Fine-grained Image Retrieval via Dual-Vision Adaptation

Xin Jiang[1*], Meiqi Cao[1*], Hao Tang[2], Fei Shen[3], Zechao Li[1†]

[1]Nanjing University of Science and Technology, China
[2]Centre for Smart Health, Hong Kong Polytechnic University, China
[3]National University of Singapore, Singapore
{zechao.li}@njust.edu.cn

Abstract

Fine-Grained Image Retrieval (FGIR) faces challenges in learning discriminative visual representations to retrieve images with similar fine-grained features. Current leading FGIR solutions typically follow two regimes: enforce pairwise similarity constraints in the semantic embedding space, or incorporate a localization sub-network to fine-tune the entire model. However, such two regimes tend to overfit the training data while forgetting the knowledge gained from large-scale pre-training, thus reducing their generalization ability. In this paper, we propose a Dual-Vision Adaptation (DVA) approach for FGIR, which guides the frozen pre-trained model to perform FGIR through collaborative sample and feature adaptation. Specifically, we design Object-Perceptual Adaptation, which modifies input samples to help the pre-trained model perceive critical objects and elements within objects that are helpful for category prediction. Meanwhile, we propose In-Context Adaptation, which introduces a small set of parameters for feature adaptation without modifying the pre-trained parameters. This makes the FGIR task using these adapted features closer to the task solved during the pre-training. Additionally, to balance retrieval efficiency and performance, we propose Discrimination Perception Transfer to transfer the discriminative knowledge in the object-perceptual adaptation to the image encoder using the knowledge distillation mechanism. Extensive experiments show that DVA performs well on three fine-grained datasets.

Introduction

Unlike general image retrieval, fine-grained image retrieval (FGIR) attempts to retrieve images belonging to the same subcategory as the query image from a database within a broader meta category (*i.e.*, birds, cars) (Wei et al. 2017). It has been extensively applied across various domains, including intelligent transportation (Ramdurai 2025) and biodiversity monitoring (Vendrow et al. 2024). However, FGIR presents significant challenges due to two inherent difficulties: i) subtle visual distinctions between different categories, and ii) significant appearance variations within the same category. This task usually requires models to simultaneously localize discriminative regions and identify minute visual differences, creating a dual optimization dilemma. Therefore, the key to FGIR is learning discriminative and generalizable embeddings to identify visually similar objects accurately.

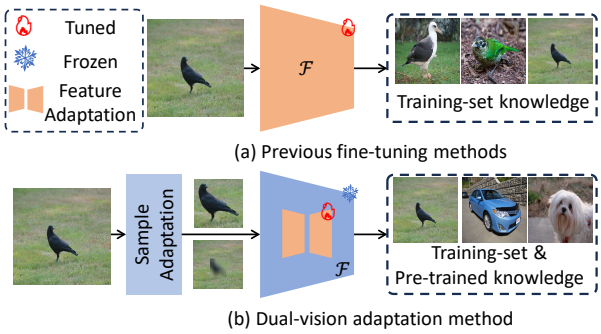

Figure 1: (a) previous fine-tuning methods. (b) our dual-vision adaptation method. Our approach designs the collaborative sample and feature adaptation to exploit category-specific differences. This dual strategy enables model to sustain broad representation capabilities from pre-training data while dynamically adjusting its adaptability to fine-grained data.

Conventional FGIR methods typically adopt two primary strategies: encoding-based (Movshovitz-Attias et al. 2017; Teh, DeVries, and Taylor 2020; Ermolov et al. 2022; Lim et al. 2022) and localization-based (Zheng et al. 2018; Wang et al. 2023c; Wei et al. 2017; Wang et al. 2022). Encoding-based methods primarily learn image-level features but struggle to suppress background and irrelevant information interference. In contrast, localization-based methods enhance feature extraction within the encoder to capture subtle differences among subcategories, often by focusing on distinctive object regions. Although both strategies capture discriminative embedding learning for fine-grained retrieval, their dependence on full-parameter fine-tuning introduces prohibitive computational costs and limits cross-task generalization, as shown in Fig. 1(a). Simultaneously, fine-tuning the entire model destroys the knowledge in the large-scale pre-trained parameters and may lead to suboptimal convergence, which ultimately limits retrieval performance (Wang et al. 2023c). This raises the question: can dis-

*Equal contribution.
†Corresponding author.

criminative representations be learned for FGIR tasks without fine-tuning the model's pre-trained parameters?

Recent advancements in parameter-efficient fine-tuning (PEFT) techniques have demonstrated the feasibility of learning representations while introducing a limited number of new parameters to adapt a frozen pre-trained model to downstream tasks. Its core idea is to redesign downstream tasks to align closely with those addressed during pre-training, ensuring that large-scale pre-trained knowledge is effectively retained. Owing to PEFT, large-scale pre-trained visual-language models (*e.g.,* CLIP (Radford et al. 2021)) have achieved impressive results in various visual tasks. However, existing PEFT methods (Zhou et al. 2022a,b) are designed for multimodal models to capture high-level category semantics rather than subtle subcategory differences, so directly applying existing PEFT techniques to FGIR without task-specific adaptation tends to be suboptimal, due to their limited focus on capturing subtle intra-class variations. Therefore, it is crucial to develop an efficient fine-tuning strategy specifically designed for fine-grained visual models, as this can prevent suboptimal convergence that may occur when fine-tuning the entire FGIR model.

