Knowledge-Based Systems 318 (2025) 113535

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Knowledge-driven prototype refinement 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
Knowledge transfer
Semantic alignment

ABSTRACT

Advancements in deep learning have made image classification rival human performance with sufficient data and supervision. However, in domains with limited visual samples and high labeling costs, enabling AI systems to learn efficiently from few examples is challenging. This challenge is compounded in fine-grained categories, where subtle differences and scarce samples hinder robust representation extraction. To address this, we propose the Knowledge-Driven Prototype Refinement (KDPR) framework, which enhances few-shot fine-grained recognition by integrating prior knowledge from text. KDPR simulates human focus on discriminative foreground regions to extract refined views, forming a dual-branch learning framework alongside original images. It also constructs an unsupervised adjacency graph among visual instances and uses graph neural networks to improve category representation robustness. Additionally, a knowledge transfer-based image recognizer integrates prior text embeddings with global semantics directly into visual recognition, providing extra semantic guidance. To optimize knowledge-to-vision mapping, an auxiliary spatial prototype calibration aligns prototype representations across multiple spaces. Extensive experiments on three fine-grained datasets and two popular backbones demonstrate the effectiveness and state-of-the-art performance of our approach, especially in 1-shot learning. The source code is available at: https://github.com/HHU-JialeChen/KDPRNet.

1. Introduction

Image classification constitutes a fundamental task in computer vision. Recent advancements in artificial intelligence have significantly enhanced its application across various fields, such as remote sensing [1–3] and medicine [4,5]. However, the excellent performance of these methods heavily rely on extensive sample acquisition and meticulous manual labeling. When focusing on subcategory recognition for fine-grained differences in downstream tasks, the costly processes of sample acquisition and annotation pose significant constraints on the deployment of previous high data-dependent methods [6,7]. Additionally, the introduction of new category necessitates the retraining of existing models, leading to substantial resource wastage [8]. Consequently, enabling models to possess rapid and efficient learning capabilities akin to humans is crucial [9,10]. This also represents a necessary step towards aligning artificial intelligence (AI) with human intelligence (HI), and few-shot fine-grained recognition (FS-FGR) serves as an exemplary validation platform for this research [11–14].

FS-FGR aims to recognize unseen categories with only fine-grained differences using a limited number of labeled samples [15,16]. Therefore, it faces dual challenges: high intra-class variability and high inter-class similarity in fine-grained images, coupled with the limitation of learnable samples. Existing methods for fine-grained recognition in few-shot scenarios can be categorized as metric-based [17–22], augmentation-based [23–25] or a combination of both [15,26–29]. These methods have demonstrated impressive performance within the prevalent episode training paradigm. However, they struggle to generalize the discrimination of visible classes to novel ones, neglecting the significant potential of accessible prior knowledge—a common aspect of human discriminative learning. Particularly in scenarios where intuitive visual cues are scarce, summarized prior semantic knowledge, such as category labels and textual descriptions, can elucidate the distinct features of each category effectively, thereby enhancing learning efficiency and recognition accuracy. As illustrated in Fig. 1, the Black-footed Albatross is visually difficult to distinguish, but is easily recognized with the aid of prior knowledge. Guided by a summary of

* 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.knosys.2025.113535
Received 16 December 2024; Received in revised form 27 March 2025; Accepted 9 April 2025
Available online 19 April 2025

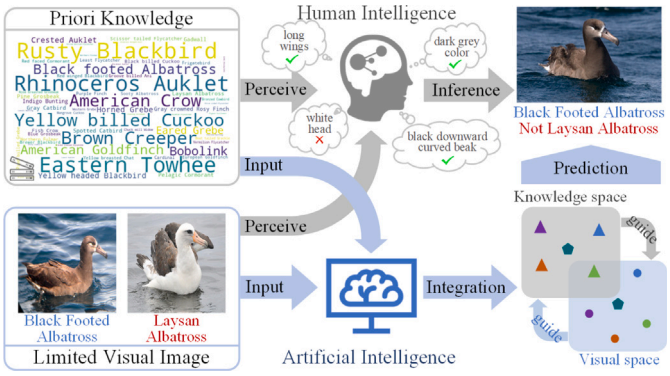

Fig. 1. Illustration of human intelligence versus artificial intelligence in understanding multi modal input: Humans possess the innate ability to automatically conceptualize knowledge from diverse modalities and match key points to draw accurate inferences. In contrast, artificial intelligence necessitates the integration of supplementary articulation mechanisms.

prior human experience, individuals can instinctively integrate conceptualized semantics with current visual perceptions to swiftly achieve cognition of novel objects.

Words and symbols embody abstractions derived from human experience accumulated over time, encapsulating semantics that are often absent in a single computer image. Consequently, seeking external experience transfer and guidance has consistently proven effective for humans in initiating the recognition of novel objects. However, for artificial intelligence, accurately integrating knowledge from diverse modalities remains a notable challenge, particularly when compared to human biological perception [12,30]. Specifically, in few-shot fine-grained recognition, category labels are readily available textual modality knowledge alongside the image modality. In contrast to the difficulties associated with fine-tuning or training a text encoder on limited text samples, we utilize publicly available text encoders that have been trained on extensive text data. These encoders pre-code category labels to extract global semantics, serving as prior knowledge. Furthermore, to enable AI to naturally process and integrate diverse data modalities, we design an image recognizer based on knowledge transfer. This approach actively engages prior knowledge in visual discrimination and loss optimization, leveraging global abstract semantics to drive model learning of visual discriminative features. Additionally, we present a vision-knowledge aligned prototype calibration module, which facilitates the reinforcement of multi-spatial knowledge and jointly drives the construction of robust prototypes and similarity measures.

Apart from being driven by prior knowledge, we devise a two-branch network architecture to mitigate the challenges associated with fine-grained recognition in scenarios where the number of learnable samples is limited. The learning of foreground discriminative regions is strengthened by the additional foreground refinement view branch, while complementing the learnable samples with diverse scales. Moreover, with the successful application of graph neural networks in computer vision tasks, there has been a surge in the exploration of graph relational networks, including category co-occurrence relations or hierarchical structure of category attributes [21,31]. In this regard, we construct category-consistent connections between labeled support set images and unlabeled query set images within episodes to augment visual representations of instances through knowledge transfer of graph structures. As a result, our approach enables the construction of category robust prototypes. And, numerous quantitative and qualitative implementations demonstrate the optimal predictive performance of our method on three mainstream fine-grained image datasets. Our contributions can be summarized as follows:

- This paper proposes a two-branch framework that leverages category-consistent relations among visual samples and prior text embeddings to boost few-shot fine-grained recognition, without requiring additional annotations.
- We propose a graph-guided instance representation augmentation approach that unsupervisedly identifies category-consistent associations among visual instances and enhances discriminative visual representations through graph-convolution based fusion.
- We propose a knowledge transfer-based image recognizer and prototype calibration module that utilizes prior text encoding to drive visual recognition and achieve multi-spatial semantic alignment, thereby compensating for the scarcity of visual information and facilitating robust prototype construction.
- Extensive experiments on three fine-grained datasets and two popular backbones demonstrate the effectiveness and outstanding performance of our proposed method.

2. Related work

2.1. Fine-grained recognition

Fine-grained recognition (FGR) aims to distinguish multiple subclasses within the same visual superclass, such as different species of birds or dogs [6,32,33]. In contrast to coarse-grained image recognition, which focuses on cross-species semantics, FGR poses challenges due to significant intra-class diversity and inter-class similarity. Even for humans without specialized training, accurately identifying some of these subtly differentiated categories is difficult. Consequently, considerable research has been dedicated to discovering key regions and matching detailed discriminative features to effectively differentiate fine-grained categories [6,7,33]. Recent advances in artificial intelligence have enabled FGR to achieve results comparable to humans with sufficient supervised training.

Despite significant advancements in FGR under fully supervised conditions, these methods are inherently reliant on extensive labeled datasets, which are frequently unattainable in many specific scenarios [9,10]. Specifically, labeling fine-grained images necessitates high expert costs to ensure accuracy. Furthermore, collecting sufficient samples for rare fine-grained categories, such as endangered birds [34], rare fish species [35], and disease samples [36], is particularly challenging. Additionally, any change in the recognition target necessitates the reorganization and retraining of existing samples and predefined labels, resulting in resource wastage [8]. Consequently, there has been a growing emphasis on few-shot learning, which involves training a model to recognize new categories not encountered during training using a minimal number of labeled samples [9,10].

2.2. Few-shot learning

Few-shot learning (FSL) seeks to train intelligent models capable of making effective predictions for unseen categories using only a limited number of labeled samples [9,10]. In contrast to traditional fully supervised learning methods, FSL's potential to operate without massive samples and heavy labeling has garnered significant research interest. Moreover, the pursuit of enabling learning systems to rapidly learn in FSL represents an important step in the development of artificial intelligence towards human-like capabilities [12,13].

Existing research on Few-Shot Learning can be broadly categorized into four types. The first is the metric-based approach, which utilizes labeled samples to refine category prototypes, embeds query set samples into the same space, and predicts based on spatial metric results [15,37–41]. The second is the optimization-based approach, which aligns with the concept of "learning to learn" and involves directly generating or updating the weight parameters of convolutional neural networks for new category prediction through fine-tuning on learnable samples [42–45]. The third is the augmentation-based approach, which

addresses the challenge of limited learnable samples by generating additional samples through augmentation operations, such as image transformations, to enhance feature learning [46–48]. The fourth is the emerging reconstruction-based approach, which reconstructs the target sample for prediction based on fragmented features of learnable samples [49–51]. These efforts have generally yielded promising results in cross-semantic coarse-grained image classification, and our work introduces a metric-based approach to expand limited samples with attention-based foreground region refinement for more challenging fine-grained recognition.

Based on metrics, incorporating additional textual prior knowledge to enhance few-shot learning performance has garnered increasing attention. Initially, Chen et al. [52] addressed the visual and semantic discriminative differences among various concepts by adaptively integrate information from both modalities, tailored to the new category being learned. Schwartz et al. [12] further demonstrated that introducing richer semantics, such as category labels, attributes, and natural language descriptions, can significantly improve the prediction accuracy of purely visual networks. These methods set a precedent for combining textual semantics with few-shot visual tasks, but sparse text similarly constrains the training of effective text encoders. In response, Peng et al. [30,53] utilized knowledge graphs sourced from Wikipedia or WordNet [54], in conjunction with word embeddings derived from a pre-trained GloVe [55] text model, to achieve visual-knowledge mapping for predictive assistance. Additionally, Zhang et al. [31] introduced a prototype-completion network designed to enhance visual prototype representations by inferring discriminative features of unseen classes based on prior knowledge of the attributes of visible classes. However, these approaches rely on accurate modeling between categories or attributes, which will be difficult to adapt effectively in fine-grained scenarios especially when confronted with unseen categories.

On the other hand, graph neural networks (GNN) have emerged as a powerful tool in few-shot learning, leveraging the relational structure of data to enhance model performance [56]. Early work by Garcia et al. [57] proposed propagating label information to unlabeled samples via graph inference to adapt to incomplete supervision. Subsequently, Ma et al. [58] expanded on this idea by treating the relationships between samples as graph nodes. Building on these foundational concepts, Yu et al. [59] introduced a hybrid model with two GNNs, including an instance GNN and a prototype GNN, that serve as adaptation modules for feature embeddings, allowing for rapid adaptation to new tasks beyond mere label propagation. In parallel, Cheng et al. [60] addressed common GNN issues of overfitting and oversmoothing by proposing a triple-attention mechanism to improve scalability in few-shot tasks. In this paper, our method leverages the powerful relational modeling capabilities of GNN to aid visual discrimination in few-shot learning as well. In this paper, we aim to utilize the robust relational modeling of GNN to facilitate the spontaneous propagation of category-related semantics among instances, thereby compensating for the insufficient visual discriminability caused by limited samples in few-shot learning.

