

CARE: Constrained Attention Refinement for Fine-Grained Visual Classification via Teacher-Student Distillation

Ruibo Wen¹[0009-0008-1671-2279], Hang Shao¹[0000-0002-1322-4789], and
Yiming Lei¹[0000-0002-1349-7074]✉

College of Computer Science and Technology, Qingdao University, Qingdao, China
leiyiming@qdu.edu.cn

Abstract. Fine-grained visual classification requires models to recognize subtle local traits while exposing the visual evidence behind their predictions. Class-specific attention pathways provide a natural basis for interpretable recognition, but their constrained prediction structure limits discriminative capacity and underuses intermediate representations from strong pretrained backbones. To address this problem, we propose CARE, a constrained attention refinement framework for interpretable fine-grained recognition via teacher-student distillation. CARE keeps the final prediction and explanation within a class-specific attention student, while introducing a training-only auxiliary query teacher that reads selected intermediate DINOv2 layers with learnable queries. The teacher fuses multi-level representations and transfers logit-standardized class-discriminative knowledge to the student. To further refine the explanation pathway, we design diversity and sparsity terms to regularize student attention heads, reducing redundancy and encouraging compact trait localization. Experiments on CUB, Oxford-IIIT Pet, Stanford Dogs, and Stanford Cars show that CARE achieves strong classification performance under an interpretable frozen-backbone setting, reaching 78.5% Top-1 accuracy on CUB. Faithfulness analysis with insertion and deletion metrics further indicates that the top-ranked attention regions retain class-relevant evidence for explanation. The source code is publicly available at <https://github.com/panpan0814/CARE>.

Keywords: Fine-grained recognition · Interpretable vision transformers · Visual prompt tuning · Knowledge distillation

1 Introduction

Fine-grained visual classification (FGVC) aims to separate categories that differ in subtle local cues, including species, breeds, and object subtypes. In biodiversity monitoring, industrial inspection, and medical image analysis, a small visual

✉ Corresponding author.

detail may determine the correct category. Models used in these settings therefore need to expose the visual evidence behind their predictions, rather than rely on background bias or holistic shortcuts [13,6].

Vision Transformers (ViTs) are strong backbones for FGVC because their token interactions capture both local parts and global context [5,14,6]. Self-supervised models such as DINO and DINOv2 further provide transferable frozen representations for downstream recognition tasks [2,16]. Post-hoc attribution methods can explain decisions after a predictive model has been trained, but they do not require the classifier to use trait-level evidence when making predictions [20,1,3]. Intrinsically interpretable models connect prediction and explanation more directly, although they often introduce additional structures that are not always suitable for lightweight adaptation of frozen ViT backbones [18,15].

Visual Prompt Tuning (VPT) adapts frozen ViTs with learnable visual prompt tokens and provides a parameter-efficient alternative to full fine-tuning [8]. Its focus, however, is mainly on recognition performance, and it does not provide class-specific visual explanations. Prompt-CAM extends prompt-based adaptation to intrinsic interpretability by using class-specific prompt outputs for prediction, with prompt-to-patch attention maps serving as class-level explanations [4]. This design is efficient and compatible with frozen ViTs, but prediction is still limited to final-layer prompt representations and a shared scoring head. As a result, it may not fully use intermediate Transformer representations that contain complementary local, part-level, and semantic evidence.

We propose CARE, a constrained attention refinement framework for interpretable fine-grained recognition. CARE keeps a class-specific prompt-attention student as the final prediction and explanation pathway, and improves it through training supervision. Diversity and sparsity losses regularize the student attention heads to reduce redundancy and encourage compact trait localization. A training-only auxiliary query teacher reads selected intermediate DINOv2 layers and transfers multi-level discriminative knowledge to the student through logit-standardized knowledge distillation (KD) [7,21]. During inference, the teacher branch is removed, so prediction and explanation are produced by the student alone. Our contributions are summarized as follows:

- We formulate constrained attention refinement for interpretable FGVC, aiming to improve discriminative capacity while keeping prediction and explanation within a class-specific attention pathway.
- We introduce a training-only auxiliary query teacher that reads multi-level DINOv2 representations and distills logit-standardized class-discriminative knowledge to the student.
- We apply diversity and sparsity regularization to student attention heads, reducing redundancy and encouraging compact localization of fine-grained traits.
- We evaluate CARE on CUB, Oxford-IIIT Pet, Stanford Dogs, and Stanford Cars using classification accuracy, insertion/deletion faithfulness, component ablations, and teacher-layer analysis.

2 Related Work

Fine-grained visual classification. FGVC distinguishes visually similar categories using subtle local traits. Early methods relied on part localization, bilinear interactions, or region modeling [28,13]. Transformers reduce the need for hand-crafted part modeling by representing images as patch tokens [5,22,14], and TransFG adapts ViTs to FGVC through discriminative token selection [6]. Furthermore, self-supervised foundation models such as DINO and DINOv2 provide strong frozen representations for downstream recognition [2,16]. Our work studies how to exploit multi-level DINOv2 representations while preserving class-specific visual evidence for interpretable fine-grained recognition.

