OPAL: Orthonormal Prototype Alignment Learning for Interpretable Image Classification

Ilán Carretero[1], Gustavo Jesús Angulo[2], Rocío del Amor[1,3], and Valery Naranjo[1,3]

[1] CVBLab, HumanTech, Universitat Politècnica de València, Valencia, Spain
{ilcarjuc,madeam2,vnaranjo}@upv.es
[2] CMA, Mines Paris, PSL University, Sophia Antipolis, France
jesus.angulo_lopez@minesparis.psl.eu
[3] Artikode Intelligence S.L., Valencia, Spain

Abstract. Prototypical part-based models provide explainable predictions by comparing input regions to learned prototypes. However, current approaches are burdened by complex, multi-stage training pipelines and heavily rely on auxiliary regularization to prevent prototype collapse. To overcome these limitations, we introduce Orthonormal Prototype Alignment Learning (OPAL), a single-stage, end-to-end framework that simplifies interpretable classification. Our approach anchors the latent space using predefined orthonormal bases, embedding each class within a dedicated subspace spanned by fixed part-prototypes. To achieve precise part localization, OPAL enforces spatial competition across feature maps. This mechanism isolates sparse, discriminative regions, directing each prototype to consistently attend to the same semantic concept across different images. By framing classification as a direct representation alignment task, our method eliminates the need for auxiliary losses. Extensive experiments on fine-grained benchmarks demonstrate that OPAL outperforms both its non-interpretable counterparts and state-of-the-art part-prototype methods, delivering granular visual explanations by explicitly revealing the specific image regions driving every prediction. Code is available at https://github.com/ilancarretero/OPAL.

Keywords: Explainable computer vision · Prototype-based classification · Representation alignment · Fine-grained image recognition

1 Introduction

Deep neural networks have revolutionized the field of computer vision, delivering unprecedented performance across a spectrum of complex recognition tasks [7, 9, 18, 19]. Despite these advancements, the inherent opacity of these architectures presents a fundamental challenge [4, 11, 20]. Beyond the critical trust issues arising in high-stakes deployment scenarios [5, 8], the inability to investigate the internal reasoning of a network limits the capacity to identify when and why the model fails, mitigate systematic errors, and potentially derive novel

Table 1: High-level comparison between a standard CNN and representative inherently interpretable models, including our contribution, OPAL. ✓ indicates that the method satisfies the corresponding property, and ✗ otherwise. "Aux-loss free" denotes training without additional auxiliary regularizers beyond the standard classification objective. "External-model free" denotes no reliance on auxiliary pretrained models or external part-discovery modules.

Method	Single-stage training	Aux-loss free	External-model free	Interpretable by design
Standard CNN	✓	✓	✓	✗
ProtoPNet NeurIPS '19 [3]	✗	✗	✓	✓
ProtoTree CVPR '21 [25]	✗	✓	✓	✓
ProtoPShare KDD '21 [34]	✗	✗	✓	✓
TesNet ICCV '21 [43]	✗	✗	✓	✓
ProtoPool ECCV '22 [33]	✗	✗	✓	✓
PIP-Net CVPR '23 [24]	✗	✗	✓	✓
MCPNet CVPR '24 [41]	✓	✗	✓	✓
LucidPPN ICLR '25 [28]	✗	✗	✗	✓
OPAL *(Ours)*	✓	✓	✓	✓

domain knowledge from learned feature representations. While post-hoc explanation methods, such as saliency maps [31, 35], attempt to interpret trained networks by highlighting relevant input regions, they have been shown to lack faithfulness to the actual decision-making process [1, 32]. Consequently, reliance on these approximations can be misleading, driving a paradigm shift towards inherently interpretable architectures where transparency is not a retrospective approximation, but a structural constraint enforced by design.

Among inherently interpretable models, Prototypical Part-based Networks (ProtoPNet) [3] and their recent variants [24, 28, 33, 34] have emerged as the dominant paradigm. These architectures ground their decision process in case-based reasoning, decomposing classification into a transparent matching process where local image regions are compared against a set of learned latent prototypes. However, despite their conceptual appeal, optimizing these frameworks involves a structural complexity significantly higher than their non-interpretable counterparts. As systematically illustrated in Tab. 1, state-of-the-art methods rely on cumbersome, multi-stage training pipelines that alternate between backbone optimization, prototype clustering, and discrete projection steps to map latent vectors to real training patches. This complexity stems fundamentally from the *"moving target"* problem [3, 43]: since prototypes are learnable parameters residing in a continuously shifting latent space, the model struggles to simultaneously learn discriminative features and align prototypes to them. To mitigate semantic collapse, where prototypes converge to identical features [34], existing approaches depend on multiple auxiliary regularization terms (e.g., separation, clustering, and diversity losses). This dependency significantly increases the complexity of hyperparameter tuning, as balancing these competing objectives requires meticulous dataset-specific calibration.