To this end, we design a novel adaptation framework for FGIR, which efficiently unleashes the power of pre-trained model, as shown in Fig. 1(b). Technically, we propose Dual-Vision Adaptation (DVA) to resolve the conflict between preserving pre-trained knowledge and acquiring fine-grained discriminative power. By maintaining the frozen backbone, DVA establishes category-sensitive perception through sample-feature co-adaptation, inherently avoiding the optimization dilemma of conventional full fine-tuning. Specifically, the Object-Perceptual Adaption (OPA) is a proposed sample-adapted strategy for the visual foundation model, encouraging the pre-trained model to capture critical object regions and locate subclass-specific differences. Meanwhile, as a feature adaptation process, the proposed In-Context Adaption (ICA) aims to focus on fine-grained features that contribute to category prediction through lightweight learnable parameters. In this way, ICA can adapt the feature content in a direction that is conducive to category recognition, which makes the FGIR task with such adapted features close to the task solved during the original pre-training. Nevertheless, a non-negligible problem is that the integration between the visual foundation model and the frozen pre-trained backbone introduces prohibitive computational overhead during inference. To address this critical efficiency bottleneck while retaining fine-grained discriminative power, we devise a Discriminative Perceptual Transfer (DPT) module that transfers the discriminative knowledge from OPA to the backbone via a distillation mechanism.

In summary, the primary contributions of this work are as follows:

- We propose a dual vision adaptation framework to enable the frozen pre-trained model to capture subtle subcategory differences without forgetting the pre-trained knowledge, balancing accuracy and efficiency.
- We propose a dynamic sample and feature co-adaptation strategy and a distillation-based sample perception transfer module to bridge the gap between high-level semantics and fine-grained distinctions, which are realized through the collaborative modification of samples to features.
- With only 3.5% of the tunable parameters compared to full fine-tuning, DVA achieves competitive performance on three fine-grained datasets.

Related Works

Fine-Grained Image Retrieval

Existing methods for fine-grained image retrieval (FGIR) can be categorized as encoding-based or localization-based. The encoding-based methods (Movshovitz-Attias et al. 2017; Teh, DeVries, and Taylor 2020; Lim et al. 2022; Ermolov et al. 2022; Jiang et al. 2024b; Jiang, Tang, and Li 2024; Tang et al. 2020a) aim to learn an embedding space in which samples of a similar subcategory are attracted and samples of different subcategories are repelled. The methods can be decomposed into roughly two components: the image encoder maps images into an embedding space, and the metric method ensures that samples from the same subcategories are grouped closely, while samples from different subcategories are separated. While these studies have achieved significant achievements, they primarily concentrate on optimizing image-level features that include numerous noisy and non-discriminatory details. Therefore, the localization-based methods (Han et al. 2018; Tang et al. 2020b; Wang et al. 2023c; Wei et al. 2017; Moskvyak et al. 2021; Tang et al. 2022) are proposed, which focus on training a subnetwork for locating discriminative regions or devising an effective strategy for extracting attractive object features to facilitate the retrieval task. Unlike these approaches, our method considers the specific characteristics of the FGIR tasks, offering guidance for designing high-performance retrieval models.

Parameter-Efficient Fine Tuning

Parameter-efficient fine-tuning (Hu et al. 2022) in NLP reformulates downstream tasks as language modeling problems, allowing pre-trained language models to adapt more efficiently to new tasks. As a result, these techniques are now widely used in various NLP applications, including language understanding and generation. Recently, parameter-efficient fine-tuning has also been applied to multimodal computer vision (Zhou et al. 2022b; Xing et al. 2025; Jiang et al. 2024a; Tang, He, and Qin 2025; Cao et al. 2024a; Gao et al. 2025a). However, existing methods primarily enhance language-based models, making them unsuitable for direct application to pre-trained vision models. Additionally, some recent approaches (Chen et al. 2022; Gao et al. 2025b) introduce a small number of learnable parameters to guide pre-trained models in general vision tasks. However, their fine-tuning strategies focus on capturing category-level semantics rather than the fine-grained visual differences needed to distinguish similar objects. To address this limitation, we propose a parameter-efficient DVA that incorporates sample

and feature adaptation, enabling the frozen pre-trained vision model to perform FGIR tasks effectively.

Knowledge Distillation

Knowledge Distillation (KD) is dedicated to compressing the informative knowledge from a large and computationally expensive model (*i.e,* teacher model) to a small and computationally efficient model (*i.e,* student model) (Hinton, Vinyals, and Dean 2015; Fang et al. 2024; Cao et al. 2025). Given the strong knowledge transfer ability, it is widely applied to natural language processing, computer vision and other fields (Wang and Yoon 2021; Tang et al. 2023, 2024). Most classification-based KD approaches explore enhancing the student model by mimicking the teacher model's predictions or output distribution. Additionally, researchers have delved into exploring KD for image retrieval by leveraging distances between samples, such as learning to rank (Chen, Wang, and Zhang 2018) and regression on quantities involving one or more pairs, like distances (Wang et al. 2023b) or angles (Park et al. 2019). In this paper, we transfer discriminative knowledge that aids fine-grained image retrieval via proxy features.

Method

Dual-Vision Adaptation

Object-Perceptual Adaption. Subtle yet discriminative discrepancies are widely recognized to be significant for fine-grained understanding (Fang et al. 2024; Wang et al. 2023c; Cao et al. 2024a). However, the vanilla ViT model is originally designed to identify different species (*e.g.*, cats, dogs, and birds), rather than to exploit subtle differences between subcategories within a species. Therefore, without task-specific adaptation, the vanilla ViT has limited focus on capturing subtle intra-class variations. To alleviate this situation, we design an Object-Perceptual Adaption (OPA) that only modifies the visual input of the image encoder (ViT) to assist the model in perceiving discriminative differences in fine-grained images. Specifically, OPA consists of two components: discriminative perception and object perception. The former enhances category discrimination by focusing on foreground regions identified by the visual foundation model, while the latter improves object perception by utilizing background information separated by the same model.

