Expert Systems With Applications 283 (2025) 127764

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Multi-view aggregation and multi-relation alignment for few-shot fine-grained recognition

Jiale Chen [a], Feng Xu [a,b,c,*], Xin Lyu [a,c], Tao Zeng [a], Xin Li [a,c], Shangjing Chen [a]

[a] College of Computer Science and Software Engineering, Hohai University, Nanjing, 210024, China
[b] School of Computer Engineering, Jiangsu Ocean University, Lianyungang, 222005, China
[c] Key Laboratory of Water Big Data Technology of Ministry of Water Resources, Hohai University, Nanjing, 211106, China

ARTICLE INFO

Keywords:
Few-shot learning
Fine-grained recognition
Background diversity
Semantic alignment

ABSTRACT

Few-shot fine-grained recognition (FS-FGR) aims to recognize nuanced categories with a limited number of labeled samples that were not encountered during training. Previous work has made significant progress by enhancing the learning of foreground refined regions and the alignment of consistent semantics. However, the detrimental impact of insufficient background diversity on constructing representative category prototypes has been overlooked. Meanwhile, the alignment of semantically consistent features has been hampered by the reliance on singular metrics, resulting in suboptimal feature extraction. To address the limitations above, a novel framework with multi-view aggregation and multi-relation alignment (MVRA) is proposed. In this framework, we strive to refine category prototypes by generating and consolidating multiple views from limited learnable samples. Specifically, we generate foreground-refined views, pinpointing discriminative regions, and background-obfuscated views, broadening the landscape of background diversity. Further, without relying on the entire prior, a global label assignment module is designed to automatically assign reliable labels to the query set samples. Finally, armed with these credible labels, the multi-relation alignment module harnesses the enriched views and their semantic congruencies, facilitating robust feature extraction. The effectiveness and outstanding performance of MVRA are evaluated through extensive experiments conducted on three fine-grained benchmark datasets.

1. Introduction

Fine-grained recognition (FGR) endeavors to distinguish subclasses of samples belonging to the same superclass, such as western gulls and ring-billed gulls, and different car brands. This task poses a significant challenge due to fine-grained categories having large intra-class variance and small inter-class variance (Chang et al., 2021; Chang, Tong et al., 2023; Han et al., 2022; Wang, Hou et al., 2023; Wei et al., 2022). Accurately distinguishing between specific categories remains challenging, even for domain experts with prior knowledge. Fortunately, with the rapid progress of deep learning, deep convolutional models have attained comparable performance when trained on a sufficient number of manually labeled samples (Han et al., 2022; Ma et al., 2024). Nevertheless, these methods heavily rely on adequate sample collection and costly manual labeling, which are prohibitively expensive to curate in domains like biology (Chang, Pang et al., 2023; Li, Chen et al., 2024; Roy et al., 2023) and medicine (Park & Ryu, 2024; Wen et al., 2023). Additionally, with new categories, the samples and labels must be reintegrated for retraining, which is not conducive to rapid response to specific task requirements (Du et al., 2023; Fang et al., 2023). The limitations mentioned above have profoundly hindered the progress of fine-grained recognition applications, prompting a shift in focus towards the more arduous task of few-shot fine-grained recognition (FS-FGR). This task circumvents the reliance on extensive samples and annotations, yet it confronts dual challenges. Firstly, FS-FGR inherits the inherent challenge of large intra-class variance and small inter-class variance in traditional fine-grained recognition. Secondly, FS-FGR is limited by the significantly restricted number of learnable samples and labels under few-shot conditions.

In the context of fine-grained recognition with few samples, capturing discriminative features of foreground objects, such as a bird's beak and feathers or a car's logo, has emerged as a widely recognized approach to address the dual challenges posed by fine-grained categories and limited samples. Numerous studies have pursued this objective, employing attention mechanisms to extract foreground objects for prediction (Liu et al., 2024; Wang et al., 2021; Yu et al., 2022), enhancing

* Corresponding author.
E-mail addresses: jialechen@hhu.edu.cn (J. Chen), xufeng@hhu.edu.cn (F. Xu), lvxin@hhu.edu.cn (X. Lyu), tzeng.nj@hhu.edu.cn (T. Zeng), li-xin@hhu.edu.cn (X. Li), hhu_csj@hhu.edu.cn (S. Chen).

https://doi.org/10.1016/j.eswa.2025.127764
Received 23 November 2024; Received in revised form 11 April 2025; Accepted 14 April 2025
Available online 25 April 2025

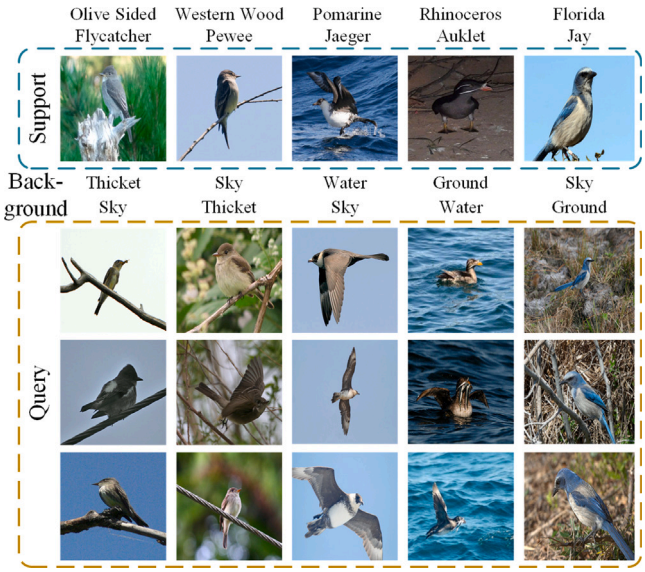

Fig. 1. An example of sampling from the CUB-200-2011 dataset with a 5-way 1-shot and 3-query episode setup. The background corresponding to the objects in the support and query sets are labeled separately, revealing a notable mismatch.

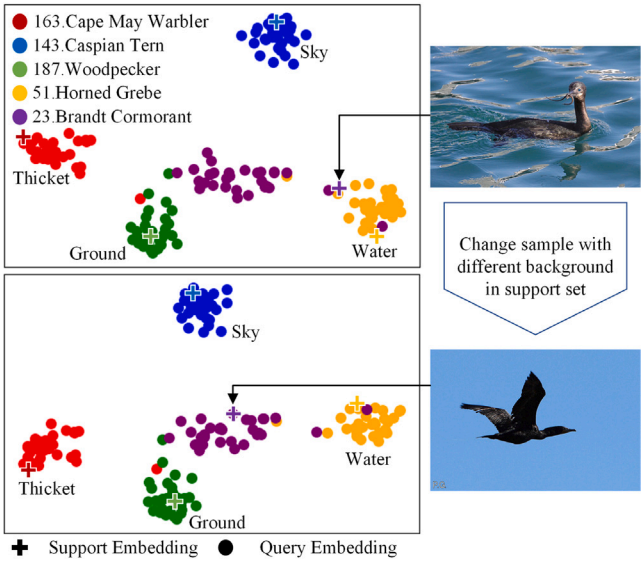

Fig. 2. Visualization of t-SNE for all samples under the 5-way 1-shot 30-query test episode in the CUB-200-2011. Four categories of samples appeared primarily in the same background: thicket, sky, ground, and water. The Brandt Cormorant appeared in multiple backgrounds.

the model's ability to learn salient features (Hou et al., 2019; Huang et al., 2022; Miao et al., 2024; Wang, Ji et al., 2024; Wu et al., 2021), or a combination of both (Cao et al., 2022; Zha et al., 2023). Additionally, some methods preprocess the dataset using manually labeled bounding boxes or additional foreground segmentation, which has been shown to significantly improve accuracy compared to using the original image directly (Li, Song et al., 2023; Liu et al., 2024; Zha et al., 2023). While considerable research has focused on enhancing the learning of foreground discriminative regions, the confounding effects of hard-to-erase backgrounds have often been overlooked. This is particularly problematic in the 1-shot scenario, where the lack of background diversity in trainable samples can adversely impact the construction of category prototypes and similarity measures.

Fig. 1 exemplifies the sampling process in the well-established 5-way 1-shot episode training paradigm for FS-FGR. In this paradigm, the model aims to predict query set samples based on support set training. It is apparent that birds belonging to the same category can be observed in diverse backgrounds, including the sky, water, and ground, demonstrating significant visual variations. However, the limited number of samples in the support set poses challenges in capturing these diverse backgrounds. Consequently, the stark differences across backgrounds, coupled with the subtle distinctions between fine-grained categories, increase the likelihood of erroneous predictions.

To further illustrate this point, we visualize the t-SNE (Maaten & Hinton, 2008) representation of samples in manually constructed test episodes, employing the advanced BSFA (Zha et al., 2023) method, as shown in Fig. 2. Notably, samples from the support and query sets sharing similar backgrounds exhibit a pronounced clustering tendency. Conversely, the Brandt Cormorant, which possesses a variety of backgrounds, displays a distinct dispersion. Moreover, when support set samples with disparate backgrounds are presented in the same water background, they yield distinct category centers and cluster closely with the Horned Grebe. This spurious correlation between foreground objects and background features leads to the learning of non-causal features during training, ultimately hindering performance.

Apart from accentuating the foreground and mitigating the influence of the background, the semantic alignment of discriminative features has garnered considerable attention in recent years. Notably in metric-based few-shot learning approaches (Snell et al., 2017), numerous studies have focused on maximizing the alignment between the embedded features of the query set and the support set (Hou et al., 2019; Snell et al., 2017). Other methods have aimed to enhance the metric by considering factors such as the spatial location of features (Huang et al., 2022; Yu et al., 2022; Zha et al., 2023). However, these approaches have been limited by their reliance on a single semantic consistency relation, overlooking the potential of exploiting diverse relations between samples.

Drawing upon the preceding considerations, we propose a novel framework for few-shot fine-grained recognition, named MVRA, which incorporates multi-view aggregation and multi-relation alignment. Within MVRA, the multi-view aggregation encompasses raw images, refined images obtained through the foreground object extraction module, and mixed samples generated by the background obfuscation module. By integrating data with diverse views, multi-view aggregation not only augments the limited sample pool but also addresses the scarcity of background diversity in 1-shot scenarios. Subsequently, we propose the multi-relation alignment module (MRA), which establishes and exploits multiple semantic consistency relations among available samples based on the results of global label assignment across episodes. MRA aggregates image embeddings and global predictive distributions from multi-view samples to construct supervised comparisons of embeddings and bi-directional metrics of distributions, fostering the alignment of consistent semantics.

Our contributions can be summarized as follows:

- We propose a multi-view aggregation and multi-relation alignment framework (MVRA) for few-shot fine-grained recognition, which pioneers the consideration of the potential impact of background diversity.
- Without additional annotations, we generate and aggregate multiple views (raw, foreground refined, background obfuscated) to expand learnable samples and enrich background diversity, thereby facilitating the creation of more representative category prototypes.
- Employing solely the minimal global label prior, our automatic global label assignment and multi-relation alignment modules boost semantic alignment by constructing numerous consistency relations among samples across episodes.
- Evaluations on three fine-grained benchmark datasets and two popular backbones demonstrate the effectiveness and outstanding performance of our proposed method.

2. Related work

2.1. Few-shot learning

Few-shot learning (FSL) strives to achieve effective recognition of unseen samples during training by learning from only a limited number of labeled instances (Hospedales et al., 2022; Li, Yang et al., 2023). Unlike fully supervised deep learning approaches, FSL does not necessitate vast quantities of labeled samples. Crucially, FSL's exploration of empowering learning systems with the ability to learn new knowledge quickly is a necessary path for artificial intelligence to evolve into human intelligence.

