PRESERVING PRETRAINED INTRA-CLASS STRUCTURE IN FINE-GRAINED VISUAL ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision foundation models often organize samples from the same category into distinct structures in feature space, reflecting variations in viewpoint, pose, appearance, and local details. Modeling such structure is crucial for fine-grained recognition, where subtle inter-class differences coexist with large intra-class variation. Yet downstream fine-tuning is typically driven by category-level supervision and does not explicitly account for this pretrained intra-class organization. We further observe that such structure is category-dependent: some categories form a single compact mode, while others contain several distinct modes. This motivates us to exploit category-specific structure already present in pretrained features. We propose Adaptive Proxy Modeling (APM), which models pretrained intra-class structure with a category-specific set of proxies. For each category, APM estimates the number of visual modes from the spectral structure of its DINOv2 feature graph and constructs corresponding proxies through expectation-maximization. During fine-tuning, these proxies are dynamically updated with the evolving representation and provide structural guidance for downstream learning. Unlike conventional proxy-based methods using a fixed number of freely learned proxies, APM adapts both proxy number and representation to each category without introducing additional trainable parameters. Across fine-grained classification, person re-identification, and vehicle re-identification, APM consistently improves downstream performance without introducing additional trainable parameters. Beyond accuracy, APM also better preserves pretrained representations after fine-tuning, suggesting that effective adaptation can benefit from retaining the category-specific structure already learned by foundation models.

1 INTRODUCTION

Vision foundation models learned rich visual representations that capture variations in viewpoint, pose, appearance, and local structure (Dünkel et al., 2026; Banani et al., 2024). For samples from the same category, these variations can form distinct modes in the pretrained feature space, revealing intra-class organization. Such structure is particularly relevant to fine-grained visual recognition (FGVR) (Qian et al., 2015; He & Peng, 2017), where categories are visually similar to each other but may exhibit substantial variations within each category (Jiang et al., 2024; Yu et al., 2022). Yet, downstream fine-tuning is typically driven by category-level supervision and does not explicitly account for how such intra-class structures are organized in the pretrained representation.

A further observation is that intra-class structure is highly category-dependent. As illustrated in Figure 1(a), samples from Identity A form a single compact mode, while Identities B and C exhibit more diverse visual patterns and form two and three modes, respectively. Therefore, different categories may require different levels of representational complexity. This suggests that effectively exploiting pretrained intra-class structure requires not only identifying representative visual modes, but also adapting their number to each category.

Prototype-based representation learning provides a natural way to model such structures (Kim et al., 2020). Multi-proxy methods (Mao et al., 2023; Zou et al., 2023; Yu et al., 2025) extend a single category prototype to multiple proxies, allowing different proxies to capture distinct intra-class patterns. However, existing methods typically assign the same number of proxies to all categories as a manually selected hyperparameter. Moreover, proxies are often treated as freely learned param-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

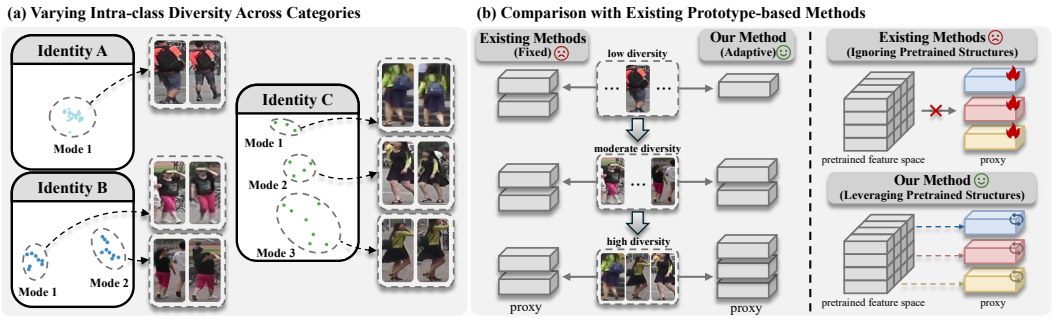


Figure 1: **Motivation of our method.** (a) Different categories exhibit different levels of intra-class diversity and may form different numbers of visual modes. (b) Our method leverages pretrained feature structure to adaptively determine the number of proxies for each category and construct the corresponding proxy representations.

eters optimized only by downstream objectives (Ren et al., 2024; Park et al., 2025; Bhatnagar & Ahuja, 2025), without explicitly using the category-specific structure already encoded in pretrained features. As illustrated in Figure 1(b), this leaves two closely related questions: how many proxies should each category use, and how should these proxies reflect its pretrained feature structure?

To address these questions, we propose Adaptive Proxy Modeling (APM), which models pretrained intra-class structure with a category-specific set of proxies. For each category, APM first constructs a feature graph from pretrained DINOv2 representations and estimates the number of visual modes from its spectral structure. Specifically, we use the largest eigengap of the graph Laplacian to determine the proxy number, and show that the spectral gap associated with the underlying number of modes remains stable under small perturbations. Given the estimated number of proxies, APM then uses Expectation-Maximization (EM) to construct proxy representations aligned with the corresponding pretrained modes. The pretrained and downstream feature spaces are not identical, so keeping these proxies fixed would limit adaptation. APM therefore dynamically updates them during fine-tuning through Exponential Moving Average (EMA), allowing the proxies to follow the evolving representation while retaining guidance from the pretrained structure. These adaptive proxies are further used in a representation preservation objective to guide downstream learning. The resulting module supports both single- and multi-class-token settings and can be integrated into different fine-tuning strategies without introducing additional trainable parameters.

We evaluate APM across fine-grained visual classification, person re-identification, and vehicle re-identification. APM consistently improves the corresponding fine-tuning baselines, while adaptive proxy modeling also outperforms single- and fixed-multi-proxy alternatives. Further analysis shows that APM better retains the intra-class organization present in pretrained features. For example, after fine-tuning DINOv2 on Market-1501, ImageNet-1K linear-probe accuracy is 52.9% with APM compared with 19.8% under standard full fine-tuning. These results suggest that category-specific pretrained structure can provide useful guidance for fine-grained vision model adaptation.

In summary, our contributions are threefold:

- We highlight the category-dependent nature of pretrained intra-class structure: different categories can exhibit different numbers of visual modes, motivating adaptive rather than fixed proxy modeling.
- We propose **Adaptive Proxy Modeling (APM)**, which estimates category-specific proxy numbers from the spectral structure of pretrained DINOv2 features, constructs the corresponding proxies through EM, and dynamically updates them during downstream fine-tuning. We further establish the stability of the spectral gap for reliable proxy number estimation under small perturbations.
- We evaluate APM across fine-grained classification, person re-identification, and vehicle re-identification, where it improves downstream performance without additional trainable parameters and better retains pretrained representations after fine-tuning.

2 Related Work

Fine-Grained Visual Recognition. FGVR methods mainly focused on learning discriminative representations through part localization and feature interactions. Bilinear CNNs (Lin et al., 2015) modeled second-order feature interactions, while attention-based approaches explored informative regions (Fu et al., 2017; Zheng et al., 2017). Subsequently, multi-granularity methods (Du et al., 2020) and Transformer-based approaches (He et al., 2022) further enhanced fine-grained representation learning by capturing multi-scale visual cues and long-range feature dependencies, respectively. With the emergence of vision foundation models (*e.g.*, DINOv2 (Oquab et al., 2024)), recent studies adapt pretrained visual representations to FGVR through part discovery and parameter-efficient strategies (Aniraj et al., 2024; Chowdhury et al., 2025; Wang et al., 2026b). However, existing methods primarily focus on enhancing downstream discrimination, while the rich intra-class structures encoded in pretrained feature spaces are not explicitly exploited during fine-tuning. This motivates us to exploit pretrained intra-class structures for adaptive proxy modeling in FGVR.

Prototype-based Representation Learning. Prototype-based representation learning represents semantic structures in feature spaces through representative prototypes. Early studies (Snell et al., 2017; Chen et al., 2019) introduce class-level proxies for classification and metric learning, while subsequent studies (Zhang et al., 2020; Yu et al., 2025; Huang et al., 2025) exploit multiple proxies or proxy-based optimization to capture complex intra-class variations. However, existing approaches (Ren et al., 2024; Park et al., 2025; Bhatnagar & Ahuja, 2025) often treat proxies as independent parameters, without explicitly associating proxies with the corresponding feature subspaces, making proxies less compatible with multi-CLS representations. Moreover, the number of proxies is typically manually defined, without considering the varying intra-class diversity across categories (Mao et al., 2023; Zou et al., 2023). This motivates us to construct proxies within the same subspaces as the corresponding features and adaptively determine the number of proxies according to category-specific structural patterns.

3 Adaptive Proxy Modeling

3.1 Preliminaries

Vision foundation models such as DINOv2 (Oquab et al., 2024) commonly adopt a Vision Transformer (ViT) (Dosovitskiy et al., 2021) backbone. Given an input image x, ViT projects P non-overlapping patches into d-dimensional embeddings $\{z_i^0\}_{i=1}^P$, where $z_i^0 \in \mathbb{R}^d$ denotes the embedding of the i-th patch. To aggregate global visual information, ViT introduces learnable class tokens together with the patch embeddings. We consider a general setting with $K \geq 1$ class tokens. The input sequence is $\mathbf{Z}_0 = [cls_1^0, \ldots, cls_K^0, z_1^0, \ldots, z_P^0]$, and after L Transformer blocks, the output sequence becomes $\mathbf{Z}_L = [cls_1^L, \ldots, cls_K^L, z_1^L, \ldots, z_P^L]$. We use the output class tokens as visual representations. For a category y, the feature set associated with the k-th class token is denoted as $\mathbf{F}_y^k = \{f_i^k \mid y_i = y\}$, where $f_i^k \in \mathbb{R}^d$ is the feature of the i-th sample extracted from the k-th class token. All features are ℓ_2-normalized before further processing, i.e., $\|f_i^k\|_2 = 1$.

