconstruction (BCR) to effectively mine and leverage the underlying label correlations, along with varying instance-importance and label-importance from both instance-to-label and label-to-instance perspectives. Our extensive experimental analysis unequivocally demonstrates that BCR, even in the absence of auxiliary information, outperforms existing ML-FSL methods by a significant margin, which unveils the efficiency of utilizing instance-importance and label-importance. We hope our efforts could be helpful for both few-shot learning and multi-label learning communities.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 62076062, 62206050, 62306070, 62125602 and 62076063), the Social Development Science and Technology Project of Jiangsu Province (No. BE2022811), the China Postdoctoral Science Foundation (No. 2021M700023), the Jiangsu Province Science Foundation for Youths (No. BK20210220) and the Young Elite Scientists Sponsorship Program of Jiangsu Association for Science and Technology (No. TJ-2022-078). Furthermore, the work was also supported by the Big Data Computing Center of Southeast University.

References

[1] Y. Xiao, V. Lepetit, and R. Marlet, "Few-shot object detection and viewpoint estimation for objects in the wild," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3090–3106, 2023.

[2] C. Chen, X. Yang, J. Zhang, B. Dong, and C. Xu, "Category knowledge-guided parameter calibration for few-shot object detection," *IEEE Trans. Image Process.*, vol. 32, pp. 1092–1107, 2023.

[3] A. Thatipelli, S. Narayan, S. Khan, R. M. Anwer, F. S. Khan, and B. Ghanem, "Spatio-temporal relation modeling for few-shot action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 926–19 935.

[4] X. Wang, S. Zhang, Z. Qing, M. Tang, Z. Zuo, C. Gao, R. Jin, and N. Sang, "Hybrid relation guided set matching for few-shot action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 916–19 925.

[5] Y. Li, H. Zhu, S. Tian, F. Feng, J. Ma, C. S. Teo, C. Xiang, P. Vadakkepat, and T. H. Lee, "Incremental few-shot object detection for robotics," in *Proceedings of the 2022 IEEE International Conference on Robotics and Automation*, 2022, pp. 8447–8453.

[6] W. Gao, M. Shao, J. Shu, and X. Zhuang, "Meta-bn net for few-shot learning," *Frontiers Comput. Sci.*, vol. 17, no. 1, p. 171302, 2023.

[7] 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," in *Proceedings of the 6th International Conference on Learning Representations*, 2018.

[8] W. Chen, Y. Liu, Z. Kira, Y. F. Wang, and J. Huang, "A closer look at few-shot classification," in *Proceedings of the 7th International Conference on Learning Representations*, 2019.

[9] Y. Jian and L. Torresani, "Label hallucination for few-shot classification," in *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, 2022, pp. 7005–7014.

[10] J. Lanchantin, T. Wang, V. Ordonez, and Y. Qi, "General multi-label image classification with transformers," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 478–16 488.

[11] W. Liu, H. Wang, X. Shen, and I. W. Tsang, "The emerging trends of multi-label learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7955–7974, 2022.

[12] Z. Chen, Q. Cui, B. Zhao, R. Song, X. Zhang, and O. Yoshie, "SST: spatial and semantic transformers for multi-label image recognition," *IEEE Trans. Image Process.*, vol. 31, pp. 2570–2583, 2022.

[13] A. Alfassy, L. Karlinsky, A. Aides, J. Shtok, S. Harary, R. S. Feris, R. Giryes, and A. M. Bronstein, "Laso: Label-set operations networks for

multi-label few-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6548–6557.

[14] T. Chen, L. Lin, R. Chen, X. Hui, and H. Wu, "Knowledge-guided multi-label few-shot learning for general image recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1371–1384, 2022.

[15] K. Yan, C. Zhang, J. Hou, P. Wang, Z. Bouraoui, S. Jameel, and S. Schockaert, "Inferring prototypes for multi-label few-shot image classification with word vector guided attention," in *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, 2022, pp. 2991–2999.

[16] C. Simon, P. Koniusz, and M. Harandi, "Meta-learning for multi-label few-shot classification," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2022, pp. 346–355.

[17] Y. Zhu, F. Zhuang, X. Zhang, Z. Qi, Z. Shi, J. Cao, and Q. He, "Combat data shift in few-shot learning with knowledge graph," *Frontiers Comput. Sci.*, vol. 17, no. 1, p. 171305, 2023.

[18] Y. An, H. Xue, X. Zhao, and L. Zhang, "Conditional self-supervised learning for few-shot classification," in *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, 2021, pp. 2140–2146.

[19] Z. Gu, W. Li, J. Huo, L. Wang, and Y. Gao, "Lofgan: Fusing local representations for few-shot image generation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 8443–8451.

[20] S. Yang, L. Liu, and M. Xu, "Free lunch for few-shot learning: Distribution calibration," in *Proceedings of the 9th International Conference on Learning Representations*, 2021.

[21] H. Ye and W. Chao, "How to train your MAML to excel in few-shot classification," in *Proceedings of the 10th International Conference on Learning Representations*, 2022.

[22] 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*, 2017, pp. 1126–1135.