Existing research in few-shot learning can be broadly categorized into four groups based on underlying technological approaches. The first category is the metric-based approach, which refines category prototype embeddings by labeled samples and embeds query set samples in the same space for category prediction and optimization based on distance (Hao et al., 2022; Jiang et al., 2021; Snell et al., 2017; Sung et al., 2018; Tang et al., 2020; Zha et al., 2023). The second category comprises optimization-based techniques, which directly generate or update the weight parameters of convolutional neural networks (Chi et al., 2022; Li, Tang et al., 2023; Rusu et al., 2019). This is achieved by a small number of support samples tailored for specific learning tasks, thereby adhering closely to the principles of learning to learn. Thirdly, augmentation-based approaches (Li et al., 2020; Xu, Liu et al., 2023; Xu et al., 2024) aim to address the scarcity of labeled samples in FSL by exploring the potential of generating additional samples that can augment and enhance the learning of model. Finally, the reconstruction-based approach uses the available representations of each category to find the best weights to reconstruct the samples and thus judge the categories (Li, Song et al., 2023; Sun et al., 2023; Wertheimer et al., 2021; Wu et al., 2023).

These efforts have yielded encouraging results in general image classification tasks while reducing the reliance on extensive samples and annotations, which is becoming a trend for numerous applications. As a result, limited shots for tricky fine-grained recognition are also receiving increasing attention and exploration. And this paper will employ a metric-based few-shot learning approach to tackle the fine-grained recognition challenge.

2.2. Fine-grained recognition

Fine-grained recognition (FGR) is a challenging task that involves distinguishing various subcategories within a shared visual superclass (Chang et al., 2021; Chang, Tong et al., 2023; Han et al., 2022; Wang, Hou et al., 2023; Wei et al., 2022). Unlike general image classification, FGR focuses on specific domains with numerous fine-grained subclasses, such as different bird species and car brands. The task is particularly difficult due to significant intra-class variations and subtle inter-class differences, to the extent that even humans require extensive training for accurate recognition. However, with the rapid advancement of deep learning, FGR has achieved human-level performance when provided with sufficient samples and annotations (Wei et al., 2022; Zhu et al., 2022).

Traditional approaches to FGR often revolve around keypoint localization, which identifies key regions for object recognition before making predictions (Han et al., 2022; Wang, Wang et al., 2023). For example, CAPR (Wang, Chang et al., 2024) enhances the discriminative power of salient regions while spreading attention to neighboring non-salient regions. In contrast, (Du et al., 2020) replace explicit part operations, encourage the network to learn at different granularities, and gradually integrate multi-granularity features. There are also methods that focus on optimizing coding and loss processes to directly model the discriminative differences between fine-grained categories (Yu et al., 2021). For instance, Chang et al. (2020) introduces a method that constrains all feature channels belonging to the same class to be both discriminative and diverse, focusing on extracting discriminative features from different parts of the object. This approach has shown promise in enhancing the model's ability to distinguish subtle differences between categories. Furthermore, the integration of vision-language modeling has opened new avenues for improving fine-grained recognition. By leveraging cross-modal descriptions, these methods exploit the rich semantic information available in language models to enhance visual recognition (Jiang et al., 2023). Additionally, CSC-Net (Li, Li et al., 2024) addresses modal differences by utilizing intra- and inter-modal multi-granularity correlations, guiding the consistent expression of semantics across different modalities.

Building on these advances, our approach emphasizes foreground discriminative regions while actively complementing background diversity to suppress background interference. In addition, broader inter- and intra-class correlations are explored to enhance the modeling of fine-grained class-discriminative features.

2.3. Few-shot fine-grained recognition

Few-shot fine-grained recognition (FS-FGR) addresses the practical challenge of recognizing novel categories with limited labeled samples. This task is critical in applications like biology and medicine, where obtaining sufficient annotated samples for rare subclasses is both costly and difficult (Li, Xu et al., 2024; Li et al., 2025; Park & Ryu, 2024; Roy et al., 2023; Wen et al., 2023). Traditional supervised learning methods struggle in such scenarios due to the high cost of expert annotation and the dynamic nature of fine-grained recognition tasks, where existing samples and labels often fail to meet evolving application demands (Fang et al., 2023).

The concept of few-shot fine-grained recognition was first introduced in Wei et al. (2019) to achieve accurate fine-grained recognition under the constraints of few-shot learning. Tackling the dual challenges of limited annotated samples and the significant intra-class variations with slight inter-class variations in fine-grained recognition has been the focus of numerous studies. Early work by CBAM (Zhu et al., 2021) employed two convolutional block-attention modules to improve feature discrimination. FOT (Wang et al., 2021) introduced foreground object transformation to capture and transform foreground regions, compensating for the lack of samples in few-shot conditions. Similarly, Liu et al. (2024) utilized a saliency detection model to exclude background regions and emphasize object-specific information. Subsequent research explored weakly supervised learning to extract and refine object features. Methods like BSFA (Zha et al., 2023) performed similarity measures on support and query sets for both raw and refined images. Considering the diversity of object scales, methods like those proposed in Tang et al. (2022) and Xu, Zhang et al. (2023) progressively extracted and enhanced foreground discriminative regions from features of different scales.

Beyond enhancing foreground features, aligning semantics and understanding inter-semantic relationships have become key focuses (Hao et al., 2022; Tang et al., 2023). Works like Huang et al. (2022) and Zha et al. (2023) consistently account for feature variations across spatial locations, embedding features of support and query set samples in a more nuanced alignment. Qi et al. (2024) reconstructed query sample features to emphasize discriminative aspects while maintaining local consistency. Wang et al. (2020) mined correlations between multiple discriminative regions to build a more powerful feature set, while Xu et al. (2022) developed a dual attention network to model part relations and capture subtle fine-grained details. Li et al. (2021) introduced a bi-similarity metric module to mitigate similarity bias in feature learning, moving beyond reliance on a single metric or feature domain. Li, Yao et al. (2023) and Wang, Zhao et al. (2023) proposed integrating multiple feature spaces, including frequency domains, to acquire diverse and discriminative features. Bi-FRN (Wu et al., 2023) addressed the limitations of unidirectional reconstruction by using query sets to reconstruct the support set, accommodating intra-class diversity.

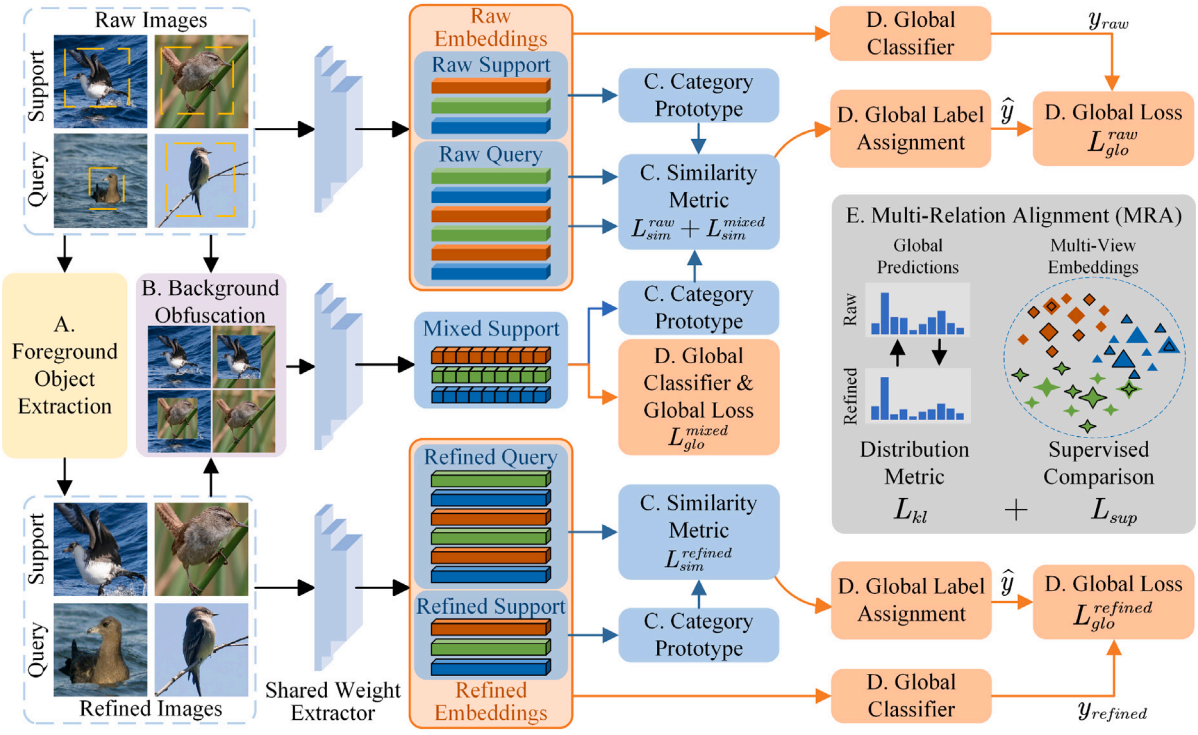

Fig. 3. The proposed FS-FGR framework begins by refining raw images through foreground extraction and background obfuscation, creating views with refined foregrounds and diverse backgrounds. These views are then encoded using a shared extractor, and similarity metrics predict for the query set. Subsequently, additional classifiers leverage the global labels assigned by the similarity prediction to compute the global loss. Finally, embeddings and distributions are aggregated, reinforcing semantic alignment across consistency relations.

Together, these approaches advance the development of few-shot fine-grained recognition, each proposing unique strategies to improve feature recognition and model performance under sample-constrained conditions. Our approach continues previous work by adding consideration of neglected background diversity and utilizing a limited number of labeling adaptations to construct more diverse semantic consistency relations that collectively contribute to the construction of discriminative prototypes.

3. Methodology

In this section, we propose a unified deep learning framework termed MVRA, comprising four principal components. Firstly, we present foreground object extraction and background obfuscation, which expands the limited sample set while focusing on the foreground identification region. Secondly, a multi-stream similarity metric is employed to extract category prototypes independently and predict query set samples. Thirdly, the global classification branch, while inspired by Zha et al. (2023), differs in that our global category labels are derived from the results of the similarity measure branch, augmented with confidence weighting. Lastly, the multi-relation alignment module (MRA) integrates multi-view sample embeddings to facilitate the convergence of consistent semantics and divergence of different semantics. Fig. 3 depicts the overall framework, and we will elaborate on the workflow of each component in the following sections.

For the problem formulation, we follow the general episode training strategy of few-shot learning (Huang et al., 2022; Zha et al., 2023), all datasets are first divided into training and test sets with no category overlap, denoted as $\langle I_{train}, C_{train}\rangle$ and $\langle I_{test}, C_{test}\rangle$, where I represents the sample image, C represents the category label and $C_{train} \cap C_{test} = \phi$. Then, the episode learning tasks T are constructed separately with the same rules, and each task contains two parts, the support set (S) and the query set (Q). Each episode learning task aims to train a robust network that accurately identifies the category of each sample in the query set by learning only from the support set. All tasks use the popular N-way K-shot setup, i.e., the samples in the support and query sets in the same task come from the same N categories, with only K samples per category in the support set and q samples per category in the query set. Thus, we can further represent the previous S and Q as $S = \left\{\left(x_i, c_i, \hat{y}_i\right)\right\}_{i=1}^{N\times K}$ and $Q = \left\{\left(x_j, c_i\right)\right\}_{j=1}^{N\times q}$, where x is the image input, c is its category label in that episode task, and $\hat{y}$ represents the global label in the full dataset. In this paper, all experiments were conducted in the most dominant 5-way 1-shot and 5-way 5-shot conditions, with q taking the value of 15.

3.1. Foreground object extraction

Inspired by human visual recognition, when confronted with a object that is difficult to recognize at a glance, the eyes are first guided to focus on the target area, and then judgments are made in conjunction with local details (Navon, 1977). Similarly, the detailed features of the foreground target are essential for the recognition of fine-grained images (Chang et al., 2020; Han et al., 2022; Liu et al., 2024). Following BSFA (Zha et al., 2023), class-independent activation maps are obtained by summing the features of raw images in the channel dimension. However, in contrast to taking the most extensive connectivity map as the refined region, we pool the activation maps in row and column dimensions separately and take the most edge position exceeding the threshold τ as the boundary coordinates of the foreground object bounding box. With such a rectangular bounding box expansion strategy, all potentially discriminative regions are maximally preserved under weakly supervised attention.

