View-Robust Backbone and Discriminative Reconstruction for Few-Shot Fine-Grained Image Classification

Jiawen Jiang[1], Jiahang Li[2], Jin Lu[1,*]
[1]School of Electronic Information and Artificial Intelligence, Shaanxi University of Science & Technology, Xi'an, China
[2]School of Artificial Intelligence, Tianjin Normal University, Tianjin, China
*Corresponding author: lujin@sust.edu.cn

***Abstract*—We study few-shot fine-grained image classification, a task that faces two key challenges: (1) the scarcity of labeled samples amplifies the model's sensitivity to viewpoint variations, resulting in feature inconsistency, and (2) reconstruction-based methods, while improving inter-class separability, inadvertently introduce intra-class variations, further complicating discrimination. To address these challenges, we propose the View-Robust Attention Selector (VRAS), a feature enhancement backbone designed to mitigate viewpoint-induced misclassifications. By integrating cross-scale feature interaction and adaptive selection mechanisms, VRAS effectively reduces spatial sensitivity arising from the limited viewpoint diversity in few-shot support sets. This approach not only preserves intra-class consistency but also enhances inter-class discriminability, ensuring robust feature representations. Furthermore, we introduce the Enhancement and Reconstruction (ER) module, designed to strengthen discriminative learning. ER achieves this by maximizing inter-class divergence while enhancing intra-class compactness through a regularized Ridge Regression optimization strategy. By dynamically suppressing low-saliency dimensions, ER maintains geometric coherence and effectively filters out semantic noise. Extensive experiments on three fine-grained datasets show that our method significantly outperforms state-of-the-art few-shot classification methods. Codes are available at *https://github.com/jiangjiawen321/VRAS* .**

***Index Terms*—Few-shot learning, Fine-grained image classification, Viewpoint-robust learning, Discriminative feature representation, Ridge regression optimization**

I. Introduction

Deep neural networks trained with large-scale datasets have dramatic advances in **F**ine **G**rained **I**mage **C**lassification (FGIC) [1], [2]. However, their success heavily depends on large-scale datasets with meticulously annotated fine-grained labels by domain experts. To circumvent the prohibitive costs of large-scale labeling, recent research has shifted focus to **F**ew-**S**hot **F**ine-**G**rained **I**mage **C**lassification (FS-FGIC) [3]–[7]. In the conventional **F**ew-**S**hot **L**earning (FSL) setting [8], [9], it requires effective model transfer using only a few support samples [5]. However, in fine-grained tasks, the model must capture fine-grained feature distinctions. Consequently, a key prerequisite for FS-FGIC tasks is ensuring model robustness to sparse samples and feature discriminability, which are crucial for achieving accurate classification under such challenging conditions.

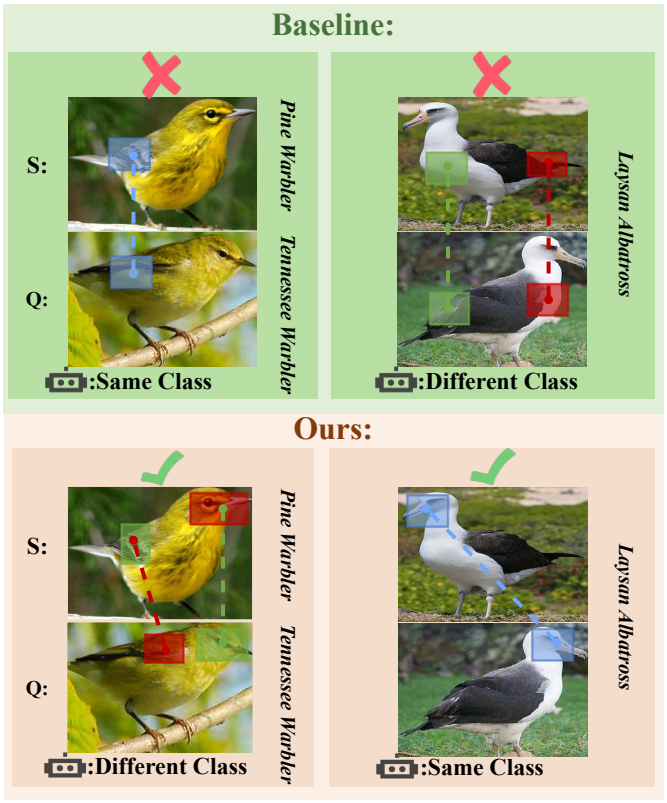


Fig. 1: A toy example illustrating feature representations and predictions by the baseline (FRN) and our View-Robust Attention Selector. Support (S) and query (Q) sets are marked; colors denote adaptive attention weights.

Motivated by this, researchers have started to focus on feature representation in FS-FGIC tasks. However, it remains challenging due to the following two reasons. The first is the difficulty of handling the scarcity of labeled samples, which makes the model overly sensitive to variations in image viewpoints. FGIC tasks require obtaining sufficiently detailed information to achieve fine-grained discriminative capability.

However, in FSL tasks, the scarcity of labeled data fails to guarantee intra-class spatial viewpoint diversity (*e.g.*, most samples in the Laysan Albatross support set share identical shooting angles). This results in the model being overly sensitive to variations in image viewpoints during fine-grained feature learning, thereby hindering the development of consistent intra-class feature representations (see Fig. 1 for baseline Laysan Albatross prediction outcomes). The second is the difficulty of enhancing inter-class fine-grained discriminability without compromising the consistency of intra-class feature representations. A fundamental conflict exists between FGIC and FSL: FGIC needs to learn distinguishable fine-grained features between subspecies, while FSL must extract inter-class discriminative features from extremely limited samples (*e.g.*, head and wing textures of Pine Warbler vs Tennessee Warbler) see Fig. 1. Simultaneously learning local and global features under limited training data may introduce additional intra-class noise which counteracts the goal of enlarging inter-class margins while reducing intra-class variations. A naïve approach to mitigate FS-FGCI issues is combining feature maps from multiple layers [3], [4], as features extracted from different backbone layers exhibit complementary properties. However, simply extracting and concatenating feature maps introduces additional intra-class noise in FSL settings, while failing to alleviate the model's hypersensitivity to viewpoint variations.