[23] M. Abbas, Q. Xiao, L. Chen, P. Chen, and T. Chen, "Sharp-maml: Sharpness-aware model-agnostic meta learning," in *Proceedings of the 39th International Conference on Machine Learning*, 2022, pp. 10–32.

[24] Y. Che, Y. An, and H. Xue, "Boosting few-shot open-set recognition with multi-relation margin loss," in *Proceedings of the 32nd International Joint Conference on Artificial Intelligence*, 2023, pp. 3505–3513.

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

[26] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical networks for few-shot learning," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 4077–4087.

[27] J. Xie, F. Long, J. Lv, Q. Wang, and P. Li, "Joint distribution matters: Deep brownian distance covariance for few-shot classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7972–7981.

[28] Y. Liu, W. Zhang, C. Xiang, T. Zheng, D. Cai, and X. He, "Learning to affiliate: Mutual centralized learning for few-shot classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 411–14 420.

[29] C. Zhang, Y. Cai, G. Lin, and C. Shen, "2020 IEEE/CVF conference on computer vision and pattern recognition, CVPR 2020, seattle, wa, usa, june 13-19, 2020," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 200–12 210.

[30] B. Zhang, X. Li, Y. Ye, Z. Huang, and L. Zhang, "Prototype completion with primitive knowledge for few-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3754–3762.

[31] W. Xue and W. Wang, "One-shot image classification by learning to restore prototypes," in *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, 2020, pp. 6558–6565.

[32] P. Li, S. Gong, C. Wang, and Y. Fu, "Ranking distance calibration for cross-domain few-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9099–9108.

[33] Y. An, H. Xue, X. Zhao, and J. Wang, "From instance to metric calibration: A unified framework for open-world few-shot learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 9757–9773, 2023.

[34] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, "Clip-adapter: Better vision-language models with feature adapters," *CoRR*, vol. abs/2110.04544, 2021. [Online]. Available: https://arxiv.org/abs/2110.04544

[35] X. Wang, S. Zhang, J. Cen, C. Gao, Y. Zhang, D. Zhao, and N. Sang, "Clip-guided prototype modulating for few-shot action recognition," *International Journal of Computer Vision*, 2023.

[36] S. Hu, L. Ke, X. Wang, and S. Lyu, "Tkml-ap: Adversarial attacks to top-k multi-label learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 7629–7637.

[37] S. Rajeswar, P. Rodríguez, S. Singhal, D. Vázquez, and A. C. Courville, "Multi-label iterated learning for image classification with label ambiguity," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4773–4783.

[38] X. Zhao, Y. An, N. Xu, and X. Geng, "Fusion label enhancement for multi-label learning," in *Proceedings of the 31st International Joint Conference on Artificial Intelligence*, 2022, pp. 3773–3779.

[39] X. Gong, D. Yuan, and W. Bao, "Understanding partial multi-label learning via mutual information," in *Advances in Neural Information Processing Systems 34*, 2021, pp. 4147–4156.

[40] M. Xie and S. Huang, "Multi-label learning with pairwise relevance ordering," in *Advances in Neural Information Processing Systems 34*, 2021, pp. 23 545–23 556.

[41] J. Bai, S. Kong, and C. P. Gomes, "Gaussian mixture variational autoencoder with contrastive learning for multi-label classification," in *Proceedings of the 39th International Conference on Machine Learning*, vol. 162, 2022, pp. 1383–1398.

[42] J. Hang and M. Zhang, "Dual perspective of label-specific feature learning for multi-label classification," in *Proceedings of the 39th International Conference on Machine Learning*, vol. 162, 2022, pp. 8375–8386.

[43] S. Wang, G. Peng, and Z. Zheng, "Capturing joint label distribution for multi-label classification through adversarial learning," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 12, pp. 2310–2321, 2020.

[44] M. Xu, Y. Li, and Z. Zhou, "Robust multi-label learning with PRO loss," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 8, pp. 1610–1624, 2020.

[45] Y. Zhu, J. T. Kwok, and Z. Zhou, "Multi-label learning with global and local label correlation," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 6, pp. 1081–1094, 2018.

[46] J. Huang, G. Li, Q. Huang, and X. Wu, "Learning label-specific features and class-dependent labels for multi-label classification," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 12, pp. 3309–3323, 2016.

[47] M. Zhang and Z. Zhou, "A review on multi-label learning algorithms," *IEEE Trans. Knowl. Data Eng.*, pp. 1819–1837, 2014.

[48] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognit.*, pp. 1757–1771, 2004.

[49] J. Fürnkranz, E. Hüllermeier, E. L. Mencía, and K. Brinker, "Multilabel classification via calibrated label ranking," *Mach. Learn.*, pp. 133–153, 2008.

[50] K. Bhatia, H. Jain, P. Kar, M. Varma, and P. Jain, "Sparse local embeddings for extreme multi-label classification," in *Advances in Neural Information Processing Systems 28*, 2015, pp. 730–738.

[51] X. Geng, "Label distribution learning," *IEEE Trans. Knowl. Data Eng.*, pp. 1734–1748, 2016.