Prompt learning and parameter-efficient adaptation. Prompt learning adapts pretrained models with a small number of trainable parameters. VPT introduces visual prompts into frozen ViTs [8], while CoOp and MaPLe adapt vision-language models with learned prompts [30,10]. Recent work on PEFT further improves prompt-based adaptation [19]. These methods mainly target recognition performance rather than intrinsic explanation. Prompt-CAM extends prompt-based adaptation by using class-specific prompts for both prediction and explanation [4]. In contrast, CARE focuses on refining the class-specific attention pathway through attention regularization and training-only teacher-student distillation.

Visual attribution and interpretability. Post-hoc methods such as Grad-CAM, Attention Rollout, and Transformer attribution explain trained classifiers after the decision function has been learned [20,1,3]. They are useful diagnostic tools, but they do not require the classifier itself to rely on interpretable visual evidence. Intrinsic methods instead build explanation structures into the model. MCTformer uses class tokens for localization [26], ProtoTree and ProtoPFormer use prototype reasoning [15,27], and INTR uses class-specific queries in an encoder-decoder architecture [18]. Class-specific prompt-attention models offer a simpler and more parameter-efficient form of intrinsic explanation, but their constrained prediction pathway can limit discriminative capacity. CARE addresses this limitation by preserving the attention-based explanation pathway while strengthening it with multi-level teacher supervision.

Knowledge distillation and intermediate representations. Knowledge distillation transfers softened outputs or intermediate features from a teacher to a student [7]. Decoupled distillation and logit standardization further refine the transferred signal [29,21]. In ViTs, intermediate layers encode complementary cues, ranging from local patterns to semantic abstractions. Visual Query Tuning shows that learnable queries can access such representations efficiently [24]. Inspired by this idea, the Auxiliary Query Teacher in CARE reads selected DINOv2 layers, produces teacher logits, and transfers multi-level discriminative knowledge to the Constrained Attention Student only during training.

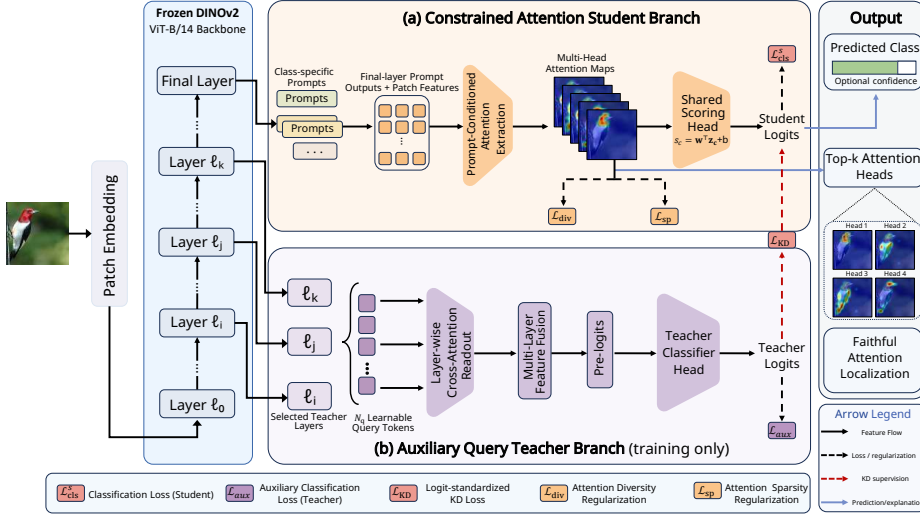


Fig. 1. Overall architecture of CARE. The framework consists of (a) a Constrained Attention Student branch for prediction and explanation, and (b) a training-only Auxiliary Query Teacher branch for multi-layer distillation.

3 Methodology

3.1 Overview

Given an input image $\mathbf{x}$ and label $y \in \{1, \dots, C\}$, our goal is to strengthen the discriminative capacity of a class-specific attention pathway while preserving it as the final prediction and explanation source. Fig. 1 shows the proposed CARE framework, which is built on a frozen DINOv2 Vision Transformer [16]. During training, CARE contains two branches: (a) a Constrained Attention Student for prediction and explanation, and (b) a training-only Auxiliary Query Teacher for multi-layer distillation.

The Constrained Attention Student follows a class-specific prompt-attention design for prediction and explanation. Class-specific prompt outputs are scored by a shared vector to produce logits, and the corresponding prompt-to-patch multi-head attention maps are used as class-specific explanations. We regularize these maps with diversity and sparsity losses to reduce redundancy across heads and encourage more compact localization. This keeps the explanation source tied to the same class-specific pathway used for prediction. The Auxiliary Query Teacher is introduced to exploit hierarchical evidence that may not be fully represented by the final student prompt outputs. It reads features from selected intermediate layers, where cues such as local structure, part boundaries, texture, and higher-level semantics are useful for fine-grained recognition.

The frozen backbone extracts both intermediate and final-layer features. The student produces logits and attention maps through the final class-specific attention pathway. In parallel, the teacher uses learnable query tokens to read selected

intermediate layers, fuses the query-conditioned readouts, and predicts teacher logits. These logits supervise the student through logit-standardized KD [7,21]. At inference, the teacher branch is removed, and the student alone provides both prediction and explanation.