2.3. Few-shot fine-grained recognition

Few-shot fine-grained recognition was initially proposed by Wei et al. [61] to address the challenges posed by the difficulty of acquiring fine-grained images and the substantial cost of annotation, garnering significant research attention. To address the challenge of discriminating fine-grained categories, Wang et al. [25] proposed capturing foreground targets extensively for learning and recognition purposes, while transforming foreground poses to mitigate the issue of limited samples in few-shot scenarios. Similarly, Zha et al. [15] employed a weakly-supervised method to extract foreground targets, followed by similarity measures for both raw and refined images. Furthermore, Hao et al. [40] underscore the importance of considering the interaction between global and local features, particularly when focusing on local detail features. In response, Tang et al. [23] and Xu et al. [62] approach incrementally enhance the learning of local detail features across various feature scales. And Yu et al. [63] filtered the most similar local patches from the support image and aggregated the patch-level similarities into class prediction scores to reinforce the focus on the discriminative region. Additionally, the alignment of coherent semantic features has garnered significant attention, with both Huang et al. [27] and Zha et al. [15] incorporating alignment supervision of feature embeddings in a richer and more detailed spatial context. These approaches work to discover subtle visual differences in response to fine-grained recognition, but neglect the reliance on prior knowledge, which is important in human perceptual learning. In this paper, our approach integrates inter-image class-consistent semantics with prior knowledge to collaboratively construct robust category prototypes, while emphasizing the enhancement of discriminative region learning.

3. Methodology

We propose a unified deep learning framework to boost few-shot fine-grained recognition through knowledge-driven prototype refinement, referred to as KDPR. Fig. 2 depicts the overall framework, and we will elaborate on the workflow of each component in the following sections.

3.1. Problem formulation

Following the general episode training strategy of few-shot learning [15,27], we divide the dataset into training and testing sets without category overlap, denoted as $\langle I_{train}, C_{train}\rangle$ and $\langle I_{test}, C_{test}\rangle$, where I represents the image, C represents the category label and $C_{train} \cap C_{test} = \phi$. Then, in the training and testing phases, mutually independent episode (D) is constructed, in which the support set (S) and query set (Q) are contained. Following the popular N-way K-shot setup, i.e., the samples in the support and query sets in the same episode 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, the previous S and Q can be further represent as $S = \left\{\left(x_i, \hat{y}_i\right)\right\}_{i=1}^{N\times K}$ and $Q = \left\{\left(x_j, \hat{y}_j\right)\right\}_{j=1}^{N\times q}$, where x is the image input after preprocessing and encoding, $\hat{y}$ is its category label in that episode. As a result, each episode learning task aims to train a robust network that requires the support of only minimal amounts of labeled samples to accurately identify the category of each sample in the query set. 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. With such an episode training paradigm, the feature encoder is forced to learn category-independent discriminative features, and the model remains efficiently adapted to learning and recognizing new categories without additional fine-tuning or training during the testing phase.

3.2. Dual-branch framework with raw image and foreground refined view

When confronted with difficult-to-recognize targets, humans instinctively direct their brain and eyes to focus on and identify potential discriminative regions [64,65]. Driven by this conditioned response, we propose an unsupervised foreground refinement module that leverages feature activation maps to effectively localize foreground regions. Unlike the original image, the refined foreground view eliminates a great deal of extraneous background and focuses only on target-related discriminative regions, which is crucial in fine-grained image recognition [33].

Specifically, Fig. 3 illustrates the workflow of the foreground refinement module. The feature maps of each raw image, encoded by the backbone network (excluding the final global average pooling layer), are utilized as inputs, summed along the channel dimensions, and

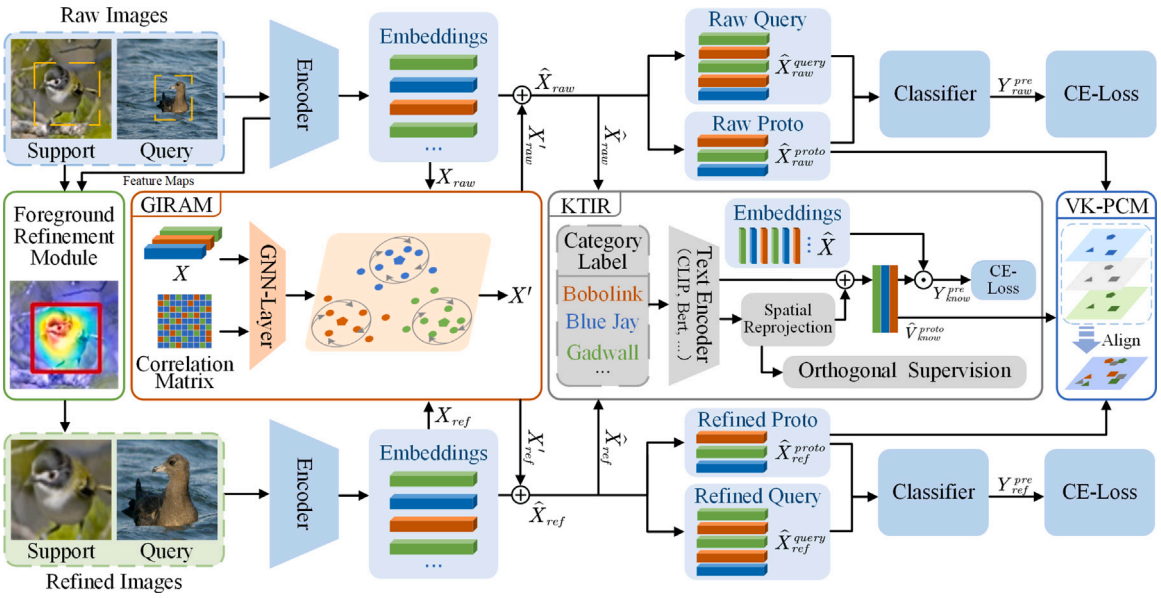

Fig. 2. Overview of the proposed method. First, the refined view obtained by the unsupervised foreground refinement module forms a two-branch network framework with the raw image. Second, the graph-guided instance representation augmentation module (GIRAM) boosts feature encoding by leveraging class-consistency relationships. Following this, metric learning is applied to extract prototypes and make predictions based on the enhanced instances. Additionally, a knowledge transfer-based image recognizer (KTIR) integrates priori textual knowledge for visual recognition. Lastly, the vision-knowledge aligned prototype calibration module (PCM) ensures semantic consistency by aligning prototypes from diverse spaces.

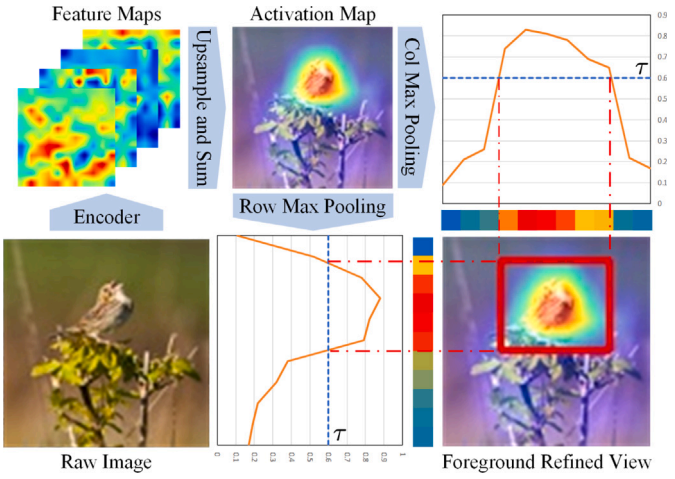

Fig. 3. Foreground refinement module. The feature maps encoded in the raw image are summed over the channel dimension to obtain a category independent activation map, and then the foreground refined region is obtained by locating the most salient regions by pooling over rows and columns.

upsampled to match the input size to generate the activation maps. This map highlights the regions of focus for the model and pinpoints the foreground areas where the target is relevant. Subsequently, the activation maps undergo pooling along both the horizontal and vertical axes, and the largest boundary exceeding the hyperparameter τ (set to 0.4 in this paper) is selected to determine the final foreground refinement region. Given the consistent semantics with the raw image, the refined foreground views are assigned the corresponding category labels.

Furthermore, in the foreground refined view branch, the weight-shared encoder, recognizer, and graph-guided representation argumentation are employed, mirroring their implementation in the raw image input branch. Distinguishing from Yu et al. [63], who disassembled the image into multiple chunks for inter-local dense metrics, this configuration constitutes a two-branch model architecture, as shown in Fig. 2. By substantially expanding the learnable sample space and leveraging semantically consistent yet visually distinct representations between the two branches, this dual-branch approach effectively enhances the model's capacity to learn robust representations.

3.3. Graph-guided Instance Representation Augmentation Module (GIRAM)

One of the primary challenges in few-shot learning is extracting discriminative feature representations with limited labeled samples. Many metric-based methods construct category prototypes from support sets and measure their similarity to query sets, often neglecting the query sets' discriminative potential [19,66]. Meanwhile, studies show that limited samples can lead to biased representations influenced by irrelevant backgrounds or ambiguous details [67,68]. To address these issues, we propose the Graph-guided Instance Representation Augmentation Module. It constructs an unsupervised relational graph for all visual instances within an episode, enhancing category-aware visual representations through the powerful relational modeling capabilities of graph neural networks.

Specifically, Fig. 4 delineates the workflow of GIRAM. The visual embeddings of all instances within an episode serve as inputs, denoted as $X \in R^{v\times d}$, where $v = N \times k + N \times q$ and d represents the encoded feature dimension. In GIRAM, each instance is treated as an autonomous node, and features are augmented by leveraging potential and displayed graph relations. Initially, in the fusion convolution, 1×1 convolutions and nonlinear transformations are executed in the node and channel dimensions, respectively. In this way, by amalgamating the transformed convolutions to discern potentially inherent relationships, even if they are not directly interconnected through the graph structure. The formal expression of the above process is outlined in Eq. (1), where f denotes the 1×1 convolution, W represents its learnable weights, and σ signifies an activation function, such as ReLU.

$$\acute{X} = \sigma\left(f_{1\times1}\left(\sigma\left(f_{1\times1}\left(X, W_{v*v}\right)\right)^T, W_{d\times d}\right)\right)^T \tag{1}$$

Furthermore, to fully exploit the class-consistent semantic connections between nodes, we introduce graph convolution. The primary advantage of graph convolution lies in its ability to leverage the structural information of the graph to update the features of a current node

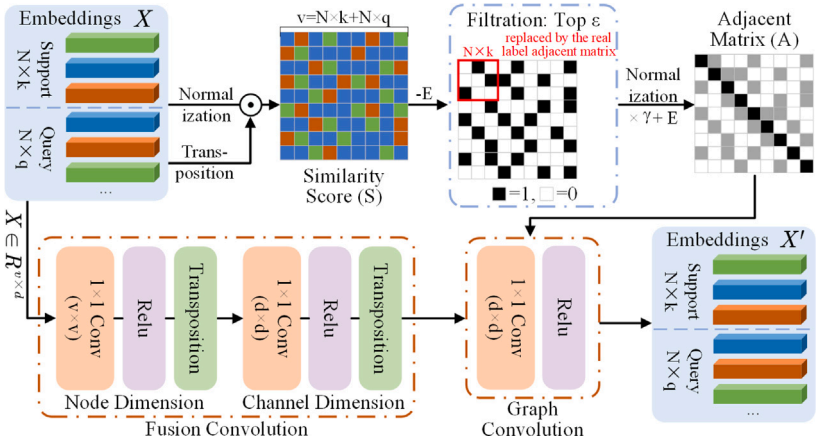

Fig. 4. The architecture of the graph-guided instance representation augmentation module, which constructs adjacency matrices among all samples within an episode and generates class-consistent visual representations through fusion convolution and graph convolution.

by utilizing the features of its neighboring nodes [21,69]. Broadly, the update of a node using a single layer of graph convolution can be formalized as follow:

$$X' = \sigma\left(A\acute{X}W\right), \tag{2}$$

where A is the unsupervised constructed inter-node adjacency matrix, and W represents the learnable weights for training.