3.2 Adaptive Proxy Modeling Module

The goal of APM is to exploit the category-specific intra-class structure already encoded in pretrained features during downstream fine-tuning. As illustrated in Figure 2, APM first estimates the number of visual modes for each category and constructs the corresponding proxy representations from the pretrained feature space. These proxies are then progressively updated during training to remain aligned with the evolving downstream representation. When multiple class tokens are used, we additionally introduce a self-diverse constraint to encourage complementary representations across tokens. For the single-class-token setting, this constraint is omitted.

3.2.1 Adaptive Proxy Construction

Adaptive Proxy Number Estimation. We construct a feature graph for each category based on the similarities among feature representations, where higher feature similarity corresponds to larger edge weights. In this similarity graph, the number of connected components corresponds to the

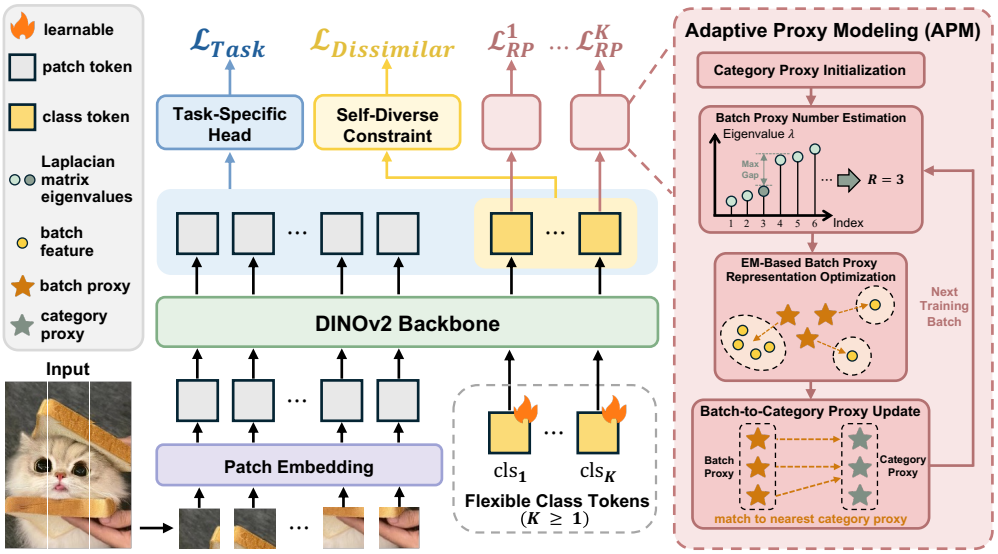


Figure 2: **Overview of Adaptive Proxy Modeling (APM).** The framework supports flexible class-token configurations and combines task supervision, representation preservation, and a self-diverse constraint when multiple class tokens are used. APM first estimates the proxy number for each category from pretrained DINOv2 features and constructs the corresponding proxy representations through EM-based optimization. At each training iteration, APM estimates the proxy number from current mini-batch and optimizes batch proxies. APM then matches batch proxies to category proxies for dynamic updating, providing structural guidance for the evolving feature space.

number of visual modes reflected in the intra-class feature structure and determines the required proxy number. To estimate the number of connected components, we establish two theoretical results based on the spectral properties of the graph Laplacian.

Theorem 1 (Spectral Property of Ideal Graph). For an ideal graph G^*, let $\{\mu_i\}$ denote the eigenvalues of the normalized Laplacian L^* sorted in ascending order. Then, G^* contains exactly R connected components if and only if the eigenvalues satisfy:

$$\mu_1 = \cdots = \mu_R = 0, \quad \mu_{R+1} = \gamma > 0.$$

Theorem 2 (Spectral Property of Perturbed Graph). For a perturbed graph G with normalized Laplacian $L = L^* + E$, where E is a perturbation matrix satisfying $\|E\|_2 \leq \epsilon$, let $\{\lambda_i\}$ denote the eigenvalues of L sorted in ascending order. If $\gamma > 2\epsilon$, then G contains exactly R intrinsic clusters if and only if the eigenvalues satisfy:

$$0 \leq \lambda_i \leq \epsilon, \quad i = 1, \cdots, R,$$

and

$$\gamma - \epsilon \leq \lambda_{R+1} \leq \gamma + \epsilon.$$

The proofs of Theorems 1 and 2 are provided in Appendix A. These spectral properties indicate that under small perturbations, the first R eigenvalues remain close to zero, while the $(R+1)$-th eigenvalue stays bounded away from zero. Therefore, a stable spectral gap is preserved between the R-th and $(R+1)$-th eigenvalues. Although real-world feature graphs may not strictly satisfy the ideal clustering assumptions, clear spectral gaps can still indicate the number of visual modes when modes remain sufficiently well separated. Motivated by this observation, we estimate the number of proxies by identifying the largest eigengap in the sorted Laplacian spectrum.

Specifically, given the L2-normalized feature set of category y corresponding to the k-th class token, denoted as $\mathbf{F}_y^k = \{f_i^k \mid y_i = y\}$, we construct a similarity graph G_y^k, where each feature is treated

as a node. The affinity matrix W is calculated using an RBF kernel:

$$W_{ij}^k = \begin{cases} \exp\left(-\frac{\|f_i^k - f_j^k\|_2}{\tau_g}\right), & i \neq j, \\ 0, & i = j, \end{cases} \tag{1}$$

where τ_g is a temperature parameter. The degree matrix D_y^k is defined as $D_{ii}^k = \sum_j W_{ij}^k$, and the normalized Laplacian matrix is formulated as:

$$L_{\text{norm}}^k = I - (D_y^k)^{-\frac{1}{2}} W_y^k (D_y^k)^{-\frac{1}{2}}, \tag{2}$$

where I denotes the identity matrix. Then, let the eigenvalues of L_{norm}^k be sorted in ascending order as $0 \leq \lambda_1^k \leq \lambda_2^k \leq \cdots \leq \lambda_n^k$. According to the spectral properties, we define the eigengap between two adjacent Laplacian eigenvalues as $g_r^k = \lambda_{r+1}^k - \lambda_r^k$, and estimate the number of proxies for category y under the k-th class token representation as:

$$R_y^k = \arg\max_r g_r^k. \tag{3}$$

The estimated R_y^k reflects the intra-class diversity of category y in the corresponding feature space and is used for the following EM-based proxy representation optimization.

EM-Based Proxy Representation Optimization. Given the estimated proxy number R_y^k, we fit an R_y^k-component von Mises–Fisher mixture to the normalized category features, where τ_p controls the sharpness of the feature-to-proxy assignment probabilities. Expectation-Maximization (EM) iteratively estimates feature-to-proxy responsibilities γ_{ir}^k and updates proxy $p_{y,r}^k$ as responsibility-weighted feature averages. This proxy construction keeps the proxies aligned with the pretrained intra-class structure. The optimized proxy set $\mathbf{P}_y^k = \{p_{y,r}^k\}_{r=1}^{R_y^k}$ is used for subsequent category proxy initialization and updating. The complete EM procedure is provided in Appendix B.

3.2.2 Category Proxy Initialization and Updating

To maintain stable category-level prototypes, we introduce category proxies that are initialized before training and progressively updated during training. Specifically, for each category y under the k-th class token representation, the category proxy set is initialized using the **adaptive proxy construction** procedure described in Section 3.2.1 with the complete category feature set, denoted as $\mathbf{P}_y^{k,\text{cat}} = \{p_{y,r}^{k,\text{cat}}\}_{r=1}^{R_y^{k,cat}}$. During training, the current mini-batch generates the batch proxy set $\mathbf{P}_y^{k,\text{bat}} = \{p_{y,r}^{k,\text{bat}}\}_{r=1}^{R_y^{k,bat}}$ through the same proxy construction procedure.

Since batch proxies are estimated from partial observations and may not directly correspond to category proxies, we match each batch proxy with its nearest category proxy based on cosine similarity:

$$r_i^* = \arg\max_r \ cos\left(p_{y,i}^{k,\text{bat}}, p_{y,r}^{k,\text{cat}}\right), \tag{4}$$

where $\cos(\cdot,\cdot)$ denotes the cosine similarity. And the matched category proxy is updated using **Exponential Moving Average (EMA)**:

$$p_{y,r_i^*}^{k,cat} \leftarrow \beta p_{y,r_i^*}^{k,cat} + (1-\beta) p_{y,i}^{k,bat}, \tag{5}$$

where β is the momentum coefficient. When multiple batch proxies are matched to the same category proxy, we first aggregate these batch proxies into a weighted proxy representation, where the weight of each batch proxy is determined by the number of associated samples. The aggregated representation is then used for the EMA update. For clarity, the category proxy initialization and updating procedures are summarized in Appendix C.

3.2.3 Representation Preservation Loss

To preserve the semantic structure of the pretrained representation space, we introduce a representation preservation loss consisting of two components.