It was only recently that reconstruction-based methods started to gain popularity [5]–[7], [10]. Wertheimer *et al.* [6] proposed a **F**eature-map **R**econstruction **N**etworks (FRN) that reconstructs a new query set from a support set using ridge regression. The metric score is then calculated based on the reconstruction error between the new query set and the original query set. However, the semantic information represented by the reconstruction methods of FRN is entirely different between any two samples belonging to the same class, as shown in Fig. 1 (*e.g.*, the FRN focuses on the physical spatial location of features for the Laysan Albatross). This suggests that reconstruction-based methods still exhibit significant intra-class variations. We hypothesize that the inability to address a fundamental conflict between FGIC and FSL prevents overcoming the model's hypersensitivity to image viewpoint variations, thereby leading to inconsistent intra-class feature representations.

To tackle these challenges, we propose a novel backbone, the **V**iew-**R**obust **A**ttention **S**elector (VRAS), along with a discriminative reconstruction method. Specifically, the VRAS architecture consists of a **C**ross-scale **F**eature **E**xtractor (CFE) that enhances the backbone and a **C**ross-scale **F**eature **S**elector (CFS) for adaptive feature selection. The CFE and CFS modules capture cross-scale interactive features and adaptively select informative ones, respectively. This strategy preserves both global context and fine-grained details, reducing the model's reliance on specific viewpoints under the few-shot setting, which often leads to misclassification. Consequently, it ensures *intra-class consistency* and *inter-class distinctiveness* (see Fig. 1, Ours). In addition, to further enhance inter-class separability and maintain intra-class compactness, we introduce an **E**nhancement and **R**econstruction (ER) module. The enhancement stage applies adaptive spatial filtering to suppress low-saliency dimensions and improve the condition of class-specific representations. The reconstruction stage employs regularized ridge regression to disentangle semantic noise and preserve the local geometric structure of the class manifold during cross-sample reconstruction. Together, these two stages refine the feature space, promoting intra-class coherence and reinforcing inter-class separation. We validate the proposed framework on three benchmark datasets, demonstrating its robustness to viewpoint variations and superior discriminative performance.

II. Related Work

A. Few-Shot Image Classification

Existing popular few-shot learning methods can be broadly classified into two categories: meta-learning based methods [11], [12] and metric-learning based methods [13], [14]. Meta-learning-based methods focus on learning meta-knowledge that enables the model to generalize well with only a few training examples per class. These methods aim to learn how to learn by optimizing a model's ability to adapt quickly to new tasks. This is typically achieved by training on a variety of tasks, where the model learns a shared representation or optimization strategy that can be applied to unseen tasks with limited data. metric-learning based methods aim to embed samples into a feature space where the distances between samples reflect their similarity. These methods use similarity metrics (*e.g.*, Euclidean distance or cosine similarity) to assess the closeness between samples, enabling the model to classify new samples based on how similar they are to known samples. In metric learning, the objective is to learn a distance function or embedding space where the features from the same class are close together and those from different classes are far apart.

B. Few-Shot Fine-grained Image Classification

Few-shot fine-grained image classification extends conventional FSL by requiring joint modeling of subtle inter-class distinctions and viewpoint-robust representations. While initial attempts leverage multi-layer feature fusion to combine local details (*e.g.*, feather textures) with global morphology, these approaches suffer from two critical limitations: naïve concatenation schemes introduce spurious intra-class variations, and inherent spatial sensitivity persists due to viewpoint scarcity in support sets. Recent work by Du *et al.* [3] demonstrates that jigsaw-driven patch mining can amplify discriminative local cues, whereas Zhu *et al.* [4] employs attention-guided layer aggregation to suppress irrelevant features. However, neither method explicitly addresses the fundamental tension between granularity preservation and spatial invariance in low-shot regimes.

Existing few-shot fine-grained methods based on feature reconstruction align the spatial positions of fine-grained image objects, enhancing the adaptability of FS-FGIC tasks. Wertheimer *et al.* [6] used ridge regression to obtain optimal