3.2 Constrained Attention Student Branch

The student branch follows the class-specific prompt-attention principle of Prompt-CAM [4]. Instead of using a generic CLS-token classifier, class-specific prompt tokens are used to compute category scores from their corresponding prompt outputs. The same prompts also query image patches through attention, so the prompt-to-patch attention maps can be used as class-specific visual evidence.

Let F_θ denote the frozen DINOv2 ViT with N Transformer layers. The input image $\mathbf{x}$ is mapped to patch tokens $\mathbf{E}_0 \in \mathbb{R}^{M \times D}$, where M is the number of image patches and D is the hidden dimension. We introduce class-specific prompt tokens $\mathbf{P} = [\mathbf{p}^1, \dots, \mathbf{p}^C]^\top \in \mathbb{R}^{C \times D}$, where $\mathbf{p}^c \in \mathbb{R}^D$ is the learnable prompt token for class c . After the final Transformer layer, the class prompt outputs are $\mathbf{Z}_N^s = [\mathbf{z}_N^{s,1}, \dots, \mathbf{z}_N^{s,C}]^\top \in \mathbb{R}^{C \times D}$.

A shared scoring vector $\mathbf{w}_s \in \mathbb{R}^D$ and bias $b_s \in \mathbb{R}$ map each class prompt to a scalar logit:

$$s_c^s = \mathbf{w}_s^\top \mathbf{z}_N^{s,c} + b_s, \quad c = 1, \dots, C. \quad (1)$$

The student logit vector and classification loss are

$$\mathbf{s}^s = [s_1^s, \dots, s_C^s]^\top \in \mathbb{R}^C, \quad \mathcal{L}_{\text{cls}}^s = -\log \frac{\exp(s_y^s)}{\sum_{c=1}^C \exp(s_c^s)}. \quad (2)$$

For class c and attention head r , let $\alpha^{c,r} \in \mathbb{R}^M$ denote the final-layer prompt-to-patch attention map over image patches. With R attention heads, the attention-map set for class c is $\mathcal{A}^c = \{\alpha^{c,r}\}_{r=1}^R$. The Constrained Attention Student keeps the class-specific prediction pathway and adds $\mathcal{L}_{\text{div}}$ and $\mathcal{L}_{\text{sp}}$ to regularize the student attention maps.

3.3 Auxiliary Query Teacher Branch

The Auxiliary Query Teacher reads features from selected layers of the frozen backbone. Let the frozen backbone contain N Transformer layers indexed as $\ell_0, \ell_1, \dots, \ell_{N-1}$. The selected teacher-layer set is denoted as $\mathcal{S} = \{\ell_i, \ell_j, \ell_k\}$, where $0 \leq i < j < k \leq N-1$. Unless otherwise stated, we use the zero-based DINOv2 layer indices $\{4, 8, 11\}$.

For layer $\ell \in \mathcal{S}$, the context tokens are $\mathbf{H}_\ell = [\mathbf{h}_\ell^1, \dots, \mathbf{h}_\ell^{M_\ell}]^\top \in \mathbb{R}^{M_\ell \times D}$, where M_ℓ denotes the number of context tokens available to the teacher branch. These tokens contain the intermediate non-prompt representations produced by the frozen backbone.

We introduce N_q learnable query tokens $\mathbf{U}_\ell \in \mathbb{R}^{N_q \times D}$ for each selected layer, with $N_q = 8$ in the main configuration. The query, key, and value matrices are

$$\mathbf{Q}_\ell = \mathbf{U}_\ell \mathbf{W}_\ell^Q, \quad \mathbf{K}_\ell = \mathbf{H}_\ell \mathbf{W}_\ell^K, \quad \mathbf{V}_\ell = \mathbf{H}_\ell \mathbf{W}_\ell^V, \quad (3)$$

where $\mathbf{W}_\ell^Q, \mathbf{W}_\ell^K, \mathbf{W}_\ell^V \in \mathbb{R}^{D \times D}$. The layer-wise cross-attention matrix is

$$\mathbf{A}_\ell = \text{softmax}_{\text{ctx}} \left(\frac{\mathbf{Q}_\ell \mathbf{K}_\ell^\top}{\sqrt{D}} \right) \in \mathbb{R}^{N_q \times M_\ell}, \quad (4)$$

where the softmax is applied over the context-token dimension. Equivalently,

$$A_{\ell,q,m} = \frac{\exp(\langle \mathbf{Q}_\ell^q, \mathbf{K}_\ell^m \rangle / \sqrt{D})}{\sum_{m'=1}^{M_\ell} \exp(\langle \mathbf{Q}_\ell^q, \mathbf{K}_\ell^{m'} \rangle / \sqrt{D})}. \quad (5)$$

The query-conditioned readout is

$$\mathbf{R}_\ell = \mathbf{A}_\ell \mathbf{V}_\ell \in \mathbb{R}^{N_q \times D}. \quad (6)$$

In implementation, this readout is realized as a residual multi-head cross-attention module followed by a lightweight feed-forward network, and the readout dimension is kept the same as the backbone embedding dimension.