As shown in Fig. 4, for any input image x, the activation map A is obtained through the encoder $f(\theta)$ by $A = \sum_{i=1}^{d} f(x, \theta)$, where $A \epsilon R^{h\times w}$. In order to more accurately localize the foreground salient regions on the original image, we up-sample the activation map to the original size using bilinear interpolation for a new activation map $A \epsilon R^{H\times W}$. Next, the row and column dimensions of A are pooled and normalized,

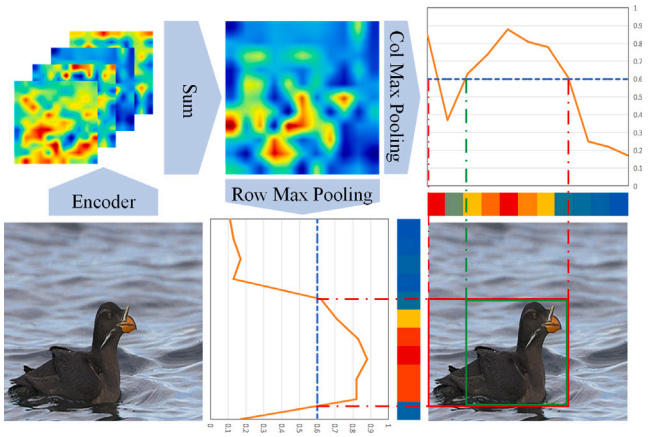

Fig. 4. Foreground object extraction module. The features encoded in the raw image are summed over the channel dimension to obtain a category-independent activation map, and then foreground refinement objects are obtained by locating the most salient regions by pooling over rows and columns. More potential regions are found compared to selecting the maximally connected region (green).

respectively, to obtain bar activation maps A_x and A_y that match the length and width of the original image as shown in Eq. (1).

$$\begin{aligned} A_x &= \text{Normalize}\left(\text{MaxPool}\left(A,(H,1)\right)\right) \\ A_y &= \text{Normalize}\left(\text{MaxPool}\left(A,(1,W)\right)\right) \end{aligned} \tag{1}$$

Then regions larger than τ are selected on the bar activation map respectively, where τ is a manually set hyperparameter, and the extreme position of the selected region is the edge of the final localized bounding box as shown in Eq. (2).

$$Bbox = \begin{cases} x_{left} = argmin\left(index\left(A_x > \tau\right)\right) \\ y_{top} = argmin\left(index\left(A_y > \tau\right)\right) \\ x_{right} = argmax\left(index\left(A_x > \tau\right)\right) \\ y_{bottom} = argmax\left(index\left(A_y > \tau\right)\right), \end{cases} \tag{2}$$

where $index$ aims at obtaining the position of the elements on the bar activation map that exceed the threshold value, $argmin$ and $argmax$, are used to locate the most bounded position of the salient region quickly. And the final obtained foreground refinement region is cropped and up-sampled on the original image as a refined image.

In fine-grained recognition, the activation maps generated by the encoder often exhibit multiple discrete vital points rather than focusing solely on a single region, such as the beak and tail of a bird (Chang et al., 2020; Han et al., 2022). In this regard, our foreground object extraction module can effectively assist in generating the candidate box containing all potential keypoints. On the other hand, we were surprised to find that different τ values yield distinct outcomes in various few-shot training setups. For further experiments and analysis, please refer to section of sensitivity analysis.

3.2. Background obfuscation

To mitigate the adverse consequences of limited background diversity, we introduce a background obfuscation module specifically designed for 1-shot scenarios, as depicted in Fig. 5. This module exclusively operates on the support set, with the objective of augmenting the background diversity of the learnable samples to facilitate the creation of more representative category prototypes.

Specifically, we leverage foreground object regions extracted from the foreground object extraction module to embed each foreground refinement view into the foreground regions of other support set samples within the same episode. Labels are then assigned to these newly generated background-obfuscated views based on the categories of

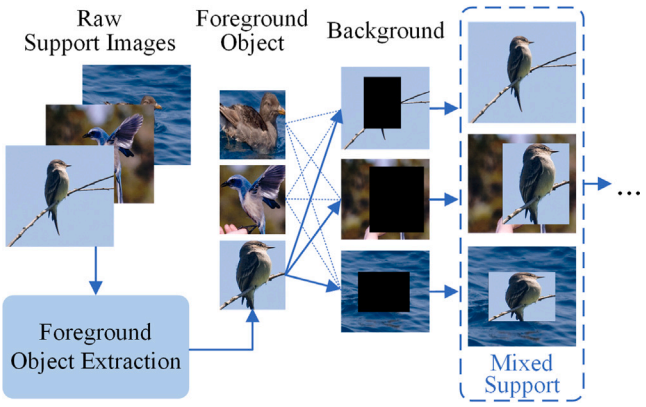

Fig. 5. Background obfuscation module. The refined views obtained from the support set samples after foreground object extraction are scaled to incorporate the foreground regions of the remaining samples to achieve diversity obfuscation of the background.

the embedded foreground objects, and prototypes are constructed by computing mean values. The detailed process for generating these obfuscated views is outlined in Algorithm 1.

This strategy ensures that foreground objects in each learnable sample are consistently represented across diverse backgrounds, thereby preventing salient background features from being misclassified as discriminative category features due to sample sparsity. Unlike enhancement methods such as MixUp (Zhang et al., 2018), which may cause detail blurring and semantic ambiguity due to indiscriminate sample mixing, our approach enriches background diversity while preserving the discriminative representations of foregrounds. Additionally, the category labels consistent with the foregrounds ensure clear semantic optimization goals, especially in metric-based few-shot learning.

Algorithm 1 Generating Mixed Support Set

$S_{raw} \leftarrow$ LoadFromRawSupportSet()
$S_{mixed} \leftarrow \emptyset$
$Bboxes \leftarrow$ ForegroundObjectExtraction(S_{raw})
for x_i in S_{raw} **do**
  $Bbox_i \leftarrow Bboxes[x_i]$
  $x_{refined} \leftarrow$ Crop(x_i, $Bbox_i$)
  for x_j in S_{raw} **do**
    $Bbox_j \leftarrow Bboxes[x_j]$
    $x_{mixed} \leftarrow$ Replace(x_j, $Bbox_j$, $x_{refined}$) ▷ Replace the region of $Bbox_j$ in x_j with $x_{refined}$
    $S_{mixed} \leftarrow S_{mixed} \cup \{x_{mixed}\}$
  end for
end for
return S_{mixed}

3.3. Two- or three-stream similarity metric

In the metric-based approach to FS-FGR, an encoder maps both support and query set samples into a shared embedding space. Within this space, samples from the same category in the support set are aggregated to form a representative prototype for that class. Subsequently, each query sample is encouraged to converge towards the prototype of the category to which it belongs. Ultimately, the prediction for a given query sample is determined by identifying the category prototype that exhibits the shortest distance metric concerning that sample.

Distinguishing our framework from the general single data stream approach to metric learning (Huang et al., 2022; Snell et al., 2017), we incorporate multiple data stream inputs, including raw, foreground refined, and background obfuscated under 1-shot. For consistent semantic matching metrics, all these data streams are encoded by the

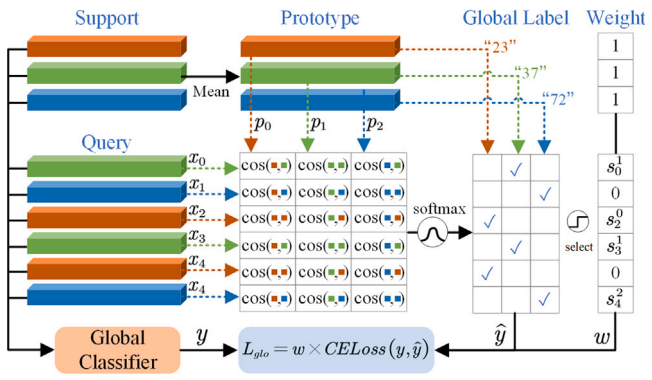

Fig. 6. Similarity metric and global label assignment. Prototypes are first constructed for the support set by category for the similarity metric of the query set. Then global labels and corresponding weights are assigned based on the global labels of the support set and the similarity discrimination results of the query set.

same encoder. The extraction of category prototypes and query set similarity metrics are then accomplished in the respective data streams. Specifically, as illustrated in Fig. 6, the mean embedding vector of samples belonging to the same category in the support set, denoted as S_c, is employed as the embedding prototype for each category. The prototype for category c is computed using the following equation:

$$p_c = \frac{1}{|S_c|} \sum_{x_i \epsilon S_c} AveragePool\left(f\left(x_i, \theta\right)\right). \tag{3}$$

Consequently, the query set vectors are compared with the prototype of each category using the cosine similarity measure. The similarity between the query sample x_j and the prototype of category c is given by:

$$sim_j^c = \cos\left(AveragePool\left(f\left(x_j, \theta\right)\right), p_c\right). \tag{4}$$

Each query set sample undergoes a similarity calculation with all the prototypes in the episode to obtain a similarity score for each category, denoted as $sim_j = \left[sim_j^c \,|c = 1, 2, \ldots, N\right]$.

Exceptionally, we construct category prototypes similarly for the additional background obfuscated data stream in the 1-shot setting, which is the enhancement of background diversity in raw images. However, since our background obfuscation module does not apply to query set samples, the query set embedding of the raw image is still used to perform sample-by-sample similarity measurements with the new category prototypes.

In each episode task, our primary objective is to accurately predict the category for each sample in the query set. To accomplish this, we aggregate the outcomes from all data branches and designate the category with the minimum average distance as the final prediction. Additionally, for each data stream, we compute the cross-entropy loss independently between the prediction results of the query set samples and their corresponding category labels for that specific task. The loss function for the similarity measure of each branch is formulated as follows:

$$L_{sim} = \sum_{j=1}^{N \times q} CrossEntropyLoss\left(sim_j, c_j\right). \tag{5}$$

where $N \times q$ represents the total number of samples in the query set. The final metric loss is expressed in Eq. (6):

$$L_{sim}^{final} = L_{sim}^{raw} + L_{sim}^{refined}\left(+L_{sim}^{mixed}\right) \tag{6}$$

Here, L_{sim}^{raw}, $L_{sim}^{refined}$, and L_{sim}^{mixed} represent the loss contributions of the similarity measurements obtained from the raw, refined, and background obfuscated image streams, respectively.

3.4. Global classification branch

In the episode training paradigm, each category is assigned a local label specific to that episode. However, extensive research has emphasized the importance of consistent global labels across episodes for enhancing performance in few-shot learning (Liu et al., 2021; Zha et al., 2023). This is primarily attributed to the ability of additional global classification branches to facilitate the modeling of discriminative features across different categories (Wang, Falk et al., 2024; Wang, Pontil et al., 2024). Therefore, our framework retains the global classification branch as shown in Fig. 3, where the global classifier predicts labels for all data stream samples. But different from previous works (Liu et al., 2021; Wang, Falk et al., 2024; Wang, Pontil et al., 2024; Xu, Liu et al., 2023; Zha et al., 2023), the global label assignment module is designed to automatically assign credible global labels and corresponding weights to query set samples based on similarity metrics.

As illustrated in Fig. 6, the similarity score sim_j of each sample in the query set undergoes a softmax operation to identify the most similar category prototypes. The corresponding score, denoted as s_j, is calculated using the equation $s_j = \text{argmax}\left(\text{softmax}\left(sim_j\right)\right)$. Subsequently, the global label of the identified prototype $\hat{y}$ is assigned to the corresponding query set sample.