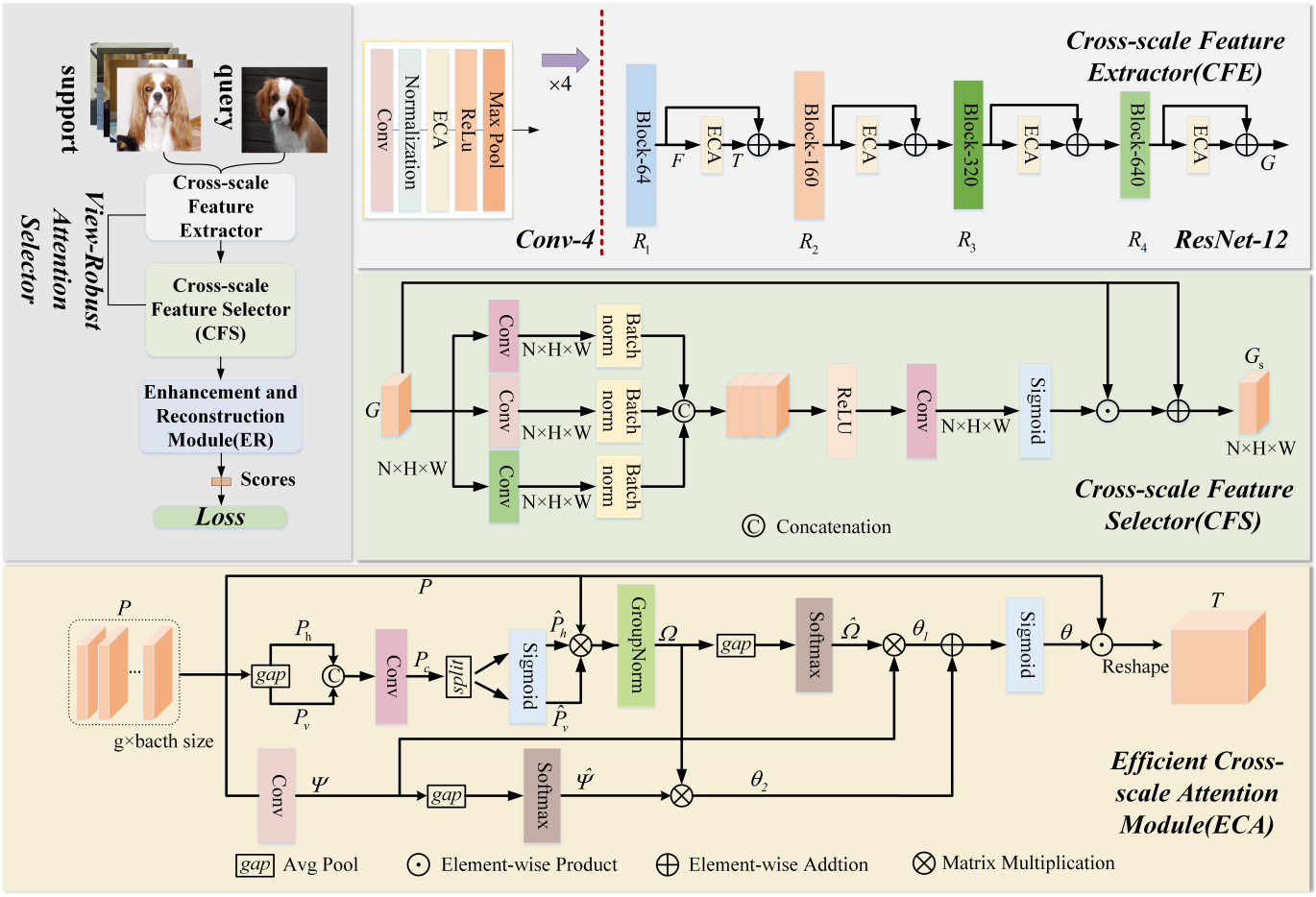


Fig. 2: Overview of our Network.

weights for the least squares method (FRN) and introduced feature reconstruction to address the few-shot image classification problem. Wu *et al.* [5] proposed a bidirectional feature reconstruction network that utilizes a self-attention mechanism to mutually reconstruct the query and support sets, overcoming the limitations of FRN's unidirectional reconstruction. Guo *et al.* [15] proposed a dual-feature reconstruction network (DFRN) for few-shot image classification, which constructs dual feature vectors through two modules, improving its generalization performance. However, these reconstruction methods focus on sample-level features, making it difficult to ensure intra-class consistency.

In contrast, our VRAS enhances the feature extractor's robustness to viewpoint variations by promoting intra-class consistency and inter-class separability. Complementarily, the ER module further refines the feature space by adaptively suppressing low-saliency dimensions and applying regularized ridge regression, which preserves local geometric structure and disentangles semantic noise.

III. Method

A. Overview

As illustrated in Fig. 2, our proposed network integrates a View-Robust Attention Selector (VRAS) backbone to address perspective variations and an Enhancement and Reconstruction (ER) module to ensure manifold geometric consistency through feature refinement and semantic alignment.

The FSFGIC task setting contains a support set A and a query set B. The support set A contains C different image classes, each with K labeled samples. The query set B consists of M unlabeled samples. Both sets share the same label space. The goal of FSFGIC is to train a model which has the ability to classify each query sample b ($b \in B$) into its corresponding class in C. Thus, the FSFGIC task is called a C-way K-shot task.

B. View-Robust Attention Selector

1) ***Basic Backbone****:* Conv-4: Conv-4 has 4 blocks, each block consists of a 3×3 convolution layer, a batch normal layer, an ECA module, a max pooling layer and a ReLu layer. ResNet-12: ResNet-12 has four residual blocks, each consisting of 1×1 convolution layer, 3×3 convolution layer, a batch normal layer, a max pool layer, an ECA module, and a relu layer.

2) ***Cross-scale Feature Extractor****:* By introducing the Efficient Cross-scale Attention (ECA) module into Conv-4 or ResNet-12, a cross-scale feature extractor is proposed. Taking the first ResNet block of ResNet-12 with ECA as an example.

For each image X, after passing through the first ResNet block R_1, the feature map is $F = R_1(X) \in \mathbb{R}^{n\times h\times w}$, where n is the number of channels after the first ResNet block, h and w are the height and width of the feature map at this time, respectively. First, the feature map F is reshaped to $P \in \mathbb{R}^{n//g\times h\times w}$. Then the feature map P is the input of the ECA module. Then, 1×1 scale branch and 3×3 scale branch are used to extract attention weight descriptors for grouped feature map, respectively.

In the 1×1 scale branch, P is decomposed into two 1D feature vectors $P_v \in \mathbb{R}^{n//g\times h\times 1}$ and $P_h \in \mathbb{R}^{n//g\times 1\times w}$ using a global average pooling operation, where P_v and P_h represent feature information along the vertical dimension and horizontal dimensions of the 1×1 scale, respectively. In the following process, P_v, P_h are connected and merged as $P_c \in \mathbb{R}^{n//g\times 1\times (h+w)}$ through a 1×1 convolution. After decomposing P_c again into two vectors, Sigmoid functions are employed to obtain the attention weights $\hat{P}_v \in \mathbb{R}^{n//g\times h\times 1}$, $\hat{P}_h \in \mathbb{R}^{n//g\times 1\times w}$. Then, the spatial feature map $\Omega \in \mathbb{R}^{n//g\times h\times w}$ is obtained by element-wise multiplication of $\hat{P}_v$, $\hat{P}_h$ and the original feature map P, followed by group normalization. The 3×3 scale branch captures spatial feature map $\Psi \in \mathbb{R}^{n//g\times h\times w}$ through 3×3 convolution.