We average the query readouts and fuse the selected layers:

$$\rho(\mathbf{R}_\ell) = \frac{1}{N_q} \sum_{q=1}^{N_q} \mathbf{R}_\ell^q, \quad \mathbf{g} = \phi(\text{Concat}[\rho(\mathbf{R}_{\ell_i}), \rho(\mathbf{R}_{\ell_j}), \rho(\mathbf{R}_{\ell_k})]), \quad (7)$$

where $\rho(\mathbf{R}_\ell) \in \mathbb{R}^D$, $\phi: \mathbb{R}^{3D} \rightarrow \mathbb{R}^D$ is a lightweight fusion MLP, and $\mathbf{g} \in \mathbb{R}^D$ is the fused teacher representation.

The teacher logits and auxiliary loss are

$$\mathbf{s}^t = \mathbf{W}_t \mathbf{g} + \mathbf{b}_t, \quad \mathcal{L}_{\text{aux}} = -\log \frac{\exp(s_y^t)}{\sum_{c=1}^C \exp(s_c^t)}. \quad (8)$$

Here, $\mathbf{W}_t \in \mathbb{R}^{C \times D}$, $\mathbf{b}_t \in \mathbb{R}^C$, and $\mathbf{s}^t \in \mathbb{R}^C$. The teacher branch is used only for training-time supervision and is not used as the explanation source.

3.4 Logit-Standardized Knowledge Distillation

Teacher and student logits may have different scales, so we standardize them before distillation [21]. For $\mathbf{s} \in \mathbb{R}^C$,

$$\hat{\mathbf{s}} = \frac{\mathbf{s} - \mu(\mathbf{s})\mathbf{1}}{\sigma(\mathbf{s}) + \epsilon}, \quad \mu(\mathbf{s}) = \frac{1}{C} \sum_{j=1}^C s_j, \quad \sigma(\mathbf{s}) = \left(\frac{1}{C} \sum_{j=1}^C (s_j - \mu(\mathbf{s}))^2 \right)^{1/2}. \quad (9)$$

Here, $\mathbf{1} \in \mathbb{R}^C$ is an all-one vector and ϵ is a small constant for numerical stability. The standardized teacher and student distributions are

$$\mathbf{q}^t = \text{softmax} \left(\frac{\text{sg}(\hat{\mathbf{s}}^t)}{T} \right), \quad \mathbf{q}^s = \text{softmax} \left(\frac{\hat{\mathbf{s}}^s}{T} \right), \quad (10)$$

where $T > 0$ is the distillation temperature, $\mathbf{q}^t, \mathbf{q}^s \in \mathbb{R}^C$ are probability distributions, and $\text{sg}(\cdot)$ denotes stop-gradient. Thus, the KD term transfers knowledge from the teacher to the student without updating the teacher through the distillation loss.

The KL divergence and KD loss are defined as:

$$\text{KL}(\mathbf{q}^t \parallel \mathbf{q}^s) = \sum_{c=1}^C q_c^t \log \left(\frac{q_c^t + \epsilon}{q_c^s + \epsilon} \right), \quad \mathcal{L}_{\text{KD}} = T^2 \text{KL}(\mathbf{q}^t \parallel \mathbf{q}^s). \quad (11)$$

This loss transfers the teacher’s relative class distribution to the Constrained Attention Student.

3.5 Attention Diversity and Sparsity Regularization

For the ground-truth class y , the normalized attention map of head r is

$$\tilde{\alpha}^{y,r} = \frac{\alpha^{y,r}}{\sum_{m=1}^M \alpha_m^{y,r} + \epsilon} \in \mathbb{R}^M, \quad (12)$$

where $\alpha_m^{y,r}$ denotes the attention value on patch m and ϵ is a small constant.

The diversity loss is the average pairwise cosine similarity between different heads:

$$\mathcal{L}_{\text{div}} = \frac{1}{R(R-1)} \sum_{\substack{r,r'=1 \\ r \neq r'}}^R \frac{(\tilde{\alpha}^{y,r})^\top \tilde{\alpha}^{y,r'}}{\|\tilde{\alpha}^{y,r}\|_2 \|\tilde{\alpha}^{y,r'}\|_2 + \epsilon}. \quad (13)$$

For sparsity, we first select the top- k attention heads of the ground-truth class and then regularize their entropy. The head-selection score is defined by the maximum prompt-to-patch attention response:

$$\eta_r^y = \max_{m \in \{1, \dots, M\}} \alpha_m^{y,r}. \quad (14)$$

The selected head index set is

$$\mathcal{I}_k^y = \text{TopK}_{r \in \{1, \dots, R\}} (\text{sg}(\eta_r^y), k), \quad (15)$$

where $\text{sg}(\cdot)$ denotes stop-gradient. The discrete top- k operation determines which heads are regularized, while gradients are back-propagated through the selected attention maps.

The sparsity loss is the normalized entropy of the selected heads:

$$\mathcal{L}_{\text{sp}} = \frac{1}{|\mathcal{I}_k^y|} \sum_{r \in \mathcal{I}_k^y} \frac{\mathcal{H}(\tilde{\alpha}^{y,r})}{\log M}, \quad (16)$$

where

$$\mathcal{H}(\tilde{\alpha}^{y,r}) = - \sum_{m=1}^M \tilde{\alpha}_m^{y,r} \log (\tilde{\alpha}_m^{y,r} + \epsilon). \quad (17)$$

Both regularizers act directly on the Constrained Attention Student attention maps.