For the proxy-based preservation loss, given a feature f_i^k with ground-truth category y_i, we encourage f_i^k to be close to the representative proxy of category y_i and separated from proxies of other

categories in the corresponding class-token feature space. Specifically, $\hat{p}_{c,i}^k$ denotes the category proxy in $\mathbf{P}_c^{k,\text{cat}}$ with the highest cosine similarity with f_i^k, where $c \in \{1, \ldots, C\}$ and C is the number of categories. The proxy-based preservation loss is defined as:

$$\mathcal{L}_{proxy}^k = \text{CE}\left(\left\{\cos(f_i^k, \hat{p}_{c,i}^k)\right\}_{c=1}^C, y_i\right). \tag{6}$$

where $\text{CE}(\cdot)$ denotes the cross-entropy loss.

For the boundary-based preservation loss, we encourage the feature f_i^k to approach the boundary representative of category y_i while pushing f_i^k away from boundary representatives of other categories. We initialize a boundary representative b^k for each category using the mean feature of all samples in that category, and update b^k during training based on the category-specific samples observed in the current mini-batch $\mathcal{B}$. Specifically, for each category c appearing in the current mini-batch $\mathcal{B}$, we randomly select one intra-class feature a_c^k from the corresponding feature set $\mathbf{F}_c^{k,\mathcal{B}} = \{f_j^k | y_j = c, (x_j, y_j) \in \mathcal{B}\}$ as a reference feature. The boundary representative b_c^k is then updated as the feature in $\mathbf{F}_c^{k,\mathcal{B}}$ that has the lowest cosine similarity to a_c^k. For categories absent from the mini-batch $\mathcal{B}$, their previous boundary representatives are retained. The boundary-based preservation loss is formulated as:

$$\mathcal{L}_{boundary}^k = \text{CE}\left(\left\{\cos(f_i^k, b_c^k)\right\}_{c=1}^C, y_i\right). \tag{7}$$

The final representation preservation loss is obtained by averaging the two components:

$$\mathcal{L}_{RP}^k = \frac{1}{2}\left(\mathcal{L}_{proxy}^k + \mathcal{L}_{boundary}^k\right). \tag{8}$$

By combining proxy-level semantic constraints and boundary-aware discriminative constraints, the loss maintains the representation structure of the pretrained feature space during fine-tuning.

3.3 Overall Optimization Objective

The proposed framework is jointly optimized with the downstream task objective, self-diverse constraint, and representation preservation objective. Specifically, $\mathcal{L}_{Task}$ is obtained from the task-specific head and provides supervision for downstream tasks to learn task-relevant discriminative features. When multiple class tokens are adopted ($K > 1$), the self-diverse constraint $\mathcal{L}_{Dissimilar}$ is introduced following Li et al. (2023b) to encourage different class tokens to capture complementary representations. The representation preservation loss is computed independently for each class token, where $\mathcal{L}_{RP}^k$ denotes the representation preservation loss corresponding to the k-th class token. The overall optimization objective is formulated as:

$$\mathcal{L} = \mathcal{L}_{Task} + \lambda_D \mathcal{L}_{Dissimilar} + \lambda_{RP} \sum_{k=1}^{K} \mathcal{L}_{RP}^k, \tag{9}$$

where λ_D and λ_{RP} are balance coefficients controlling relative contributions of individual objectives. When $K = 1$, we set $\lambda_D = 0$ since the self-diverse constraint is unnecessary.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our method on three representative FGVR tasks: Fine-Grained Visual Classification (FGVC), Person Re-Identification (Person ReID), and Vehicle Re-Identification (Vehicle ReID). For FGVC, we conduct experiments on the CUB-200-2011 (Wah et al., 2011), Stanford Cars (Krause et al., 2013), Stanford Dogs (Khosla et al., 2011), and NABirds (Horn et al., 2015), reporting Top-1 accuracy (%) on each dataset. For Person ReID, we evaluate our method on Market-1501 (Zheng et al., 2015) and MSMT17 (Wei et al., 2018), reporting mean Average Precision (mAP) and Rank-1 accuracy. For Vehicle ReID, we evaluate our method on VeRi-776 (Liu et al., 2016), reporting mAP and Rank-1 accuracy.

Table 1: Performance comparison on FGVC across CUB-200-2011 (CUB), Stanford Cars (Cars), Stanford Dogs (Dogs), and NABirds. † denotes VPT-based methods. "Full FT" denotes full fine-tuning of the backbone. "Extra Params" indicates additional parameterized modules beyond the adopted fine-tuning strategy. "Avg." is computed over the datasets with available results. The best performance is highlighted in **bold**, and the second-best result is underlined.

Methods	Backbone	Extra Params	CUB	Cars	Dogs	NABirds	Avg.
Full FT	ViT-B/16	✗	87.3	84.5	89.4	82.7	86.0
BPT† (Wang et al., 2025b)	ViT-B/16	✗	89.1	82.8	90.5	84.6	86.8
BPT+PAE† (Wang et al., 2026a)	ViT-B/16	✓	91.2	89.4	**91.9**	89.0	90.4
Full FT	DINOv2-B/14	✗	89.2	90.2	88.7	89.1	89.3
VPT-Shallow† (Jia et al., 2022)	DINOv2-B/14	✗	88.5	86.5	87.9	89.6	88.1
VPT-Deep† (Jia et al., 2022)	DINOv2-B/14	✗	89.1	89.0	89.2	91.2	89.6
EvalProtoPNet (Huang et al., 2023)	DINOv2-B/14	✓	89.7	91.2	87.9	-	89.6
LOFF-TA (Konuk et al., 2024)	DINOv2-B/14	✗	-	89.3	85.5	83.8	86.2
PDiscoFormer (Aniraj et al., 2024)	DINOv2-B/14	✗	89.1	-	-	-	89.1
MDC (Chen et al., 2025)	DINOv2-B/14	✓	86.0	-	86.0	85.6	85.9
Zhu et al. (Zhu et al., 2025)	DINOv2-B/14	✗	90.4	91.3	87.1	-	89.6
IVPT† (Wang et al., 2026b)	DINOv2-B/14	✗	90.8	-	-	-	90.8
VPT-Shallow+Ours†	DINOv2-B/14	✗	90.0	87.3	88.4	90.0	88.9
VPT-Deep+Ours†	DINOv2-B/14	✗	91.2	90.9	89.5	**91.9**	90.9
Full FT+Ours	DINOv2-B/14	✗	**91.4**	**92.0**	89.4	91.3	**91.0**

Table 2: Performance comparison on Person ReID across Market-1501 and MSMT17. "Full FT" denotes full fine-tuning of the backbone. The best performance is highlighted in **bold**, and the second-best result is underlined.

Method	Multi-CLS	Market-1501		MSMT17	
		mAP	Rank-1	mAP	Rank-1
TransReID (He et al., 2021)	✗	88.9	95.2	67.4	85.3
DC-Former (Li et al., 2023b)	✓	90.6	96.0	70.7	86.9
CLIP-ReID (Li et al., 2023a)	✗	89.6	95.5	73.4	88.7
DINOv2 Full FT (Oquab et al., 2024)	✗	85.0	93.0	70.9	86.9
FlexiReID (Sun et al., 2025)	✗	92.1	96.0	67.5	83.7
CLIMB-ReID (Yu et al., 2025)	✗	92.6	96.8	77.8	90.5
HPL (Zhou et al., 2026)	✗	89.8	95.9	79.0	91.0
DINOv2 Full FT + Ours$_{1\text{-cls}}$	✗	93.3	96.4	79.5	91.5
DINOv2 Full FT + Ours$_{2\text{-cls}}$	✓	**93.5**	**96.9**	**80.3**	**91.6**

Table 3: Ablation studies on Market-1501 with two learnable class tokens. $\mathcal{L}_{base}$ denotes the original objective containing $\mathcal{L}_{Task}$ and $\mathcal{L}_{Dissimilar}$. The best performance is highlighted in **bold**.

Configuration	mAP	Rank-1
(a) Framework Ablation		
Baseline	92.0	96.0
+ Single-Proxy Modeling ($R = 1$)	92.5	96.1
+ Fixed Multi-Proxy Modeling ($R = 3$)	92.5	96.6
+ Adaptive Multi-Proxy Modeling (Ours)	**93.5**	**96.9**
(b) Loss Ablation		
$\mathcal{L}_{base}$	92.0	96.0
$\mathcal{L}_{base} + \mathcal{L}_{proxy}$	92.3	96.3
$\mathcal{L}_{base} + \mathcal{L}_{boundary}$	90.9	96.0
$\mathcal{L}_{base} + \mathcal{L}_{proxy} + \mathcal{L}_{boundary}$ (Ours)	**93.5**	**96.9**

Implementation Details. We use DINOv2-B/14 (Oquab et al., 2024) as the backbone for all experiments, which is initialized with pretrained weights and fine-tuned on each downstream dataset. The input resolutions are set to 518×518 for FGVC, 256×128 for Person ReID, and 256×256 for Vehicle ReID, following the standard protocols of each benchmark. We use AdamW for FGVC and Vehicle ReID, and SGD for Person ReID. We set $\lambda_{\text{RP}} = 3$ for FGVC and $\lambda_{\text{RP}} = 4$ for ReID, while setting $\lambda_D = 1$ for multi-CLS setting and $\lambda_D = 0$ for single-CLS setting. We set the momentum coefficient $\beta = 0.2$, the graph affinity temperature $\tau_g = 0.3$, the proxy assignment temperature $\tau_p = 0.05$, and the number of EM iterations $T_{EM} = 50$. For parameter-efficient fine-tuning, we use 100 VPT prompt tokens; LoRA with $r = 64$ and $\alpha = 128$; and an Adapter bottleneck dimension of 64. The hyperparameter experiments are presented in Appendix D.