Furthermore, we introduce soft weights for all samples to mitigate the potential overfitting to categories during training. Specifically, as defined in Eq. (7), if the similarity score s_j exceeds a predefined threshold κ, it is assigned as the weight for the corresponding sample. Conversely, zero weight is applied to mitigate the potential impact of erroneous similarities on the global classifier's predictions, which is particularly crucial during the initial stages of training. In this paper, we set κ to 0.5 for all experiments.

$$w_j = \begin{cases} s_j^c & s_j^c \geq \kappa \\ 0 & s_j^c < \kappa \end{cases} \tag{7}$$

In particular, due to the sparse number of support set samples and the fact that their global labels are accessible, their weights are uniformly assigned to 1.

After acquiring the global labels and weights for the entire query set, all samples are collectively fed into the global classifier, denoted as $f_{cls}()$, which is implemented using a single layer of 1×1 convolution, and no weights are shared between different branches. Subsequently, the prediction results corresponding to the global labels are derived using Eq. (8):

$$y_i = f_{cls}\left(AveragePool\left(f\left(x_i, \theta\right)\right)\right). \tag{8}$$

The cross-entropy loss is calculated by incorporating the global label of the support set with the global label and sample weight w obtained from the query set following the global label assignment module, as shown in Eq. (9).

$$L_{glo} = \sum_{x_i \in S \cup Q} w_i \times CrossEntropyLoss\left(y_i, \hat{y}_i\right) \tag{9}$$

Consequently, the overall global classification loss of our framework can be formulated as:

$$L_{glo}^{final} = L_{glo}^{raw} + L_{glo}^{refined}\left(+L_{glo}^{mixed}\right). \tag{10}$$

Here, L_{glo}^{raw} and $L_{glo}^{refined}$ represent the global loss for the samples of the raw and the refined image stream, respectively, while L_{glo}^{mixed} calculates the global loss specifically for the samples of the third support set generated by the background obfuscation module in 1-shot setting.

3.5. Multi-relation alignment module

Although the previous similarity metric and global classification approaches have effectively enhanced the model's ability to extract discriminative features for fine-grained category differentiation, they

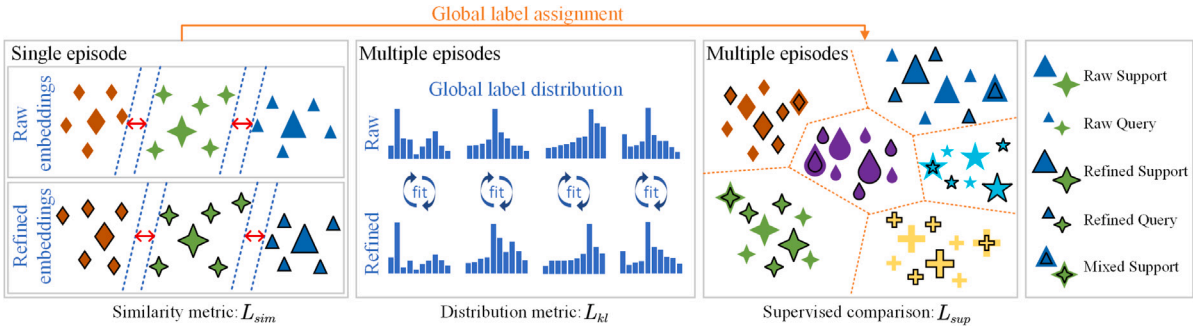

Fig. 7. Based on the intra-episode similarity metric, the one-to-one distribution metrics of raw and refined embeddings are further extended, and supervised contrasts between total samples of multiple episodes are implemented with global labels.

are not without limitations. Specifically, similarity metrics are inherently confined to single-episode tasks, whereas global categorization focuses solely on individual samples, overlooking potential semantic connections among multiple samples within a broader context (Ling et al., 2022). To overcome this limitation, we propose the multi-relation alignment module (MRA), which aims to establish potential connections among as many samples as possible. As depicted in Fig. 7, we incorporate one-to-one distribution metrics and many-to-many supervised comparisons beyond leveraging the correlation between the samples in the query set and their respective category prototypes.

Firstly, in the one-to-one distribution metric, we utilize the matching relationship between the raw and refined views and the global classifier's high-dimensional prediction for all sample embeddings. These predictions naturally constitute distinct distributional representations of the same semantic (Leng et al., 2024; Yang et al., 2022). To quantify the disparity between these distributions, we employ the Kullback–Leibler ($D_{kl}()$) divergence, which measures the difference in the global predicted distribution between paired raw and refined images, such as y_i^{raw} and $y_i^{refined}$. As illustrated in Eqs. (11) and (12), the two distributions guide each other in the computation of the Kullback–Leibler divergence loss, which is subsequently aggregated to form the final loss, denoted L_{kl}.

$$D_{kl}\left(y^{raw}|y^{refined}\right) = \sum_i y_i^{raw} \log\left(\frac{y_i^{raw}}{y_i^{refined}}\right) \tag{11}$$

$$L_{kl} = D_{kl}\left(y^{raw}|y^{refined}\right) + D_{kl}\left(y^{refined}|y^{raw}\right) \tag{12}$$

Secondly, as depicted in the rightmost section of Fig. 7, for the many-to-many supervised comparison, we make full use of the output from the global label assignment module. This allows us to compare all sample embeddings (raw, refined and background obfuscated views) within multiple episodes of the same batch. This approach eliminates the isolation of samples imposed by episode learning in few-shot scenarios and substantially increases the number of comparison pairs. Furthermore, it facilitates learning semantically consistent features across a broader scope. The specific formulation of this loss function is presented in Eq. (13).

$$L_{sup} = \sum_{i \in I}\left(\frac{-1}{|P(i)|}\sum_{p \in P(i)} \log \frac{\exp\left(z_i \cdot z_p\right)}{\sum_{a \in A(i)} \exp\left(z_i \cdot z_a\right)}\right) \tag{13}$$

In this equation, I represents the set of all samples in each batch, $A(i)$ comprises all samples excluding sample i, and $P(i)$ contains only those samples that belong to the same class as sample i. The variable z represents the normalized embedding vector of each sample, and $\cdot$ denotes the inner product operation between vectors. This supervised comparison loss function, as designed in Khosla et al. (2020), strives to enhance the clustering of intra-class samples while promoting the dispersion of inter-class samples.

In summary, our framework elegantly integrates several crucial components. Firstly, it incorporates general similarity metric constraints for both query set and support set samples. Secondly, it leverages these similarity metric results to generate global labels, perform global category prediction, and compute corresponding global predictive distribution metrics for both the raw and refined images. Lastly, it executes supervised comparisons of sample embeddings aggregated from multiple data streams. The overall loss function that dominates this framework is given in Eq. (14).

$$L = L_{sim} + L_{glo} + \lambda_1 L_{kl} + \lambda_2 L_{sup} \tag{14}$$

In our work, λ_1 and λ_2 are empirically fixed to 0.1 and 0.2, respectively.

4. Experiments

4.1. Implementation details

For a fair comparison, we adhered to the prevalent few-shot episode learning strategy, encompassing 5-way 1-shot and 5-way 5-shot support set sample allocation settings while maintaining a query set count of 15 for each category (Huang et al., 2022; Zha et al., 2023). All images underwent random flipping before being used as input, resizing to 96 × 96, and subsequent random cropping to 84 × 84. For the backbone, we employed ResNet-12 and Conv-64, which are commonplace in few-shot fine-grained recognition, albeit with the final pooling layer omitted to facilitate the extraction of foreground refinement regions.

During the training phase, we adopted classical stochastic gradient descent (SGD) with a momentum of 0.9 and a weight decay of $5e^{-4}$ as the optimizer. The learning rate was initially set to 0.1, following the learning rate update strategy of Zha et al. (2023), decreasing to 0.06 at the 60th epoch and multiplying by 0.2 every ten epochs. In the testing phase, we adhered to the same episode delineation strategy used during training and recorded the average accuracy within the 95% confidence intervals of 2000 episodes as the final result.

4.2. Datasets

We conducted experiments on three mainstream fine-grained classification datasets, including CUB-200-2011 (Wah et al., 2011), Stanford Dogs (Khosla et al., 2011), and Stanford Cars (Krause et al., 2013), which are widely accepted as few-shot fine-grained recognition validation. Furthermore, although bounding boxes that label the locations of foreground objects are present in these datasets, and some works rely directly on them to filter out background interference, our work is solely implemented in the context of image-level category labeling.

(1) CUB-200-2011 (Wah et al., 2011): contains 11,788 images from 200 subcategories of birds. Following the widely recognized paradigm, 100 categories were randomly selected for training, 50 for validation, and 50 for testing.

(2) Stanford Dogs (Khosla et al., 2011): contains 20,580 images from 120 subcategories of dogs. Following the widely recognized paradigm, 70 categories were randomly selected for training, 20 for validation, and 30 for testing.

(3) Stanford Cars (Krause et al., 2013): contains 16,185 images from 196 subcategories of cars. Following the widely recognized paradigm, 130 categories were randomly selected for training, 17 for validation, and 49 for testing.

Table 1
Average accuracy (%) and 95% confidence intervals for 2000 test episodes under the 5-way 1-shot and 5-way 5-shot setups on three fine-grained datasets with backbone based on **ResNet-12**. The best results are shown in bold. The remaining results are reproduced with the open-source code.

Method	CUB-200–2011		Stanford dogs		Stanford cars	
	5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
ProtoNet (Snell et al., 2017)	63.44 ± 0.56	83.17 ± 0.35	41.61 ± 0.50	76.78 ± 0.36	45.01 ± 0.49	87.19 ± 0.31
TOAN[a] (Huang et al., 2022)	66.10 ± 0.86	82.27 ± 0.60	49.77 ± 0.86	69.29 ± 0.70	75.28 ± 0.73	87.45 ± 0.48
RelationNet (Sung et al., 2018)	70.92 ± 0.54	84.90 ± 0.35	61.21 ± 0.51	80.27 ± 0.37	78.04 ± 0.53	90.03 ± 0.30
BSNet(R&C)[a] (Li et al., 2021)	73.48 ± 0.92	83.84 ± 0.59	61.95 ± 0.97	79.62 ± 0.63	71.07 ± 1.03	88.38 ± 0.62
CAN (Hou et al., 2019)	76.98 ± 0.48	87.77 ± 0.30	64.73 ± 0.52	77.93 ± 0.35	86.90 ± 0.42	93.93 ± 0.22
OLSA[a] (Wu et al., 2021)	77.77 ± 0.44	89.87 ± 0.24	64.15 ± 0.49	78.28 ± 0.32	77.03 ± 0.46	88.85 ± 0.46
AGPF[a] (Tang et al., 2022)	78.73 ± 0.84	89.77 ± 0.47	72.34 ± 0.86	84.02 ± 0.57	85.34 ± 0.74	94.79 ± 0.35
MPHAM[a] (Li, Yao et al., 2023)	79.48 ± 0.47	85.25 ± 0.25	–	–	–	–
BSFA[a] (Zha et al., 2023)	82.27 ± 0.46	90.76 ± 0.26	69.58 ± 0.50	82.59 ± 0.33	**88.93 ± 0.38**	**95.20 ± 0.20**
RSaD[a] (Liu et al., 2024)	82.45 ± 0.79	92.02 ± 0.44	73.75 ± 0.93	**86.65 ± 0.54**	87.27 ± 0.70	95.01 ± 0.49
Ours	**83.74 ± 0.42**	**92.20 ± 0.23**	**74.82 ± 0.47**	85.54 ± 0.31	87.56 ± 0.34	94.92 ± 0.17

[a] Indicates that results are obtained from the original paper.

Table 2
Average accuracy (%) and 95% confidence intervals for 2000 test episodes under the 5-way 1-shot and 5-way 5-shot setups on three fine-grained datasets with backbone based on **Conv-64**. All of results are obtained from the original paper and the best results are shown in bold.