In the 1×1 scale branch, a 2D global average pooling operation followed by the Softmax is applied to the feature map Ω to obtain the attention weight $\hat{\Omega} \in \mathbb{R}^{1\times n//g}$. Then by performing matrix multiplication on Ψ and $\hat{\Omega}$, we obtain a spatial attention map $\Theta_1 \in \mathbb{R}^{1\times h\times w}$. Similarly, in the 3×3 scale branch, the Ψ is sent through 2D global average pooling and Softmax to obtain the attention weight $\hat{\Psi} \in \mathbb{R}^{1\times n//g}$. And through the matrix multiplication of $\hat{\Psi}$ and Ω, a spatial attention map $\Theta_2 \in \mathbb{R}^{1\times h\times w}$ is derived. The cross-scale attention weights Θ are obtained by aggregating the spatial attention map Θ_1 and Θ_2, followed by the Sigmoid function (σ). Finally, cross-scale attention weights Θ is element-wise multiplied with P and reshaped to obtain the final output $T \in \mathbb{R}^{n\times h\times w}$. The main computation process of the ECA module is given by the following formula:

$$T = \sigma((\Psi \otimes \hat{\Omega}) + (\hat{\Psi} \otimes \Omega)) \odot P. \tag{1}$$

*3) **Cross-scale Feature Selector**:* To adaptively select cross-scale feature information from the feature extractor, a new cross-scale feature selector is designed. Specifically, three types of convolutions with different receptive fields are used: 1×1 convolution , 3×3 convolution and 3×3 dilated convolution. The output feature map is then concatenated and directly activated using the ReLU function. Finally, the most valuable feature information in different scale spaces is adaptively selected through another 1×1 convolution, as shown in the following formula:

$$\begin{aligned} G_{Concat} = \sigma_1(Concat\{&BN(Conv_{1\times1}(G)) \\ &BN(Conv_{3\times3}(G)) \\ &BN(Conv_{3\times3}(G))\}), \end{aligned} \tag{2}$$

$$G_s = G \times \sigma_2\left(Conv_{1\times1}\left(G_{Concat}\right)\right) + G, \tag{3}$$

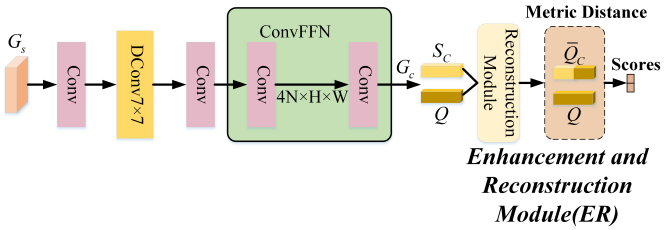


Fig. 3: The structure of Enhancement and Reconstruction Module Module.

where G represents the cross-scale feature information obtained through the feature extractor, G_{Concat} is the concatenated feature. G_s is the output feature map of the cross-scale feature selection module. σ_1 and σ_2 respectively represent the ReLU function and Sigmoid function. The experiment on viewpoint robustness is presented in Section IV-D2.

C. Enhancement and Reconstruction Module

To enhance inter-class separation and reduce intra-class variation, we introduce an enhancement and reconstruction module, as illustrated in Fig. 3. The process of this module is as follows:

$$\begin{aligned} G_C &= \text{ConvFFN}\left(\text{Conv}_{1\times1}\left(\text{DConv}_{7\times7}\left(\text{Conv}_{1\times1}(G_s)\right)\right)\right), \\ \overline{W} &= \arg\min_W \|Q - WS_C\|^2 + \lambda\|W\|^2, \\ \overline{W} &= QS_C^\top\left(S_CS_C^\top + \lambda I\right)^{-1}, \\ \overline{Q}_C &= \overline{W}S_C, \end{aligned} \tag{4}$$

where $Conv_{1\times1}$ represents 1×1 convolution, $DConv_{7\times7}$ represents a depth separable 7×7 convolution, $\|\cdot\|$ is Frobenius norm, λ is the regularization parameter and $\overline{Q}_C$ is the optimal reconstruction. The image representation G_C is divided into support set feature S_C and query set feature Q according to the experimental settings. Specifically, feature reconstruction mainly reconstructs query set features from the support set features through linear changes, such that $W \cdot S_C \approx Q$.

Theorem 1 (Manifold Geometric Integrity of the ER Module)**.** *Assume that the Enhancement and Reconstruction (ER) module in Eq. 4 preserves the intrinsic geometric structure of the support set via its joint enhancement (spatial filtering) and reconstruction (ridge regression) operations. Then:*

*1) **Condition Number Bound:** The support set feature matrix S_C satisfies*

$$\kappa(S_C) = \frac{\sigma_{\max}(S_C)}{\sigma_{\min}(S_C)} \le \sqrt{\frac{1+\epsilon}{1-\epsilon}}, \tag{5}$$

where $\sigma_{\max}(S_C)$ and $\sigma_{\min}(S_C)$ denote the largest and smallest singular values of S_C, respectively, and $\epsilon > 0$ is a small constant.

2) ***Local Isometry:*** *The reconstruction mapping $f : S_C \to Q_C$, defined by $f(x) = Wx$ from the ridge solution, is ϵ-isometric on local neighborhoods $B_\delta(p)$, i.e.,*

$$(1-\epsilon)\|x-y\| \le \|f(x)-f(y)\| \le (1+\epsilon)\|x-y\|, \quad \forall\, x, y \in B_\delta(p). \tag{6}$$

Proof. **(i) Spatial Filtering.** The 7×7 depthwise separable convolution in the enhancement stage acts as a spatial low-pass filter that attenuates high-frequency noise. This ensures that S_C mainly contains signal-dominated components, thereby preventing the smallest singular values from approaching zero.

(ii) Ridge Regression. In the reconstruction stage, the ridge regression

$$W = \arg\min_W \|Q - WS_C\|_F^2 + \lambda\|W\|_F^2, \tag{7}$$

leads to a closed-form solution that, via the SVD $S_C = U\Sigma V^T$, effectively scales the singular values as

$$\sigma_i' = \frac{\sigma_i}{\sigma_i^2 + \lambda}. \tag{8}$$

For sufficiently large σ_i, one has $\sigma_i' \approx 1/\sigma_i$, whereas for small σ_i, σ_i' is strongly damped. With an appropriate choice of λ, this modulation guarantees that

$$\kappa(S_C) = \frac{\sigma_{\max}(S_C)}{\sigma_{\min}(S_C)} \le \sqrt{\frac{1+\epsilon}{1-\epsilon}}, \tag{9}$$

thus ensuring a well-conditioned S_C.

(iii) Local Isometry. Owing to the effective preservation of principal components by ridge regression, the mapping $f(x) = Wx$ behaves nearly as an isometry on the signal subspace of S_C. Specifically, for any $x, y \in B_\delta(p)$,

$$(1-\epsilon)\|x-y\| \le \|W(x-y)\| \le (1+\epsilon)\|x-y\|, \tag{10}$$

which establishes the local isometry property. □

Theorem 1 elucidates how the ER module optimizes the feature space for efficient reconstruction through spectral decoupling and singular value modulation. Consequently, it enhances intra-class feature coherence while reinforcing distinct inter-class separation.