4.2 Experimental Comparisons

Comparison on Fine-Grained Visual Classification. Table 1 shows the experimental results on four FGVC benchmarks. "Full FT+Ours" achieves the highest average Top-1 accuracy of 91.0% without introducing additional parameterized modules beyond the adopted fine-tuning strategy. "VPT-Deep+Ours" also achieves a competitive average accuracy of 90.9% and achieves the best result of 91.9% on NABirds. Although ViT-B/16-based "BPT+PAE" obtains the highest accuracy on Stanford Dogs, our method reaches 89.5% and outperforms all competing methods using

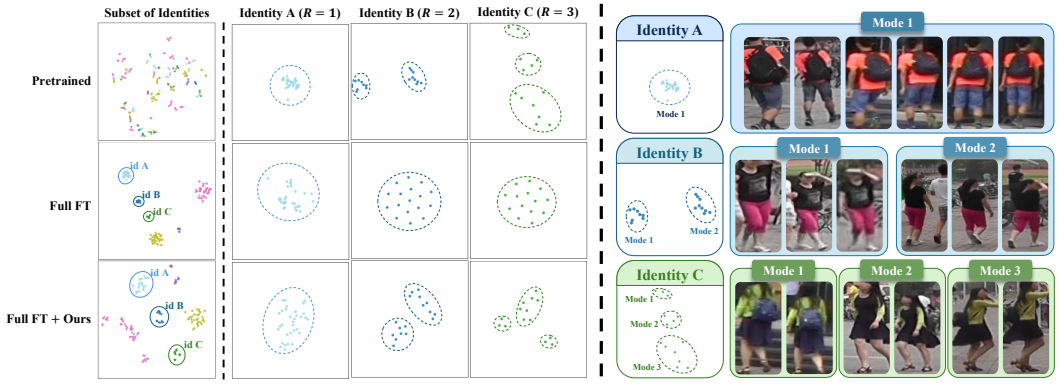


Figure 3: Visualization of pretrained intra-class structures and structure preservation during fine-tuning on Market-1501. **Left:** t-SNE of feature distributions before and after fine-tuning. R denotes the proxy number automatically estimated from the pretrained feature structures. **Right:** Representative images corresponding to the visual modes observed in the pretrained feature space.

the same DINOv2-B/14 backbone. These results demonstrate the effectiveness of our method across different fine-tuning strategies and FGVC benchmarks.

Comparison on Person Re-Identification. Table 2 compares our method with existing Person ReID approaches on Market-1501 and MSMT17. In particular, "Ours$_{\text{2-cls}}$" achieves 93.5% mAP and 96.9% Rank-1 accuracy on Market-1501, and 80.3% mAP and 91.6% Rank-1 accuracy on MSMT17. "Ours$_{\text{2-cls}}$" also consistently outperforms "Ours$_{\text{1-cls}}$", demonstrating the effectiveness of using multiple class tokens for Person ReID task.

Comparison on Vehicle Re-Identification. Our method achieves 88.3% mAP and 98.4% Rank-1 accuracy on VeRi-776, outperforming existing Vehicle ReID approaches on both metrics. Detailed comparisons are provided in Appendix E.

4.3 Ablation Studies and Further Analysis

4.3.1 Component Ablations

Table 3 evaluates the key components of our method on Market-1501. Table 3(a) shows that single-proxy modeling raises mAP from 92.0% to 92.5%, while fixed multi-proxy modeling ($R = 3$, the largest estimated value) improves Rank-1 to 96.6%. Adaptive multi-proxy modeling performs best, achieving 93.5% mAP and 96.9% Rank-1, confirming the benefit of allocating proxies according to intra-class diversity. Notably, the learnable proxy setting is not considered in this ablation, as it is not compatible with the multi-CLS setting. These results confirm the importance of allocating multiple proxies according to intra-class diversity. Table 3(b) evaluates the two components of the representation preservation loss. Adding $\mathcal{L}_{\text{proxy}}$ alone improves the baseline, while adding $\mathcal{L}_{\text{boundary}}$ alone fails to bring performance gains. Combining both losses achieves the best performance, improving the baseline by 1.5% in mAP and 0.9% in Rank-1 accuracy.

Further evaluations across different backbone scales and fine-tuning strategies are provided in Appendix F, demonstrating the generalizability of our method.

4.3.2 Pretrained Structure Preservation

Qualitative Analysis. Figure 3 compares the feature distributions produced by pretrained DINOv2, vanilla full fine-tuning, and our method on Market-1501, together with representative images from different visual modes. Although vanilla full fine-tuning improves inter-identity separation, it tends to collapse distinct intra-class visual modes into a single cluster. In contrast, our method preserves discriminative identity representations while maintaining diverse intra-class structures. Moreover,

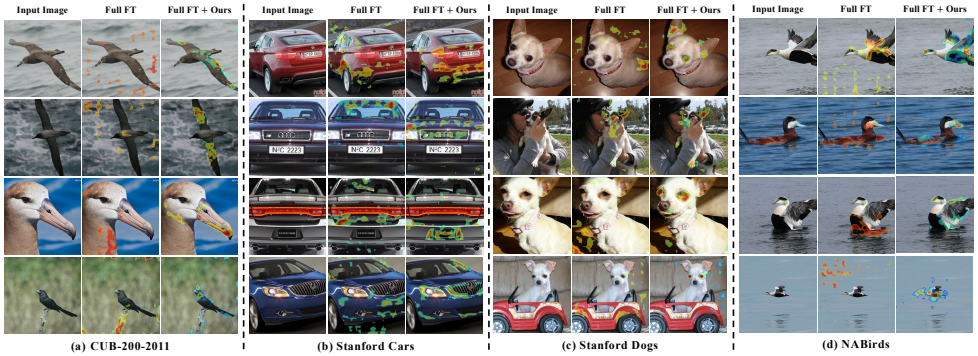


Figure 4: **Grad-CAM visualizations on four FGVC datasets.** Compared with Full FT, our method better captures fine-grained discriminative regions and suppresses responses to irrelevant backgrounds. Our method produces attention maps that focus more consistently on subtle visual cues distinguishing highly similar categories, demonstrating improved fine-grained feature learning.

the estimated proxy number R aligns well with the observed number of visual modes, supporting the effectiveness of our adaptive proxy number estimation.

Quantitative Analysis. To evaluate representation preservation, we perform ImageNet-1K (Deng et al., 2009) linear probing after fine-tuning DINOv2-B/14 on Market-1501. Table 4 shows that full fine-tuning reduces Top-1 accuracy from 83.6% to 19.8%, whereas our method retains 52.9% while simultaneously improving downstream performance. This indicates that our method better preserves pretrained representations while adapting to downstream task variations during fine-tuning.

Table 4: Linear probing on ImageNet-1K.

Method	Top-1 Acc (%)
Pretrained	83.6
Full FT	19.8 (↓63.8)
Full FT+Ours	52.9 (↓30.7)

4.4 Visualization of Fine-Grained Attention

Figure 4 presents Grad-CAM visualizations on four FGVC datasets. Vanilla full fine-tuning often results in dispersed feature responses and attention to irrelevant background regions. In contrast, our method more effectively localizes category-specific discriminative regions, such as the heads, beaks, and wings of birds; the headlights and brand emblems of vehicles; and the faces, ears, and paws of dogs. This suggests that our method better captures subtle visual cues that are crucial for fine-grained recognition. Additional attention visualizations on Person ReID are provided in Appendix G.

5 Conclusion

Vision foundation models often exhibit meaningful intra-class organization in their pretrained feature space, but this structure is not explicitly considered during downstream fine-tuning. In this paper, we introduce Adaptive Proxy Modeling (APM) to better exploit such pretrained structure for fine-grained vision model adaptation. APM adaptively determines the number of proxies for each category, constructs them from pretrained features, and updates them during fine-tuning. Across FGVC, Person ReID, and Vehicle ReID, APM consistently improves downstream performance without introducing additional trainable parameters. Our analysis further shows that APM better preserves pretrained representations during adaptation.

A current limitation is that APM relies on the quality of the pretrained feature structure used for proxy estimation, which may become less reliable under large domain shifts. In addition, a discrete set of proxies may not fully capture more continuous or hierarchical intra-class variations. Future work could explore more robust structure estimation, more flexible structural representations, and broader applications beyond the fine-grained recognition settings considered here.

AI USE STATEMENT

In this work, we used generative AI tools to aid or polish writing. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Ananthu Aniraj, Cássio Fraga Dantas, Dino Ienco, and Diego Marcos. Pdiscoformer: Relaxing part discovery constraints with vision transformers. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *European Conference on Computer Vision*, volume 15143 of *Lecture Notes in Computer Science*, pp. 256–272. Springer, 2024.

Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas J. Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21795–21806. IEEE, 2024.

Shubhang Bhatnagar and Narendra Ahuja. Potential field based deep metric learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25549–25559. Computer Vision Foundation / IEEE, 2025.