3.6 Overall Objective and Training

The total objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}}^s + \lambda_{\text{aux}}\mathcal{L}_{\text{aux}} + \lambda_{\text{KD}}\mathcal{L}_{\text{KD}} + \lambda_{\text{div}}\mathcal{L}_{\text{div}} + \lambda_{\text{sp}}\mathcal{L}_{\text{sp}}. \quad (18)$$

Here, λ_{aux} , λ_{KD} , λ_{div} , and λ_{sp} are scalar loss weights. The attention regularizers do not introduce additional trainable parameters; they only impose constraints on student attention maps. The trainable parameters include the class-specific prompt tokens, the shared student scoring head, and the Auxiliary Query Teacher branch, while the DINOv2 backbone remains frozen. During inference, the final prediction and explanation are produced by the Constrained Attention Student branch, and the Auxiliary Query Teacher is removed.

4 Experiments

4.1 Datasets and Implementation Details

Datasets We evaluate CARE on CUB-200-2011 [23], Stanford Dogs [11], Oxford-IIIT Pet [17], and Stanford Cars [12]. CUB is used for ablation studies, teacher-layer selection, and hyperparameter analysis. The remaining datasets test whether the observed behavior transfers across different fine-grained domains.

Experimental settings All prompt-based models use a frozen DINOv2 ViT-B/14 backbone [16] with an input resolution of 224×224 . Unless otherwise stated, models are trained for 60 epochs with SGD, momentum 0.9, weight decay 0.001, batch size 64, learning rate 0.005, and 10 warm-up epochs. For CARE, we use zero-based teacher-layer indices $\{4, 8, 11\}$, $N_q = 8$, $\lambda_{\text{aux}} = 0.5$, $\lambda_{\text{KD}} = 1.0$, $T = 2.0$, $\lambda_{\text{div}} = 0.05$, $\lambda_{\text{sp}} = 0.01$, and $k = 4$ for sparsity regularization.

Evaluation Metrics Classification performance is measured by Top-1 accuracy. For explanation faithfulness, we report insertion and deletion scores. Class-specific prompt-attention methods are evaluated using their intrinsic attention maps, while Grad-CAM, Layer-CAM, and Attention Rollout are evaluated on the DINOv2 Linear Probing classifier [20,9,1].

Baselines We compare CARE with both recognition-oriented and interpretable baselines. DINOv2 Linear Probing is included as a recognition-oriented reference with a frozen ViT-B/14 backbone and a linear head. For interpretable fine-grained recognition, we compare with INTR [18], ProtoTree [15], TesNet [25], and the class-specific prompt-attention baseline Prompt-CAM [4]. Unless otherwise stated, Linear Probing, Prompt-CAM, and CARE are evaluated under the same frozen DINOv2 setting, while the remaining interpretable baseline results follow the corresponding published reports. Prompt-CAM is the closest baseline, while CARE further introduces constrained attention refinement with training-only teacher-student distillation.

Table 1. Top-1 classification accuracy (%) on four fine-grained recognition benchmarks. Gray values denote recognition-oriented Linear Probing results provided for reference. Bold indicates the best result among the compared interpretable methods.

Dataset	Linear Probing	INTR	ProtoTree	TesNet	Prompt-CAM	CARE
CUB	89.6	71.8	68.5	77.7	74.1	78.5
Dog	87.3	72.5	68.4	76.5	81.3	81.9
Pet	96.2	90.4	–	–	92.7	93.6
Cars	88.2	86.8	70.0	84.6	84.3	87.1

4.2 Main Results

The classification results show that CARE consistently improves the class-specific prompt-attention baseline across all four fine-grained benchmarks, as reported in Table 1. Compared with Prompt-CAM, CARE increases Top-1 accuracy from 74.1% to 78.5% on CUB, from 81.3% to 81.9% on Stanford Dogs, from 92.7% to 93.6% on Oxford-IIIT Pet, and from 84.3% to 87.1% on Stanford Cars.

These gains suggest that the auxiliary query teacher and logit-standardized distillation improve the discriminative ability of the class-specific attention pathway while preserving its intrinsic explanation mechanism.

4.3 Faithfulness Analysis

Explanation faithfulness is evaluated with insertion and deletion scores under the two settings defined above: post-hoc explanations on the DINOv2 Linear Probing classifier and intrinsic attention maps from class-specific prompt-attention methods. All methods follow the same perturbation protocol. Saliency maps are upsampled to the image resolution, pixels are ranked by descending saliency, and each score is computed as the area under the target-class confidence curve over 50 perturbation steps. The quantitative results are reported in Table 2.

Under the intrinsic prompt-attention setting, CARE improves insertion over Prompt-CAM on CUB, Oxford-IIIT Pet, and Stanford Dogs. This suggests that the top-ranked attention regions selected by CARE contain more prediction-relevant evidence during the insertion process. On CUB, the regularization-only variant also improves insertion, indicating that diversity and sparsity constraints can improve the quality of selected-head evidence. Deletion results are more mixed: CARE is slightly worse than Prompt-CAM on CUB, but performs better on Oxford-IIIT Pet and Stanford Dogs. The corresponding insertion and deletion curves are shown in Fig. 2.