Method	CUB-200–2011		Stanford dogs		Stanford cars	
	5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
MPHAM (Li, Yao et al., 2023)	62.15 ± 0.44	68.05 ± 0.33	–	–	–	–
DLG (Cao et al., 2022)	64.77 ± 0.90	83.31 ± 0.55	47.77 ± 0.86	67.07 ± 0.72	62.56 ± 0.82	88.98 ± 0.47
TOAN (Huang et al., 2022)	65.34 ± 0.75	80.43 ± 0.60	49.30 ± 0.77	67.16 ± 0.49	65.90 ± 0.72	84.24 ± 0.48
BSFA (Zha et al., 2023)	65.48 ± 0.51	76.01 ± 0.41	–	–	–	–
BSNET(R&C) (Li et al., 2021)	65.89 ± 1.00	80.99 ± 0.63	51.06 ± 0.94	68.60 ± 0.73	54.12 ± 0.96	73.47 ± 0.75
MattML (Zhu et al., 2021)	66.29 ± 0.56	80.34 ± 0.30	54.84 ± 0.53	71.34 ± 0.38	66.11 ± 0.54	82.80 ± 0.28
FOT (Wang et al., 2021)	67.46 ± 0.68	83.19 ± 0.43	49.32 ± 0.74	68.18 ± 0.69	54.55 ± 0.73	73.69 ± 0.65
LSANET (Yu et al., 2022)	67.75	82.76	55.85	71.78	68.65	87.23
BAMM (Wang, Ji et al., 2024)	68.55 ± 1.12	84.77 ± 0.79	55.76 ± 1.11	72.59 ± 0.88	63.38 ± 1.09	84.42 ± 0.84
OLSA (Wu et al., 2021)	73.07 ± 0.46	86.24 ± 0.29	55.53 ± 0.45	71.68 ± 0.36	70.13 ± 0.48	84.29 ± 0.31
AttNet (Xu et al., 2022)	72.89 ± 0.50	86.60 ± 0.31	**59.81 ± 0.50**	**77.19 ± 0.35**	70.21 ± 0.50	85.55 ± 0.31
Ours	**73.67 ± 0.44**	**88.87 ± 0.29**	58.12 ± 0.51	76.14 ± 0.38	**70.44 ± 0.46**	**88.17 ± 0.27**

4.3. Results

(1) Comparison with methods based ResNet-12.

Table 1 presents a comprehensive comparison of the accuracy achieved by our method in comparison to ten representative few-shot learning approaches, utilizing ResNet-12 as the backbone. Notably, our method demonstrates superior performance on the CUB-200-2011 and consistently lower confidence intervals indicate its robustness and stability. Additionally, as evidenced in Table 1, our method surpasses BSFA (Zha et al., 2023), another approach with multiple views and global classification branches, by over one point on CUB-200-2011 and by more than two points on Stanford Dogs, with a marginal deficit only on Stanford Cars. On the Stanford Cars, our approach lags slightly behind, primarily owing to the ability of BSFA to precisely align objects based on their scale and spatial location. This capability is particularly advantageous in managing car designs' intricate diversity and complexity. Furthermore, when incorporating the enhanced performance achieved by replacing the weaker backbone, as shown in Table 2, BSFA's performance degrades substantially across all tested datasets, whereas our method remains consistently competitive, thereby highlighting the effectiveness of our improvement in constructing robust prototypes. On the other hand, our method underperforms compared to RSaD (Liu et al., 2024) on the 5-way 5-shot of Standford Dogs and Standford Cars. This is due to the fact that RSaD introduces additional pre-trained salient target detection templates to completely exclude background interference, which greatly simplifies the prediction difficulty of the model. Comparatively, our superior performance on 1-shot also side-steps the necessity of considering background diversity for building robust prototypes, which has been previously overlooked. And how to improve the performance of our method in the 5-shot test will also be the goal of our subsequent ongoing work.

(2) Comparison with methods based Conv-64. We replaced the backbone network with Conv-64, a shallower and simpler architecture, and compared its performance against existing methods using the same backbone for FS-FGR. The results, presented in Table 2, align with those in Table 1, demonstrating that our model maintains consistent competitiveness and robustness across different backbone architectures. Specifically, our approach outperformed the closely related BSFA (Zha et al., 2023) by achieving improvements of 10.9% and 12.9% under 1-shot and 5-shot conditions on CUB-200-2011, respectively. This highlights the effectiveness of our method in constructing discriminative prototypes and maintaining consistent semantic alignment, even when backbone architectures vary. Notably, while BSFA (Zha et al., 2023) and OLSA (Wu et al., 2021) exhibit performance fluctuations across different backbones, our method demonstrates stable and competitive results, further underscoring its reliability. However, when compared to AttNet (Xu et al., 2022), our approach shows sub-optimal performance on Stanford Dogs. This can be attributed to AttNet's dual-branch attention mechanism, which localizes multiple target parts and models inter-part dependencies through multi-instance packages. This finegrained modeling is particularly beneficial for Stanford Dogs because it involves more subtle intraclass variation and is more challenging to identify than CUB-200-2011 and Stanford Cars. In contrast, our method focuses on separating foreground targets from category instances, which may limit its ability to capture such detailed interactions. Despite this limitation, our approach consistently outperformed AttNet on the remaining two datasets. Future work will explore enhanced representation learning techniques to address this limitation and further refine semantic alignment capabilities.

4.4. Ablation studies

In this section, we conduct comprehensive experiments aimed at quantitatively and qualitatively validating the effectiveness of the modules within our approach. Unless specified, all experiments employ

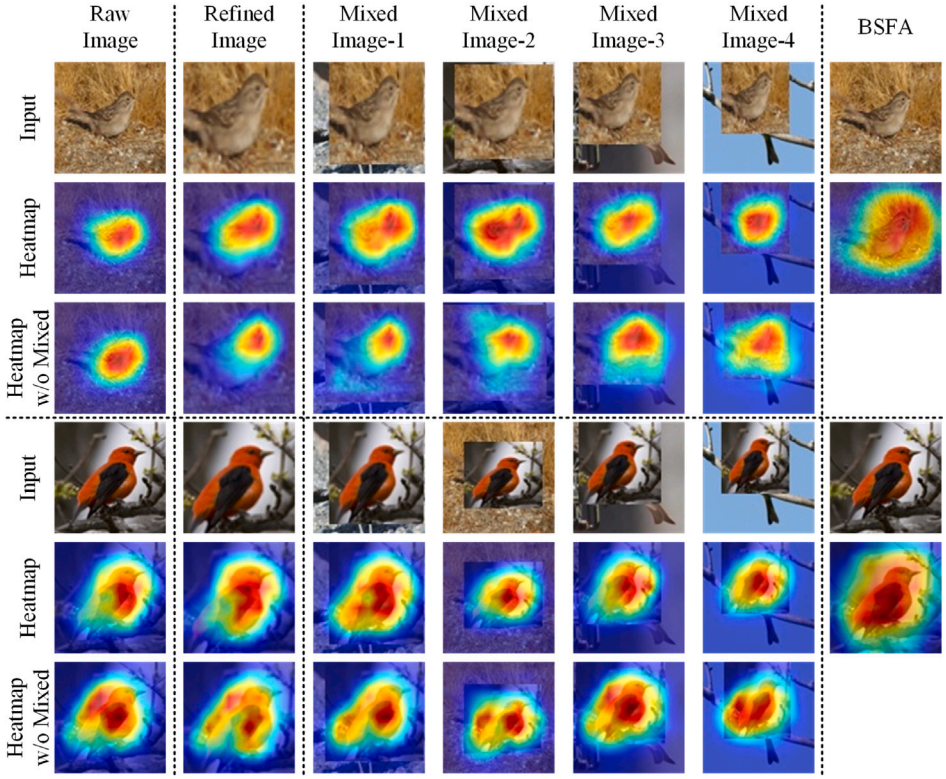

Fig. 8. Heatmap visualization comparing Raw, Refined, and background-confused (Mixed) images. Our method, which incorporates a background confusion branch, demonstrates more precise focus on the foreground target region during training compared to the BSFA and our method without mixing (w/o Mixed).

ResNet-12 as the backbone and follow the same setup outlined in section of implementation details.

(1) Visualization of foreground object extraction and background obfuscation. Fig. 8 illustrates the foreground refinement and background obfuscation processes for two support-set samples within the same episode, alongside their corresponding attentional activation maps after processing through the backbone network. Comparing the raw images with the foreground refined views, it is evident that the lower threshold design and rectangular box expansion strategy in our foreground extraction module effectively capture nearly the entire foreground region. However, due to the absence of precise positional supervision and the inherent bias of attention mechanisms toward salient regions, certain details, such as the bird's tail in Mixed Images 3 and 4, remain challenging to detect and retain. Nevertheless, compared to undifferentiated image augmentation methods such as MixUp (Zhang et al., 2018), our method greatly enriches the diversity of the background while maintaining an accurate rendering of details in the foreground region. This is further validated by comparing attentional activation maps trained with and without the background confusion branch (w/o Mixed). The inclusion of background confusion samples leads to a model that focuses more sharply on foreground discriminative regions. Additionally, compared to similar methods such as BSFA (Zha et al., 2023), our method demonstrates superior focus on foreground regions, attributed to our emphasis on background diversity and semantic consistency alignment across multiple relations. This highlights the effectiveness of our proposed strategy in enhancing both representational accuracy and robustness.

(2) Effectiveness of two-branch data streams as well as global classification branching. In Table 3, we have quantitatively documented the validity of the global classification branch under different combinations of data streams, although it is widely recognized (Liu et al., 2021; Wu et al., 2021; Zha et al., 2023). A comparative analysis of the results from (*a-b*), (*c-d*), (*e-f*), and (*g-h*) reveals that the incorporation of the global classification branch significantly enhances the model's similarity measure. This enhancement is primarily attributable to the introduction of additional global labels, which facilitates the optimization of the backbone encoder by the global classifier and classification loss. Consequently, the model gains access to more beyond-episode insights into the category distribution, ultimately leading to a more precise similarity metric among sample embeddings.

Table 3
Comparison of accuracy on the CUB-200–2011 dataset with/without foreground refined images (Refined), background obfuscated images (Mixed), and the global classification branch (Global Classifier).

ID	Refined	Mixed	Global Classifier	CUB-200–2011	
				5-way 1-shot	5-way 5-shot
a				71.50 ± 0.52	81.50 ± 0.34
b			✓	77.21 ± 0.40	88.07 ± 0.29
c	✓			75.37 ± 0.50	85.35 ± 0.31
d	✓		✓	79.32 ± 0.46	90.64 ± 0.25
e		✓		72.56 ± 0.52	79.27 ± 0.41
f		✓	✓	81.20 ± 0.43	87.14 ± 0.31
g	✓	✓		78.47 ± 0.45	85.50 ± 0.31
h	✓	✓	✓	82.77 ± 0.44	90.07 ± 0.29

Under uniform global classification branching conditions, it is discernible that various combinations of views elicit distinct impacts on model predictions. Specifically, in the absence of global classification branching (a, c, e, g), the introduction of foreground refined view (c) brings more significant improvement compared to the introduction of background obfuscation view (e). This is attributed to the fact that the foreground refined view focuses on discriminative regions and has a scale difference compared to the raw image, which effectively enhance the learning of discriminative features, especially in the case of weak model modeling capabilities. In addition, the co-introduction of the two views (g) brings further performance improvement, effectively demonstrating that the pooling of multiple views can effectively facilitate robust prototype refinement.

Table 4
Comparison of results with/without the distribution metric (KL) and the supervised contrast (SUP) in the multi-relation alignment module (MRA).

MRA		CUB-200–2011		Stanford dogs	
KL	SUP	5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
		82.77 ± 0.44	90.64 ± 0.25	71.26 ± 0.48	84.12 ± 0.31
✓		83.51 ± 0.43	91.40 ± 0.24	73.76 ± 0.46	85.41 ± 0.31
	✓	83.37 ± 0.42	90.96 ± 0.24	73.79 ± 0.47	84.85 ± 0.31
✓	✓	**83.74 ± 0.42**	**92.20 ± 0.23**	**74.82 ± 0.47**	**85.54 ± 0.30**

Table 5
Comparison of results for different prediction-target combinations in the MRA distribution metric. "Raw" and "Refined" represent the global prediction distributions for the original and foreground refinement images, respectively.