Jihao Cai, Zhiqiang He, Zhi Liu, and Yangjie Cao. Diperceivenet: A bidirectional cross-scale perception network for vehicle re-identification. *Pattern Recognit.*, 178:113476, 2026.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan Su. This looks like that: Deep learning for interpretable image recognition. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems*, pp. 8928–8939, 2019.

Xuanbai Chen, Xiang Xu, Zhihua Li, Tianchen Zhao, Pietro Perona, Qin Zhang, and Yifan Xing. Model diagnosis and correction via linguistic and implicit attribute editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14281–14292. Computer Vision Foundation / IEEE, 2025.

Arpita Chowdhury, Dipanjyoti Paul, Zheda Mai, Jianyang Gu, Ziheng Zhang, Kazi Sajeed Mehrab, Elizabeth G. Campolongo, Daniel I. Rubenstein, Charles V. Stewart, Anuj Karpatne, Tanya Y. Berger-Wolf, Yu Su, and Wei-Lun Chao. Prompt-cam: Making vision transformers interpretable for fine-grained analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4375–4385. Computer Vision Foundation / IEEE, 2025.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE Computer Society, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations*. OpenReview.net, 2021.

Ruoyi Du, Dongliang Chang, Ayan Kumar Bhunia, Jiyang Xie, Zhanyu Ma, Yi-Zhe Song, and Jun Guo. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *European Conference on Computer Vision*, volume 12365 of *Lecture Notes in Computer Science*, pp. 153–168. Springer, 2020.

Olaf Dünkel, Basavaraj Sunagad, Haoran Wang, David T. Hoffmann, Christian Theobalt, and Adam Kortylewski. SOCO: benchmarking semantic object correspondence in vision foundation models. *CoRR*, abs/2605.31597, 2026.

Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4476–4484. IEEE Computer Society, 2017.

Ju He, Jieneng Chen, Shuai Liu, Adam Kortylewski, Cheng Yang, Yutong Bai, and Changhu Wang. Transfg: A transformer architecture for fine-grained recognition. In *Thirty-Sixth Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, pp. 852–860. AAAI Press, 2022.

Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *2021 IEEE/CVF International Conference on Computer Vision*, pp. 14993–15002. IEEE, 2021.

Xiangteng He and Yuxin Peng. Fine-grained image classification via combining vision and language. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7332–7340. IEEE Computer Society, 2017.

Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge J. Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 595–604. IEEE Computer Society, 2015.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799. PMLR, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations*. OpenReview.net, 2022.

Chen Huang, Skyler Seto, Hadi Pouransari, Mehrdad Farajtabar, Raviteja Vemulapalli, Fartash Faghri, Oncel Tuzel, Barry-John Theobald, and Joshua M. Susskind. Proxy-fda: Proxy-based feature distribution alignment for fine-tuning vision foundation models without forgetting. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025.

Qihan Huang, Mengqi Xue, Wenqi Huang, Haofei Zhang, Jie Song, Yongcheng Jing, and Mingli Song. Evaluation and improvement of interpretability for self-explainable part-prototype networks. In *IEEE/CVF International Conference on Computer Vision*, pp. 2011–2020. IEEE, 2023.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge J. Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *European Conference on Computer Vision*, volume 13693 of *Lecture Notes in Computer Science*, pp. 709–727. Springer, 2022.

Xin Jiang, Hao Tang, Junyao Gao, Xiaoyu Du, Shengfeng He, and Zechao Li. Delving into multimodal prompting for fine-grained visual classification. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (eds.), *Thirty-Eighth Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, pp. 2570–2578. AAAI Press, 2024.

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proceedings of the CVPR Workshop on Fine-Grained Visual Categorization*, 2011.

Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Proxy anchor loss for deep metric learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3235–3244. Computer Vision Foundation / IEEE, 2020.

Emir Konuk, Christos Matsoukas, Moein Sorkhei, Phitchapha Lertsiravarameth, and Kevin Smith. Learning from offline foundation features with tensor augmentations. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024*, 2024.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pp. 554–561. IEEE Computer Society, 2013.

Baolu Li, Ping Liu, Lan Fu, Jinlong Li, Jianwu Fang, Zhigang Xu, and Hongkai Yu. Vehiclegan: Pair-flexible pose guided image synthesis for vehicle re-identification. In *IEEE Intelligent Vehicles Symposium, IV 2024*, pp. 447–453. IEEE, 2024.

Siyuan Li, Li Sun, and Qingli Li. Clip-reid: Exploiting vision-language model for image re-identification without concrete text labels. In Brian Williams, Yiling Chen, and Jennifer Neville (eds.), *Thirty-Seventh Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, pp. 1405–1413. AAAI Press, 2023a.

Wen Li, Cheng Zou, Meng Wang, Furong Xu, Jianan Zhao, Ruobing Zheng, Yuan Cheng, and Wei Chu. Dc-former: Diverse and compact transformer for person re-identification. In Brian Williams, Yiling Chen, and Jennifer Neville (eds.), *Thirty-Seventh Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, pp. 1415–1423. AAAI Press, 2023b.

Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear CNN models for fine-grained visual recognition. In *2015 IEEE International Conference on Computer Vision*, pp. 1449–1457. IEEE Computer Society, 2015.

Xinchen Liu, Wu Liu, Huadong Ma, and Huiyuan Fu. Large-scale vehicle re-identification in urban surveillance videos. In *IEEE International Conference on Multimedia and Expo*, pp. 1–6. IEEE Computer Society, 2016.

Shunan Mao, Yaowei Wang, Xiaoyu Wang, and Shiliang Zhang. Multi-proxy feature learning for robust fine-grained visual recognition. *Pattern Recognit.*, 143:109779, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024, 2024.

Jinhee Park, Jisoo Park, Dagyeong Na, and Junseok Kwon. Deep disentangled metric learning. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *Thirty-Ninth Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, pp. 19830–19838. AAAI Press, 2025.

Qi Qian, Rong Jin, Shenghuo Zhu, and Yuanqing Lin. Fine-grained visual categorization via multi-stage metric learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3716–3724. IEEE Computer Society, 2015.

Li Ren, Chen Chen, Liqiang Wang, and Kien A. Hua. Learning semantic proxies from visual prompts for parameter-efficient fine-tuning in deep metric learning. In *The Twelfth International Conference on Learning Representations*. OpenReview.net, 2024.

Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*, pp. 4077–4087, 2017.

Zhen Sun, Lei Tan, Yunhang Shen, Chengmao Cai, Xing Sun, Pingyang Dai, Liujuan Cao, and Rongrong Ji. Flexireid: Adaptive mixture of expert for multi-modal person re-identification. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan

Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Junze Wang, Lei Fan, Dezheng Zhang, Weipeng Jing, Donglin Di, Yang Song, Sidong Liu, and Cong Cong. Visual prompt-agnostic evolution. *CoRR*, abs/2601.20232, 2026a.

Qi Wang, Zeyu Zhang, Dong Wang, Di Gai, Xin Xiong, Jiyang Xu, and Ruihua Zhou. Vehiclemae: View-asymmetry mutual learning for vehicle re-identification pre-training via masked autoencoders. In *IEEE/CVF International Conference on Computer Vision*, pp. 4701–4711. IEEE, 2025a.

Yubin Wang, Xinyang Jiang, De Cheng, Xiangqian Zhao, Zilong Wang, Dongsheng Li, and Cai Zhao. Exploring interpretability for visual prompt tuning with cross-layer concepts. In *International Conference on Learning Representations*, volume 2026, pp. 32636–32656, 2026b.

Yuzhu Wang, Manni Duan, and Shu Kong. Attention to the burstiness in visual prompt tuning! In *IEEE/CVF International Conference on Computer Vision*, pp. 4253–4263. IEEE, 2025b.

Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer GAN to bridge domain gap for person re-identification. In *2018 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 79–88. Computer Vision Foundation / IEEE Computer Society, 2018.

Xuezhi Xiang, Zhushan Ma, Xiaoheng Li, Lei Zhang, and Xiantong Zhen. Vehicle re-identification with large separable kernel attention and hybrid channel attention. *Image Vis. Comput.*, 155: 105442, 2025a.

Xuezhi Xiang, Zhushan Ma, Lei Zhang, Denis Ombati, Himaloy Himu, and Xiantong Zhen. Lkareid: Vehicle re-identification with large kernel attention. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025b.

Chenyang Yu, Xuehu Liu, Jiawen Zhu, Yuhao Wang, Pingping Zhang, and Huchuan Lu. Climb-reid: A hybrid clip-mamba framework for person re-identification. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *Thirty-Ninth Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, pp. 9589–9597. AAAI Press, 2025.

Yunlong Yu, Dingyi Zhang, Yingming Li, and Zhongfei Zhang. Multi-proxy learning from an entropy optimization perspective. In Luc De Raedt (ed.), *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 1594–1600. ijcai.org, 2022.

Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12200–12210. Computer Vision Foundation / IEEE, 2020.

Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *IEEE International Conference on Computer Vision*, pp. 5219–5227. IEEE Computer Society, 2017.

Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *2015 IEEE International Conference on Computer Vision*, pp. 1116–1124. IEEE Computer Society, 2015.

Linhan Zhou, Shuang Li, Neng Dong, Yonghang Tai, Yafei Zhang, and Huafeng Li. Hierarchical prompt learning for image- and text-based person re-identification. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, pp. 13728–13736. AAAI Press, 2026.