4.4 Ablation Study

Component ablation. The component ablation on CUB examines how each part of CARE contributes to the final performance, including the Auxiliary Query Teacher, logit-standardized KD, and attention regularization, as reported in Table 3. The teacher-distillation variant improves the Prompt-CAM baseline from

Table 2. Insertion/deletion faithfulness evaluation under post-hoc and class-specific prompt-attention explanation settings. Prompt-CAM + Reg. denotes the student-only CARE variant with $\mathcal{L}_{\text{div}} + \mathcal{L}_{\text{sp}}$, without the Auxiliary Query Teacher or KD.

(a) Post-hoc methods on CUB			(b) Class-specific prompt-attention methods					
Method	Ins.↑	Del.↓	CUB		Pet		Dog	
			Ins.↑	Del.↓	Ins.↑	Del.↓	Ins.↑	Del.↓
Grad-CAM	0.61	0.17						
Layer-CAM	0.77	0.28						
Att. Rollout	0.73	0.28						
			CUB		Pet		Dog	
Prompt-CAM	0.78	0.13	0.79	0.40	0.68	0.31		
Prompt-CAM + Reg.	0.85	0.15	—	—	—	—		
CARE	0.86	0.16	0.86	0.39	0.83	0.29		

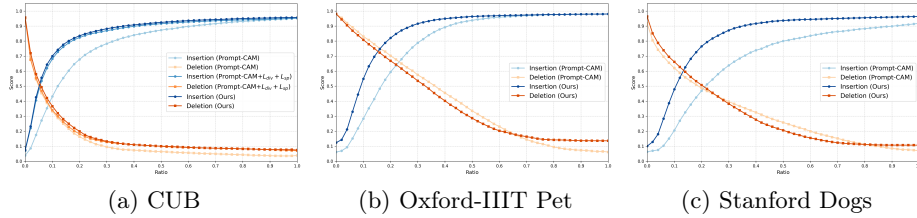


Fig. 2. Insertion and deletion curve comparisons on CUB, Oxford-IIIT Pet, and Stanford Dogs. The CUB curves include Prompt-CAM, Prompt-CAM with diversity and sparsity attention regularization, and CARE. The Oxford-IIIT Pet and Stanford Dogs curves include Prompt-CAM and CARE.

74.1% to 78.3%, indicating that multi-level teacher supervision is the main source of the accuracy gain. In contrast, applying $\mathcal{L}_{\text{div}}$ and $\mathcal{L}_{\text{sp}}$ without teacher supervision reduces accuracy to 73.4%, suggesting that attention regularization alone can over-constrain the class-specific attention pathway. When teacher-guided distillation and attention regularization are combined, accuracy reaches 78.5%, outperforming both the Prompt-CAM baseline and the teacher-only variant. This result suggests that multi-level teacher supervision and attention regularization play complementary roles in improving the Constrained Attention Student.

KD-weight sensitivity. Although the framework contains several loss weights, we focus the sensitivity analysis on λ_{KD} because it directly controls the strength of knowledge transfer from the Auxiliary Query Teacher to the Constrained Attention Student. The accuracy curve in Fig. 3(a) shows that the best result is obtained at $\lambda_{\text{KD}} = 1.0$, reaching 78.55% Top-1 accuracy on CUB. The faithfulness curves in Fig. 3(b) show that insertion and deletion scores vary only slightly across the tested KD weights. We therefore use $\lambda_{\text{KD}} = 1.0$ as the default setting in the main experiments.

4.5 Query Token Analysis

The number of learnable query tokens controls the readout capacity of the Auxiliary Query Teacher for multi-level feature aggregation. We analyze this factor on CUB by fixing the selected teacher-layer indices to $\{4, 8, 11\}$ and changing only

Table 3. Component ablation study of CARE on the CUB dataset. The teacher branch includes $\mathcal{L}_{\text{aux}}$ by default. CARE w/o Teacher/KD removes both the Auxiliary Query Teacher and KD, leaving only attention regularization on the student.

Setting	Teacher KD Reg. Top-1			
Prompt-CAM	–	–	–	74.1
CARE w/o Reg.	✓	✓	–	78.3
CARE w/o Teacher/KD	–	–	✓	73.4
CARE	✓	✓	✓	78.5

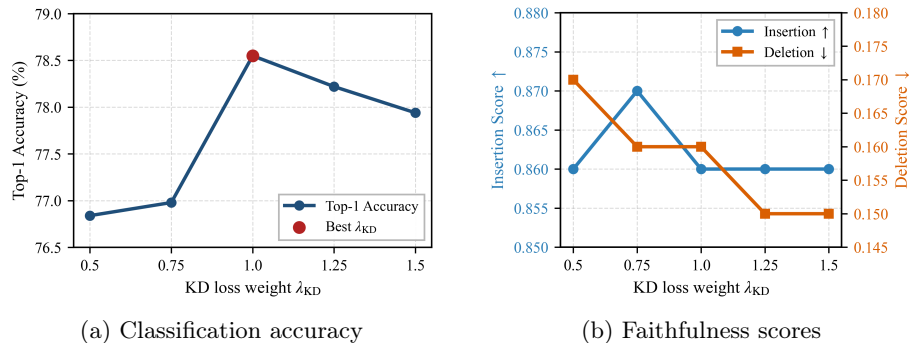


Fig. 3. Sensitivity analysis of the KD loss weight λ_{KD} on CUB. (a) reports Top-1 classification accuracy. (b) reports insertion and deletion faithfulness scores. All other hyperparameters are kept identical to the default CARE configuration.