Distribution metric		CUB-200–2011	
Prediction	Target	5-way 1-shot	5-way 5-shot
Raw	Refined	83.47 ± 0.43	91.44 ± 0.25
Refined	Raw	83.40 ± 0.43	91.12 ± 0.25
Raw,Refined	Refined,Raw	**83.74 ± 0.42**	**92.20 ± 0.23**

On the other hand, in the case with global classification branching (b, d, f, h), the introduction of the background obfuscation view (f) brings significantly better performance improvement in the 5-way 1-shot test. This demonstrates that in the 1-shot condition with extremely sparse samples, the prototype incorporating diverse backgrounds possesses superior discriminative properties compared to removing backgrounds (d). However, it is worth noting that the introduction of the background obfuscation view (f) exhibits unstable gains in the 5-way 5-shot test. This is due to the fact that the effect of missing background diversity is significantly attenuated as the number of learnable samples increases. And too many hybrids generated samples interfere with feature learning leading to sub-optimal performance. For this reason, our method introduces background obfuscated views only in 5-way 1-shot training. And obviously, the simultaneous introduction of the raw image, the foreground refined view, and the background obfuscated view (h) is consistently optimal, maximizing the expansion and enrichment of the learnable views in the 1-shot condition.

(3) Effectiveness of the multi-relation alignment. We conduct an ablation study under our optimal framework by eliminating the one-to-one distribution metric (KL) and the many-to-many supervised comparison (SUP). The results, presented in Table 4, indicate that removing either component results in a consistent decrement in accuracy. This is attributed to the fact that both KL and SUP aim to enhance the accuracy of similarity measures by drawing embeddings with similar semantics closer together, facilitated by additional guidance (Yang et al., 2022; Zhang et al., 2022). Moreover, the combined implementation of the two modules exhibited superior performance, attributed to the KL metric acting on the high-dimensional global predictive distribution while SUP supervises the sample embeddings in feature space (Ling et al., 2022). On a finer granularity, for distribution metrics and supervised comparisons, we perform more comparisons of different implementations to explore the optimal implementation, respectively.

Firstly, concerning the Kullback–Leibler scatter for the global classification distribution, we experimented with different combinations of predictions and the target distribution to identify the optimal approach. As shown in Table 5, solely aligning the global prediction distribution of the raw image to the refined image, or vice versa, yields limited gains. However, a bidirectional KL metric approach significantly outperforms these unidirectional methods, particularly in the 5-shot scenario with more samples. This improvement is attributed to aligning the refined foreground of the samples with the raw image on the global predictive distribution, further encouraging embeddings with the same semantics to be aligned, improving predictive performance.

Secondly, for the supervised comparison, we explored various scopes of comparison. As depicted in Table 6, we conducted supervised comparisons within individual episodes and cross multi episodes between a batch, testing performance under different batch sizes. The results indicate that broader comparisons across Batch tend to yield superior performance, leveraging the diversity between semantic embeddings to optimize the encoder. Moreover, in the 1-shot setting, a larger batch size (4) is preferable due to the limited sample number, whereas in the 5-shot condition, a smaller batch size (2) achieves the best results. This disparity also underscores the effectiveness of appropriate cross-episode information introduction in FSL (Wang, Falk et al., 2024; Wang, Pontil et al., 2024).

Table 6
Comparison of results across different ranges and batch sizes for supervised contrast in MRA. "Episode" and "Batch" in the range refer to the supervised contrast between samples within a single episode and between samples within all episodes in a batch, respectively.

Supervised Contrast		CUB-200–2011	
Range	Batch size	5-way 1-shot	5-way 5-shot
Batch	2	83.59 ± 0.43	**92.20 ± 0.23**
Batch	4	**83.74 ± 0.42**	91.45 ± 0.24
Batch	8	81.28 ± 0.44	88.73 ± 0.28
Episode	2	83.37 ± 0.43	91.72 ± 0.23
Episode	4	83.43 ± 0.43	91.51 ± 0.24
Episode	8	81.12 ± 0.45	88.21 ± 0.28

(4) Effectiveness of prediction combinations for different branches.

Table 7 provides a detailed analysis of our method's 5-way 1-shot prediction performance, considering various data stream branches and their combinations across diverse backbones and datasets. In cases represented by a and c in the table, we observe that the Refined branch achieves optimal performance, followed closely by the Mixed branch. This superiority is attributed to the Refined branch's exclusive focus on the foreground region, while the Mixed branch enhances background diversity while maintaining the foreground regions. Both approaches outperform the original image branch. Furthermore, averaging the predictions from any two or three distinct branches leads to further improvements, indicating that complementary knowledge still exists between these branches. Investigating the reasons behind the complementary advantages of diverse branches is a compelling endeavor, and identifying strategies to optimally harness their combined potential will continually guide our ongoing research endeavors.

When the background obfuscated branch is excluded from our method, as demonstrated in b and d, the prediction performance of the raw and refined image branches decreases consequently. This result emphasizes the importance of diverse backgrounds in establishing robust category prototypes. Furthermore, the heatmap visualization of various input samples presented in Fig. 8 underscores the effectiveness of incorporating background obfuscation branches in enabling the model to concentrate more precisely on the semantic discriminative regions of each category. Consequently, we choose the Raw-Refined combination as the final prediction due to its consistent and superior performance in all conditions.

(5) Sensitivity analysis. In the foreground object extraction and background obfuscation module, we introduce the hyperparameter τ, which plays a pivotal role in determining the foreground refined region in the raw image and affects the relative ratio of foreground to background in the obfuscation module. As depicted in Fig. 9, we visualize refined regions and background-obfuscated samples generated with varying values of τ, ranging from 0.1 to 0.6. Through the bounding box visualization of different samples, it can be intuitively found that our foreground object extraction can accurately locate the salient feature regions of foreground objects under any hyperparameters. In addition, by comparing the bounding boxes generated from the same sample under different hyperparameters, it can be found that the bounding boxes keep decreasing in size with the increase of τ-value and keep focusing on the feature-rich regions such as the bird's head. Observing the mixing performance of sample-1 and other samples, it effectively

Table 7
Accuracy (%) of our method on different branch combination approaches on CUB-200–2011 (CUB) and Stanford Dogs (DOGS) datasets under 5-way 1-shot conditions where *w/o mixed* represents the removal of background obfuscated data streams.

ID	Method	Backbone	Raw	Refined	Mixed	Raw-refined-mixed	Raw-refined	Raw-mixed	Refined-mixed
a	CUB	ResNet-12	81.24	82.78	81.70	83.51	83.74	81.84	**83.79**
b	CUB *w/o mixed*	ResNet-12	79.79	81.19	–	–	**82.57**	–	–
c	DOGS	ResNet-12	72.36	72.96	72.41	74.64	**74.82**	72.99	74.57
d	DOGS *w/o mixed*	ResNet-12	70.18	71.51	–	–	**72.33**	–	–

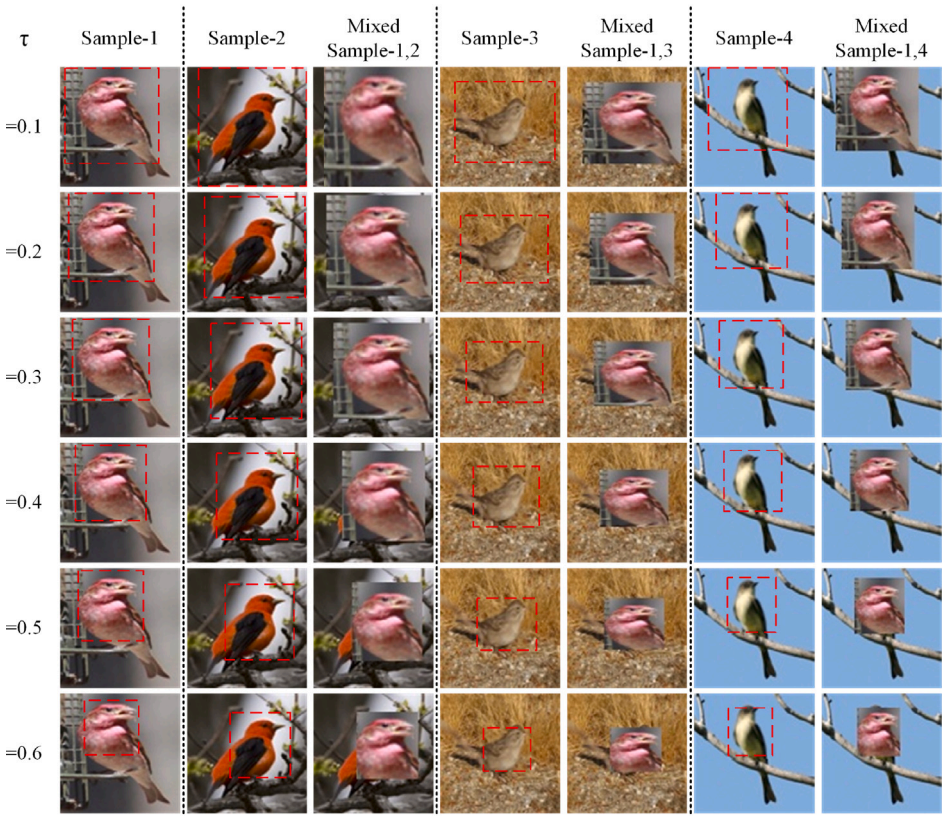

Fig. 9. Visualization of foreground object extraction regions and obfuscated background views for randomly selected samples at different τ-values.

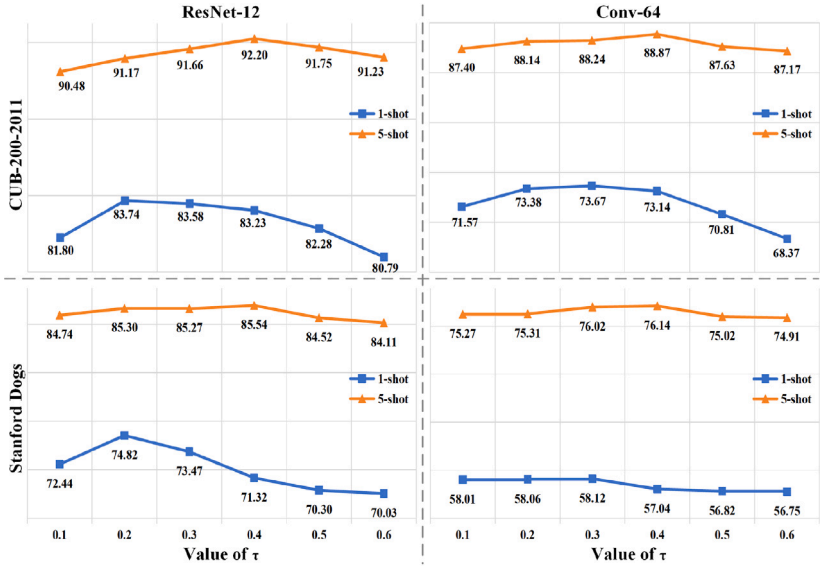

Fig. 10. Performance of our method on different backbone networks (ResNet-12, Conv-64) and different datasets (CUB-200-2011, Stanford Dogs) with different τ-values in 5-way 1-shot and 5-way 5-shot settings.

Table 8
Comparing the efficiency of various few-shot learning methods based on ResNet-12 where *w/o mixed* represents the removal of background obfuscated data streams.

Method	CUB-200–2011	Params	FLOPs
	5-way 1-shot	(G)	(M)
CAN (Hou et al., 2019)	76.98 ± 0.48	8.04	12.75
AGPF (Tang et al., 2022)	78.73 ± 0.84	8.77	51.53
RENet (Kang et al., 2021)	79.49 ± 0.44	12.63	35.70
BSFA (Zha et al., 2023)	82.27 ± 0.46	8.03	33.76
Ours *w/o mixed*	82.57 ± 0.43	8.03	33.76
Ours	83.74 ± 0.42	8.04	75.96

blends the background of other samples while retaining the category discriminative region.