Xiaoying Zhou, Xi Li, Houren Zhou, Xiyu Pang, Jiachen Tian, Xiushan Nie, Cheng Wang, and Yilong Yin. Adaptive division and priori reinforcement part learning network for vehicle re-identification. *Pattern Recognit.*, 163:111453, 2025.

Zhijie Zhu, Lei Fan, Maurice Pagnucco, and Yang Song. Interpretable image classification via non-parametric part prototype learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9762–9771. Computer Vision Foundation / IEEE, 2025.

Chang Zou, Zeqi Chen, Zhichao Cui, Yuehu Liu, and Chi Zhang. Discrepant and multi-instance proxies for unsupervised person re-identification. In *IEEE/CVF International Conference on Computer Vision*, pp. 11024–11034. IEEE, 2023.

A Proofs of the Spectral Properties

A.1 Proof of Theorem 1

Proof. Let $G^* = (V, W)$ be the ideal category-specific similarity graph, where $W = [w_{ij}]$ is a symmetric nonnegative adjacency matrix. Let

$$d_i = \sum_j w_{ij}, \qquad D = \mathrm{diag}(d_1, \ldots, d_n),$$

and recall that the normalized graph Laplacian is

$$L^* = I - D^{-1/2} W D^{-1/2}.$$

For simplicity, we first assume that $d_i > 0$ for all vertices. An isolated vertex can be treated separately and contributes exactly one additional zero eigenvalue, so the conclusion remains unchanged.

For an arbitrary vector $z \in \mathbb{R}^n$, the quadratic form associated with L^* can be written as

$$\begin{aligned} z^\top L^* z &= z^\top z - z^\top D^{-1/2} W D^{-1/2} z \\ &= \frac{1}{2} \sum_{i,j} w_{ij} \left(\frac{z_i}{\sqrt{d_i}} - \frac{z_j}{\sqrt{d_j}} \right)^2 \geq 0. \end{aligned} \tag{10}$$

Hence, L^* is positive semidefinite and all its eigenvalues are nonnegative.

Moreover, equation 10 is equal to zero if and only if

$$\frac{z_i}{\sqrt{d_i}} = \frac{z_j}{\sqrt{d_j}} \qquad \text{for every edge } (i, j) \text{ with } w_{ij} > 0.$$

Therefore, the quantity $z_i/\sqrt{d_i}$ must be constant over every connected component of G^*.

Suppose first that G^* contains exactly R connected components, denoted by

$$\mathcal{C}_1, \ldots, \mathcal{C}_R.$$

For each component $\mathcal{C}_r$, define a vector $q_r \in \mathbb{R}^n$ by

$$(q_r)_i = \begin{cases} \sqrt{d_i}, & i \in \mathcal{C}_r, \\ 0, & i \notin \mathcal{C}_r. \end{cases}$$

Since no edge with positive weight connects two different connected components, we have

$$L^* q_r = 0, \qquad r = 1, \ldots, R.$$

The supports of $q_1, \ldots, q_R$ are mutually disjoint, and therefore these vectors are linearly independent. Consequently,

$$\dim \ker(L^*) \geq R.$$

Conversely, let $z \in \ker(L^*)$. By equation 10, $z_i/\sqrt{d_i}$ is constant within every connected component. Hence, there exist constants $a_1, \ldots, a_R$ such that

$$z = \sum_{r=1}^{R} a_r q_r.$$

Thus,

$$\ker(L^*) = \mathrm{span}\{q_1, \ldots, q_R\},$$

and consequently

$$\dim \ker(L^*) = R.$$

Since L^* is real symmetric and positive semidefinite, the dimension of its null space is exactly the algebraic multiplicity of the eigenvalue zero. Therefore, if its eigenvalues are sorted in ascending order,

$$0 \leq \mu_1 \leq \mu_2 \leq \cdots \leq \mu_n,$$

then a graph with exactly R connected components satisfies

$$\mu_1 = \cdots = \mu_R = 0, \qquad \mu_{R+1} > 0.$$

Defining

$$\gamma := \mu_{R+1},$$

we obtain

$$\mu_1 = \cdots = \mu_R = 0, \qquad \mu_{R+1} = \gamma > 0.$$

For the reverse direction, suppose that

$$\mu_1 = \cdots = \mu_R = 0, \qquad \mu_{R+1} = \gamma > 0.$$

Then the multiplicity of the eigenvalue zero is exactly R, i.e.,

$$\dim \ker(L^*) = R.$$

As established above, the nullity of the normalized graph Laplacian is equal to the number of connected components of the graph. Hence, G^* contains exactly R connected components.

This proves both directions of the theorem. □

A.2 Proof of Theorem 2

Proof. Let

$$L = L^* + E, \qquad \|E\|_2 \le \epsilon,$$

and denote the eigenvalues of L^* and L, respectively, by

$$\mu_1 \le \mu_2 \le \cdots \le \mu_n$$

and

$$\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n.$$

Since both L and L^* are symmetric normalized Laplacians, E is also symmetric.

We first establish the standard eigenvalue perturbation bound. For every unit vector x,

$$|x^\top E x| \le \|E\|_2 \le \epsilon.$$

Therefore,

$$x^\top L^* x - \epsilon \le x^\top L x \le x^\top L^* x + \epsilon.$$

Applying the Courant–Fischer min–max characterization,

$$\lambda_i(A) = \min_{\substack{S \subseteq \mathbb{R}^n \\ \dim(S) = i}} \max_{\substack{x \in S \\ \|x\|_2 = 1}} x^\top A x,$$

to L and L^* yields

$$\mu_i - \epsilon \le \lambda_i \le \mu_i + \epsilon, \qquad i = 1, \ldots, n.$$

Equivalently,

$$|\lambda_i - \mu_i| \le \epsilon, \qquad i = 1, \ldots, n. \tag{11}$$

This is precisely Weyl's eigenvalue perturbation bound in the present setting.

We now prove the forward direction. Suppose that G contains exactly R intrinsic clusters. Under the ideal-graph perturbation model, this means that the corresponding ideal graph G^* contains exactly R connected components. By Theorem 1,

$$\mu_1 = \cdots = \mu_R = 0, \qquad \mu_{R+1} = \gamma > 0.$$

For $i = 1, \ldots, R$, equation 11 gives

$$-\epsilon \le \lambda_i \le \epsilon.$$

Since L is itself a normalized graph Laplacian, it is positive semidefinite and hence $\lambda_i \ge 0$. It follows that

$$0 \le \lambda_i \le \epsilon, \qquad i = 1, \ldots, R.$$

For the $(R+1)$-th eigenvalue, again using equation 11,

$$|\lambda_{R+1} - \mu_{R+1}| \leq \epsilon.$$

Since $\mu_{R+1} = \gamma$, we obtain

$$\gamma - \epsilon \leq \lambda_{R+1} \leq \gamma + \epsilon.$$

Thus the claimed spectral conditions hold.

Furthermore, when $\gamma > 2\epsilon$,

$$\lambda_{R+1} - \lambda_R \geq (\gamma - \epsilon) - \epsilon = \gamma - 2\epsilon > 0.$$

Hence the perturbation cannot close the eigengap separating the first R eigenvalues from the remaining spectrum. In particular, the R-dimensional low-frequency eigenspace remains spectrally isolated.

For completeness, we next prove the reverse implication under the standard ideal-cluster spectral-separation assumption implicit in the perturbation model: every nonzero eigenvalue of L^* is bounded below by γ, namely,

$$\mathrm{spec}(L^*) \subseteq \{0\} \cup [\gamma, 2]. \tag{12}$$

Suppose that the eigenvalues of L satisfy

$$0 \leq \lambda_i \leq \epsilon, \qquad i = 1, \ldots, R,$$

and

$$\gamma - \epsilon \leq \lambda_{R+1} \leq \gamma + \epsilon.$$

For each $i \leq R$, the perturbation bound equation 11 implies

$$\mu_i \leq \lambda_i + \epsilon \leq 2\epsilon.$$

Because $\gamma > 2\epsilon$, we have

$$\mu_i < \gamma.$$

According to the spectral-separation condition equation 12, an eigenvalue of L^* that is strictly smaller than γ must be zero. Therefore,

$$\mu_1 = \cdots = \mu_R = 0.$$

For the next eigenvalue, equation 11 gives

$$\mu_{R+1} \geq \lambda_{R+1} - \epsilon \geq \gamma - 2\epsilon > 0.$$

Consequently,

$$\mu_1 = \cdots = \mu_R = 0, \qquad \mu_{R+1} > 0.$$

By Theorem 1, the ideal graph G^* therefore contains exactly R connected components. Equivalently, under the adopted perturbation model, the observed graph G contains exactly R intrinsic clusters.

Hence,

$$0 \leq \lambda_i \leq \epsilon, \qquad i = 1, \ldots, R,$$

together with

$$\gamma - \epsilon \leq \lambda_{R+1} \leq \gamma + \epsilon$$

provides a stable spectral characterization of the R intrinsic clusters whenever $\gamma > 2\epsilon$. □

Remark on the Reverse Implication. The reverse implication in Theorem 2 relies on the usual ideal-cluster assumption that the nonzero spectrum of L^* is separated from zero, i.e.,

$$\mathrm{spec}(L^*) \subseteq \{0\} \cup [\gamma, 2].$$

Without such an assumption, the perturbation bound $\|E\|_2 \leq \epsilon$ alone does not exclude the possibility that L^* has small but strictly positive eigenvalues below γ, and hence the stated "if and only if" characterization would in general reduce to a one-way perturbation guarantee.