Therefore, to derive the inter-node adjacency matrix, we normalize and compute the cosine similarity between all instances, resulting in the similarity scoring matrix S. Subsequently, we set the main diagonal elements of S to zero and select the largest ε values on an instance-wise basis to identify instances belonging to the same class, setting these to 1 to form the binary correlation matrix. It is noteworthy that Fig. 4, indicated by the red box markers, demonstrates how the instances in the support set benefit from the available category labels. Specifically, these labels replace the corresponding portion of the previously constructed binary adjacency matrix, which was based on similarity, with adjacencies constructed using the actual labels.

Furthermore, as Eq. (2) demonstrates, post-graph convolution, node features integrate more information from neighboring nodes. Excessively large fusion weights lead to a loss of diversity due to over-smoothing of features within the same category [69]. To mitigate this issue, we reduce the weight of each neighboring node through additional normalization and weighting, supplementing the identity matrix to preserve the criticality of the instances' inherent features. In summary, the graph matrix representing the class-consistent relationship between nodes can be expressed as follows:

$$A = \gamma \times Filtration(S - E)/\varepsilon + E, \tag{3}$$

where $Filtration()$ is a filtering operation that selects the most similar instances for each node, γ as the weighting parameter empirically set to 0.5, and E denotes the identity matrix.

Finally, the features processed through the graph neural network are residually connected to the original features to mitigate the loss of basic information. The resultant augmented representation, denoted as $\hat{X} = X + X'$, replaces the initial visual embedding and is utilized in subsequent processes. In contrast to FGFD [21], which employs graph neural networks in conjunction with fine-grained descriptions to extract structured information from support set samples, our approach unsupervisedly identifies semantically consistent relationships within full-episode visual samples, thereby enhancing their representational capacity.

3.4. Knowledge Transfer-based Image Recognizer (KTIR)

While GIRAM enhances the visual representation of original instances under unsupervised conditions, generalizing to novel categories through limited visual samples remains challenging. Consequently, much research has focused on incorporating prior knowledge beyond visual images to facilitate few-shot learning [12,21,30,52]. However, these approaches often involve training word embedding networks, which can be difficult and yield minimal gains when using limited samples. Moreover, they require additional retraining or fine-tuning to address biased text encodings when encountering novel categories. To address these issues, we propose the Knowledge Transfer-based Image Recognizer. As shown in Fig. 2, this method embeds category labels using readily available text models like CLIP [70], without additional training or fine-tuning. Our approach transforms these embeddings into visual recognizers that participate in visual discrimination and supervised training for all instances.

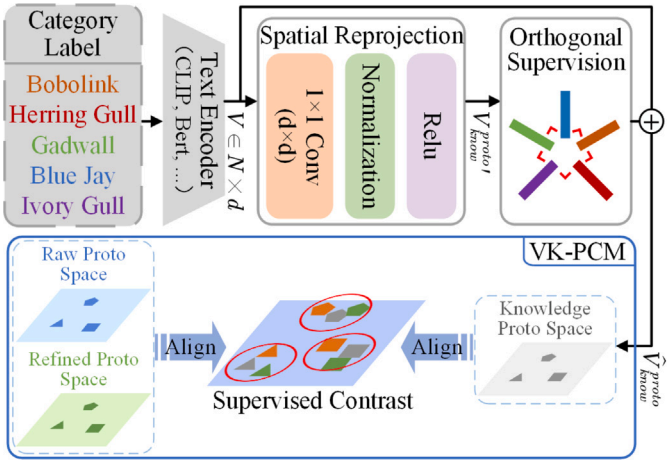

Fig. 5. Illustration of spatial transformation and supervision for knowledge embedding. Our approach executes a nonlinear projection of prior coding and facilitates cross-modality semantic consistency matching through the utilization of orthogonal supervision and contrastive supervision.

To make the text encoder work effectively in our KTIR framework, we input the entire set of category labels from the dataset into the pre-trained text encoder (e.g., CLIP) to obtain prior embeddings that encapsulate global category abstraction semantics, denoted as $V \in R^{C\times d}$, where C is the number of categories and d is the embedding dimension. To ensure compatibility with the visual encoding, we use Principal Component Analysis (PCA) to adjust the embedding dimension of V to match that of the visual encoding. And these category encodings will remain consistent throughout any stage of episode training and testing, ensuring stability and reliability in the representation of category semantics.

Due to the inherent bias in the distribution of embeddings between visual and knowledge modalities, we introduce a supplementary learnable spatial reprojection process. This process, detailed in Eq. (4), involves the application of 1×1 convolution, batch normalization (BN), and activation functions to achieve a learnable nonlinear transformation of the label embeddings.

$$\acute{V}_{know}^{proto} = \sigma\left(BN\left(f_{1\times1}\left(V, W_{1\times1}\right)\right)\right) \tag{4}$$

Following this, Eq. (5) is used to compute the loss of orthogonality between those reprojected prior knowledge embeddings within the same episode to obtain a more independent representation.

$$L_{orth} = \frac{1}{N \times N - N} \cdot \sum_{v \in \acute{V}_{know}^{proto}, i \neq j} \left(v_i \cdot v_j^T\right)^2 \tag{5}$$

To prevent the loss of basic information, we complement the reprojection results with residual connections to the original input, as described in the Eq. (6).

$$\hat{V}_{know}^{proto} = V + \acute{V}_{know}^{proto} \tag{6}$$

Based on the obtained category prototype in the knowledge space, we further transpose it as recognizer weights directly involved in the prediction of visual samples, as outlined in Eq. (7).

$$Y_{know}^{pre} = \hat{X} \cdot \left(\hat{V}_{know}^{proto}\right)^T \tag{7}$$

To ensure effective supervision and fine-tuning of the visual encoder and spatial reprojection, we use cross-entropy loss to supervise the knowledge-based prediction results:

$$L_{know} = -\sum_{D_{raw} \cup D_{ref}} \hat{y} \log\left(softmax\left(y_{know}^{pre}\right)\right), \tag{8}$$

where D_{raw} and D_{ref} denote all the available image samples in the episode under the raw and foreground refined view branches, respectively. And $\hat{y}$ is their corresponding label, which is fully available in the base class training phase. Numerous experiments have demonstrated that reprojection of prior knowledge not only effectively assumes the role of a recognizer. Moreover, benefiting from the stable and robust global semantic energy of prior text encoding, it can provide positive guidance on abstract semantics for learning visual representations.

3.5. Prototype calibration module

In the aforementioned recognizer construction based on knowledge transfer, we derive prototypes of categories within the prior knowledge space ($\hat{V}_{know}^{proto}$). Additionally, with the mean-based prototype construction method of Eq. (9), we can obtain visuospatial prototypes corresponding to the raw image ($\hat{X}_{raw}^{proto}$) and the foreground refinement view ($\hat{X}_{ref}^{proto}$), respectively.

$$\hat{X}_c^{proto} = \frac{1}{|S_c|} \sum_{\hat{X}_i \in S_c} \hat{X}_i \tag{9}$$

Despite the consistent semantics among prototypes belonging to the same category, significant distributional biases are observed across different spaces. Therefore, we propose a visual-knowledge aligned prototype calibration module to unify category prototypes in different visual and knowledge spaces into a single space. At the same time, prior knowledge with global abstract semantics is further exploited to help the model generalize to novel categories and suppress the interference of base category visual representations.

Specifically, in the PCM module illustrated in Fig. 5, the input consists of category prototypes from three distinct spaces within the same episode: the raw image space, the foreground refinement view space, and the prior knowledge space. And the computation of the supervised contrast loss is formalized as follows.

$$L_{sup} = \sum_{i \epsilon H}\left(\frac{-1}{|P(i)|} \sum_{p \epsilon P(i)} \log \frac{\exp\left(z_i \cdot z_p\right)}{\sum_{a \epsilon A(i)} \exp\left(z_i \cdot z_a\right)}\right) \tag{10}$$

In this equation, $H = Set\left(\hat{X}_{raw}^{proto}, \hat{X}_{ref}^{proto}, \hat{V}_{know}^{proto}\right)$ represents the set of all prototypes under different spaces within the episode, $A(i)$ comprises all prototypes excluding prototype i, and $P(i)$ contains only those prototypes that belong to the same class as prototype i. The variable z represents the normalized embedding vector of each prototype, and $\cdot$ denotes the inner product operation between vectors.

The supervised contrast loss function, as presented in [71], aims to enhance consistency among various representations of the same class while facilitating semantic distinctions. In contrast to Song et al. [72] and Liu et al. [73], who use additional image argumentation operations to generate differentiated views, our approach leverages the inherent multi-branching structure to further align the representations of category prototypes across different spaces. Furthermore, instead of comparing multiple views of the same instance, our method solely performs comparisons between category prototypes across spaces, thereby significantly reducing computation while focusing on enhancing the construction of robust prototypes. Specifically, it fosters generalization to scale diversity and background noise suppression in the alignment of raw-refined views, and mitigates cross-modality semantic bias in the alignment of visual-knowledge spaces.

3.6. Prediction and overall loss function

As depicted in Fig. 2, the twinned raw image and foreground refined view branches adhere to a consistent metric learning paradigm. Following the widely used few-shot learning metric approach, the category prediction for each query set sample is achieved by calculating the cosine similarity ($Cos()$) with each mean-based category prototype (Eq. (9)). The formula is

$$y^{pre} = Cos\left(\hat{x}, \hat{x}_c^{proto}\right). \tag{11}$$

During the training phase, the cross-entropy loss is calculated for both the raw image branch and the foreground-refined image branch predictions, utilizing the true labels of the samples, as illustrated in Eqs. (12), (13).

$$L_{raw} = -\sum_{Q} \hat{y}_i \log\left(softmax\left(y_{raw}^{pre}\right)\right) \tag{12}$$

$$L_{ref} = -\sum_{Q} \hat{y}_i \log\left(softmax\left(y_{ref}^{pre}\right)\right) \tag{13}$$

Combined with Eq. (8) for knowledge transfer-based recognizer supervision, the complete prediction loss of our method is $L_{cls} = L_{raw} + L_{ref} + L_{know}$. In addition, by incorporating the orthogonality constraints on spatial knowledge transfer discussed in Section 3.4, and the visual-knowledge prototype corrections described in Section 3.5, the complete supervised expression of our method during the training phase is expressed as Eq. (14).

$$L = L_{cls} + \alpha L_{orth} + L_{sup} \tag{14}$$

In this paper, α is a weighting factor empirically determined through experiments to be 0.4.