For C-class classification, the similarity score between query features Q and reconstructed features $\overline{Q}_C$ is defined as the normalized squared Euclidean distance:

$$d(Q, \overline{Q}_C) = \frac{1}{r}\|Q - \overline{Q}_C\|^2, \tag{11}$$

where r denotes the spatial resolution of Q. The class prediction probability for the i-th query sample is then formulated through softmax normalization:

$$d_i^c = \frac{1}{r}\|Q_i - \overline{Q}_{(c,i)}\|^2, \quad \hat{d}_i^c = \frac{e^{-\gamma d_i^c}}{\sum_{c' \in C} e^{-\gamma d_i^{c'}}}, \tag{12}$$

where $\gamma > 0$ is a learnable scale parameter.

D. Loss Function

After the enhancement and reconstruction module, we learned how to measure the similarity between two images and calculated the prediction probability $\hat{d}_i^c$. Therefore, the total loss of the C-way K-shot few-shot classification task can be defined as:

$$Loss = -\frac{1}{M \times C} \sum_{i=1}^{M \times C} \sum_{c=1}^{C} 1(y_i = c) \log(\hat{d}_i^c), \tag{13}$$

Where M is the number of query set images during training, C is the number of categories of support set images and $1(y_i == c)$ means that the formula is equal to 1 if y_i and c are equal, and otherwise 0. During the training process, for the C-way K-shot classification task, we minimize the $loss$ to update the proposed network and repeat this process on all randomly generated tasks. During the testing process, we calculated the prediction probabilities of the query set images for all categories using Eq. (13), and then classified them into the category with the highest probability.

IV. Experiments

To evaluate the performance of the proposed method, we conducted experiments on three benchmark fine-grained datasets: CUB-200-2011 [27], Stanford Dogs [28], and Stanford Cars [29]. The CUB-200-2011 (CUB) dataset contains $11,788$ images of 200 bird species for fine-grained visual categorization. Following standard protocols [30], all images were preprocessed using manually annotated bounding boxes for region cropping. We adopt the same data split strategy as [6]. Stanford Dogs (Dogs) is a dataset developed by Stanford University for fine-grained dog breed classification. It consists of $20,580$ images spanning 120 different dog breeds. We follow the same data split as in [31]. Stanford Cars (Cars) is a dataset designed for fine-grained vehicle classification, containing $16,185$ images of 196 car categories. Again, we adopt the same data split as in [31]. All images were resized to 84×84 pixels as input to our network, maintaining consistency across experimental settings.

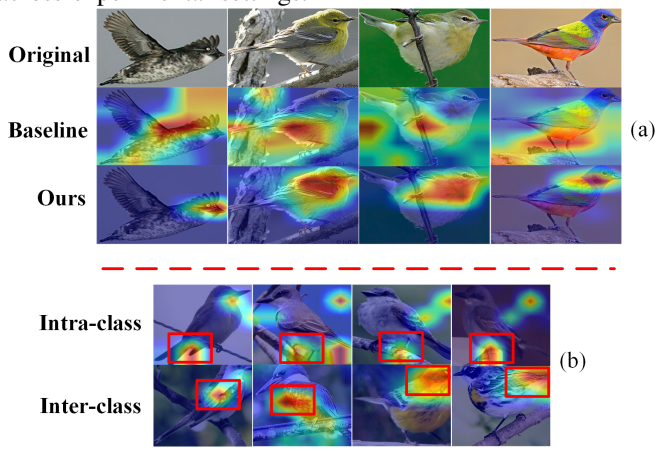


Fig. 4: Qualitative Comparison on the CUB-200-2011 Dataset.

TABLE I: The 5-way few-shot classification accuracy (%) comparison on fine-grained datasets. **Bold** denotes the best performance. Underline denotes the second best performance.

Method	Publication	CUB		Dogs		Cars	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Conv-4 Backbone							
ProtoNet [16]	NeurIPS 2017	64.67±0.23	85.67±0.14	46.27±0.21	70.45±0.16	50.74±0.22	74.78±0.17
Relation [17]	CVPR 2018	63.94±0.92	77.87±0.64	47.35±0.88	66.20±0.74	46.04±0.91	68.52±0.78
DN4 [18]	CVPR 2019	57.45±0.89	84.41±0.58	39.08±0.76	69.81±0.69	34.12±0.68	87.47±0.47
DeepEMD [19]	CVPR 2020	64.08±0.50	80.55±0.71	46.73±0.49	65.74±0.63	61.63±0.27	72.95±0.38
BSNet [20]	TIP 2021	62.84±0.95	85.39±0.56	43.42±0.86	71.90±0.68	40.89±0.77	86.88±0.50
LRPABN [21]	TMM 2021	63.63±0.77	76.06±0.58	45.72±0.75	60.94±0.66	60.28±0.76	73.29±0.58
FRN [6]	CVPR 2021	73.95±0.21	88.67±0.12	60.27±0.22	79.17±0.14	66.09±0.22	86.77±0.12
FRN+TDM [22]	CVPR 2022	77.41±0.21	90.70±0.11	62.77±0.22	79.71±0.14	72.26±0.21	89.55±0.10
AGPF [23]	PR 2022	74.03±0.90	86.54±0.50	60.89±0.98	78.14±0.62	78.14±0.84	87.42±0.57
BSFA [24]	TCSVT 2023	59.40±0.97	74.42±0.62	49.13±0.84	63.27±0.73	56.06±0.89	73.28±0.68
BiFRN [5]	AAAI 2023	78.66±0.20	91.13±0.11	64.48±0.22	81.07±0.14	74.90±0.20	90.63±0.10
LCCRN [25]	TCSVT 2023	77.83±0.21	89.94±0.12	64.30±0.22	80.36±0.14	74.07±0.20	89.95±0.10
C2-Net [7]	AAAI 2024	78.63±0.46	89.48±0.26	66.42±0.50	81.23±0.34	**81.29±0.45**	91.08±0.26
VRAS-Conv-4	IJCNN 2025	**79.20±0.20**	**91.28±0.11**	**68.06±0.21**	**83.26±0.13**	76.17±0.21	**92.10±0.09**
Improvement		↑ 0.54	↑ 0.15	↑ 1.64	↑ 2.03	↓ 5.12	↑ 1.02
ResNet-12 Backbone							
ProtoNet [16]	NeurIPS 2017	69.90±0.20	90.65±0.11	73.00±0.22	86.47±0.13	84.56±0.20	93.36±0.10
DeepEMD [19]	CVPR 2020	75.59±0.30	88.23±0.18	70.38±0.30	85.24±0.18	80.62±0.26	92.63±0.13
FRN [6]	CVPR 2021	83.16±0.19	92.42±0.11	75.93±0.22	88.72±0.13	86.82±0.18	94.77±0.09
FRN+TDM [22]	CVPR 2022	83.26±0.20	92.80±0.11	75.98±0.22	88.70±0.13	86.91±0.17	96.11±0.07
AGPF [23]	PR 2022	78.73±0.84	89.77±0.47	72.34±0.86	84.02±0.57	85.34±0.74	94.79±0.35
BSFA [24]	TCSVT 2023	82.27±0.46	90.76±0.26	69.58±0.50	82.59±0.33	88.93±0.38	95.20±0.20
LCCRN [25]	TCSVT 2023	82.38±0.20	93.11±0.10	75.32±0.22	88.33±0.12	85.76±0.18	96.01±0.07
BiFRN [5]	AAAI 2023	82.13±0.20	93.12±0.11	75.89±0.22	88.60±0.12	87.11±0.17	96.06±0.07
SRN [26]	PR 2024	83.82±0.18	93.45±0.10	76.54±0.21	88.52±0.12	88.02±0.16	96.23±0.07
C2-Net [7]	AAAI 2024	83.65±0.20	92.57±0.10	77.72±0.46	89.59±0.24	86.48±0.40	94.07±0.22
VRAS-ResNet-12	IJCNN 2025	**84.54±0.17**	**93.87±0.09**	**78.56±0.19**	**89.84±0.11**	**88.21±0.16**	**97.06±0.06**
Improvement		↑ 0.72	↑ 0.42	↑ 0.84	↑ 0.25	↑ 0.19	↑ 0.83