B Details of EM-Based Proxy Representation Optimization

After determining the proxy number R_y^k for category y under the k-th class token representation, we further optimize the proxies to capture diverse intra-category visual modes. Instead of directly optimizing free proxy vectors, we parameterize each proxy as a convex combination of category-specific features, ensuring that the proxies remain within the convex hull of the category-specific features. Specifically, given the feature set $\mathbf{F}_y^k = \{f_i^k \mid y_i = y\}$, the r-th proxy is defined as:

$$p_{y,r}^k = \sum_{i=1}^{n} \alpha_{ri}^k f_i^k, \tag{13}$$

where α_{ri}^k denotes the contribution weight of the i-th feature to the r-th proxy. The weights satisfy $\alpha_{ri}^k \geq 0$ and $\sum_{i=1}^{n} \alpha_{ri}^k = 1$, ensuring that each proxy remains within the feature subspace of the corresponding category.

To optimize the proxy representations, we formulate the problem as a **von Mises-Fisher (vMF)** mixture optimization objective:

$$\min_{\{p_{y,r}^k, \pi_{y,r}^k\}} \mathcal{L}_{\text{vMF}} = -\frac{\tau_p}{n} \sum_{i=1}^{n} \log \left[\sum_{r=1}^{R_y^k} \pi_{y,r}^k \exp\left(\frac{(p_{y,r}^k)^T f_i^k}{\tau_p} \right) \right], \tag{14}$$

where $\pi_{y,r}^k$ denotes the proportion of the r-th proxy component, satisfying $\pi_{y,r}^k \geq 0$ and $\sum_{r=1}^{R_y^k} \pi_{y,r}^k = 1$. τ_p is a temperature parameter that controls the sharpness of the feature-to-proxy assignment probabilities. n denotes the number of samples belonging to category y.

We employ an **Expectation-Maximization (EM)** procedure to optimize the above objective. Before the first EM iteration, the mixture weights are uniformly initialized as $\pi_{y,r}^k = 1/R_y^k$, and the initial proxy set is constructed by randomly selecting R_y^k from $\mathbf{F}_y^k$ as the initial proxy representations.

In the **E-step**, the responsibility of feature f_i^k assigned to the r-th proxy is computed based on the current proxy set $\{p_{y,r}^k\}_{r=1}^{R_y^k}$ and mixture weights $\{\pi_{y,r}^k\}_{r=1}^{R_y^k}$ as:

$$\gamma_{ir}^k = \frac{\pi_{y,r}^k \exp\left(\frac{(p_{y,r}^k)^T f_i^k}{\tau_p} \right)}{\sum_{j=1}^{R_y^k} \pi_{y,j}^k \exp\left(\frac{(p_{y,j}^k)^T f_i^k}{\tau_p} \right)}, \tag{15}$$

where γ_{ir}^k denotes the soft assignment probability of feature f_i^k to the r-th proxy.

In the **M-step**, we update the proxy representations according to the estimated assignments γ_{ir}^k. Specifically, the effective number of features associated with the r-th proxy is computed as $N_r^k = \sum_{i=1}^{n} \gamma_{ir}^k$. Then, the mixture proportion is updated as $\pi_{y,r}^k = N_r^k / n$, and the normalized weight for constructing the r-th proxy is calculated as $\alpha_{ri}^k = \gamma_{ir}^k / N_r^k$. The proxy representation is subsequently updated according to Equation 13:

$$p_{y,r}^k = \frac{\sum_{i=1}^{n} \gamma_{ir}^k f_i^k}{\sum_{i=1}^{n} \gamma_{ir}^k}. \tag{16}$$

We iteratively perform the E-step and M-step to refine the feature-proxy assignments and proxy representations. After convergence, we obtain the optimized proxy set $\mathbf{P}_y^k = \{p_{y,r}^k\}_{r=1}^{R_y^k}$, which is used for subsequent category proxy initialization and updating.

C Category Proxy Initialization and Updating Procedures

Algorithms 1 and 2 summarize category proxy initialization and batch-wise updating, respectively. Algorithm 1 estimates $R_y^{k,cat}$ from the largest Laplacian spectral gap and constructs the category proxies from normalized pretrained features via EM. Algorithm 2 applies the same procedure to batch features, matches batch proxies to category proxies by cosine similarity, and performs proportion-weighted aggregation followed by EMA updating. Here, $\mathbf{S}_r$ denotes the batch proxies matched to the r-th category proxy, and $p_{y,r}^{k,cat}$ remains unchanged if $\mathbf{S}_r$ is empty.

Algorithm 1 Category Proxy Initialization

Require: features $\mathbf{F}_y^k$, graph affinity temperature τ_g, proxy assignment temperature τ_p, and the number of EM iterations T_{EM}.
Output: proxy number $R_y^{k,cat}$, category proxy set $\mathbf{P}_y^{k,cat}$, and proxy component proportion $\mathbf{\Pi}_y^{k,cat}$.

for each class token k and category y **do**
  $\mathbf{F}_y^k = \{f_i^k \mid y_i = y\} \leftarrow \mathrm{Normalize}(\mathbf{F}_y^k)$
  Compute affinity matrix W_y^k
  Compute normalized Laplacian matrix L_{norm}^k
  Perform eigen-decomposition and sort eigenvalues: $0 \le \lambda_1^k \le \lambda_2^k \le \cdots \le \lambda_n^k$
  $R_y^{k,cat} = \arg\max_r(\lambda_{r+1}^k - \lambda_r^k)$
  $\{p_{y,r}^{k,cat}\}_{r=1}^{R_y^{k,cat}} \leftarrow \mathrm{RandomlySelect}(F_y^k, R_y^{k,cat}), \quad \{\pi_{y,r}^k\}_{r=1}^{R_y^{k,cat}} \leftarrow \mathrm{Uniform}(R_y^{k,cat})$
  for $t = 1, \ldots, T_{EM}$ **do**
    $\gamma_{ir}^k = \frac{\pi_{y,r}^{k,cat} \exp((\mathbf{p}_{y,r}^{k,cat})^T \mathbf{f}_i^k / \tau_p)}{\sum_{j=1}^{R_y^{k,cat}} \pi_{y,j}^{k,cat} \exp((\mathbf{p}_{y,j}^{k,cat})^T \mathbf{f}_i^k / \tau_p)}$
    $\pi_{y,r}^{k,cat} = \frac{1}{n}\sum_{i=1}^n \gamma_{ir}^k, \quad p_{y,r}^{k,cat} = \frac{\sum_{i=1}^n \gamma_{ir}^k f_i^k}{\sum_{i=1}^n \gamma_{ir}^k}$
  end for
  $\mathbf{P}_y^{k,cat} \leftarrow \{p_{y,r}^{k,cat}\}_{r=1}^{R_y^{k,cat}}, \quad \mathbf{\Pi}_y^{k,cat} \leftarrow \{\pi_{y,r}^{k,cat}\}_{r=1}^{R_y^{k,cat}}$
end for
return $R_y^{k,cat}$, $\mathbf{P}_y^{k,cat}$ and $\mathbf{\Pi}_y^{k,cat}$

Algorithm 2 Batch Proxy Estimation and Category Proxy Updating

Require: batch features $\mathbf{F}_y^{k,bat}$, category proxy set $\mathbf{P}_y^{k,cat}$, graph affinity temperature τ_g, proxy assignment temperature τ_p, the number of EM iterations T_{EM}, and momentum coefficient β.
Output: updated category proxy set $\mathbf{P}_y^{k,cat}$.

▷ computed from batch features rather than the complete feature set
$R_y^{k,bat}, \mathbf{P}_y^{k,bat}, \mathbf{\Pi}_y^{k,bat} = \mathrm{Algorithm1}\left(\mathbf{F}_y^{k,bat}, \tau_g, \tau_p, T_{EM}\right)$
for each class token k and category y **do**
  ▷ match each batch proxy to its nearest category proxy and collect the matched indices.
  $r_i^* = \arg\max_r \ cos\left(p_{y,i}^{k,\mathrm{bat}}, p_{y,r}^{k,\mathrm{cat}}\right), \quad \mathbf{S}_r = \{1 \le i \le R_y^{k,\mathrm{bat}} : r_i^* = r\}$
  $\bar{p}_{y,r}^{k,\mathrm{bat}} = \frac{\sum_{s\in\mathbf{S}_r} \pi_{y,s}^{k,\mathrm{bat}} p_{y,s}^{k,\mathrm{bat}}}{\sum_{s\in\mathbf{S}_r} \pi_{y,s}^{k,\mathrm{bat}}}$
  $p_{y,r}^{k,\mathrm{cat}} \leftarrow \beta p_{y,r}^{k,\mathrm{cat}} + (1-\beta)\bar{p}_{y,r}^{k,\mathrm{bat}}$
  $\mathbf{P}_y^{k,cat} \leftarrow \{p_{y,r}^{k,cat}\}_{r=1}^{R_y^k}$
end for
return $\mathbf{P}_y^{k,cat}$

D Selection of Key Hyperparameters

Figure 5(a) and Figure 5(b) evaluate the effect of λ_{RP} on CUB-200-2011 and Market-1501, respectively. On CUB-200-2011, Top-1 accuracy increases as λ_{RP} increases from 1 to 3, reaching the highest accuracy of 91.4% at $\lambda_{\mathrm{RP}} = 3$, and then decreases for larger values. On Market-1501, both mAP and Rank-1 accuracy reach the best performance at $\lambda_{\mathrm{RP}} = 4$. Accordingly, we set $\lambda_{\mathrm{RP}} = 3$ for FGVC and $\lambda_{\mathrm{RP}} = 4$ for ReID.