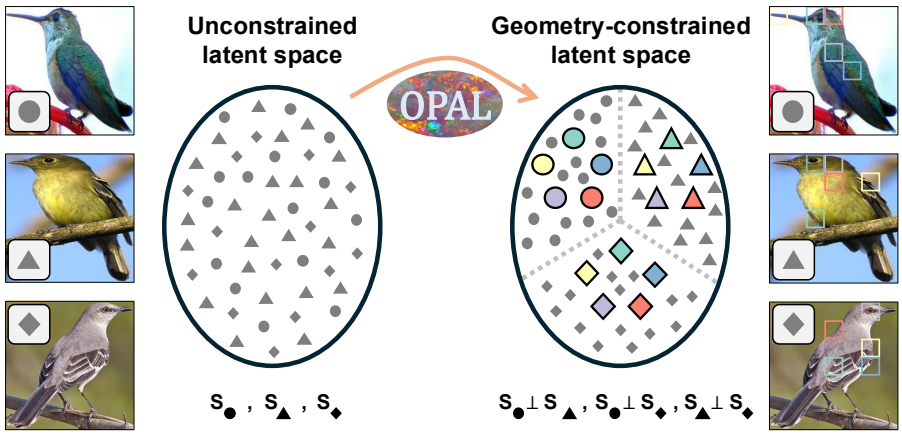

Fig. 1: Conceptual overview of OPAL. Prior inherently interpretable approaches typically learn prototypes in an (*unconstrained latent space*), where class evidence may overlap in a shared representation. OPAL instead imposes a (*geometry-constrained latent space*) by assigning each class c to a dedicated orthonormal subspace $\mathcal{S}_c \subset \mathbb{R}^K$, shown as $\mathcal{S}_\circ, \mathcal{S}_\triangle, \mathcal{S}_\diamond$ in the figure. This structure encourages different prototype coordinates within $\mathcal{S}_c$ to capture distinct semantic attributes and supports faithful part-based evidence, visualized as the image regions selected by the corresponding prototype activations.

In this work, we introduce a geometry-driven formulation for interpretable prototype-based classification that replaces unconstrained prototype discovery with structured subspace alignment. Leveraging the principle that fixed, maximally separable geometric constraints induce robust and transferable feature representations [6, 10, 21, 29], we introduce Orthonormal Prototype Alignment Learning (OPAL). As illustrated in Fig. 1, OPAL anchors each class to a preassigned orthonormal subspace and represents interpretability as alignment within this fixed geometry, rather than as the optimization of free prototype locations. This design decouples prototype geometry from representation learning and turns classification into a direct alignment task in a stable, theoretically grounded coordinate system.

To translate this geometric alignment into granular visual explanations, OPAL incorporates an intrinsic spatial competition mechanism. Unlike traditional approaches that rely on soft attention maps or auxiliary concentration losses to localize parts [16, 45], our method enforces exclusiveness through channel-wise normalization followed by global feature selection. By compelling feature maps to compete for activation at each spatial location, OPAL prevents semantic concepts from overlapping. Consequently, a specific image region cannot simultaneously represent multiple distinct prototypes. This "winner-takes-all" dynamic, combined with the structural stability of the orthonormal basis, naturally drives

the network to discover sparse and highly discriminative object parts without requiring part-level annotations or complex bounding-box supervision [27, 46].

The main contributions of this work are summarized as follows:

- We introduce OPAL, a geometry-constrained approach to interpretable prototype based classification that moves from unconstrained prototype discovery to structured alignment in a predefined orthonormal space. This design enables a single-stage, end-to-end training pipeline without warm-up phases or alternating optimization steps.
- We identify that imposing strict orthogonality on the output space, combined with an explicit spatial competition mechanism, is sufficient to induce semantic sparsity, eliminating the need for the auxiliary regularization terms (e.g., separation, clustering, or diversity losses) required by prior art to prevent prototype collapse. This results in a minimalist architecture that naturally isolates discriminative image regions by projecting them onto exclusive, preassigned semantic axes.
- We conduct an extensive evaluation across five fine-grained benchmarks, where OPAL attains leading performance among inherently interpretable models on most datasets. Moreover, OPAL consistently outperforms the corresponding non-interpretable CNN backbones across diverse architectures and model capacities. Qualitative results further indicate that prototype slots align with distinct semantic parts, yielding granular and conceptually meaningful visual explanations.