N_q , with Prompt-CAM included as the baseline without teacher query tokens. The results are reported in Table 4.

With $N_q = 4$, CARE reaches 77.1% Top-1 accuracy, improving over Prompt-CAM but remaining below the default setting. This suggests that four query tokens may provide insufficient capacity to summarize intermediate representations. Increasing N_q to 8 raises accuracy to 78.5%, indicating that moderate query-token capacity helps the teacher extract multi-level evidence. Increasing N_q to 16 lowers accuracy to 76.7%, suggesting that excessive query tokens may introduce redundant or unstable readouts for distillation.

The parameter counts change only marginally across different values of N_q , so the accuracy differences mainly reflect teacher readout capacity rather than parameter scaling. We therefore use $N_q = 8$ as the default setting.

4.6 Selecting Complementary Teacher Layers

The selected teacher-layer indices control the depth distribution of the multi-level supervision provided by the Auxiliary Query Teacher. As reported in Table 5, the zero-based index combination $\{4, 8, 11\}$ achieves the best accuracy, while changing the early, middle, or late layer reduces performance. This suggests that

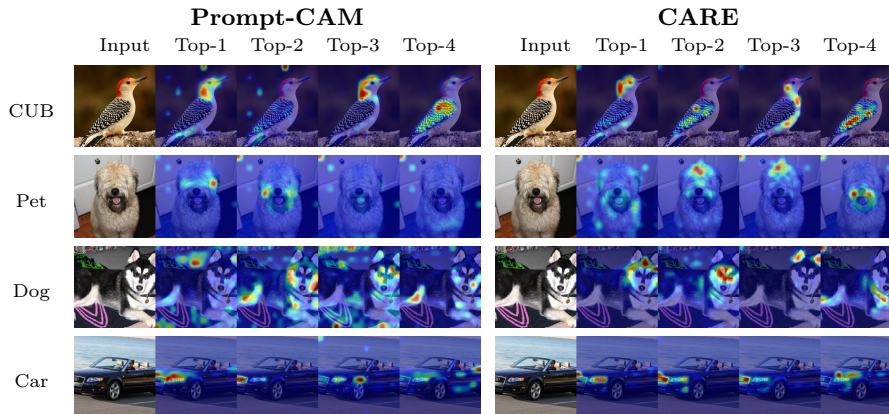


Fig. 4. Visualization of class-specific attention maps from the top-4 attention heads on CUB, Oxford-IIIT Pet, Stanford Dogs, and Stanford Cars.

Table 4. Query-token analysis on the CUB dataset.

Setting	N_q	Top-1	Train. Params	Total Params
Prompt-CAM	—	74.1	1.84M	87.56M
CARE	4	77.1	25.05M	110.77M
CARE	8	78.5	25.06M	110.78M
CARE	16	76.7	25.07M	110.79M

Table 5. Layer combination study on the CUB dataset.

Teacher indices	Top-1
4, 8, 11	78.5
5, 8, 11	76.9
4, 7, 11	76.7
4, 9, 11	76.6
3, 7, 10	77.1

the teacher branch benefits from complementary representations across different Transformer depths.

This pattern is consistent with the hierarchy of ViT features: earlier layers preserve local appearance cues, middle layers provide part-level structure, and deeper layers encode semantic information. The $\{4, 8, 11\}$ combination therefore provides a more balanced teacher signal than alternatives at similar depths.

4.7 Qualitative Discussion

Qualitative visualizations are used to examine whether the refined class-specific attention maps highlight meaningful fine-grained regions across CUB, Oxford-IIIT Pet, Stanford Dogs, and Stanford Cars, as shown in Fig. 4. Prompt-CAM and CARE follow the same visualization protocol for selecting the displayed attention heads. Specifically, we adopt the iterative head-pruning procedure used by the Prompt-CAM baseline: at each step, one candidate head is replaced by a uniform attention pattern, the model is re-evaluated, and the head whose removal causes the smallest decrease in target-class confidence is pruned. The four remaining heads are visualized as trait-level attention maps. This visualization-time selection is separate from the training-time sparsity regularization in Sec. 3.5.

The selected heads generally correspond to semantically meaningful fine-grained regions, such as the head, beak, neck, ears, or other object-specific local parts. Since the heads are selected according to their contribution to target-class confidence rather than human visual preference, some maps may still contain contextual or diffuse responses. Compared with Prompt-CAM, CARE produces more compact and less redundant attention patterns in many examples. This qualitative behavior is consistent with the insertion/deletion results and supports the role of attention regularization in refining class-specific localization.