To quantitatively assess the impact of varying τ, Fig. 10 presents the prediction accuracy of our method for different τ-value. It can be observed that the effect of τ-values is relatively weak in the 5-shot condition, primarily stemming from the absence of the additional background obfuscation module. Conversely, in the 1-shot setting, the influence of τ-values is more pronounced, with smaller values (around 0.2) tending to yield superior performance. This observation also demonstrates the positive impact of rich background diversity on category differentiation. In addition, how to follow the training to dynamically adjust the threshold to avoid heavy manual intervention will also be one of the focuses of our subsequent work.

(6) Model complexity analysis. Apart from prediction accuracy, floating-point operations (FLOPs) and parameters (Params) are also pivotal metrics for assessing model performance. To facilitate a fair comparison, we utilized the default FLOPs and Params computation criteria of the thop libraries in PyTorch to evaluate our proposed approach. Subsequently, Table 8 compares our method's model complexity and accuracy performance to several publicly available fine-grained recognition methods. This analysis was conducted under the identical ResNet-12 backbone and 5-way 1-shot input conditions for all methods.

The results in Table 8 demonstrate that our method achieves a balanced trade-off between model parameters and computational complexity while delivering higher accuracy compared to other methods. Specifically, the additional computational overhead in our approach primarily stems from the introduction of the background confusion branch. This branch enhances model accuracy by expanding the support set with samples of diverse backgrounds. When compared to the BSFA (Zha et al., 2023), which relies solely on the raw and refined views, our method achieves significantly improved accuracy. Notably, even when the background confusion branch is removed (*w/o mixed*), our method continues to perform exceptionally well. By constructing richer semantic consistency constraints to align discriminative semantics, our approach maintains superior performance over methods like BSFA, despite operating with nearly identical computational costs and without additional background confusion data streams. This outcome validates the effectiveness of our proposed multi-relational semantic alignment mechanism, highlighting its low overhead and practical efficiency.

5. Conclusion

In this paper, we propose a multi-view aggregation and multi-relation alignment framework (MVRA) for few-shot fine-grained recognition. To overcome the difficulties posed by fine-grained categories and limited learnable samples, we aggregate samples with multiple views, including raw and foreground refined images, thereby enhancing the learning of discriminative regions. Besides, to mitigate the negative impact of limited background diversity in single-shot, the background obfuscation module is designed to facilitate the development of representative category prototypes. Furthermore, combined with the credible labels assigned by the global label assignment, the multi-relation alignment module integrates embeddings and distributions of samples from multiple views, strengthening consistent semantic alignment via supervised comparison and distribution metrics. Extensive experiments on three datasets with two backbones validate the efficacy of our proposed method.

CRediT authorship contribution statement

Jiale Chen: Conceptualization, Methodology, Validation, Investigation, Writing – original draft, Writing – review & editing. **Feng Xu:** Supervision, Project administration, Funding acquisition, Writing – review & editing. **Xin Lyu:** Supervision, Investigation, Writing – review & editing. **Tao Zeng:** Investigation. **Xin Li:** Investigation. **Shangjing Chen:** Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Key Research and Development Program of China [grant number 2023YFC3209301, 2023YFC3209201]; the Natural Science Foundation of Jiangsu Province [grant number BK20241508]; the National Natural Science Foundation of China [grant number 62401196]; the Fundamental Research Funds for the Central Universities, China [grant number B230204009, B230201007, B220206006].

Data availability

Data will be made available on request.

References

Cao, S., Wang, W., Zhang, J., Zheng, M., & Li, Q. (2022). A few-shot fine-grained image classification method leveraging global and local structures. *International Journal of Machine Learning and Cybernetics*, *13*(8), 2273–2281. http://dx.doi.org/10.1007/s13042-022-01522-w.

Chang, D., Ding, Y., Xie, J., Bhunia, A. K., Li, X., Ma, Z., Wu, M., Guo, J., & Song, Y.-Z. (2020). The devil is in the channels: Mutual-channel loss for fine-grained image classification. *IEEE Transactions on Image Processing*, *29*, 4683–4695. http://dx.doi.org/10.1109/TIP.2020.2973812.

Chang, D., Pang, K., Du, R., Tong, Y., Song, Y.-Z., Ma, Z., & Guo, J. (2023). Making a bird AI expert work for you and me. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(10), 12068–12084. http://dx.doi.org/10.1109/TPAMI.2023.3274593.

Chang, D., Pang, K., Zheng, Y., Ma, Z., Song, Y.-Z., & Guo, J. (2021). Your "flamingo" is my "bird": Fine-grained, or not. In *Conference on computer vision and pattern recognition* (pp. 11471–11480). http://dx.doi.org/10.1109/CVPR46437.2021.01131.

Chang, D., Tong, Y., Du, R., Hospedales, T., Song, Y.-Z., & Ma, Z. (2023). An erudite fine-grained visual classification model. In *Conference on computer vision and pattern recognition* (pp. 7268–7277). http://dx.doi.org/10.1109/CVPR52729.2023.00702.

Chi, Z., Gu, L., Liu, H., Wang, Y., Yu, Y., & Tang, J. (2022). MetaFSCIL: A meta-learning approach for few-shot class incremental learning. In *Proc. IEEE conf. comput. vis. pattern recognit.* (pp. 14146–14155). http://dx.doi.org/10.1109/CVPR52688.2022.01377.

Du, R., Chang, D., Bhunia, A. K., Xie, J., Ma, Z., Song, Y.-Z., & Guo, J. (2020). Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In *Computer vision* (pp. 153–168). Cham.

Du, R., Chang, D., Liang, K., Hospedales, T., Song, Y.-Z., & Ma, Z. (2023). On-the-fly category discovery. In *Conference on computer vision and pattern recognition* (pp. 11691–11700). http://dx.doi.org/10.1109/CVPR52729.2023.01125.

Fang, Z., Wang, X., Li, H., Liu, J., Hu, Q., & Xiao, J. (2023). FastRecon: Few-shot industrial anomaly detection via fast feature reconstruction. In *Proc. IEEE int. conf. comput. vis.* (pp. 17435–17444). http://dx.doi.org/10.1109/ICCV51070.2023.01603.

Han, J., Yao, X., Cheng, G., Feng, X., & Xu, D. (2022). P-CNN: Part-based convolutional neural networks for fine-grained visual categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(2), 579–590. http://dx.doi.org/10.1109/TPAMI.2019.2933510.

Hao, F., He, F., Cheng, J., & Tao, D. (2022). Global-local interplay in semantic alignment for few-shot learning. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(7), 4351–4363. http://dx.doi.org/10.1109/TCSVT.2021.3132912.

Hospedales, T., Antoniou, A., Micaelli, P., & Storkey, A. (2022). Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(9), 5149–5169. http://dx.doi.org/10.1109/TPAMI.2021.3079209.

Hou, R., Chang, H., Ma, B., Shan, S., & Chen, X. (2019). Cross attention network for few-shot classification. In *Proc. adv. neural inf. process. syst.* (pp. 4003–4014). Red Hook, NY, USA: Curran Associates Inc., http://dx.doi.org/10.5555/3454287.3454647.

Huang, H., Zhang, J., Yu, L., Zhang, J., Wu, Q., & Xu, C. (2022). TOAN: Target-oriented alignment network for fine-grained image categorization with few labeled samples. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(2), 853–866. http://dx.doi.org/10.1109/TCSVT.2021.3065693.

Jiang, W., Huang, K., Geng, J., & Deng, X. (2021). Multi-scale metric learning for few-shot learning. *IEEE Transactions on Circuits and Systems for Video Technology*, *31*(3), 1091–1102. http://dx.doi.org/10.1109/TCSVT.2020.2995754.

Jiang, X., Tang, H., Gao, J., Du, X., He, S., & Li, Z. (2023). Delving into multimodal prompting for fine-grained visual classification. In *AAAI conf. artif. intell.*

Kang, D., Kwon, H., Min, J., & Cho, M. (2021). Relational embedding for few-shot classification. In *Proc. IEEE int. conf. comput. vis.* (pp. 8802–8813). http://dx.doi.org/10.1109/ICCV48922.2021.00870.

Khosla, A., Jayadevaprakash, N., Yao, B., & Fei-Fei, L. (2011). Novel dataset for fine-grained image categorization : Stanford dogs. In *Proc. IEEE conf. comput. vis. pattern recognit.* (pp. 554–561).

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., & Krishnan, D. (2020). Supervised contrastive learning. In *Proc. adv. neural inf. process. syst.*. Red Hook, NY, USA: Curran Associates Inc., http://dx.doi.org/10.5555/3495724.3497291.

Krause, J., Stark, M., Deng, J., & Fei-Fei, L. (2013). 3D object representations for fine-grained categorization. In *Proc. IEEE int. conf. comput. vis. workshop* (pp. 554–561). http://dx.doi.org/10.1109/ICCVW.2013.77.

Leng, Z., Wang, M., Wan, Q., Xu, Y., Yan, B., & Sun, S. (2024). Meta-learning of feature distribution alignment for enhanced feature sharing. *Knowledge-Based Systems*, *296*, Article 111875. http://dx.doi.org/10.1016/j.knosys.2024.111875.

Li, Y., Chen, L., Li, W., & Wang, N. (2024). Few-shot fine-grained classification with rotation-invariant feature map complementary reconstruction network. *IEEE Transactions on Geoscience and Remote Sensing*, *62*, 1–12. http://dx.doi.org/10.1109/TGRS.2024.3361501.

Li, H., Li, M., Peng, Q., Wang, S., Yu, H., & Wang, Z. (2024). Correlation-guided semantic consistency network for visible-infrared person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, *34*(6), 4503–4515. http://dx.doi.org/10.1109/TCSVT.2023.3340225.

Li, X., Song, Q., Wu, J., Zhu, R., Ma, Z., & Xue, J.-H. (2023). Locally-enriched cross-reconstruction for few-shot fine-grained image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, *33*(12), 7530–7540. http://dx.doi.org/10.1109/TCSVT.2023.3275382.

Li, Z., Tang, H., Peng, Z., Qi, G.-J., & Tang, J. (2023). Knowledge-guided semantic transfer network for few-shot image recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 1–15. http://dx.doi.org/10.1109/TNNLS.2023.3240195.

Li, X., Wu, J., Sun, Z., Ma, Z., Cao, J., & Xue, J.-H. (2021). BSNet: Bi-similarity network for few-shot fine-grained image classification. *IEEE Transactions on Image Processing*, *30*, 1318–1331. http://dx.doi.org/10.1109/TIP.2020.3043128.

Li, X., Xu, F., Liu, F., Tong, Y., Lyu, X., & Zhou, J. (2024). Semantic segmentation of remote sensing images by interactive representation refinement and geometric prior-guided inference. *IEEE Transactions on Geoscience and Remote Sensing*, *62*, 1–18. http://dx.doi.org/10.1109/TGRS.2023.3339291.

Li, X., Xu, F., Yu, A., Lyu, X., Gao, H., & Zhou, J. (2025). A frequency decoupling network for semantic segmentation of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, *63*, 1–21. http://dx.doi.org/10.1109/TGRS.2025.3531879.

Li, X., Yang, X., Ma, Z., & Xue, J.-H. (2023). Deep metric learning for few-shot image classification: A review of recent developments. *Pattern Recognition*, *138*, Article 109381. http://dx.doi.org/10.1016/j.patcog.2023.109381.

Li, M., Yao, H., & Wang, Y. (2023). Focus nuance and toward diversity: exploring domain-specific fine-grained few-shot recognition. *Neural Computing and Applications*, *35*(28), 21275–21290. http://dx.doi.org/10.1007/s00521-023-08787-4.

Li, K., Zhang, Y., Li, K., & Fu, Y. (2020). Adversarial feature hallucination networks for few-shot learning. In *Proc. IEEE conf. comput. vis. pattern recognit.* (pp. 13467–13476). http://dx.doi.org/10.1109/CVPR42600.2020.01348.