Discriminative Perception. For fine-grained understanding, there have been many works (Xing et al. 2024; Jiang et al. 2024b) demonstrating that the background can perturb the model's perception of discriminative regions. Therefore, we design discriminative perception to help the model focus on objects in the image, thereby exploring discriminative features through critical regions. We use visual foundation model to extract object regions from training images, keep the core object foreground, obtaining the discriminative image $\mathbf{I}_{\mathrm{d}}$. Specifically, each training image along with the super-class name (*e.g.,* bird for CUB-200-2011 (Branson et al. 2014)) of the image is fed to open-vocabulary detector GroundingDINO (Liu et al. 2023) to obtain the bounding box that contains the foreground object of the category. Therefore, there is no risk of information leakage for FGIR. Then, the region containing the foreground object is extracted from the image as the discriminative image $\mathbf{I}_{\mathrm{d}}$. To ensure objects are fully contained in the image and free from deformation, we filter out low-confidence detection results, and finally apply padding on the short sides of the image to preserve the original shape of the object.

Object Perception. In fine-grained tasks, the background often provides contextual information that is complementary to the foreground and can help distinguish similar objects. For example, the foreground features of the "Red-bellied Tit" and the "Golden Pheasant" are similar, but the former often appears in bushes, while the latter often appears in rocky areas. Therefore, we aim to extract background regions from training images and establish an explicit background category to enhance model discrimination. Specifically, we reuse the detection boxes of foreground objects obtained from discriminative perception. Unlike previous use, we apply a mean filter to blur the image region within the bounding box, resulting in a background image that excludes object information. For those training images whose foreground objects occupy the majority of the whole image region, background images would contain much less background information after blurring foreground regions. Therefore, only when the area of the foreground region is smaller than a predefined proportion $\alpha\%$ (*e.g.*, 50%) of the whole training image, the training image is used to construct the background category image $\mathbf{I}_{\mathrm{b}}$ as described above.

In-Context Adaptation. Conventional methods (Ermolov et al. 2022; Fang et al. 2024; Cao et al. 2024b) usually require fine-tuning the entire network, but this approach ignores the large-scale pre-trained knowledge present in the pre-trained parameters. We hypothesize that FGIR on downstream datasets can be learned and understood by specific modules within the network. To verify this hypothesis, we conducte an in-depth analysis of the fine-tune preferences of ViT.

We first divide the trainable parameters of transformer layers into three parts: the attention projector, output projector and MLP. Subsequently, we separately fine-tune these three parts and ViT as a whole and evaluate on the CUB-200-2011 (Branson et al. 2014) dataset. As shown in Table 1, we found that by fine-tuning only specific modules within ViT, it is possible to achieve performance comparable to that of fine-tuning the entire model. Meanwhile, fine-tuning the attention projector yields the best performance. This is intuitive, as the self-attention mechanism is central to ViT, and the attention projector enhances its ability to encode contextual information. This experiment further validates our hypothesis that the FGIR task can be learned by a specific network module, namely attention projector.

Based on this insight, we attempt to design a in-context adaptation module with tune only the attention projector of the ViT. However, this approach still destroys the knowledge in the pre-trained parameters and still requires learning a large number of parameters (22.5M). To address this issue, we keep the attention projector frozen and propose a lighter-weight learnable in-context adaptation module in parallel with the attention projector to learn the specific knowledge

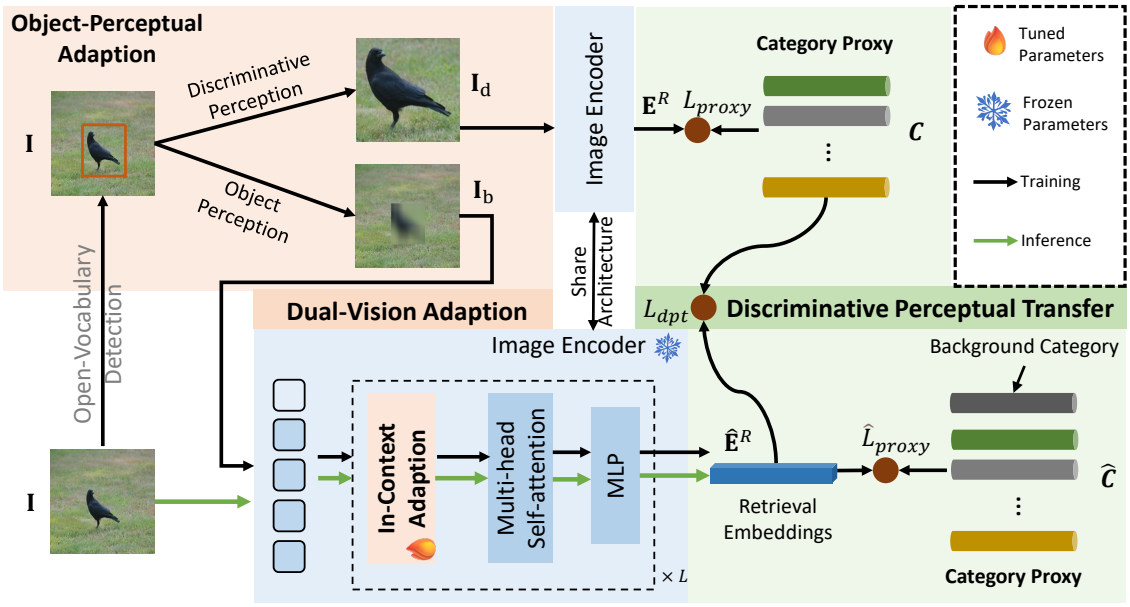

Figure 2: DVA consists of three essential modules: the Object-Perceptual Adaptation module to enhance the encoder's ability to focus on discriminative object regions, the In-Context Adaptation module to dynamically refine fine-grained features while suppressing irrelevant background retained in frozen representations, and the Discriminative Perceptual Transfer module to distill discriminative awareness into the encoder, enabling auxiliary-free inference while preserving pre-trained knowledge.