During the testing phase, due to the differentiated views and spatially embedded but consistent semantic relations among the three branches, each branch's predictive performance complements the others. Consequently, our final prediction result, which combines the predictions of all branches. As shown in Eq. (15), contains the prediction results for the raw images and foreground refined views based on traditional similarity metrics ($Y_{raw}^{pre}, Y_{ref}^{pre}$) and knowledge transfer based recognizers ($Y_{know-raw}^{pre}, Y_{know-ref}^{pre}$), respectively.

$$Y = Avg\left(Y_{raw}^{pre}, Y_{ref}^{pre}, Y_{know-raw}^{pre}, Y_{know-ref}^{pre}\right) \tag{15}$$

The final predictions following linear integration consistently outperform any single-branch decision, further underscoring the significance of incorporating diverse views and prior knowledge in novel category recognition. In future work, we will explore more optimal multi-branch supervision and fusion techniques to fully leverage the distinct contributions of various views and modalities in supporting few-shot learning.

Table 1
Average accuracy (%) and 95% confidence intervals for 600 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 and ResNet-12. The best results are shown in bold and the suboptimal results are underlined.

Method	Backbone	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 [24]	Conv-64	62.15 ± 0.44	68.05 ± 0.33	–	–	–	–
DLG [26]	Conv-64	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 [27]	Conv-64	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 [15]	Conv-64	65.48 ± 0.51	76.01 ± 0.41	–	–	–	–
BSNET(R&C) [22]	Conv-64	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 [74]	Conv-64	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 [25]	Conv-64	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 [63]	Conv-64	67.75	82.76	55.85	71.78	68.65	87.23
BAMM [19]	Conv-64	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 [28]	Conv-64	73.07 ± 0.46	86.24 ± 0.29	55.53 ± 0.45	71.68 ± 0.36	70.13 ± 0.48	84.29 ± 0.31
RaPSPNet [16]	Conv-64	73.53 ± 0.88	**91.21 ± 0.42**	55.77 ± 0.89	73.58 ± 0.66	71.39 ± 0.87	**92.60 ± 0.81**
Our	Conv-64	**78.85 ± 0.79**	90.72 ± 0.68	**68.88 ± 0.77**	**78.08 ± 0.63**	**75.86 ± 0.76**	89.10 ± 0.46
ProtoNet [37]	ResNet-12	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 [27]	ResNet-12	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 [41]	ResNet-12	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) [22]	ResNet-12	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 [75]	ResNet-12	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 [28]	ResNet-12	77.77 ± 0.44	89.87 ± 0.24	64.15 ± 0.49	78.28 ± 0.32	77.03 ± 0.46	88.85 ± 0.46
FGFD [21]	ResNet-34	–	–	65.88 ± 0.15	79.74 ± 0.46	81.55 ± 0.21	92.75 ± 0.10
AGPF [23]	ResNet-12	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 [24]	ResNet-12	79.48 ± 0.47	85.25 ± 0.25	–	–	–	–
BSFA [15]	ResNet-12	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 [66]	ResNet-12	82.45 ± 0.79	**92.02 ± 0.44**	73.75 ± 0.93	**86.65 ± 0.54**	87.27 ± 0.70	95.01 ± 0.49
RWFE [29]	ResNet-12	84.62 ± 0.89	90.08 ± 0.97	71.02 ± 0.84	83.44 ± 0.63	–	–
Our	ResNet-12	**86.56 ± 0.68**	91.70 ± 0.51	**80.44 ± 0.71**	84.97 ± 0.57	88.03 ± 0.59	93.17 ± 0.41

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 settings while maintaining a query set count of 15 for each category [15,27]. 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. [15], decreasing to 0.06 at the 45th epoch and multiplying by 0.2 every ten epochs. In the testing phase, we froze the learnable weights associated with the feature extractor and spatial reprojection, adopted the same episode delineation strategy as in training, and recorded the average accuracy within the 95% confidence intervals of 600 episodes as the final result.

4.2. Datasets

We conducted experiments on three mainstream fine-grained classification datasets, including CUB-200-2011 [76], Stanford Dogs [77], and Stanford Cars [78], which are widely accepted as few-shot fine-grained recognition validation.

(1) CUB-200-2011 [76]*:* 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 [77]*:* 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 [78]*:* 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.

4.3. Results

Table 1 presents the results of our method, utilizing the widely adopted Conv-64 and ResNet-12 backbones, in comparison with other state-of-the-art methods across three mainstream fine-grained datasets. Our method shows consistently best performance on most metrics across two backbones and multiple datasets, demonstrating robust representation learning capabilities. Especially in the 1-shot condition, where the visual sample is significantly limited, our method shows remarkable advantages due to the effective utilization of prior knowledge. Subsequently, we will provide detailed comparisons of the specific performance under the Conv-64 and ResNet-12, respectively.

(1) Comparison with methods based Conv-64. The upper part of Table 1 documents the accuracy of our method compared to eleven representative few-shot fine-grained recognition methods with Cove-64 as the backbone. Our method demonstrates significantly superior performance in the 1-shot condition, achieving improvements of 7.3%, 23.5%, and 6.3% over the next best method on the CUB-200-2011, Stanford Dogs, and Stanford Cars datasets, respectively. However, in the 5-shot condition, with more visual samples, our method's performance improvement is slightly less impressive but remains competitive, trailing slightly behind RaPSPNet [16] on the CUB-200-2011 and Stanford Cars. This is attributed to the observation that, as shown in Figs. 6 and 7, the discriminative performance gains of each branch under the 5-shot condition indicate a diminishing return from additional prior knowledge as the number of visual samples increases in few-shot learning. Furthermore, when compared to BSFA [15], which also employs a two-branch structure and introduces additional global category labeling, our approach demonstrates significantly better and more stable performance. This is accomplished by encoding the inherent category labels using an accessible text encoder and integrating visual recognition capabilities. This integration compensates for the limited comprehension of global abstractions that a few-shot learner has regarding the recognized target.

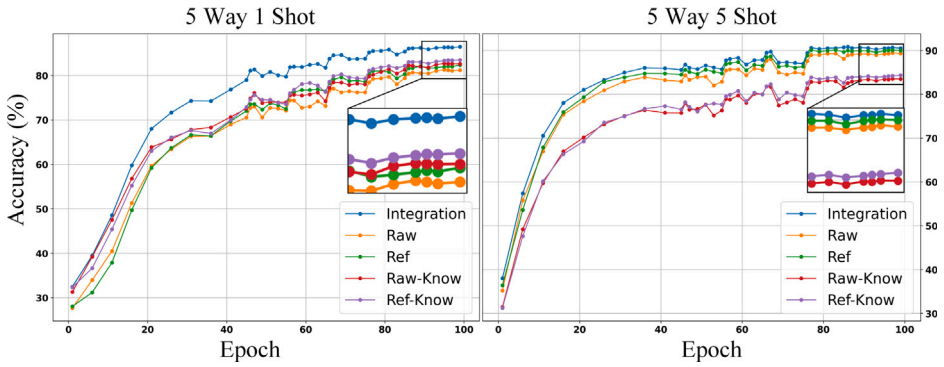

Fig. 6. Accuracy curves for each prediction branch for 5-way 1-shot and 5-way 5-shot tests on CUB-200-2011. Obviously, the different branches exhibit significant stratification, and the prediction of fusing all branches (Integration) exhibits significantly optimal accuracy. Specifically, foreground refinement views outperform raw images for classification by either vision (Ref-Vis) or knowledge (Ref-Know); prediction by vision (Raw-Vis,Ref-Vis) improves significantly as the number of learnable samples increases.

(2) Comparison with methods based ResNet-12. The second half of Table 1 presents the performance of our method using the more structurally complex ResNet-12 backbone, in comparison to twelve representative few-shot fine-grained recognition methods. Thanks to the enhanced coding and learning capabilities of the backbone network, all studies demonstrate significant improvements in both 1-shot and 5-shot recognition accuracy. Consistent with the Conv-64 backbone, our methods achieve competitive performance across most metrics, particularly exhibiting notable advantages in 1-shot scenarios. Specifically, in contrast to RWFE [29], which also emphasizes foreground target regions and prototype correction, our method eliminates the need for single modality and additional large-scale pre-training, achieving superior performance on all metrics by simply incorporating a priori text encoding to facilitate the learning of robust representations. Additionally, our method exhibits a significant disadvantage in 5-shot comparisons with RSaD [66]. This is attributed to the introduction of additional salient target detection in RSaD, which masks the entire background. Consequently, it simplifies the model's prediction task, particularly in 5-shot scenarios with a relatively larger number of visual samples. Conversely, RSaD's performance in 1-shot scenarios is less advantageous due to the scarcity of visual samples, whereas our method benefits from the incorporation of a priori knowledge, leading to significant advantages. When compared to BSFA [15], our approach performs poorly on the Stanford Cars, a dataset that contains a large number of artificial images and lacks explicit abstract semantics. However, considering all three datasets and both backbones, our method exhibits a more stable and consistent state-of-the-art performance. In comparison to FGFD [21], which also leverages graph neural networks, our method demonstrates significantly superior performance, even with a weaker backbone. This is due to our design, which facilitates knowledge transfer across multiple dimensions beyond adjacency, including spatial and channel dimensions. Additionally, our method pays special attention to foreground refinement regions and incorporates prior knowledge, collectively contributing to effective gains in accurate recognition. To address the shortcomings of our approach in 5-shot scenarios, we are exploring more optimal visual and textual knowledge fusion methods that balance the growing visual and knowledge space in few-shot learning, aiming to fully exploit the potential of a priori abstract semantic gains for visual discrimination.

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

(1) Visualization of Activation Maps and the Foreground Refinement. Fig. 8 visualizes our method for predicting episodes in a novel category following 5-way 1-shot training. The first row of each dataset shows the activation map of raw images after backbone encoding and feature map stacking. The highlighted thermal regions indicate the areas the model focuses on during prediction. Comparing the attention heatmaps across multiple datasets, it is clear that our method effectively captures almost all discriminative regions, such as the local details of a bird, a dog's face, and a car's head. Specifically, taking the performance in the CUB-200-2011 dataset as an example, the model primarily focuses on the foreground target region, which is a small fraction of the raw image, as seen in samples (a), (c), and (f). And targets initially obscured in diverse backgrounds are extracted and enlarged to form the second row of foreground refinement views.

A comparison of the foreground refined views of different samples reveals that targets of various categories are further unified in scale and location, with significant suppression of irrelevant background interference, thereby enhancing discriminative feature learning. Additionally, a comparison of the activation maps of raw images and foreground refined views of the same sample demonstrates that the model dynamically adjusts its attention region based on target scale changes, with notable examples shown in Fig. 8(a, b, g). Notably, under the foreground refined view, the model places greater emphasis on discriminative features, such as the head of a bird in (c) and (e), due to larger and sharper visual representations. These local details are crucial for fine-grained recognition. Even when focusing on the target as a whole, the activation maps on fine-grained views exhibit more precise contour boundaries, as exemplified by (a) and (g). Furthermore, the consistently superior accuracy performance of the foreground refined view branch in Fig. 6 compared to the raw image branch underscores the significance of foreground refinement.