TABLE II: Ablation experiments using Conv-4 and ResNet-12 backbones on the CUB, Dogs, and Cars datasets.

VRAS		ER	CUB		Dogs		Cars	
CFS	ECA		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Conv-4 Backbone								
✘	✘	✘	73.73±0.21	88.46±0.13	60.53±0.21	79.29±0.15	67.68±0.22	87.24±0.12
✔	✘	✘	76.26±0.20 ↑ 2.53	89.82±0.12 ↑ 1.36	63.93±0.22 ↑ 3.40	81.33±0.14 ↑ 2.04	72.42±0.21 ↑ 4.74	89.91±0.10 ↑ 2.67
✔	✔	✘	78.29±0.20 ↑ 2.03	91.00±0.11 ↑ 1.18	65.93±0.21 ↑ 2.00	83.04±0.13 ↑ 1.71	74.31±0.22 ↑ 1.89	91.47±0.09 ↑ 1.56
✔	✔	✔	**79.20±0.20** ↑ 0.91	**91.28±0.11** ↑ 0.28	**68.06±0.21** ↑ 2.13	**83.26±0.13** ↑ 0.22	**76.17±0.21** ↑ 1.86	**92.10±0.09** ↑ 0.63
ResNet-12 Backbone								
✘	✘	✘	82.86±0.19	92.41±0.10	76.76±0.21	88.74±0.12	87.90±0.17	95.69±0.07
✔	✘	✘	83.53±0.18 ↑ 0.67	93.32±0.09 ↑ 0.91	77.27±0.20 ↑ 0.51	89.12±0.11 ↑ 0.38	87.74±0.16 ↓ 0.16	96.57±0.06 ↑ 0.88
✔	✔	✘	83.98±0.18 ↑ 0.45	93.42±0.09 ↑ 0.10	78.39±0.19 ↑ 1.12	89.52±0.10 ↑ 0.40	87.96±0.17 ↑ 0.22	97.01±0.06 ↑ 0.44
✔	✔	✔	**84.54±0.17** ↑ 0.56	**93.87±0.09** ↑ 0.45	**78.56±0.19** ↑ 0.17	**89.84±0.11** ↑ 0.32	**88.21±0.16** ↑ 0.25	**97.06±0.06** ↑ 0.05

A. Implementation Details

We evaluate our method on two standard backbones: Conv-4 and ResNet-12. Given an input of size 84×84, Conv-4 and ResNet-12 produce feature maps of size $64 \times 5 \times 5$ and $640 \times 5 \times 5$, respectively. The proposed VRAS module is applied on top of these backbones without altering output dimensions. For training, we adopt standard data augmentation techniques, including random cropping, horizontal flipping, and color jittering [6]. All experiments are implemented in PyTorch and conducted on a single NVIDIA 3090 Ti GPU. We use SGD with an initial learning rate of 0.1, weight decay of 5×10^{-4}, and a step-wise learning rate decay by a factor of 10 every 400 epochs. Models are trained for 1200 epochs, with validation performed every 20 epochs to select the best model. Training episode configurations follow prior works: for Conv-4, we use 20-way 5-shot episodes on CUB and 30-way 5-shot episodes on Dogs and Cars; for ResNet-12, 10-way 5-shot episodes on CUB and 15-way 5-shot episodes on Dogs and Cars. All evaluations are conducted under standard 5-way 1-shot and 5-shot settings, reporting the mean accuracy over 10,000 randomly sampled tasks with 95% confidence intervals.

B. Comparison with State-of-the-Arts

To evaluate the effectiveness of our method on few-shot fine-grained image classification, we conduct experiments on three widely used fine-grained datasets. TABLE I summarizes the comparative results against 14 representative baselines, including five classical few-shot classification methods (ProtoNet [16], RelationNet [17], DN4 [18], DeepEMD [19], and

TABLE III: Ablation study on viewpoint robustness with the Conv-4 backbone on CUB-200-2011. Identically colored blocks highlight comparisons, with darker shades indicating better performance.