required for FGIR. Specifically, the in-context adaptation module is designed to be a bottleneck structure for limiting the number of parameters purpose, which includes a down-projection layer with parameters $\mathbf{W}_{down} \in \mathbb{R}^{D\times d}$, an up-projection layer with parameters $\mathbf{W}_{up} \in \mathbb{R}^{d\times D}$, in this work, we set $d = 16$. For a specific input feature $\mathbf{x}$, the in-context adaptation module produces the in-context features, $\mathbf{x}_{ic}$, formally via:

$$\mathbf{x}_{ic} = \mathrm{LN}(\mathbf{x}) \cdot \mathbf{W}_{down} \cdot \mathbf{W}_{up}. \tag{1}$$

Then, we can replace the original attention projector output with a feature that contains both pre-trained knowledge and task-specific knowledge through residual connections:

$$\hat{\mathbf{Q}} = \mathrm{AttentionProjector}_q(\mathrm{LN}(\mathbf{E})) + \mathrm{IC}_q(\mathbf{E}), \tag{2}$$

where $\mathrm{IC}_q(\cdot)$ represents the in-context adaptation module of $\mathbf{E}$ corresponding to the $\mathrm{AttentionProjector}_q(\cdot)$ that generate query $\mathbf{Q}$. In this paper, we only assign in-context adaptation modules to the attention projectors that generate $\mathbf{Q}$ and $\mathbf{K}$. For experimental analysis, please see Experimental section.

Discriminative Perceptual Transfer

With the cooperation of discriminative perception, the fine-grained understanding ability of the ViT model is enhanced. However, the extra visual foundation model is time-consuming and memory-demanding for retrieval evaluation. In the classification field (Hinton, Vinyals, and Dean 2015), network distillation has been shown to be one of the solutions to this problem. Inspired by network distillation, we propose a discriminative perceptual transfer to extend the knowledge distillation theory to retrieval tasks to achieve an efficiency-performance tradeoff. Specifically, we separate discriminative perception into a separate branch, exploit a category proxy learning strategy to learn discriminative category proxy as carriers of discriminative knowledge, and transfer their discriminative knowledge.

Methods	R@1	R@2	R@4	Prams (M)
Non-fine-tuned	72.5	82.6	89.1	0.0
Fine-tuning	82.7	88.7	92.5	85.6
Attention Projector	82.7	89.0	92.8	21.5
Output Projector	82.6	88.8	92.6	7.3
MLP	82.0	88.9	92.9	56.8

Table 1: Fine-Tuning Preferences Analysis of ViT on the CUB-200-2011 dataset.

We encode both the discrimination and distribution of labelled instances via proxy-guided learning (Kim et al. 2020). Then, for the retrieval embedding $\mathbf{E}^R$, we compute the proxy learning loss as follow:

$$\mathcal{L}_{proxy} = -\log\left(\frac{\exp(-d(\|\mathbf{E}^R\|, \|\mathbf{c}\|))}{\sum_{\mathbf{c}\in\mathbf{C}}\exp(-d(\|\mathbf{E}^R\|, \|\mathbf{c}\|))}\right), \tag{3}$$

where $d(\|\mathbf{E}^R\|, \|\mathbf{c}\|)$ represents the distance between $\|\mathbf{E}^R\|$ and $\|\mathbf{c}\|$, $\mathbf{c}$ denotes the category proxy corresponding to retrieval embedding $\mathbf{E}^R$, $\mathbf{C}$ is the category proxy set of the discriminative perception images $\mathbf{I}_d$, and $\|\cdot\|$ denotes the L^2-Norm.

With the discriminative category proxy set $\mathbf{C}$, and the retrieval embeddings $\hat{\mathbf{E}}^R$ of the origin training images are used

Method	CUB-200-2011				Stanford Cars			
	R@1	R@2	R@4	R@8	R@1	R@2	R@4	R@8
Proxy-Anchor (Kim et al. 2020)	80.4	85.7	89.3	92.3	77.2	83.0	87.2	90.2
HIST (Lim et al. 2022)	75.6	83.0	88.3	91.9	**89.2**	**93.4**	95.9	97.6
PNCA++ (Teh, DeVries, and Taylor 2020)	80.4	85.7	89.3	92.3	86.4	92.3	<u>96.0</u>	<u>97.8</u>
Hyp-ViT (Ermolov et al. 2022)	84.2	91.0	94.3	96.0	76.7	85.2	90.8	94.7
DVA (Ours)	**85.2**	**91.4**	**94.6**	**96.2**	<u>88.0</u>	<u>93.2</u>	**97.0**	**98.5**

Table 2: Comparison with state-of-the-art methods in the closed-set setting on CUB-200-2011 and Stanford Cars. The best result is shown in **bold**, and the second-best result is <u>underlined</u>.

for distillation:

$$\mathcal{L}_{dpt} = -\log\left(\frac{\exp(-d(\|\hat{\mathbf{E}}^R\|, \|\mathbf{c}\|))}{\sum_{\mathbf{c}\in\mathbf{C}}\exp(-d(\|\hat{\mathbf{E}}^R\|, \|\mathbf{c}\|))}\right). \quad (4)$$

After optimizing $\mathcal{L}_{dpt}$, the image encoder adapts to capture subtle inter-category differences transferred from the discriminative branch. This enhancement enables effective retrieval of visually similar items without relying on the computationally expensive discriminative perception module.