(2) Effectiveness of Multi-Branch Inference and Integration. Fig. 6 illustrates the predictive accuracy of various inferencing branches on the CUB-200-2011 with ResNet-12. In scenarios with limited image samples (1-shot), knowledge-based predictions exhibit subtle yet discernible advantages over visual spatial metrics, highlighting the importance of incorporating additional knowledge priors for category discrimination when visual information is insufficient. As the number of visually learnable samples increases (5-shot), visually-based predictions show significant improvements, while knowledge-based predictions only exhibit subtle enhancements. This is because the text encoding model is frozen, and the introduction of additional samples contributes little to the knowledge transfer-based recognizer. On the other hand, refined foreground views outperform raw images in predictive performance, benefiting from the removal of irrelevant background information and enhanced learning of foreground discriminative regions. The ablations presented in Table 2(*a–c*, *b–d*) quantify the contribution contributed by the foreground refinement branch to fine-grained recognition. Lastly, the straightforward fusion of prediction results from the four branches yields consistent gains, particularly under the 1-shot condition where

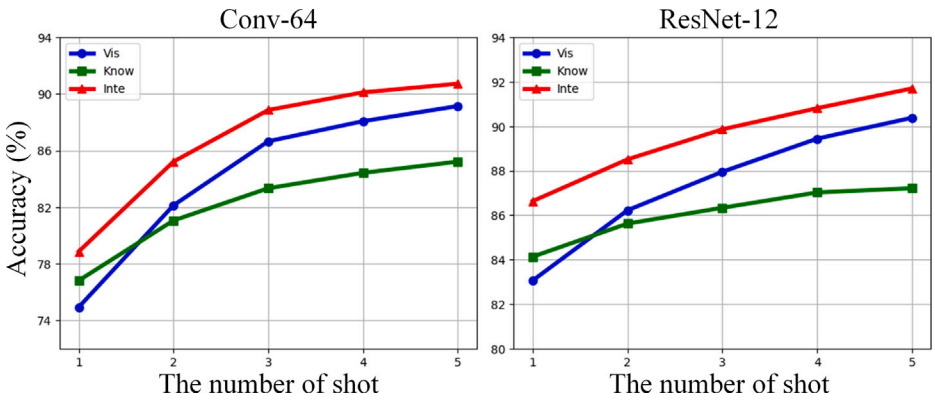

Fig. 7. Comparison of accuracy from 5-way 1-shot to 5-way 5-shot under two backbones on CUB-200-2011. Analysis of the performance trends reveals that the accuracy of vision-based (Vis) classification improves significantly as the number of shots increases, whereas the performance of knowledge transfer-based (Know) classification stabilizes. The integration of vision and knowledge (Inte) results in consistently superior performance, albeit with gradually diminishing benefits.

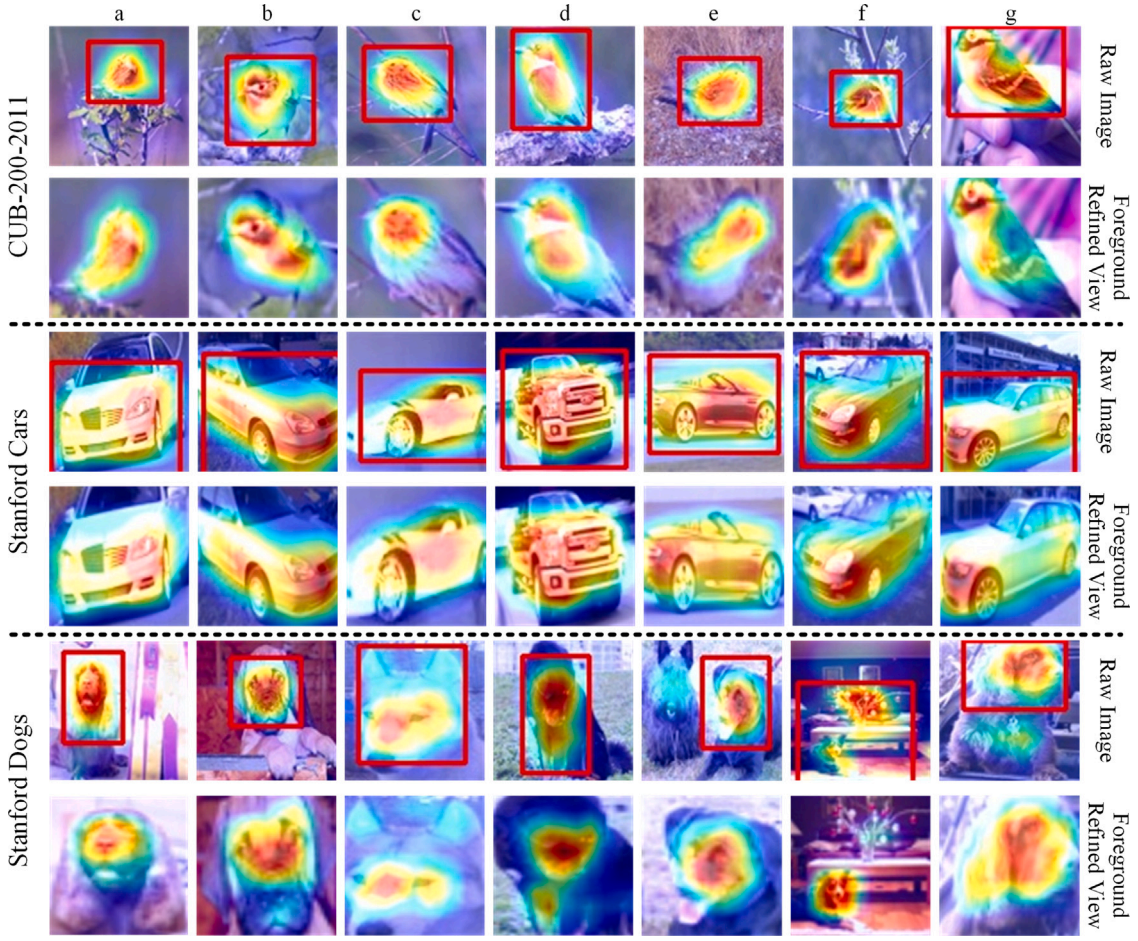

Fig. 8. The visualization of activation maps for both the raw image and the foreground-refined view in the 5-way, 1-shot condition, along with the region proposal (red box) generated by the foreground refinement module.

visual samples are scarce. This underscores the complementarity of original images and refined foreground views in terms of scale diversity and global-local perspectives, as well as the effective utilization of prior knowledge for visual category discrimination.

Additionally, Fig. 7 combines the dual-view predictions for visual and knowledge-based branching separately and presents finer-grained accuracy curves that illustrate changes in the number of learnable samples. The notable enhancement in visual prediction performance as the number of learnable samples increases, coupled with the stable performance of the knowledge branch and the consistently superior performance when both branches are combined, further support the aforementioned analysis.

(3) Effectiveness of Graph-guided Instance Representation Augmentation Module (GIRAM). Table 2 quantitatively demonstrates the impact of introducing the GIRAM on the model's prediction accuracy. In this experiment, to avoid unnecessary interference, we omitted the knowledge transferred image recognition and prototype calibration. We evaluated GIRAM's effectiveness by supplementing it with a single raw image branch and a raw-refinement dual branch, respectively. It can

Table 2
Accuracy comparison for the CUB-200–2011 dataset using ResNet-12, assessing the impact of foreground refinement view branching (Ref) and Graph-Guided Instance Representation Augmentation Module (GIRAM).

No	Raw	Ref	GIRAM	CUB-200–2011	
				5-way 1-shot	5-way 5-shot
a	✓	✗	✗	72.60 ± 1.02	81.00 ± 0.62
b	✓	✗	✓	80.44 ± 0.88	88.26 ± 0.43
c	✓	✓	✗	75.14 ± 0.92	83.68 ± 0.65
d	✓	✓	✓	**82.26 ± 0.86**	**90.54 ± 0.55**

Table 3
Accuracy comparison for the CUB-200–2011 dataset using ResNet-12, assessing the impact of Fusion Convolution (FC) and Graph Convolution (GC) and their order in the Graph-guided Instance Representation Augmentation Module (GIRAM).

GIRAM		CUB-200–2011	
FC	GC	5-way 1-shot	5-way 5-shot
✗	✗	75.14 ± 0.92	83.68 ± 0.65
✗	✓	80.23 ± 0.69	89.17 ± 0.53
✓	✗	78.57 ± 0.73	85.33 ± 0.58
✓(First)	✓(Second)	**82.26 ± 0.86**	**90.54 ± 0.55**
✓(Second)	✓(First)	81.03 ± 0.89	89.73 ± 0.55

Table 4
Accuracy and parameter comparison for the CUB-200–2011 dataset using ResNet-12 to assess the impact of the Knowledge-Transferred Image Recognizer (KTIR) and Prototype Calibration Module (PCM). Additionally, the effects of various loss supervision methods in KTIR are examined, including distance (Dis), cosine similarity (Sim), and orthogonal loss (Orth).

KTIR	PCM	CUB-200–2011		Params
		5-way 1-shot	5-way 5-shot	(M)
✗	✗	82.26 ± 0.86	90.54 ± 0.55	8.57
✗	✓	82.71 ± 0.81	90.89 ± 0.55	8.57
✓(Orth)	✗	86.46 ± 0.69	91.22 ± 0.51	8.83
✓(Orth)	✓	**86.56 ± 0.68**	**91.70 ± 0.51**	8.83
✓(Dis)	✓	85.27 ± 0.71	90.86 ± 0.55	8.83
✓(Sim)	✓	86.25 ± 0.69	91.33 ± 0.51	8.83

be intuitively seen that the introduction of GIRAM significantly improves the prediction performance based on the visuospatial similarity metric. This improvement stems from GIRAM's ability to aggregate all visual samples for category-consistent knowledge transfer, enabling the utilization of discriminative visual representations in more query set samples. Parts c and d of Fig. 9 illustrate the tighter aggregation of visual representations after GIRAM processing, further validating its utility.

Furthermore, we validate the roles of fusion convolution and graph convolution within GIRAM, as well as the impact of their order, in Table 3. It is evident that the prediction accuracies are enhanced due to the transfer and fusion of additional visual representations from the query set, which are aligned with category-consistent semantics. In contrast to FC, which adaptively fuses features in both node and channel dimensions, GC, guided by the adjacency matrix, facilitates better convergence of category-consistent visual representations among nodes, resulting in a notable improvement. Additionally, the semantic transfer process, which involves initial fusion followed by graph-based class guidance, yields superior results. We attribute this to the effective enhancement of instance representation through the prior fusion of node and channel dimensions, which allows for the subsequent derivation of more discriminative category representations through graph-based guidance, thereby facilitating recognition.

(4) Effectiveness of Knowledge Transfer-Based Image Recognizer (KTIR) and Prototype Calibration Module (PCM). Based on the dual-branch network architecture and graph-guided instance representation augmentation, Table 4 presents an ablation study on the image recognizer with knowledge transfer and the prototype calibration module. Intuitively, the integration of recognizers based on prior abstract semantics, alongside traditional visuospatial metrics, leads to substantial gains in predictive performance while incurring minimal parameter costs. This improvement is particularly pronounced in the 1-shot scenario, where visual samples are scarce. Furthermore, as illustrated in Fig. 6, the prediction of the recognizer based on knowledge transfer demonstrate consistent or even superior accuracy compared to visual spatial similarity metrics, for both raw images and foreground refined views. This demonstrates that robust and abstractly prior knowledge can assist in the extraction of category-robust visual features and achieve category discrimination when involved in visual recognition feedback.

On the other hand, the PCM further enhances the overall predictive performance by performing supervised contrastive learning on category prototypes from different spaces. As depicted in the t-SNE visualizations transitioning from c, d to f in Fig. 9, it is evident that the mutual supervision and rectification of these prototypes lead to clear clustering of samples belonging to different categories across various spaces.