Ling, J., Liao, L., Yang, M., & Shuai, J. (2022). Semi-supervised few-shot learning via multi-factor clustering. In *Proc. IEEE conf. comput. vis. pattern recognit.* (pp. 14544–14553). http://dx.doi.org/10.1109/CVPR52688.2022.01416.

Liu, H., Chen, C. L. P., Gong, X., & Zhang, T. (2024). Robust saliency-aware distillation for few-shot fine-grained visual recognition. *IEEE Transactions on Multimedia*, *26*, 7529–7542. http://dx.doi.org/10.1109/TMM.2024.3369870.

Liu, C., Fu, Y., Xu, C., Yang, S., Li, J., Wang, C., & Zhang, L. (2021). Learning a few-shot embedding model with contrastive learning. *vol. 35*, In *AAAI conf. artif. intell* (pp. 8635–8643). http://dx.doi.org/10.1609/aaai.v35i10.17047.

Ma, L., Hong, H., Meng, F., Wu, Q., & Wu, J. (2024). Deep progressive asymmetric quantization based on causal intervention for fine-grained image retrieval. *IEEE Transactions on Multimedia*, *26*, 1306–1318. http://dx.doi.org/10.1109/TMM.2023.3279990.

Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579–2605.

Miao, B., Bennamoun, M., Gao, Y., Shah, M., & Mian, A. (2024). Temporally consistent referring video object segmentation with hybrid memory. *IEEE Transactions on Circuits and Systems for Video Technology*, http://dx.doi.org/10.1109/TCSVT.2024.3419119, 1–1.

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, *9*(3), 353–383. http://dx.doi.org/10.1016/0010-0285(77)90012-3.

Park, W., & Ryu, J. (2024). Fine-grained self-supervised learning with Jigsaw puzzles for medical image classification. *Computers in Biology and Medicine*, *174*, Article 108460. http://dx.doi.org/10.1016/j.compbiomed.2024.108460.

Qi, R., Ning, S., & Jiang, Y. (2024). Prototype rectification with region-wise foreground enhancement for few-shot classification. In *Pattern recognit. comput. vis.* (pp. 15–26). Singapore: http://dx.doi.org/10.1007/978-981-99-8462-6_2.

Roy, A. M., Bhaduri, J., Kumar, T., & Raj, K. (2023). WilDect-YOLO: An efficient and robust computer vision-based accurate object localization model for automated endangered wildlife detection. *Ecol. Inform.*, *75*, Article 101919. http://dx.doi.org/10.1016/j.ecoinf.2022.101919.

Rusu, A. A., Rao, D., Sygnowski, J., Vinyals, O., Pascanu, R., Osindero, S., & Hadsell, R. (2019). Meta-learning with latent embedding optimization. In *Proc. int. conf. learn. representations*.

Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. In *Proc. adv. neural inf. process. syst.* (pp. 4080–4090). Red Hook, NY, USA: Curran Associates Inc., http://dx.doi.org/10.5555/3294996.3295163.

Sun, J., Shen, X., & Sun, Q. (2023). Efficient feature reconstruction via l2,1-norm regularization for few-shot classification. *IEEE Transactions on Circuits and Systems for Video Technology*, *33*(12), 7452–7465. http://dx.doi.org/10.1109/TCSVT.2023.3274168.

Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H., & Hospedales, T. M. (2018). Learning to compare: Relation network for few-shot learning. In *Proc. IEEE conf. comput. vis. pattern recognit.* (pp. 1199–1208). http://dx.doi.org/10.1109/CVPR.2018.00131.

Tang, H., Li, Z., Peng, Z., & Tang, J. (2020). BlockMix: Meta regularization and self-calibrated inference for metric-based meta-learning. In *Proc. ACM int. conf. multimedia* (pp. 610–618). http://dx.doi.org/10.1145/3394171.3413884.

Tang, H., Liu, J., Yan, S., Yan, R., Li, Z., & Tang, J. (2023). M3net: Multi-view encoding, matching, and fusion for few-shot fine-grained action recognition. In *Proc. ACM int. conf. multimedia* (pp. 1719–1728). http://dx.doi.org/10.1145/3581783.3612221.

Tang, H., Yuan, C., Li, Z., & Tang, J. (2022). Learning attention-guided pyramidal features for few-shot fine-grained recognition. *Pattern Recognition*, *130*, Article 108792. http://dx.doi.org/10.1016/j.patcog.2022.108792.

Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. (2011). The caltech-UCSD birds-200–2011 dataset. In *California inst. technol.*.

Wang, S., Chang, J., Wang, Z., Li, H., Ouyang, W., & Tian, Q. (2024). Content-aware rectified activation for zero-shot fine-grained image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(6), 4366–4380. http://dx.doi.org/10.1109/TPAMI.2024.3355461.

Wang, R., Falk, J. I. T., Pontil, M., & Ciliberto, C. (2024). Robust meta-representation learning via global label inference and classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(4), 1996–2010. http://dx.doi.org/10.1109/TPAMI.2023.3328184.

Wang, R., Hou, F., Cahan, S. F., Chen, L., Jia, X., & Ji, W. (2023). Fine-grained entity typing with a type taxonomy: A systematic review. *IEEE Transactions on Knowledge and Data Engineering*, *35*(5), 4794–4812. http://dx.doi.org/10.1109/TKDE.2022.3148980.

Wang, Y., Ji, Y., Wang, W., & Wang, B. (2024). Bi-channel attention meta learning for few-shot fine-grained image recognition. *Expert Systems with Applications*, *242*, Article 122741. http://dx.doi.org/10.1016/j.eswa.2023.122741.

Wang, R., Pontil, M., & Ciliberto, C. (2024). The role of global labels in few-shot classification and how to infer them. In *Proc. adv. neural inf. process. syst.*. Red Hook, NY, USA: Curran Associates Inc..

Wang, C., Song, S., Yang, Q., Li, X., & Huang, G. (2021). Fine-grained few shot learning with foreground object transformation. *Neurocomputing*, *466*, 16–26. http://dx.doi.org/10.1016/j.neucom.2021.09.016.

Wang, S., Wang, Z., Li, H., Chang, J., Ouyang, W., & Tian, Q. (2023). Semantic-guided information alignment network for fine-grained image recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, *33*(11), 6558–6570. http://dx.doi.org/10.1109/TCSVT.2023.3263870.

Wang, Z., Wang, S., Li, H., Dou, Z., & Li, J. (2020). Graph-propagation based correlation learning for weakly supervised fine-grained image classification. *AAAI Conference on Artificial Intelligence*, *34*(07), 12289–12296. http://dx.doi.org/10.1609/aaai.v34i07.6912.

Wang, M., Zhao, P., Lu, X., Min, F., & Wang, X. (2023). Fine-grained visual categorization: A spatial–frequency feature fusion perspective. *IEEE Transactions on Circuits and Systems for Video Technology*, *33*(6), 2798–2812. http://dx.doi.org/10.1109/TCSVT.2022.3227737.

Wei, X.-S., Song, Y.-Z., Aodha, O. M., Wu, J., Peng, Y., Tang, J., Yang, J., & Belongie, S. (2022). Fine-grained image analysis with deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(12), 8927–8948. http://dx.doi.org/10.1109/TPAMI.2021.3126648.

Wei, X.-S., Wang, P., Liu, L., Shen, C., & Wu, J. (2019). Piecewise classifier mappings: Learning fine-grained learners for novel categories with few examples. *IEEE Transactions on Image Processing*, *28*(12), 6116–6125. http://dx.doi.org/10.1109/TIP.2019.2924811.

Wen, J., Li, Y., Fang, M., Zhu, L., Feng, D. D., & Li, P. (2023). Fine-grained and multiple classification for alzheimer's disease with wavelet convolution unit network. *IEEE Transactions on Biomedical Engineering*, *70*(9), 2592–2603. http://dx.doi.org/10.1109/TBME.2023.3256042.

Wertheimer, D., Tang, L., & Hariharan, B. (2021). Few-shot classification with feature map reconstruction networks. In *Proc. IEEE conf. comput. vis. pattern recognit.* (pp. 8008–8017). http://dx.doi.org/10.1109/CVPR46437.2021.00792.

Wu, J., Chang, D., Sain, A., Li, X., Ma, Z., Cao, J., Guo, J., & Song, Y.-Z. (2023). Bi-directional feature reconstruction network for fine-grained few-shot image classification. In *AAAI Conf. Artif. Intell.*. http://dx.doi.org/10.1609/aaai.v37i3.25383.

Wu, Y., Zhang, B., Yu, G., Zhang, W., Wang, B., Chen, T., & Fan, J. (2021). Object-aware long-short-range spatial alignment for few-shot fine-grained image classification. In *Proc. ACM int. conf. multimedia* (pp. 107–115). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3474085.3475532.

Xu, C., Liu, C., Sun, X., Yang, S., Wang, Y., Wang, C., & Fu, Y. (2023). PatchMix augmentation to identify causal features in few-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(6), 7639–7653. http://dx.doi.org/10.1109/TPAMI.2022.3223784.

Xu, J., Liu, B., & Xiao, Y. (2024). A multitask latent feature augmentation method for few-shot learning. *IEEE Transactions on Neural Networks and Learning Systems*, *35*(5), 6976–6990. http://dx.doi.org/10.1109/TNNLS.2022.3213576.

Xu, Q., Zhang, M., Li, Y., & Tao, Z. (2023). Learning more discriminative clues with gradual attention for fine-grained visual categorization. *Image and Vision Computing*, *136*, Article 104753. http://dx.doi.org/10.1016/j.imavis.2023.104753.

Xu, S.-L., Zhang, F., Wei, X.-S., & Wang, J. (2022). Dual attention networks for few-shot fine-grained recognition. *AAAI Conference on Artificial Intelligence*, *36*(3), 2911–2919. http://dx.doi.org/10.1609/aaai.v36i3.20196.

Yang, S., Wu, S., Liu, T., & Xu, M. (2022). Bridging the gap between few-shot and many-shot learning via distribution calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(12), 9830–9843. http://dx.doi.org/10.1109/TPAMI.2021.3132021.

Yu, Y., Zhang, D., Wang, S., Ji, Z., & Zhang, Z. (2022). Local spatial alignment network for few-shot learning. *Neurocomputing*, *497*, 182–190. http://dx.doi.org/10.1016/j.neucom.2022.05.020.

Yu, X., Zhao, Y., Gao, Y., & Xiong, S. (2021). MaskCOV: A random mask covariance network for ultra-fine-grained visual categorization. *Pattern Recognition*, *119*, Article 108067. http://dx.doi.org/10.1016/j.patcog.2021.108067.

Zha, Z., Tang, H., Sun, Y., & Tang, J. (2023). Boosting few-shot fine-grained recognition with background suppression and foreground alignment. *IEEE Transactions on Circuits and Systems for Video Technology*, *33*(8), 3947–3961. http://dx.doi.org/10.1109/TCSVT.2023.3236636.

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2018). Mixup: Beyond empirical risk minimization. In *International conference on learning representations*.

Zhang, B., Ye, H., Yu, G., Wang, B., Wu, Y., Fan, J., & Chen, T. (2022). Sample-centric feature generation for semi-supervised few-shot learning. *IEEE Transactions on Image Processing*, *31*, 2309–2320. http://dx.doi.org/10.1109/TIP.2022.3154938.

Zhu, H., Ke, W., Li, D., Liu, J., Tian, L., & Shan, Y. (2022). Dual cross-attention learning for fine-grained visual categorization and object re-identification. In *Proc. IEEE conf. comput. vis. pattern recognit.* (pp. 4682–4692). http://dx.doi.org/10.1109/CVPR52688.2022.00465.

Zhu, Y., Liu, C., & Jiang, S. (2021). Multi-attention meta learning for few-shot fine-grained image recognition. In *Proc. int. joint conf. artif. intell.*.