2 Related work

2.1 Intepretable Image Classification

The domain of interpretable image classification builds on the Prototypical Part Network (ProtoPNet) [3], which introduced the case-based reasoning paradigm. However, its original formulation imposed a heavy optimization burden, relying on a multi-stage training pipeline and auxiliary regularization terms to align latent prototypes with training patches. Subsequent works aimed to mitigate prototype redundancy and improve reasoning efficiency: ProtoTree [25] structured prototypes hierarchically, ProtoPShare [34] and ProtoPool [33] introduced mechanisms for prototype merging and differentiable assignment, and TesNet [43] represented each class through dedicated subspaces of a transparent embedding space. Despite their efficacy in compacting the latent space, these methods retain, and in cases like ProtoPool even increase, the complexity of the training objective by requiring extensive hyperparameter tuning for multiple competing auxiliary losses.

More recent state-of-the-art approaches prioritize the semantic grounding of prototypes to ensure they align with human-interpretable features. PIP-Net [24] improves upon the visual coherence of explanations by framing classification as a sparse scoring task. By leveraging a self-supervised part-discovery objective, it ensures that prototypes correspond to consistent semantic patches, effectively

identifying specific object parts rather than abstract latent vectors. However, achieving this alignment requires a decoupled two-stage training schedule and two auxiliary loss functions. MCPNet [41] captures various semantic granularities through a multi-level hierarchy. However, its reliance on abstract concept prototypes rather than distinct prototypical parts compromises the interpretability of high-level attributes, making it difficult to differentiate the specific localized regions the model is attending to. Most recently, LucidPPN [28] seeks to disentangle features such as color and shape, yet it crucially depends on an external part-discovery model, PDiscoNet [39], for initial prototype alignment. This introduces an additional layer of auxiliary pre-training and parameter tuning. A separate line pursues prototype-based interpretability post-hoc, on top of frozen classifiers. Tan *et al.* [36] decompose a trained classification head into part-prototypes using non-negative matrix factorization, while EPIC [2] disentangles the final-layer channels with a learnable invertible transform, both producing prototype explanations without altering the original predictions. By operating on a frozen backbone, however, these methods explain a representation they cannot reshape, and the quality of their prototypes remains bounded by the features the model has already learned. In contrast, OPAL is interpretable by design while remaining single-stage. It removes the structural and procedural dependencies of prior ante-hoc methods and, unlike post-hoc approaches, jointly learns the representation and its explanation within a fixed, maximally separable class geometry. This matches the end-to-end simplicity of a standard CNN while providing robust, semantically meaningful interpretability.

2.2 Representation Learning and Metric Alignment

A complementary line of work studies how discriminative representations emerge from geometric constraints on the embedding space [6, 14]. Normalizing features to lie on the unit hypersphere and optimizing angular objectives has been highly effective in practice, particularly through margin-based formulations that operate directly in cosine space [17, 42]. In parallel, contrastive learning provides a geometric view of representation quality through alignment and distributional spread on the hypersphere [44], and supervised contrastive objectives further strengthen class-level clustering in the normalized embedding space [13]. These perspectives motivate OPAL's use of a fixed, structured geometry where classification can be expressed as alignment in a normalized space.

Metric learning methods also commonly replace standard linear classifiers with distance-based objectives defined over class representatives. Proxy-based approaches introduce class proxies to accelerate and stabilize metric learning, optimizing distances to representative points instead of relying on expensive sample mining [14, 22, 38]. Related observations further suggest that fixing the classifier geometry can be surprisingly effective, reducing the reliance on learning dense classification heads [10]. OPAL builds on these insights, but departs from prior proxy formulations by fixing the class geometry a priori through orthonormal subspaces and anchors, and by coupling this structure with competitive