Overall Function

As illustrated in Fig. 2, our objective function comprises two components: the proxy learning loss $\hat{\mathcal{L}}_{proxy}$ and the distillation loss $\mathcal{L}_{dpt}$. Finally, the total loss for our framework can be defined as:

$$\hat{\mathcal{L}}_{proxy} = -\log\left(\frac{\exp(-d(\|\mathbf{E}^R\|, \|\mathbf{c}\|))}{\sum_{\mathbf{c}\in\hat{\mathbf{C}}}\exp(-d(\|\mathbf{E}^R\|, \|\mathbf{c}\|))}\right), \quad (5)$$

$$\mathcal{L} = \hat{\mathcal{L}}_{proxy} + \beta\mathcal{L}_{dpt},$$

where $\hat{\mathbf{C}} \in \{\mathbf{C}^o, \mathbf{C}^{\mathrm{b}}\}$ contains original training categories $\mathbf{C}^o$ and background category $\mathbf{C}^{\mathrm{b}}$, and β is the hyper-parameter to weight the $\mathcal{L}_{dpt}$.

Experiments

Experiment Setup

Datasets. **CUB-200-2011** (Branson et al. 2014) comprises 11, 788 bird images from 200 bird species. In the closed-set setting, the dataset is divided into training and testing subsets comprising 5,994 and 5,794 images, respectively, out of a total of 11,788 images. For the open-set setting, we employ the first 100 subcategories (comprising 5,864 images) for training, and the remaining subcategories (comprising 5,924 images) are used for testing.

Stanford Cars (Krause et al. 2013) consists of 16,185 images depicting 196 car variants. Similarly, these images were split into 8,144 training images and 8,041 test images in the closed-set setting. For the open-set setting, we utilize the first 98 subcategories (comprising 8,054 images) for training and the remaining 98 subcategories (comprising 8,131 images) for testing.

Stanford Dogs (Dataset 2011) contains 20,580 images showcasing dogs across 120 subcategories. We use the 120 subcategories with 12,000 images for training and the remaining 120 subcategories with 8580 images for testing for the closed-set setting. For the open-set setting, we utilize the first 60 subcategories (comprising 10,651 images) for training and the remaining 60 subcategories (comprising 9,929 images) for testing.

Evaluation protocols. To evaluate retrieval performance, we adopt Recall@K with cosine distance in previous work (Song et al. 2016), which calculates the recall scores of all query images in the test set. For each query image, the top **K** similar images are returned. A recall score of 1 is assigned if at least one positive image among the top **K** images; otherwise, it is 0.

Implementation details. In experiments, we employ the ViT-B-16 (Dosovitskiy et al. 2021) pre-trained on ImageNet21K (Russakovsky et al. 2015) as our image encoder. All input images are resized to 256 $\times$ 256, and crop them into 224 $\times$ 224. In the training stage, we utilize the Adam optimizer with weight decay of 1e-4, and employ cosine annealing as the optimization scheduler. The learning rate for all datasets is initialized to 1e-1 except for Stanford Dogs, which has a learning rate of 1e-2. The number of training epochs for all datasets is set to 10 and the batch size is set to 32. We train our model on a single NVIDIA 3090 GPU to accelerate the training process.

Comparison with State-of-the-Art Methods

Closed-set Setting. We first compare our proposed DVA with previous competitive methods under closed-set setting. Quantitative comparison results for the CUB-200-2011 and Stanford Cars datasets are presented in Table 2, and the results for the Stanford Dogs dataset can be found in Table 4. From these tables, it can be observed that our proposed DVA outperforms other state-of-the-art methods on CUB-200-2011 and Stanford Dogs, and achieves competitive performance on Stanford Cars, which demonstrates the enhanced discriminative capability of our DVA for fine-grained visual retrieval. Specifically, in comparison with Hyp-ViT (Ermolov et al. 2022), the current state-of-the-art on CUB-200-2011, our DVA demonstrates a 1.0% improvement in Recall@1. The experimental results on the Stanford Cars indicate that our method outperforms the most of existing meth-

Method	CUB-200-2011				Stanford Cars			
	R@1	R@2	R@4	R@8	R@1	R@2	R@4	R@8
DAS (Liu et al. 2022)	69.2	79.3	87.1	92.6	87.8	93.2	96.0	97.9
IBC (Seidenschwarz, Elezi, and Leal-Taixé 2021)	70.3	80.3	87.6	92.7	88.1	93.3	96.2	98.2
Proxy-Anchor (Kim et al. 2020)	71.1	80.4	87.4	92.5	88.3	93.1	95.7	97.0
HIST (Lim et al. 2022)	71.4	81.1	88.1	-	89.6	93.9	96.4	-
PNCA++ (Teh, DeVries, and Taylor 2020)	70.1	80.8	88.7	93.6	90.1	94.5	97.0	98.4
FRPT (Wang et al. 2023c)	74.3	83.7	89.8	94.3	**91.1**	**95.1**	**97.3**	**98.6**
DFML (Wang et al. 2023a)	79.1	86.8	-	-	89.5	93.9	-	-
Hyp-ViT (Ermolov et al. 2022)	84.0	90.2	94.2	96.4	86.0	91.9	95.2	97.2
DVA (Ours)	**84.9**	**90.6**	**94.5**	**96.7**	90.7	94.8	97.1	98.4

Table 3: Comparison with state-of-the-art methods in the open-set setting on CUB-200-2011 and Stanford Cars.

Method	Open-set Setting				Closed-set Setting			
	R@1	R@2	R@4	R@8	R@1	R@2	R@4	R@8
Proxy-Anchor (Kim et al. 2020)	86.1	91.9	95.7	97.6	83.7	89.9	93.9	96.5
HIST (Lim et al. 2022)	86.2	92.3	95.6	97.5	84.7	90.1	93.4	95.6
PNCA++ (Teh, DeVries, and Taylor 2020)	86.0	92.3	95.6	97.7	83.8	90.1	94.7	96.8
Hyp-ViT (Ermolov et al. 2022)	87.8	92.8	95.9	97.6	79.2	86.9	91.8	95.2
DVA (Ours)	**90.4**	**94.6**	**97.0**	**98.3**	**87.7**	**93.0**	**95.8**	**97.3**

Table 4: Comparison with state-of-the-art methods on Stanford Dogs.

ods but falls slightly behind HIST (Lim et al. 2022). We argue the possible reason may be attributed to the relatively regular and simpler shape of the cars.