[52] B. Gao, C. Xing, C. Xie, J. Wu, and X. Geng, "Deep label distribution learning with label ambiguity," *IEEE Trans. Image Process.*, pp. 2825–2838, 2017.

[53] C. Tan, S. Chen, G. Ji, and X. Geng, "A novel probabilistic label enhancement algorithm for multi-label distribution learning," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 11, pp. 5098–5113, 2022.

[54] M. Zhang, Q. Zhang, J. Fang, Y. Li, and X. Geng, "Leveraging implicit relative labeling-importance information for effective multi-label learning," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 5, pp. 2057–2070, 2021.

[55] T. Wu, Q. Huang, Z. Liu, Y. Wang, and D. Lin, "Distribution-balanced loss for multi-label classification in long-tailed datasets," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 162–178.

[56] T. Ridnik, E. B. Baruch, N. Zamir, A. Noy, I. Friedman, M. Protter, and L. Zelnik-Manor, "Asymmetric loss for multi-label classification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 82–91.

[57] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 318–327, 2020.

[58] Y. Huang, B. Giledereli, A. Köksal, A. Özgür, and E. Ozkirimli, "Balancing methods for multi-label text classification with long-tailed class distribution," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 8153–8161.

[59] Y. Cui, M. Jia, T. Lin, Y. Song, and S. J. Belongie, "Class-balanced loss based on effective number of samples," in *CVPR*, 2019.

[60] N. Zhang, X. Chen, X. Xie, S. Deng, C. Tan, M. Chen, F. Huang, L. Si, and H. Chen, "Document-level relation extraction as semantic segmentation," in *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, 2021, pp. 3999–4006.

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

[62] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft COCO: common objects in context,” in *Proceedings of the European Conference on Computer Vision*, vol. 8693, 2014, pp. 740–755.

[63] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, “The caltech-ucsd birds-200-2011 dataset,” California Institute of Technology, Tech. Rep. CNS-TR-2010-001, 2011.

[64] T. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, “NUS-WIDE: a real-world web image database from national university of singapore,” in *Proceedings of the 8th ACM International Conference on Image and Video Retrieval*, 2009.

[65] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, “Visual genome: Connecting language and vision using crowdsourced dense image annotations,” *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, 2017.

[66] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

[67] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, “Scene graph generation by iterative message passing,” in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3097–3106.

[68] H. Ye, L. Han, and D. Zhan, “Revisiting unsupervised meta-learning via the characteristics of few-shot tasks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3721–3737, 2023.

[69] Y. Wang, W. Chao, K. Q. Weinberger, and L. van der Maaten, “Simpleshot: Revisiting nearest-neighbor classification for few-shot learning,” *CoRR*, vol. abs/1911.04623, 2019. [Online]. Available: http://arxiv.org/abs/1911.04623

[70] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2921–2929.

Yuexuan An received the B.Sc. in computer science and technology from Jiangsu Normal University in 2015 and M.Sc. degree in computer application technology in China University of Mining and Technology in 2019. She is currently pursuing the Ph.D. degree in School of computer science and engineering, Southeast University. Her research interest includes machine learning and pattern recognition.

Hui Xue (Member, IEEE) is currently a professor of School of Computer Science and Engineering at Southeast University, China. She received the B.Sc. degree in Mathematics from Nanjing Norm University in 2002. In 2005, she received the M.Sc. degree in Mathematics from Nanjing University of Aeronautics & Astronautics (NUAA). And she also received the Ph.D. degree in Computer Application Technology at NUAA in 2008. Her research interests include pattern recognition and machine learning.

Xingyu Zhao received the B.Sc. and M.Sc. degrees in School of Computer Science and Technology, China University of Mining and Technology in 2016 and 2019, respectively. He is currently pursuing the Ph.D. degree in School of Computer Science and Engineering, Southeast University. His research interests mainly include pattern recognition and machine learning.

Ning Xu received the B.Sc. and M.Sc. degrees from University of Science and Technology of China and Chinese Academy of Sciences, China, respectively, and the Ph.D. degree from Southeast University, China. He is now an assistant professor in the School of Computer Science and Engineering at Southeast University, China. His research interests mainly include pattern recognition and machine learning.

Pengfei Fang is an Associate Professor at the School of Computer Science and Engineering, Southeast University (SEU), China. Before joining SEU, he was a post-doctoral fellow at Monash University in 2022. He received the Ph.D. degree from the Australian National University and DATA61-CSIRO in 2022, and the M.E. degree from the Australian National University in 2017. His research interests include computer vision and machine learning.

Xin Geng (Senior Member, IEEE) is a chair professor of Southeast University, China. His research interests include machine learning, pattern recognition, and computer vision. He has published over 100 refereed papers in these areas. He has been an Associate Editor of IEEE T-MM, FCS and MFC, a Steering Committee Member of PRICAI, a Program Committee Chair for conferences such as PRICAI’18, VALSE’13, etc., an Area Chair for conferences such as IJCAI, CVPR, ACMMM, ICPR, and a Senior Program Committee Member for conferences such as IJCAI, AAAI, ECAI, etc. He is a Distinguished Fellow of IETI and a Senior Member of IEEE.