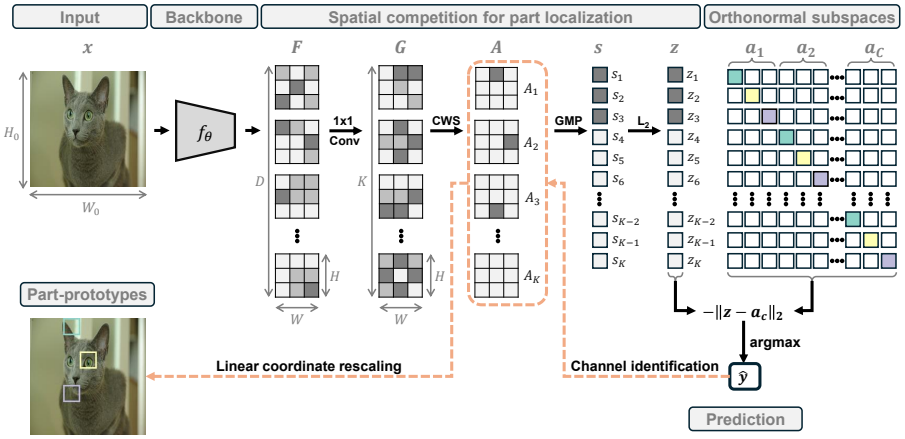

Fig. 2: Overview of the OPAL framework. Given an input image $x \in \mathbb{R}^{3\times H_0\times W_0}$, a convolutional backbone f_θ produces a feature map $F(x) \in \mathbb{R}^{D\times H\times W}$, which is projected via a 1×1 convolution to $G(x) \in \mathbb{R}^{K\times H\times W}$ with $K = C \cdot m$ channels ($m = 3$ shown as a representative example). OPAL enforces part selection through (*channel-wise softmax*, CWS), followed by (*global max pooling*, GMP) and L_2 normalization to obtain the embedding $z \in \mathbb{R}^K$. Classification is performed by distance-based alignment to fixed class anchors $\{a_c\}_{c=1}^C$ in (*orthonormal subspaces*), yielding the prediction $\hat{y}$. For interpretability, OPAL identifies the channels associated with $\hat{y}$, and maps their spatial maximizers back to the input via (*linear coordinate rescaling*) to produce part-prototype evidence.

spatial evidence aggregation to obtain interpretable, part-based decisions under image-level supervision.

3 Orthonormal Prototype Alignment Learning (OPAL)

An overview of OPAL is provided in Fig. 2. The following subsections present the problem formulation and the key components of the method.

3.1 Preliminaries and Problem Formulation

We address supervised fine-grained image classification under image-level supervision. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ denote a dataset of N labeled images, where each image is $x_i \in \mathbb{R}^{3\times H_0\times W_0}$ with height H_0 and width W_0, and its label is $y_i \in \{1, \ldots, C\}$. The supervision is limited to the global class label. No additional annotations are assumed, including part labels, segmentation masks, or any other region-level information. Our objective is to learn a predictor that is accurate while remaining inherently interpretable by design. In addition to a predicted label $\hat{y}$, the model must provide a transparent account of the evidence supporting its decision in terms of localized image regions.

We build on a convolutional feature extractor f_θ that maps an input image $x \in \mathbb{R}^{3\times H_0\times W_0}$ to a final convolutional feature map $F = f_\theta(x) \in \mathbb{R}^{D\times H\times W}$, where D is the number of channels and $H \times W$ is the spatial resolution at the last convolutional stage. OPAL then refines this representation through a lightweight reparameterization and a competitive aggregation mechanism, and uses the resulting features to impose a structured output geometry that enables part-based reasoning under image-level supervision.

3.2 Predefined Orthonormal Subspaces

OPAL fixes the geometry of the output space by assigning each class to a dedicated orthonormal subspace. Let $m \in \mathbb{N}$ denote the number of prototype slots per class and define the structured output dimensionality as $K = C \cdot m$. The coordinate set $\{1, \ldots, K\}$ is arranged into C disjoint blocks $\{B_c\}_{c=1}^{C}$, where $B_c = \{(c-1)m + 1, \ldots, c \cdot m\}$ contains the m coordinates reserved for class c. These blocks induce class-specific subspaces in $\mathbb{R}^K$. Let $\{e_k\}_{k=1}^{K}$ be the canonical basis of $\mathbb{R}^K$. The subspace associated with class c is $\mathcal{S}_c = \mathrm{span}\{e_k : k \in B_c\}$. By construction, $\mathcal{S}_c \perp \mathcal{S}_{c'}$ for any $c \neq c'$, and every vector $z \in \mathbb{R}^K$ admits a unique decomposition into components lying in these mutually orthogonal subspaces.