Open-set Setting. The open-set setting poses greater challenges compared to the closed-set setting due to the unknown subcategories in the test data. The experimental results for the CUB-200-2011 and Stanford Cars datasets are shown in Table 3, while the results for the Stanford Dogs dataset are provided in Table 4. The results reveal a consistent trend in both open and closed-set settings: our method outperforms other state-of-the-art approaches on CUB-200-2011 and Stanford Dogs, while delivering competitive results on the Stanford Cars dataset. To be specific, in comparison with Hyp-ViT (Ermolov et al. 2022) on CUB-200-2011, our DVA exhibits a 0.9% improvement in Recall@1. Experimental results on the Stanford Cars dataset show that our method outperforms most existing methods but lags behind FRPT (Wang et al. 2023c) by a slight margin. We attribute FRPT's performance gains to its carefully designed but computationally expensive modules. In contrast, our method captures subtle differences through dual-vision adaptation while maintaining computational efficiency. The results of experiments on Stanford Dogs are shown in Table 4. Our DVA demonstrates a clear performance advantage over Hyp-ViT, further indicating its superior generalizability.

Ablation Studies and Analysis

Efficacy of various components. The proposed DVA comprises three essential components: Object-Perceptual Adaption (OPA), In-Context Adaptation (ICA), Discriminative Perception (DP) of the OPA, and Discriminative Perceptual Transfer (DPT). We conducted ablation experiments on these components, and the results are reported in Table 5, with the Base representing the pure ViT. The introduction of ICA improves the Recall@1 accuracy by 11.1% on CUB-200-2011, which shows that ICA effectively guides the FGIR task to the task solved in the original process through feature adaptation. OPA further improves the ability of the pre-trained model to distinguish similar subcategories on FGIR, bringing a gain of 3.2% Recall@1 on CUB-200-2011, but it brings huge computational overhead. DPT effectively removes the huge computational overhead brought by OPA and achieves a balance between performance and retrieval efficiency. In particular, when only DP is used but no Object Perception, the performance decreases. This shows that Object Perception effectively helps DVA perceive objects and thus better distinguish similar objects.

Setting	R@1	Latency (ms)
Base	72.5	1.51
Base + ICA	83.6	1.51
Base + ICA + OPA	86.8	3.65
Base + ICA + DP + DPT	84.1	1.51
Base + ICA + OPA + DPT	84.9	1.51

Table 5: The Recall@K results (%) of component ablation study on CUB-200-2011.

Analysis of In-Context Adaptation. To validate the effec-

Setting	CUB-200-2011		Stanford Cars		Prams (M)
	Recall@1	Recall@2	Recall@1	Recall@2	
$\mathbf{IC}_q$	82.7	89.1	87.2	92.9	0.295
$\mathbf{IC}_k$	81.8	88.8	85.6	91.7	0.295
$\mathbf{IC}_v$	84.0	90.0	86.4	92.3	0.295
$\mathbf{IC}_q$, $\mathbf{IC}_k$	84.9	90.6	90.7	94.8	0.590
$\mathbf{IC}_q$, $\mathbf{IC}_v$	84.5	90.5	90.3	94.5	0.590
$\mathbf{IC}_k$, $\mathbf{IC}_v$	82.6	89.0	89.5	94.1	0.590
$\mathbf{IC}_q$, $\mathbf{IC}_k$, $\mathbf{IC}_v$	84.4	90.2	90.4	94.6	0.885

Table 6: The Recall@K results (%) of component ablation study on CUB-200-2011 and Stanford Cars.

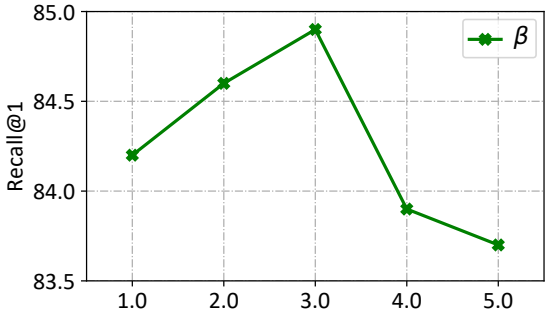

Figure 3: Analyses of hyper-parameter β on CUB-200-2011.

tiveness of the proposed ICA components, we further divide the ICA module into three parts: $\mathbf{IC}_q$, $\mathbf{IC}_k$, and $\mathbf{IC}_v$, corresponding to the Q, K, and V projectors in the Attention Projector, respectively. We perform ablation studies on these components, and the results are presented in Table 6. The results show that using $\mathbf{IC}_q$ alone or in combination with $\mathbf{IC}_k$ and $\mathbf{IC}_v$ leads to better performance. Notably, the best results are achieved when $\mathbf{IC}_q$ and $\mathbf{IC}_k$ are used together. Therefore, we adopt this combination to construct the final ICA module in our DVA.

Sensitivity analysis of parameter β. As illustrated in Fig. 3, we investigate the impact of varying values of β in Eq. (5). The performance gradually increases with the value but decreases when the value exceeds 3. The drop in performance can be attributed to the fact that as the value of β increases, the learned feature space tends to receive samples adapted by the OPA module, while the input at inference is not processed by the OPA, resulting in a difference in the input space that causes performance degradation. Consequently, we set β to 3 for all datasets.