Figures 5(c)-(e) examine the momentum coefficient β, graph affinity temperature τ_g, and proxy assignment temperature τ_p on CUB-200-2011. The highest accuracy is obtained with $\beta = 0.2$, $\tau_g = 0.3$, and $\tau_p = 0.05$. Therefore, we adopt these values as the default settings.

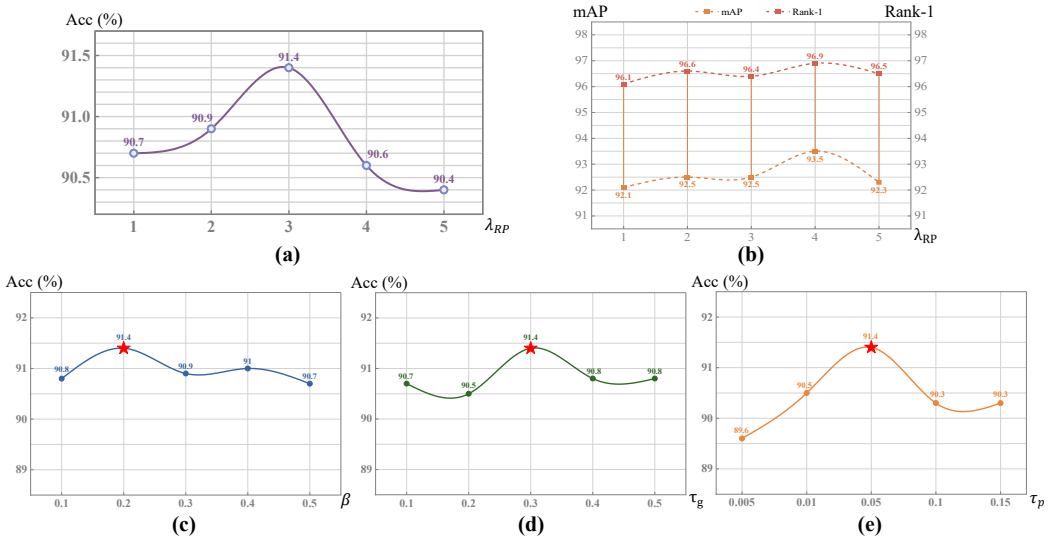


Figure 5: Selection of key hyperparameters. Effects of λ_{RP} on (a) Top-1 accuracy on CUB-200-2011 and (b) mAP and Rank-1 accuracy on Market-1501. (c)-(e) Effects of the momentum coefficient β, graph affinity temperature τ_g, and proxy assignment temperature τ_p on CUB-200-2011, respectively. Based on these results, we set $\lambda_{\text{RP}} = 3$ for FGVC and $\lambda_{\text{RP}} = 4$ for ReID, with $\beta = 0.2$, $\tau_g = 0.3$, and $\tau_p = 0.05$.

E Comparison on Vehicle Re-Identification

Table 5 compares our method with existing Vehicle ReID approaches on VeRi-776. Our method achieves the best performance on both evaluation metrics, reaching 88.3% mAP and 98.4% Rank-1 accuracy. Compared with the strongest existing results, our method improves mAP by 0.7% over VehicleMAE and Rank-1 accuracy by 0.4% over LSKA-ReID. These improvements demonstrate that our method effectively captures discriminative vehicle-specific cues, which are crucial for distinguishing vehicles with similar appearances. By leveraging adaptive proxy learning, our method better models diverse intra-class patterns and improves vehicle identity matching under fine-grained variations.

Table 5: Performance comparison on Vehicle ReID across VeRi-776. The best performance is highlighted in **bold**.

Method	VeRi-776 mAP	VeRi-776 Rank-1
TransReID (He et al., 2021)	82.0	97.1
CLIP-ReID (Li et al., 2023a)	83.3	97.4
VehicleGAN (Li et al., 2024)	79.4	94.7
LKA-ReID (Xiang et al., 2025b)	84.1	96.1
ADPRP-Net (Zhou et al., 2025)	82.8	95.6
LSKA-ReID (Xiang et al., 2025a)	85.6	98.0
VehicleMAE (Wang et al., 2025a)	87.6	97.4
DiPerceiveNet (Cai et al., 2026)	85.0	97.7
DINOv2 Full FT + Ours$_{\text{1-cls}}$	**88.3**	**98.4**

F Scalability and Compatibility

Effectiveness Across Different Backbone Scales. Table 6 evaluates our method with DINOv2 ViT-S/14, ViT-B/14, and ViT-L/14 on CUB-200-2011. Our method consistently improves full fine-tuning across all backbone scales, achieving improvements of 1.1%, 2.2%, and 1.0% on ViT-S/14, ViT-B/14, and ViT-L/14, respectively. These consistent improvements demonstrate that our method is effective across models with different capacities and is not limited to a specific backbone scale.

Compatibility Across Different Fine-Tuning Strategies. Table 7 evaluates our method with different fine-tuning strategies on four FGVC datasets. Our method consistently improves every fine-

Table 6: Effectiveness of the proposed method across different DINOv2 backbone scales on CUB-200-2011. "Full FT" denotes full fine-tuning of the backbone.

Methods	DINOv2 ViT-S/14	DINOv2 ViT-B/14	DINOv2 ViT-L/14
Full FT	89.0	89.2	91.0
Full FT+Ours	90.1 (↑1.1)	91.4 (↑2.2)	92.0 (↑1.0)

Table 7: Applicability of the proposed method across different fine-tuning strategies on FGVC datasets using DINOv2-B/14. "Full FT" denotes full fine-tuning of the backbone. "Params." denotes the number of trainable parameters, excluding the task-specific classification head. The best performance within each fine-tuning strategy is highlighted in **bold**.

Method	Params.(M)	CUB	Cars	Dogs	NABirds	Avg.
Adapter (Houlsby et al., 2019)	2.38	88.9	90.4	89.3	90.5	89.8
Adapter+Ours	2.38	**89.8**(↑0.9)	**91.5**(↑1.1)	**89.7**(↑0.4)	**91.8**(↑1.3)	**90.7**(↑0.9)
LoRA (Hu et al., 2022)	2.36	91.1	90.2	89.2	90.6	90.3
LoRA+Ours	2.36	**91.6**(↑0.5)	**91.0**(↑0.8)	**89.7**(↑0.5)	**91.7**(↑1.1)	**91.0**(↑0.7)
VPT-Shallow (Jia et al., 2022)	0.08	88.5	86.5	87.9	89.6	88.1
VPT-Shallow+Ours	0.08	**90.0**(↑1.5)	**87.3**(↑0.8)	**88.4**(↑0.5)	**90.0**(↑0.4)	**88.9**(↑0.8)
VPT-Deep (Jia et al., 2022)	0.92	89.1	89.0	89.2	91.2	89.6
VPT-Deep+Ours	0.92	**91.2**(↑2.1)	**90.9**(↑1.9)	**89.5**(↑0.3)	**91.9**(↑0.7)	**90.9**(↑1.3)
Full FT	86.58	89.2	90.2	88.7	89.1	89.3
Full FT+Ours	86.58	**91.4**(↑2.2)	**92.0**(↑1.8)	**89.4**(↑0.7)	**91.3**(↑2.2)	**91.0**(↑1.7)

tuning strategy on all datasets without increasing the number of trainable parameters. For average accuracy, our method achieves improvements of 0.9%, 0.7%, 0.8%, 1.3%, and 1.7% over Adapter, LoRA, VPT-Shallow, VPT-Deep, and full fine-tuning, respectively. These consistent improvements demonstrate that our method can be flexibly integrated into diverse fine-tuning pipelines.

G Visualization of Fine-Grained Attention on Person ReID

Figure 6 further visualizes the responses of two class tokens on Market-1501. Under vanilla full fine-tuning, one class token typically focuses on the pedestrian, while the other responds to background or less informative regions. This indicates that the representational capacity of multiple class tokens is not fully exploited. After incorporating our method, both class tokens consistently focus on the foreground person and capture complementary body regions and appearance cues. For example, different class tokens attend to distinct regions such as the upper body, lower body, and clothing patterns, providing diverse identity-related information. This suggests that our method enables different class tokens to learn identity-relevant representations and promotes complementary feature learning in the multi-CLS setting.

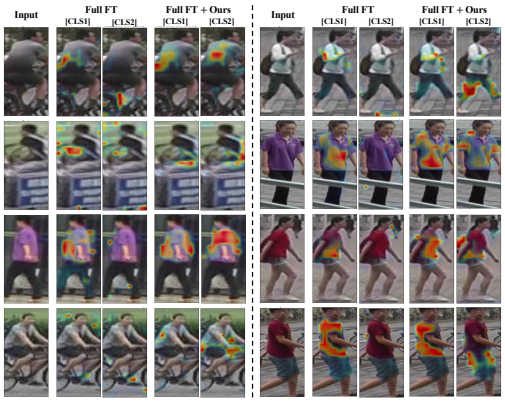


Figure 6: Grad-CAM visualizations on Market-1501 with two learnable class tokens.