While retaining KTIR and PCM, we explored alternative metric supervision methods for the spatial reprojection of text encodings within KTIR, specifically Euclidean distance (Dis) and similarity (Sim) metrics. Unlike the orthogonal supervision (Orth) outlined in Eq. (5), the euclidean distance metric loss Eq. (16) seeks to maximize inter-category vector distances within euclidean space, thereby sharpening categorical distinctions. Conversely, the cosine similarity metric Eq. (17) minimizes directional similarity between vectors, ensuring divergent orientations for different categories to prevent overlap. Although both euclidean distance and cosine similarity metrics yield performance improvements, orthogonal supervision excels in promoting category separation, outperforming similarity metrics, with distance metrics demonstrating the least effectiveness.

$$L_{Dis} = \frac{1}{N \times N - N} \cdot \sum_{v \in \hat{V}_{know}^{proto}, i \neq j} \left| v_i - v_j \right|^2 \tag{16}$$

$$L_{Sim} = \frac{1}{N \times N - N} \cdot \sum_{v \in \hat{V}_{know}^{proto}, i \neq j} \left(1 + \frac{v_i \cdot v_j}{|v_i||v_j|} \right) \tag{17}$$

(5) T-SNE Visualization of Embedding Spaces Across Various Branches and Stages. Fig. 8 presents the visualization of t-SNE [79] for our approach to sample encoding across various branches and stages within the same episode test. Specifically, subfigures (a) and (b) depict the spatial distributions of raw images and foreground-refined views, respectively, after encoding using shared weights and mean-based prototype construction. Notably, there is a clear aggregation of individual class prototypes with query set samples, while different classes maintain a distinct separation. Additionally, a comparison between (a–c) and (b–d) reveals a tighter aggregation among samples of the same class following representation enhancement. This enhancement is facilitated by graph-guided propagation of semantic knowledge between instances, enabling the condensation of more generalized features.

Moreover, subfigure (e) illustrates the visual embedding of the augmented raw image, foreground-refined view, and prior knowledge embedding that has not undergone spatial reprojection into the same space. Compared to the spatial distribution of single raw images and foreground-refined views, the category distribution after multi-space convergence exhibits more significant differentiation. This accounts for the stability observed in the simple fusion of multibranch prediction results compared to single-branch predictions. However, an apparent inconsistency in the correspondence between the spatial distribution of views and prior knowledge is also evident. This inconsistency arises because the prior knowledge is encoded from an external, independent knowledge model and is not adaptively fine-tuned during the training process of our model. Despite its use as a recognizer in our model training to influence the learning of the visual coder, it is challenging to bridge the substantial semantic distribution gap between the prior linguistic space and our visual embedding space. In this context, the introduction of nonlinear transformations of prior knowledge can significantly aid its alignment with the visual space, as demonstrated by the contrast in (e–g). In this respect, we argue that the potential role of

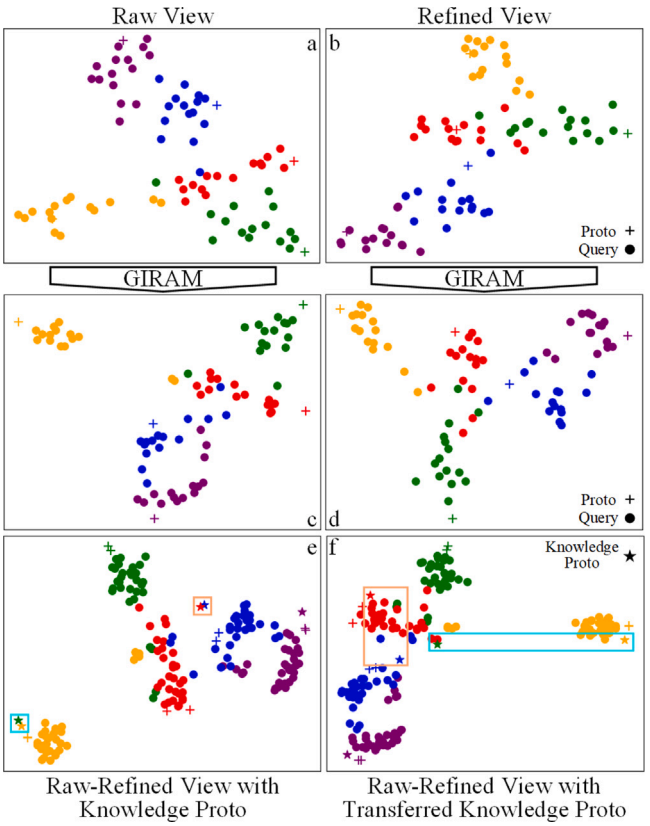

Fig. 9. Visualization of t-SNE for different branches under the same episode prediction, with the same colors representing the same categories. (a) and (b) contain only visual prototypes and query set sample embeddings for the raw image and foreground refined view, respectively. (c) and (d) are distributions that exhibit further clustering after graph-guided representation augmentation. (e) embeds the raw image, the foreground refined view, and prior knowledge that has not undergone spatial transfer in the same space; (f) replaces the spatially transferred prior knowledge compared to that of (e). Better for color.

prior semantic knowledge extends beyond this, and its more effective and graceful exploitation will continue to guide our future explorations.

(6) Sensitivity Analysis. The GIRAM described in Section 3.3 introduces an additional hyperparameter, ϵ, which governs the number of nodes associated in the construction of the adjacency matrix. Fig. 10 presents a comparison of our test results using various ϵ values for few-shot learning on the CUB-200-2011. Notably, the effect of ϵ is more pronounced in the 1-shot condition. This is because the increased number of visual samples in the 5-shot results in improved baseline accuracy, making it easier to obtain a more accurate adjacency matrix that effectively guides the enhancement process. Conversely, the model achieves optimal performance when ϵ adopts moderately centered values (e.g., 8 and 12). This is due to the fact that too few neighboring nodes can hinder the propagation of knowledge between nodes, whereas an excessively large number may introduce more misclassifications, resulting in suboptimal representation learning.

Furthermore, we evaluate the impact of various weight allocations of orthogonal losses within the overall loss function in Fig. 10. Specifically, optimal performance is attained at 1-shot and 5-shot settings with weights of 0.4 and 0.6, respectively. Regarding the trend of the curve, the effect of different orthogonal loss weights is relatively minor at the 5-shot setting. In contrast, visual samples, which are substantially constrained under the 1-shot condition, exhibit sensitivity to orthogonal loss, highlighting the importance of aligning a priori knowledge semantics with visual representations.

(7) Model Complexity Analysis. Apart from prediction accuracy, floating-point operations (FLOPs) and parameters (Params) are also

Table 5
Comparison of accuracy, parameters (Params) and floating point operations (FLOPs) of our individual modules based on ResNet-12 on CUB-200–2011 dataset in 5-way 1-shot 15-query setting.

Raw	Ref	GIRAM	KTIR	PCM	CUB-200–2011	Params	FLOPs
					5-way 1-shot	(M)	(G)
✓					72.60 ± 1.02	8.03	135.089
✓	✓				75.14 ± 0.92	8.03	270.087
✓	✓	✓			82.26 ± 0.86	8.57	270.177
✓	✓	✓	✓		86.46 ± 0.69	8.83	270.179
✓	✓	✓	✓	✓	86.56 ± 0.68	8.83	270.179

Table 6
Comparison of accuracy using Conv-64 and CUB-200–2011 datasets with different text types (Word: Category label, Description: Text description) and various pre-trained text encoders.

Knowledge	Encoder	Dimension	CUB-200–2011	
type			5-way 1-shot	5-way 5-shot
Word	CLIP	512	78.85 ± 0.79	90.72 ± 0.68
Word	GLOVE 6B	300	70.14 ± 0.92	76.31 ± 0.84
Word	GLOVE 42B	300	70.56 ± 0.94	78.72 ± 0.80
Word	BERT	768	72.37 ± 0.95	82.44 ± 0.76
Description	CLIP	512	78.93 ± 0.76	90.66 ± 0.68
Description	GLOVE 6B	300	65.24 ± 1.19	71.44 ± 0.97
Description	GLOVE 42B	300	68.36 ± 1.02	73.51 ± 0.89
Description	BERT	768	65.72 ± 1.00	76.23 ± 0.82

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 5 documents the accuracy, parameter, and floating-point operation performance at 5-way 1-shot as we increment our method module by module, starting from a base network with a single raw image branch.

It is evident that the introduction of the foreground refinement view (Ref) significantly enhances accuracy but nearly doubles the floating-point computation compared to the basic single-branch, purely visual few-shot image recognition baseline (Raw). To address this, we will continue to explore more efficient two-branch architectures that retain the accuracy gains while reducing computational demands. Additionally, the introduction of GIRAM markedly improves the model's accuracy with only a modest increase in graph convolution-related optimization parameters and computations. Similarly, the simple predictor implementation of KTIR, which introduces prior knowledge without recoding the text, achieves significant accuracy gains with minimal overhead. Orthogonal supervision (PCM) of the reprojection further enhances overall model performance with negligible added complexity. Overall, these enhancements achieve substantial performance improvements while maintaining computational efficiency.

5. Discussion of different text embeddings

In this study, we examine the impact of two forms of prior knowledge: text descriptions and category labels. We assess their performance using various text encoders. For our analysis, we employ the CUB-200-2011, with text descriptions sourced from Scott et al. [80]. Specifically, we select the longest text description per category as the prior knowledge. Fig. 11 illustrates these descriptions, which encompass detailed category-related attributes.

We utilize three text encoders: CLIP [70], BERT [81], and GLOVE [55], and all text encoding weights are sourced from the official public

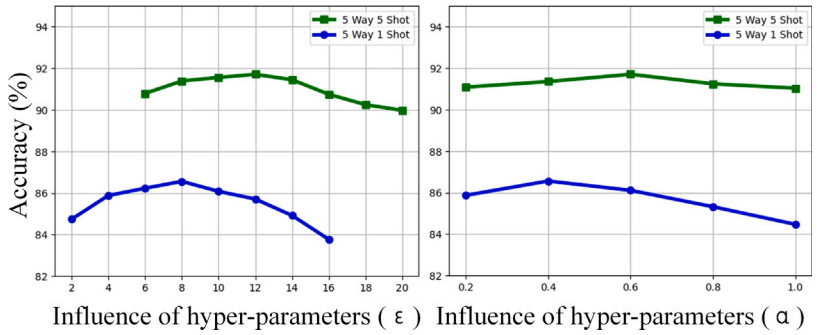

Fig. 10. Analysis of the hyperparameters of our method under Resnet-12 and CUB-200-2011: the number of neighboring nodes in the construction of the adjacency matrix (ϵ) and the hyperparameters in the final loss function (α).

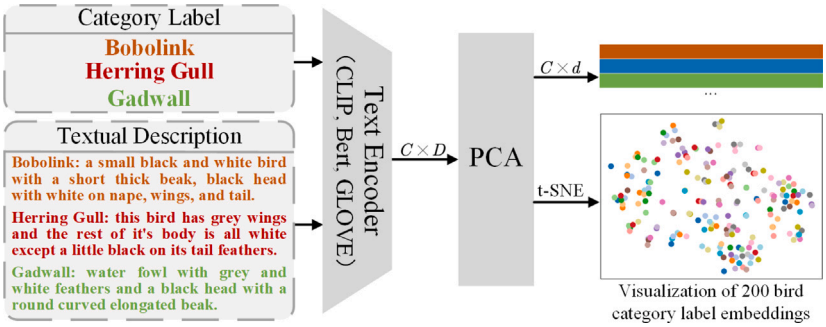

Fig. 11. Illustration of different types of text inputs and our simple prior knowledge processing. The visualization of t-SNE with uniform distribution across categories demonstrates that the encoding of pre-trained text encoders can quickly endow individuals with prior knowledge that implies global abstract semantics.

versions.[1] CLIP is a state-of-the-art model that leverages a Vision Transformer (ViT) for image encoding and a masked self-attention Transformer for text encoding, trained via a contrastive loss to align visual and textual information effectively. BERT is a powerful transformer-based model pretrained on a large corpus of English text using a masked language modeling objective, capturing bidirectional context in text but lacking cross-modal understanding. GLOVE is an unsupervised learning algorithm that generates word embeddings by training on aggregated global word–word co-occurrence statistics, capturing semantic relationships between words but lacking deep contextual understanding. To align with the dimensionality of the visual coding branch, we uniformly reduce the output dimensions of these encoders using PCA.