Visualization. To explore the superiority of the designed DVA, We also conduct visualization as shown in Fig. 4. The first column displays the input images. The second column shows the class activation maps generated by GradCAM for the baseline model, which processes the input images. The third column presents the class activation maps for DVA using the same input. The results indicate that DVA effectively reduces attention to background regions and improves the baseline model's focus on critical areas, capturing finer details in subcategories such as the head, wings, and tail.

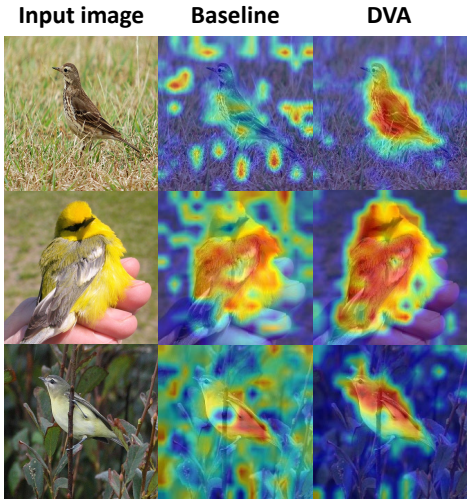

Figure 4: Class activation visualizations on CUB-200-2011. For each sample, we show the input image, the class activation map of the baseline model, and the proposed DVA.

Conclusion

In this paper, we propose Dual Visual Adaptation (DVA), a novel framework that mitigates overfitting in fine-grained image retrieval by preserving pre-trained knowledge. DVA comprises three components: (1) Object-Perceptual Adaptation, which introduces an auxiliary background category and emphasizes discriminative regions via a visual foundation model; (2) In-Context Adaptation, which injects lightweight parameters to refine fine-grained features within a frozen backbone; and (3) Discriminative Perceptual Transfer, which distills task-specific knowledge into the encoder to enable efficient, auxiliary-free inference. Extensive experiments on three fine-grained benchmarks validate the effectiveness of DVA, demonstrating competitive performance with minimal parameter overhead.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62425603) and the

Basic Research Program of Jiangsu Province (Grant No. BK20240011).

References

Branson, S.; Horn, G. V.; Belongie, S. J.; and Perona, P. 2014. Bird Species Categorization Using Pose Normalized Deep Convolutional Nets. *CoRR*, abs/1406.2952.

Cao, M.; Shu, X.; Jiang, X.; Yan, R.; Yao, Y.; and Tang, J. 2025. Exploiting Frequency Dynamics for Enhanced Multimodal Event-based Action Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5969–5979.

Cao, M.; Shu, X.; Zhang, J.; Yan, R.; Li, Z.; and Tang, J. 2024a. EventCrab: Harnessing Frame and Point Synergy for Event-based Action Recognition and Beyond. *CoRR*, abs/2411.18328.

Cao, M.; Yan, R.; Shu, X.; Dai, G.; Yao, Y.; and Xie, G.-S. 2024b. AdaFPP: Adapt-Focused Bi-Propagating Prototype Learning for Panoramic Activity Recognition. In *Proceedings of the ACM International Conference on Multimedia*, 691–700.

Chen, S.; Ge, C.; Tong, Z.; Wang, J.; Song, Y.; Wang, J.; and Luo, P. 2022. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35: 16664–16678.

Chen, Y.; Wang, N.; and Zhang, Z. 2018. Darkrank: Accelerating deep metric learning via cross sample similarities transfer. In *Proceedings of the AAAI conference on artificial intelligence*.

Dataset, E. 2011. Novel datasets for fine-grained image categorization. In *First Workshop on Fine Grained Visual Categorization, CVPR*. Citeseer.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the International Conference on Learning Representations*.

Ermolov, A.; Mirvakhabova, L.; Khrulkov, V.; Sebe, N.; and Oseledets, I. V. 2022. Hyperbolic Vision Transformers: Combining Improvements in Metric Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7399–7409.

Fang, Z.; Jiang, X.; Tang, H.; and Li, Z. 2024. Learning Contrastive Self-Distillation for Ultra-Fine-Grained Visual Categorization Targeting Limited Samples. *IEEE Transactions. on Circuits and Systems for Video Technology*.

Gao, J.; Sun, Y.; Liu, Y.; Tang, Y.; Zeng, Y.; Qi, D.; Chen, K.; and Zhao, C. 2025a. Styleshot: A snapshot on any style. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Gao, J.; Sun, Y.; Shen, F.; Jiang, X.; Xing, Z.; Chen, K.; and Zhao, C. 2025b. FaceShot: Bring Any Character into Life. *arXiv preprint arXiv:2503.00740*.

Han, K.; Guo, J.; Zhang, C.; and Zhu, M. 2018. Attribute-Aware Attention Model for Fine-grained Representation Learning. In *Proceedings of the ACM International Conference on Multimedia*, 2040–2048.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *Proceedings of the International Conference on Learning Representations*.

Jiang, X.; Tang, H.; Gao, J.; Du, X.; He, S.; and Li, Z. 2024a. Delving into Multimodal Prompting for Fine-Grained Visual Classification. In *Proceedings of the AAAI conference on artificial intelligence*, 2570–2578.

Jiang, X.; Tang, H.; and Li, Z. 2024. Global meets local: Dual activation hashing network for large-scale fine-grained image retrieval. *IEEE Transactions on Knowledge and Data Engineering*.

Jiang, X.; Tang, H.; Yan, R.; Tang, J.; and Li, Z. 2024b. DVF: Advancing Robust and Accurate Fine-Grained Image Retrieval with Retrieval Guidelines. In *Proceedings of the ACM International Conference on Multimedia*, 2379–2388.

Kim, S.; Kim, D.; Cho, M.; and Kwak, S. 2020. Proxy Anchor Loss for Deep Metric Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3235–3244.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3D Object Representations for Fine-Grained Categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 554–561.