5 Limitations and Discussion

CARE has two main limitations. First, the Auxiliary Query Teacher introduces additional training-time parameters and computational overhead, mainly from the learnable query tokens, multi-layer query readout, feature fusion, and teacher classifier. However, the teacher branch is discarded during inference, so the additional cost does not affect the deployed student model. Second, our experiments are conducted primarily on natural fine-grained datasets. Applying CARE to domains such as medical or industrial imagery requires further evaluation under domain-specific variations, robustness requirements, and reliability constraints.

6 Conclusion

We presented CARE, a constrained attention refinement framework for faithful fine-grained visual classification. CARE uses a Constrained Attention Student and a training-only Auxiliary Query Teacher to transfer multi-level discriminative knowledge while preserving class-specific attention as the prediction and explanation source. Experiments on four fine-grained benchmarks show consistent accuracy gains over the class-specific prompt-attention baseline and competitive faithfulness results, indicating that teacher-student distillation and attention regularization can strengthen the interpretable attention pathway without replacing its intrinsic localization mechanism.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No. 62306075) and the Natural Science Foundation of Shandong Province (No. ZR2026QC0725).

References

1. Abnar, S., Zuidema, W.: Quantifying attention flow in transformers. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 4190–4197 (2020)

2. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9650–9660 (2021)
3. Chefer, H., Gur, S., Wolf, L.: Transformer interpretability beyond attention visualization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 782–791 (2021)
4. Chowdhury, A., Paul, D., Mai, Z., Gu, J., Zhang, Z., Mehrab, K.S., Campolongo, E.G., Rubenstein, D., Stewart, C.V., Karpatne, A., Berger-Wolf, T., Su, Y., Chao, W.L.: Prompt-CAM: Making vision transformers interpretable for fine-grained analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4375–4385 (2025)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021)
6. He, J., Chen, J.N., Liu, S., Kortylewski, A., Yang, C., Bai, Y., Wang, C.: Transfg: A transformer architecture for fine-grained recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 852–860 (2022)
7. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: NIPS Deep Learning and Representation Learning Workshop (2015)
8. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 709–727 (2022)
9. Jiang, P.T., Zhang, C.B., Hou, Q., Cheng, M.M., Wei, Y.: Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing* **30**, 5875–5888 (2021)
10. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19113–19122 (2023)
11. Khosla, A., Jayadevaprakash, N., Yao, B., Fei-Fei, L.: Novel dataset for fine-grained image categorization. In: Proceedings of the CVPR Workshop on Fine-Grained Visual Categorization (2011)
12. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3D object representations for fine-grained categorization. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 554–561 (2013)
13. Lin, T.Y., RoyChowdhury, A., Maji, S.: Bilinear CNN models for fine-grained visual recognition. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 1449–1457 (2015)
14. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10012–10022 (2021)
15. Nauta, M., van Bree, R., Seifert, C.: Neural prototype trees for interpretable fine-grained image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14933–14943 (2021)
16. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.:

- Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research* (2024)
17. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.V.: Cats and dogs. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 3498–3505 (2012)
 18. Paul, D., Chowdhury, A., Xiong, X., Chang, F.J., Carlyn, D., Stevens, S., Provost, K.L., Karpatne, A., Carstens, B., Rubenstein, D., Stewart, C., Berger-Wolf, T., Su, Y., Chao, W.L.: A simple interpretable transformer for fine-grained image classification and analysis. In: *International Conference on Learning Representations (ICLR)* (2024)
 19. Ren, L., Chen, C., Wang, L., Hua, K.A.: Da-VPT: Semantic-guided visual prompt tuning for vision transformers. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2025)
 20. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. pp. 618–626 (2017)
 21. Sun, S., Ren, W., Li, J., Wang, R., Cao, X.: Logit standardization in knowledge distillation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 15731–15740 (2024)
 22. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers and distillation through attention. In: *Proceedings of the International Conference on Machine Learning (ICML)*. pp. 10347–10357 (2021)
 23. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset. *Tech. Rep. CNS-TR-2011-001*, California Institute of Technology (2011)
 24. Wang, J., Huang, Y., Wu, S., Chen, C., Tan, M.: Visual query tuning: Towards effective usage of intermediate representations for parameter and memory efficient transfer learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 7725–7735 (2023)
 25. Wang, J., Liu, H., Wang, X., Jing, L.: Interpretable image recognition by constructing transparent embedding space. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2021)
 26. Xu, L., Ouyang, W., Bennamoun, M., Boussaid, F., Xu, D.: Multi-class token transformer for weakly supervised semantic segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 4310–4319 (2022)
 27. Xue, M., Huang, Q., Zhang, H., Meng, L.Z., Song, J., Song, M.: Protopformer: Concentrating on prototypical parts in vision transformers for interpretable image recognition. In: *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. pp. 1517–1525 (2024)
 28. Zhang, N., Donahue, J., Girshick, R., Darrell, T.: Part-based R-CNNs for fine-grained category detection. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. pp. 834–849 (2014)
 29. Zhao, B., Cui, Q., Song, R., Qiu, Y., Liang, J.: Decoupled knowledge distillation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 11953–11962 (2022)
 30. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. *International Journal of Computer Vision* **130**, 2337–2348 (2022)