With Conv-64 serving as the backbone, Table 6 illustrates the few-shot learning performance of our method across various knowledge types and text encoders. It is evident that, when comparing different encoders, CLIP achieves significantly optimal performance regardless of the input type, owing to its cross-modal comprehension capability. Conversely, text-only encoders such as BERT and GLOVE perform poorly in meeting visual discrimination requirements due to their lack of cross-modal generalization. Furthermore, using more detailed textual descriptions in both the 1-shot setting and with CLIP resulted in additional accuracy improvements, reinforcing the substantial benefit of prior semantics for few-shot visual recognition. However, other text-only encoders surprisingly exhibited counterproductive results when encoding text descriptions. Given the significant intra-class diversity and inter-class similarity in fine-grained categories, brief text descriptions (approximately 15 words) may introduce noise, interfering with accurate category embedding compared to conceptualized category labels [30]. Additionally, simple and readily available category annotations are more economical and versatile than additional expert annotations, especially considering the limited accuracy improvement. Therefore, this paper focuses on utilizing category labels as a carrier of prior knowledge to facilitate few-shot learning in computer vision.

[1] CLIP: https://huggingface.co/openai/clip-vit-base-patch32; BERT: https://huggingface.co/google-bert/bert-base-cased; GLOVE: https://nlp.stanford.edu/projects/glove.

6. Conclusion

This paper proposes a knowledge-driven prototype refinement framework (KDPR) for boosting few-shot fine-grained recognition. To overcome the challenges posed by fine-grained categories and limited learnable samples, we initially employ visual attention to extract discriminative foreground regions and establish a foreground refinement view branch. This approach not only diversifies the scale of learnable samples but also augments the learning of fine-grained discriminative features. Further, we utilize graph neural networks to convey category-consistent semantics in episodes, mining potential knowledge from the visual representation of query sets to facilitate the construction of more robust prototypes and query instances. To fully exploit available prior knowledge, we propose a knowledge transfer-based image recognizer and prototype calibration module, directly applying text encoding results to visual classification and semantic alignment. Extensive experiments and visualizations conducted on three datasets and two backbones validate the effectiveness of our proposed method. Notably, under the highly constrained 1-shot setting with minimal learnable samples, our knowledge-driven approach demonstrates substantial improvements. Considering the shortcomings in the 5-shot condition, we will continue to explore more optimal visual and prior knowledge fusion methods to balance the semantic guidance gain under increased visual samples.

CRediT authorship contribution statement

Jiale Chen: Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Conceptualization. **Feng Xu:**

Writing – review & editing, Supervision, Project administration, Funding acquisition. **Xin Lyu:** Writing – review & editing, Supervision, Investigation. **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 Natural Science Foundation of China [grant number 62401196]; the Natural Science Foundation of Jiangsu Province, China [grant number BK20241508].

Data availability

Data will be made available on request.

References

[1] J. Wang, W. Li, M. Zhang, R. Tao, J. Chanussot, Remote-sensing scene classification via multistage self-guided separation network, IEEE Trans. Geosci. Remote Sens. 61 (2023) 1–12, http://dx.doi.org/10.1109/TGRS.2023.3295797.

[2] I. Dimitrovski, I. Kitanovski, D. Kocev, N. Simidjievski, Current trends in deep learning for earth observation: An open-source benchmark arena for image classification, J. Photogramm. Remote. Sens. 197 (2023) 18–35, http://dx.doi.org/10.1016/j.isprsjprs.2023.01.014.

[3] S. Liu, A. Ma, S. Pan, Y. Zhong, An effective task sampling strategy based on category generation for fine-grained few-shot object recognition, Remote. Sens. 15 (6) (2023) http://dx.doi.org/10.3390/rs15061552.

[4] M. Abdar, M.A. Fahami, L. Rundo, P. Radeva, A.F. Frangi, U.R. Acharya, A. Khosravi, H.-K. Lam, A. Jung, S. Nahavandi, Hercules: Deep hierarchical attentive multilevel fusion model with uncertainty quantification for medical image classification, IEEE Trans. Ind. Inform. 19 (1) (2023) 274–285, http://dx.doi.org/10.1109/TII.2022.3168887.

[5] H. Zhang, R. Song, L. Wang, L. Zhang, D. Wang, C. Wang, W. Zhang, Classification of brain disorders in rs-fMRI via local-to-global graph neural networks, IEEE Trans. Med. Imaging 42 (2) (2023) 444–455, http://dx.doi.org/10.1109/TMI.2022.3219260.

[6] X.-S. Wei, Y.-Z. Song, O.M. Aodha, J. Wu, Y. Peng, J. Tang, J. Yang, S. Belongie, Fine-grained image analysis with deep learning: A survey, IEEE Trans. Pattern Anal. Mach. Intell. 44 (12) (2022) 8927–8948, http://dx.doi.org/10.1109/TPAMI.2021.3126648.

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

[8] Z. Fang, X. Wang, H. Li, J. Liu, Q. Hu, J. Xiao, FastRecon: Few-shot industrial anomaly detection via fast feature reconstruction, in: International Conference on Computer Vision, ICCV, 2023, pp. 17435–17444, http://dx.doi.org/10.1109/ICCV51070.2023.01603.

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

[10] Y. Song, T. Wang, P. Cai, S.K. Mondal, J.P. Sahoo, A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities, ACM Comput. Surv. 55 (13) (2023) http://dx.doi.org/10.1145/3582688.

[11] B. Hariharan, R. Girshick, Low-shot visual recognition by shrinking and hallucinating features, in: International Conference on Computer Vision, ICCV, 2017, pp. 3037–3046, http://dx.doi.org/10.1109/ICCV.2017.328.

[12] E. Schwartz, L. Karlinsky, R. Feris, R. Giryes, A. Bronstein, Baby steps towards few-shot learning with multiple semantics, Pattern Recognit. Lett. 160 (C) (2022) 142–147, http://dx.doi.org/10.1016/j.patrec.2022.06.012.

[13] J. Lu, P. Gong, J. Ye, J. Zhang, C. Zhang, A survey on machine learning from few samples, Pattern Recognit. 139 (2023) 109480, http://dx.doi.org/10.1016/j.patcog.2023.109480.

[14] J. Ren, C. Li, Y. An, W. Zhang, C. Sun, Few-shot fine-grained image classification: A comprehensive review, AI 5 (1) (2024) 405–425, http://dx.doi.org/10.3390/ai5010020.

[15] Z. Zha, H. Tang, Y. Sun, J. Tang, Boosting few-shot fine-grained recognition with background suppression and foreground alignment, IEEE Trans. Circuits Syst. Video Technol. 33 (8) (2023) 3947–3961, http://dx.doi.org/10.1109/TCSVT.2023.3236636.

[16] W. Zhang, Y. Zhao, Y. Gao, C. Sun, Re-abstraction and perturbing support pair network for few-shot fine-grained image classification, Pattern Recognit. 148 (2024) 110158, http://dx.doi.org/10.1016/j.patcog.2023.110158.

[17] L. Li, J. Deng, Y. Huang, Y. Chen, W. Luo, Structural subspace learning for few-shot fine-grained recognition, in: International Conference on Machine Learning and Computing, Association for Computing Machinery, New York, NY, USA, 2024, pp. 693–699, http://dx.doi.org/10.1145/3651671.3651676.

[18] B. Zhang, J. Yuan, B. Li, T. Chen, J. Fan, B. Shi, Learning cross-image object semantic relation in transformer for few-shot fine-grained image classification, in: ACM International Conference on Multimedia, Association for Computing Machinery, New York, NY, USA, 2022, pp. 2135–2144, http://dx.doi.org/10.1145/3503161.3547961.

[19] Y. Wang, Y. Ji, W. Wang, B. Wang, Bi-channel attention meta learning for few-shot fine-grained image recognition, Expert Syst. Appl. 242 (2024) 122741, http://dx.doi.org/10.1016/j.eswa.2023.122741.

[20] H. Huang, J. Zhang, J. Zhang, J. Xu, Q. Wu, Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification, IEEE Trans. Multimed. 23 (2021) 1666–1680, http://dx.doi.org/10.1109/TMM.2020.3001510.

[21] P. Ganesan, S.K. Jagatheesaperumal, M.M. Hassan, F. Pupo, G. Fortino, Few-shot image classification using graph neural network with fine-grained feature descriptors, Neurocomputing 610 (2024) 128448, http://dx.doi.org/10.1016/j.neucom.2024.128448.

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

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

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

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

[26] S. Cao, W. Wang, J. Zhang, M. Zheng, Q. Li, A few-shot fine-grained image classification method leveraging global and local structures, Int. J. Mach. Learn. Cybern. 13 (8) (2022) 2273–2281, http://dx.doi.org/10.1007/s13042-022-01522-w.

[27] H. Huang, J. Zhang, L. Yu, J. Zhang, Q. Wu, C. Xu, TOAN: Target-oriented alignment network for fine-grained image categorization with few labeled samples, IEEE Trans. Circuits Syst. Video Technol. 32 (2) (2022) 853–866, http://dx.doi.org/10.1109/TCSVT.2021.3065693.

[28] Y. Wu, B. Zhang, G. Yu, W. Zhang, B. Wang, T. Chen, J. Fan, Object-aware long-short-range spatial alignment for few-shot fine-grained image classification, in: ACM International Conference on Multimedia, New York, NY, USA, 2021, pp. 107–115, http://dx.doi.org/10.1145/3474085.3475532.

[29] R. Qi, S. Ning, Y. Jiang, Prototype rectification with region-wise foreground enhancement for few-shot classification, in: Pattern Recognition and Computer Vision, Singapore, 2024, pp. 15–26, http://dx.doi.org/10.1007/978-981-99-8462-6_2.

[30] Z. Li, H. Tang, Z. Peng, G.-J. Qi, J. Tang, Knowledge-guided semantic transfer network for few-shot image recognition, IEEE Trans. Neural Netw. Learn. Syst. (2023) 1–15, http://dx.doi.org/10.1109/TNNLS.2023.3240195.

[31] B. Zhang, X. Li, Y. Ye, S. Feng, Prototype completion for few-shot learning, IEEE Trans. Pattern Anal. Mach. Intell. 45 (10) (2023) 12250–12268, http://dx.doi.org/10.1109/TPAMI.2023.3277881.

[32] R. Wang, F. Hou, S.F. Cahan, L. Chen, X. Jia, W. Ji, Fine-grained entity typing with a type taxonomy: A systematic review, IEEE Trans. Knowl. Database Eng. 35 (5) (2023) 4794–4812, http://dx.doi.org/10.1109/TKDE.2022.3148980.

[33] J. Han, X. Yao, G. Cheng, X. Feng, D. Xu, P-CNN: Part-based convolutional neural networks for fine-grained visual categorization, IEEE Trans. Pattern Anal. Mach. Intell. 44 (2) (2022) 579–590, http://dx.doi.org/10.1109/TPAMI.2019.2933510.