Flip	VRAS		Accuracy	
	ECA	CFS	1-shot↑	5-shot↑
Baseline (FRN)				
✘	—	—	71.85 ± 0.21	87.28 ± 0.13
✔	—	—	$73.95 \pm 0.21_{(+2.10)}$	$88.67 \pm 0.12_{(+1.39)}$
Ours				
✘	✔	✘	$77.81 \pm 0.21_{(+2.19)}$	$90.26 \pm 0.05_{(+1.35)}$
✘	✔	✔	77.92 ± 0.21	90.18 ± 0.12
✔	✘	✘	75.62 ± 0.21	88.91 ± 0.12
✔	✔	✘	78.94 ± 0.20	90.78 ± 0.11
✔	✔	✔	$79.20 \pm 0.20_{(+1.28)}$	$91.28 \pm 0.11_{(+1.10)}$

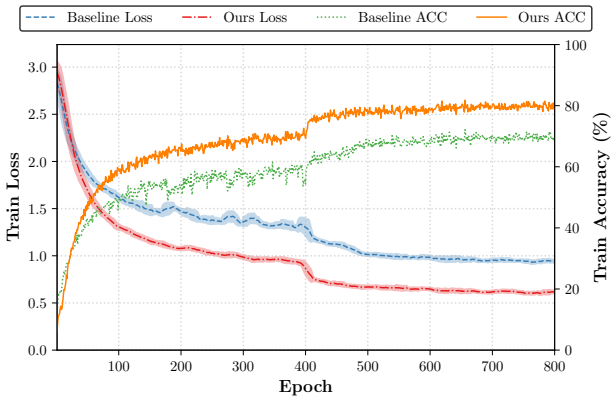
FRN [6]) and nine state-of-the-art approaches specifically designed for few-shot fine-grained image classification (LR-PABN [21], BSNet [20], TDM [22], BiFRN [5], LCCRN [25], C2-Net [7], SRN [26], AGPF [23], and BSFA [24]).

We evaluate our method using Conv-4 and ResNet-12 as base backbones for all comparison approaches, with VRAS built upon these architectures. Compared to both classical few-shot classification methods and those specifically designed for few-shot fine-grained classification, VRAS consistently achieves state-of-the-art performance across multiple datasets. To facilitate comparison, we highlight performance gains (purple) and declines (green) relative to the second-best method. As shown in TABLE I, when using Conv-4 as the base backbone, VRAS surpasses the second-best method on the Dogs dataset by **1.64%** in the 1-shot setting and **2.03%** in the 5-shot setting. On the Cars dataset, VRAS achieves a **1.02%** improvement in the 5-shot setting. The only exception occurs in the 5-way 1-shot setting on the Cars dataset, where VRAS ranks as the second-best method. With ResNet-12 as the base backbone, VRAS continues to demonstrate superior performance, outperforming the second-best method by **0.72%** on CUB (1-shot), **0.84%** on Dogs (1-shot), and **0.83%** on Cars (5-shot). Notably, VRAS achieves an accuracy of **97.06%** on the Cars dataset, underscoring its effectiveness in few-shot fine-grained classification.

C. Qualitative Results

To further evaluate our method, we present qualitative results on the CUB dataset in Fig. 4 and Fig. 5. The Fig. 4 (a) shows the attention regions of both the method FRN and our network. While the baseline captures part of the main information, it also focuses on irrelevant details. In contrast, our network focuses on the discriminative local regions of birds and extracts more compact feature information. In Fig. 4 (b), the intra-class row demonstrates our feature extractor's robustness to viewpoint changes. Despite varying viewing angles, our network consistently attends to the same discriminative features. The inter-class row highlights our model's ability to capture distinctive features unique to different bird subspecies.

Fig. 5: Training loss and average accuracy on the CUB-200-2011 dataset.



As shown in Fig. 5, on the CUB dataset with the Conv-4 backbone, our method achieves lower loss than the baseline by the 30th epoch and surpasses it in average accuracy by the 25th epoch. These results demonstrate that our approach not only improves accuracy but also enhances convergence speed and robustness.

D. Ablation Experiment

1) ***Impact of VRAS and ER****:* In TABLE II, we systematically analyze the effect of removing CFS, ECA, and ER on classification accuracy under consistent experimental settings. We evaluate Conv-4 and ResNet-12 as backbone networks across three fine-grained datasets, reporting results for both 5-way 1-shot and 5-way 5-shot scenarios. To facilitate comparison, we highlight performance gains (purple) and declines (green) as each module is incrementally added. Notably, on the Cars dataset with ResNet-12, adding CFS led to a slight drop of **0.16%**. However, when all three modules (CFS, ECA, and ER) were combined, performance exceeded the baseline by **0.31%**, demonstrating their complementary benefits. In all other cases, each module consistently contributed to improved accuracy, with the best results obtained when all were integrated. These results confirm that VRAS and ER enhance inter-class separability while improving intra-class compactness.

2) ***Robustness to Viewpoint Variations****:* As shown in TABLE III, the experimental results indicate that the preprocessing flip operation effectively mitigates performance degradation caused by limited viewpoint diversity in the support set. With flipping, the baseline achieves a **2.10%** improvement in 1-shot accuracy, while our method improves by **1.28%**. Furthermore, even without flipping, employing the VRAS backbone (orange) in our method still outperforms our approach with flipping but without the VRAS backbone (green), yielding gains of **2.30%** in 1-shot and **1.27%** in 5-shot accuracy. This underscores the effectiveness of VRAS in enhancing viewpoint robustness. Additionally, the integration

of ECA and CFS modules in VRAS consistently improves 5-way few-shot classification performance, regardless of whether flipping is applied. These findings confirm the strong viewpoint robustness of our VRAS backbone, contributing to more accurate classification.

V. Conclusion

This paper addresses viewpoint sensitivity and inter-class indistinguishability in few-shot fine-grained classification. We introduce VRAS, a feature extraction backbone that leverages cross-scale interaction and adaptive selection to enhance viewpoint robustness while maintaining intra-class consistency. Additionally, the ER module reinforces inter-class separability and intra-class compactness through reconstruction-based enhancement. Extensive experiments confirm the effectiveness of our approach, demonstrating significant gains, particularly under high viewpoint variability. These results highlight the robustness and adaptability of our framework in decoupling discriminative and invariant feature learning. In future work, we will investigate a cross-view contrastive unification strategy [32], [33] to effectively optimize global representations and local descriptors, thereby enhancing fine-grained classification performance.