Lim, J.; Yun, S.; Park, S.; and Choi, J. Y. 2022. Hypergraph-Induced Semantic Tuplet Loss for Deep Metric Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 212–222.

Liu, L.; Huang, S.; Zhuang, Z.; Yang, R.; Tan, M.; and Wang, Y. 2022. DAS: Densely-Anchored Sampling for Deep Metric Learning. In *Proceedings of the European Conference on Computer Vision*, 399–417.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; and Zhang, L. 2023. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. *CoRR*, abs/2303.05499.

Moskvyak, O.; Maire, F.; Dayoub, F.; and Baktashmotlagh, M. 2021. Keypoint-aligned embeddings for image retrieval and re-identification. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 676–685.

Movshovitz-Attias, Y.; Toshev, A.; Leung, T. K.; Ioffe, S.; and Singh, S. 2017. No Fuss Distance Metric Learning Using Proxies. In *Proceedings of the IEEE International Conference on Computer Vision*, 360–368.

Park, W.; Kim, D.; Lu, Y.; and Cho, M. 2019. Relational knowledge distillation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 3967–3976.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable

Visual Models From Natural Language Supervision. In *Proceedings of the International Conference on Machine Learning*, 8748–8763.

Ramdurai, B. 2025. Large Language Models (LLMs), Retrieval-Augmented Generation (RAG) systems, and Convolutional Neural Networks (CNNs) in Application systems. *International Journal of Marketing and Technology*, 15(01).

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. S.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3): 211–252.

Seidenschwarz, J. D.; Elezi, I.; and Leal-Taixé, L. 2021. Learning Intra-Batch Connections for Deep Metric Learning. In *Proceedings of the International Conference on Machine Learning*, 9410–9421.

Song, H. O.; Xiang, Y.; Jegelka, S.; and Savarese, S. 2016. Deep Metric Learning via Lifted Structured Feature Embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4004–4012.

Tang, H.; He, S.; and Qin, J. 2025. Connecting Giants: Synergistic Knowledge Transfer of Large Multimodal Models for Few-Shot Learning. *arXiv preprint arXiv:2510.11115*.

Tang, H.; Li, Z.; Peng, Z.; and Tang, J. 2020a. Blockmix: meta regularization and self-calibrated inference for metric-based meta-learning. In *Proceedings of the 28th ACM international conference on multimedia*, 610–618.

Tang, H.; Li, Z.; Peng, Z.; and Tang, J. 2020b. BlockMix: Meta Regularization and Self-Calibrated Inference for Metric-Based Meta-Learning. In *Proceedings of the ACM International Conference on Multimedia*, 610–618.

Tang, H.; Li, Z.; Zhang, D.; He, S.; and Tang, J. 2024. Divide-and-conquer: Confluent triple-flow network for RGB-T salient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Tang, H.; Liu, J.; Yan, S.; Yan, R.; Li, Z.; and Tang, J. 2023. M3net: multi-view encoding, matching, and fusion for few-shot fine-grained action recognition. In *Proceedings of the 31st ACM international conference on multimedia*, 1719–1728.

Tang, H.; Yuan, C.; Li, Z.; and Tang, J. 2022. Learning attention-guided pyramidal features for few-shot fine-grained recognition. *Pattern Recognition*, 130: 108792.

Teh, E. W.; DeVries, T.; and Taylor, G. W. 2020. ProxyNCA++: Revisiting and Revitalizing Proxy Neighborhood Component Analysis. In *Proceedings of the European Conference on Computer Vision*, 448–464.

Vendrow, E.; Pantazis, O.; Shepard, A.; Brostow, G.; Jones, K.; Mac Aodha, O.; Beery, S.; and Van Horn, G. 2024. INQUIRE: A natural world text-to-image retrieval benchmark. *Advances in Neural Information Processing Systems*, 37: 126500–126514.

Wang, C.; Zheng, W.; Li, J.; Zhou, J.; and Lu, J. 2023a. Deep Factorized Metric Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7672–7682.

Wang, L.; and Yoon, K.-J. 2021. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6): 3048–3068.

Wang, S.; Chang, J.; Li, H.; Wang, Z.; Ouyang, W.; and Tian, Q. 2023b. Open-Set Fine-Grained Retrieval via Prompting Vision-Language Evaluator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 19381–19391.

Wang, S.; Chang, J.; Wang, Z.; Li, H.; Ouyang, W.; and Tian, Q. 2023c. Fine-Grained Retrieval Prompt Tuning. In *Proceedings of the AAAI conference on artificial intelligence*, 2644–2652.

Wang, S.; Wang, Z.; Wang, N.; Wang, H.; and Li, H. 2022. From coarse to fine: multi-level feature fusion network for fine-grained image retrieval. *Multimedia Systems*, 28(4): 1515–1528.

Wei, X.; Luo, J.; Wu, J.; and Zhou, Z. 2017. Selective Convolutional Descriptor Aggregation for Fine-Grained Image Retrieval. *IEEE Transactions on Image Processing*, 26(6): 2868–2881.

Xing, P.; Wang, H.; Sun, Y.; Wang, Q.; Bai, X.; Ai, H.; Huang, R.; and Li, Z. 2024. Csgo: Content-style composition in text-to-image generation. *arXiv preprint arXiv:2408.16766*.

Xing, P.; Wang, N.; Ouyang, J.; and Li, Z. 2025. Inv-Adapter: ID Customization Generation via Image Inversion and Lightweight Parameter Adapter. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zheng, X.; Ji, R.; Sun, X.; Wu, Y.; Huang, F.; and Yang, Y. 2018. Centralized Ranking Loss with Weakly Supervised Localization for Fine-Grained Object Retrieval. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 1226–1233.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional Prompt Learning for Vision-Language Models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to Prompt for Vision-Language Models. *International Journal of Computer Vision*.