[34] C. Mou, A. Liang, C. Hu, F. Meng, B. Han, F. Xu, Monitoring endangered and rare wildlife in the field: A foundation deep learning model integrating human knowledge for incremental recognition with few data and low cost, Animals 13 (20) (2023) http://dx.doi.org/10.3390/ani13203168.

[35] J. Lu, S. Zhang, S. Zhao, D. Li, R. Zhao, A metric-based few-shot learning method for fish species identification with limited samples, Animals 14 (5) (2024) http://dx.doi.org/10.3390/ani14050755.

[36] Y. Chen, X. Guo, Y. Pan, Y. Xia, Y. Yuan, Dynamic feature splicing for few-shot rare disease diagnosis, Med. Image Anal. 90 (2023) 102959, http://dx.doi.org/10.1016/j.media.2023.102959.

[37] J. Snell, K. Swersky, R. Zemel, Prototypical networks for few-shot learning, in: Advances in Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA, 2017, pp. 4080–4090, http://dx.doi.org/10.5555/3294996.3295163.

[38] H. Tang, Z. Li, Z. Peng, J. Tang, BlockMix: Meta regularization and self-calibrated inference for metric-based meta-learning, in: ACM International Conference on Multimedia, 2020, pp. 610–618, http://dx.doi.org/10.1145/3394171.3413884.
[39] W. Jiang, K. Huang, J. Geng, X. Deng, Multi-scale metric learning for few-shot learning, IEEE Trans. Circuits Syst. Video Technol. 31 (3) (2021) 1091–1102, http://dx.doi.org/10.1109/TCSVT.2020.2995754.
[40] F. Hao, F. He, J. Cheng, D. Tao, Global-local interplay in semantic alignment for few-shot learning, IEEE Trans. Circuits Syst. Video Technol. 32 (7) (2022) 4351–4363, http://dx.doi.org/10.1109/TCSVT.2021.3132912.
[41] F. Sung, Y. Yang, L. Zhang, T. Xiang, P.H. Torr, T.M. Hospedales, Learning to compare: Relation network for few-shot learning, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 1199–1208, http://dx.doi.org/10.1109/CVPR.2018.00131.
[42] A.A. Rusu, D. Rao, J. Sygnowski, O. Vinyals, R. Pascanu, S. Osindero, R. Hadsell, Meta-learning with latent embedding optimization, in: International Conference on Learning Representations, 2019.
[43] N. Mishra, M. Rohaninejad, X. Chen, P. Abbeel, A simple neural attentive meta-learner, in: International Conference on Learning Representations, 2017.
[44] Q. Sun, Y. Liu, T.-S. Chua, B. Schiele, Meta-transfer learning for few-shot learning, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 403–412, http://dx.doi.org/10.1109/CVPR.2019.00049.
[45] Z. Chi, L. Gu, H. Liu, Y. Wang, Y. Yu, J. Tang, Metafscil: A meta-learning approach for few-shot class incremental learning, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2022, pp. 14146–14155, http://dx.doi.org/10.1109/CVPR52688.2022.01377.
[46] K. Li, Y. Zhang, K. Li, Y. Fu, Adversarial feature hallucination networks for few-shot learning, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2020, pp. 13467–13476, http://dx.doi.org/10.1109/CVPR42600.2020.01348.
[47] C. Xu, C. Liu, X. Sun, S. Yang, Y. Wang, C. Wang, Y. Fu, PatchMix augmentation to identify causal features in few-shot learning, IEEE Trans. Pattern Anal. Mach. Intell. 45 (6) (2023) 7639–7653, http://dx.doi.org/10.1109/TPAMI.2022.3223784.
[48] J. Xu, B. Liu, Y. Xiao, A multitask latent feature augmentation method for few-shot learning, IEEE Trans. Neural Netw. Learn. Syst. 35 (5) (2024) 6976–6990, http://dx.doi.org/10.1109/TNNLS.2022.3213576.
[49] D. Wertheimer, L. Tang, B. Hariharan, Few-shot classification with feature map reconstruction networks, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2021, pp. 8008–8017, http://dx.doi.org/10.1109/CVPR46437.2021.00792.
[50] X. Li, Q. Song, J. Wu, R. Zhu, Z. Ma, J.-H. Xue, Locally-enriched cross-reconstruction for few-shot fine-grained image classification, IEEE Trans. Circuits Syst. Video Technol. 33 (12) (2023) 7530–7540, http://dx.doi.org/10.1109/TCSVT.2023.3275382.
[51] J. Sun, X. Shen, Q. Sun, Efficient feature reconstruction via l2,1-norm regularization for few-shot classification, IEEE Trans. Circuits Syst. Video Technol. 33 (12) (2023) 7452–7465, http://dx.doi.org/10.1109/TCSVT.2023.3274168.
[52] C. Xing, N. Rostamzadeh, B.N. Oreshkin, P.O. Pinheiro, Adaptive cross-modal few-shot learning, in: International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA, 2019.
[53] Z. Peng, Z. Li, J. Zhang, Y. Li, G.-J. Qi, J. Tang, Few-shot image recognition with knowledge transfer, in: International Conference on Computer Vision, ICCV, 2019, pp. 441–449, http://dx.doi.org/10.1109/ICCV.2019.00053.
[54] C. Fellbaum, WordNet: An Electronic Lexical Database, The MIT Press, 1998, http://dx.doi.org/10.7551/mitpress/7287.001.0001.
[55] J. Pennington, R. Socher, C.D. Manning, GloVe: Global vectors for word representation, in: Conference on Empirical Methods in Natural Language Processing, 2014.
[56] X. Li, X. Yang, Z. Ma, J.-H. Xue, Deep metric learning for few-shot image classification: A review of recent developments, Pattern Recognit. 138 (2023) http://dx.doi.org/10.1016/j.patcog.2023.109381.
[57] V.G. Satorras, J.B. Estrach, Few-shot learning with graph neural networks, in: International Conference on Learning Representations, Vancouver, BC, Canada, 2018.
[58] Y. Ma, S. Bai, S. An, W. Liu, A. Liu, X. Zhen, X. Liu, Transductive relation-propagation network for few-shot learning, in: International Joint Conference on Artificial Intelligence, Virtual, Montreal, 2020, pp. 804–810, http://dx.doi.org/10.24963/ijcai.2020/112.
[59] T. Yu, S. He, Y.-Z. Song, T. Xiang, Hybrid graph neural networks for few-shot learning, AAAI Conf. Artif. Intell. 36 (3) (2022) 3179–3187, http://dx.doi.org/10.1609/aaai.v36i3.20226.
[60] H. Cheng, J.T. Zhou, W.P. Tay, B. Wen, Graph neural networks with triple attention for few-shot learning, IEEE Trans. Multimed. 25 (2023) 8225–8239, http://dx.doi.org/10.1109/TMM.2022.3233442.
[61] X.-S. Wei, P. Wang, L. Liu, C. Shen, J. Wu, Piecewise classifier mappings: Learning fine-grained learners for novel categories with few examples, IEEE Trans. Image Process. 28 (12) (2019) 6116–6125, http://dx.doi.org/10.1109/TIP.2019.2924811.
[62] Q. Xu, M. Zhang, Y. Li, Z. Tao, Learning more discriminative clues with gradual attention for fine-grained visual categorization, Image Vis. Comput. 136 (2023) 104753, http://dx.doi.org/10.1016/j.imavis.2023.104753.
[63] Y. Yu, D. Zhang, S. Wang, Z. Ji, Z. Zhang, Local spatial alignment network for few-shot learning, Neurocomputing 497 (2022) 182–190, http://dx.doi.org/10.1016/j.neucom.2022.05.020.
[64] A.V. Flevaris, A. Martínez, S.A. Hillyard, Attending to global versus local stimulus features modulates neural processing of low versus high spatial frequencies: an analysis with event-related brain potentials, Front. Psychol. 5 (2014) http://dx.doi.org/10.3389/fpsyg.2014.00277.
[65] H. Tang, Z. Li, D. Zhang, S. He, J. Tang, Divide-and-conquer: Confluent triple-flow network for RGB-T salient object detection, IEEE Trans. Pattern Anal. Mach. Intell. (2024) 1–17, http://dx.doi.org/10.1109/TPAMI.2024.3511621.
[66] H. Liu, C.L.P. Chen, X. Gong, T. Zhang, Robust saliency-aware distillation for few-shot fine-grained visual recognition, IEEE Trans. Multimed. 26 (2024) 7529–7542, http://dx.doi.org/10.1109/TMM.2024.3369870.
[67] M. Han, Y. Zhan, Y. Luo, B. Du, H. Hu, Y. Wen, D. Tao, Not all instances contribute equally: Instance-adaptive class representation learning for few-shot visual recognition, IEEE Trans. Neural Netw. Learn. Syst. 35 (4) (2024) 5447–5460, http://dx.doi.org/10.1109/TNNLS.2022.3204684.
[68] Y. Zhao, T. Zhang, J. Li, Y. Tian, Dual adaptive representation alignment for cross-domain few-shot learning, IEEE Trans. Pattern Anal. Mach. Intell. 45 (10) (2023) 11720–11732, http://dx.doi.org/10.1109/TPAMI.2023.3272697.
[69] Z.-M. Chen, X.-S. Wei, P. Wang, Y. Guo, Multi-label image recognition with graph convolutional networks, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 5172–5181, http://dx.doi.org/10.1109/CVPR.2019.00532.
[70] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, 2021.
[71] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan, Supervised contrastive learning, in: Advances in Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA, 2020, http://dx.doi.org/10.5555/3495724.3497291.
[72] Z. Song, Y. Zhao, Y. Shi, P. Peng, L. Yuan, Y. Tian, Learning with fantasy: Semantic-aware virtual contrastive constraint for few-shot class-incremental learning, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2023, pp. 24183–24192, http://dx.doi.org/10.1109/CVPR52729.2023.02316.
[73] C. Liu, Y. Fu, C. Xu, S. Yang, J. Li, C. Wang, L. Zhang, Learning a few-shot embedding model with contrastive learning, in: AAAI Conference on Artificial Intelligence, 2021.
[74] Y. Zhu, C. Liu, S. Jiang, Multi-attention meta learning for few-shot fine-grained image recognition, in: International Joint Conference on Artificial Intelligence, 2021.
[75] R. Hou, H. Chang, B. Ma, S. Shan, X. Chen, Cross attention network for few-shot classification, in: Proc. Adv. Neural Inf. Process. Syst., Curran Associates Inc., Red Hook, NY, USA, 2019, pp. 4003–4014, http://dx.doi.org/10.5555/3454287.3454647.
[76] C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie, The Caltech-UCSD Birds-200–2011 Dataset, California Institute of Technology, 2011, no. CNS-TR-2011-001.
[77] A. Khosla, N. Jayadevaprakash, B. Yao, L. Fei-Fei, Novel dataset for fine-grained image categorization : Stanford dogs, in: IEEE Conference on Computer Vision and Pattern Recognition Workshop, 2011, pp. 554–561.
[78] J. Krause, M. Stark, J. Deng, L. Fei-Fei, 3D object representations for fine-grained categorization, in: IEEE International Conference on Computer Vision Workshop, 2013, pp. 554–561, http://dx.doi.org/10.1109/ICCVW.2013.77.
[79] L.v.d. Maaten, G. Hinton, Visualizing Data using t-SNE, J. Mach. Learn. Res. 9 (86) (2008) 2579–2605.
[80] S. Reed, Z. Akata, H. Lee, B. Schiele, Learning deep representations of fine-grained visual descriptions, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 49–58, http://dx.doi.org/10.1109/CVPR.2016.13.
[81] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186, http://dx.doi.org/10.18653/v1/N19-1423.