References

[1] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear convolutional neural networks for fine-grained visual recognition," *ECCV*, vol. 40, no. 6, pp. 1309–1322, 2017. 1

[2] X.-S. Wei, Y.-Z. Song, O. Mac Aodha, J. Wu, Y. Peng, J. Tang, J. Yang, and S. Belongie, "Fine-grained image analysis with deep learning: A survey," *ECCV*, vol. 44, no. 12, pp. 8927–8948, 2021. 1

[3] R. Du, D. Chang, A. K. Bhunia, J. Xie, Z. Ma, Y.-Z. Song, and J. Guo, "Fine-grained visual classification via progressive multi-granularity training of jigsaw patches," in *ECCV*, pp. 153–168, Springer, 2020. 1, 2

[4] Y. Zhu, C. Liu, S. Jiang, *et al.*, "Multi-attention meta learning for few-shot fine-grained image recognition.," in *IJCAI*, pp. 1090–1096, Beijing, 2020. 1, 2

[5] J. Wu, D. Chang, A. Sain, X. Li, Z. Ma, J. Cao, J. Guo, and Y.-Z. Song, "Bi-directional feature reconstruction network for fine-grained few-shot image classification," in *AAAI*, vol. 37, pp. 2821–2829, 2023. 1, 2, 3, 6, 7

[6] D. Wertheimer, L. Tang, and B. Hariharan, "Few-shot classification with feature map reconstruction networks," in *CVPR*, pp. 8012–8021, 2021. 1, 2, 5, 6, 7

[7] Z.-X. Ma, Z.-D. Chen, L.-J. Zhao, Z.-C. Zhang, X. Luo, and X.-S. Xu, "Cross-layer and cross-sample feature optimization network for few-shot fine-grained image classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 4136–4144, 2024. 1, 2, 6, 7

[8] Y. Song, T. Wang, P. Cai, S. K. Mondal, and J. P. Sahoo, "A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities," *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1–40, 2023. 1

[9] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM computing surveys (csur)*, vol. 53, no. 3, pp. 1–34, 2020. 1

[10] J. Li, S. Xue, and Y. Su, "Gaze-guided learning: Avoiding shortcut bias in visual classification," *arXiv preprint arXiv:2504.05583*, 2025. 2

[11] Q. Sun, Y. Liu, T.-S. Chua, and B. Schiele, "Meta-transfer learning for few-shot learning," in *CVPR*, pp. 403–412, 2019. 2

[12] Y. Chen, Z. Liu, H. Xu, T. Darrell, and X. Wang, "Meta-baseline: Exploring simple meta-learning for few-shot learning," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9062–9071, 2021. 2

[13] X. Li, L. Yu, C.-W. Fu, M. Fang, and P.-A. Heng, "Revisiting metric learning for few-shot image classification," *Neurocomputing*, vol. 406, pp. 49–58, 2020. 2

[14] W. Jiang, K. Huang, J. Geng, and X. Deng, "Multi-scale metric learning for few-shot learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 3, pp. 1091–1102, 2020. 2

[15] X. Guo, J. Wu, K. Ren, Q. Song, and X. Li, "Dual feature reconstruction network for few-shot image classification," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1579–1584, IEEE, 2023. 3

[16] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in neural information processing systems*, vol. 30, 2017. 6

[17] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1199–1208, 2018. 6

[18] W. Li, L. Wang, J. Xu, J. Huo, Y. Gao, and J. Luo, "Revisiting local descriptor based image-to-class measure for few-shot learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7260–7268, 2019. 6

[19] C. Zhang, Y. Cai, G. Lin, and C. Shen, "Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12203–12213, 2020. 6

[20] X. Li, J. Wu, Z. Sun, Z. Ma, J. Cao, and J.-H. Xue, "Bsnet: Bi-similarity network for few-shot fine-grained image classification," *IEEE Transactions on Image Processing*, vol. 30, pp. 1318–1331, 2020. 6, 7

[21] H. Huang, J. Zhang, J. Zhang, J. Xu, and Q. Wu, "Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification," *IEEE Transactions on Multimedia*, vol. 23, pp. 1666–1680, 2020. 6, 7

[22] S. Lee, W. Moon, and J.-P. Heo, "Task discrepancy maximization for fine-grained few-shot classification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5331–5340, 2022. 6, 7

[23] H. Tang, C. Yuan, Z. Li, and J. Tang, "Learning attention-guided pyramidal features for few-shot fine-grained recognition," *Pattern Recognition*, vol. 130, p. 108792, 2022. 6, 7

[24] Z. Zha, H. Tang, Y. Sun, and J. Tang, "Boosting few-shot fine-grained recognition with background suppression and foreground alignment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 8, pp. 3947–3961, 2023. 6, 7

[25] X. Li, Q. Song, J. Wu, R. Zhu, Z. Ma, and J.-H. Xue, "Locally-enriched cross-reconstruction for few-shot fine-grained image classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 12, pp. 7530–7540, 2023. 6, 7

[26] X. Li, Z. Li, J. Xie, X. Yang, J.-H. Xue, and Z. Ma, "Self-reconstruction network for fine-grained few-shot classification," *Pattern Recognition*, vol. 153, p. 110485, 2024. 6, 7

[27] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011. 5

[28] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li, "Novel dataset for fine-grained image categorization: Stanford dogs," in *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, vol. 2, 2011. 5

[29] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013. 5

[30] C. Zhang, Y. Cai, G. Lin, and C. Shen, "Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12203–12213, 2020. 5

[31] W. Li, L. Wang, J. Xu, J. Huo, Y. Gao, and J. Luo, "Revisiting local descriptor based image-to-class measure for few-shot learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7260–7268, 2019. 5

[32] X. Liu, Y. Zhu, L. Liu, J. Tian, and L. Wang, "Feature-suppressed contrast for self-supervised food pre-training," in *ACM MM*, pp. 4359–4367, 2023. 8

[33] J. Lin, Y. Zheng, X. Chen, Y. Ren, X. Pu, and J. He, "Cross-view contrastive unification guides generative pretraining for molecular property prediction," in *ACM MM*, pp. 2108–2116, 2024. 8