Within each $\mathcal{S}_c$, we define m fixed part-prototypes for class c, denoted by $\{p_{c,j}\}_{j=1}^{m} \subset \mathbb{R}^K$, which act as canonical prototype directions (one per slot). Concretely, the set $\{p_{c,j}\}_{j=1}^{m}$ is given by $p_{c,j} = e_{(c-1)m+j}$. These canonical directions are unit-norm and mutually orthogonal. Therefore, $\{p_{c,j}\}$ forms an orthonormal family, yielding a maximally separated coordinate system for part-based evidence. For class-level decisions, each block is represented by a single class anchor $a_c \in \mathbb{R}^K$, defined as the normalized block-uniform vector

$$a_c \;=\; \frac{1}{\sqrt{m}} \sum_{j=1}^{m} p_{c,j}. \tag{1}$$

This choice satisfies $\|a_c\|_2 = 1$ and $\langle a_c, a_{c'} \rangle = 0$ for $c \neq c'$, anchoring each class to a unit vector and enforcing maximal angular separation between class representatives. The block-uniform anchor also introduces an inductive bias that favors distributing evidence across the m slots of the correct class block rather than concentrating it on a single coordinate. A formal justification is provided in Sec. A.1 of the supplementary material.

3.3 Spatial Competition for Part Localization

OPAL promotes part-based evidence by enforcing explicit competition across channels at the last convolutional stage. Starting from the backbone feature map $F \in \mathbb{R}^{D\times H\times W}$, a 1×1 convolutional projection maps D channels to $K = C \cdot m$ channels, aligned with the K prototype coordinates introduced in the previous subsection. This produces a projected activation tensor $G \in \mathbb{R}^{K\times H\times W}$. The spatial grid is indexed by $\Omega = \{1, \ldots, H\} \times \{1, \ldots, W\}$, and $G_k(u, v)$ denotes

the activation of channel k at location $(u, v) \in \Omega$. Competition is enforced by applying a channel-wise softmax (CWS) over the K channels independently at every spatial location (u, v). This yields normalized maps $A \in [0, 1]^{K \times H \times W}$ defined as:

$$A_k(u, v) = \frac{\exp(G_k(u, v))}{\sum_{j=1}^{K} \exp(G_j(u, v))}, \qquad (u, v) \in \Omega. \tag{2}$$

By construction, $\sum_{k=1}^{K} A_k(u, v) = 1$ for every (u, v), coupling all channels through direct competition. As a result, a single spatial location cannot simultaneously support multiple channels, which encourages different channels to specialize in distinct discriminative regions.

A global representation is obtained via global max pooling (GMP) per channel. For each channel k, OPAL aggregates spatial evidence and records the corresponding maximizer according to:

$$s_k = \max_{(u,v)\in\Omega} A_k(u, v), \qquad (u_k^*, v_k^*) = \arg\max_{(u,v)\in\Omega} A_k(u, v). \tag{3}$$

Stacking $\{s_k\}_{k=1}^{K}$ yields $s \in \mathbb{R}^K$, and the final embedding is obtained by L_2 normalization, $z = s/\|s\|_2 \in \mathbb{R}^K$. This design directly ties representation learning to localized evidence. Each coordinate z_k is sourced from a single spatial location (u_k^*, v_k^*) selected by max pooling, and the corresponding image patch constitutes the exact evidence that determines the channel contribution used by the classifier.

3.4 Single-stage Optimization and Inference

OPAL is trained end-to-end in a single stage using a standard discriminative objective. The forward pass produces the normalized embedding $z \in \mathbb{R}^K$ described in Sec. 3.3. Classification is reformulated as metric alignment in the fixed output geometry defined by the orthonormal class anchors $\{a_c\}_{c=1}^{C}$ introduced in Sec. 3.2. Rather than learning a dense classifier, OPAL directly uses distances to these anchors as predicted class-logit values. In particular, the logit for class c and the prediction are:

$$\ell_c = -\|z - a_c\|_2, \qquad \hat{y} = \arg\max_{c\in\{1,\dots,C\}} \ell_c. \tag{4}$$

Since both z and a_c are L_2-normalized, Euclidean distances depend only on the inter-vector angle. We analyze $\|z - a_c\|_2^2$ since it induces the same ordering as $\|z - a_c\|_2$ on $\mathbb{R}_{\geq 0}$, and obtain $\|z - a_c\|_2^2 = 2 - 2\langle z, a_c\rangle$. Thus, distance-based alignment is equivalent to maximizing cosine similarity on the unit hypersphere, with class separation enforced by anchor orthogonality. Parameter update is guided by the minimization of the cross-entropy (CE) with respect to distance-based logits, mathematically formulated as:

$$\mathcal{L}_{\mathrm{CE}} = -\log \frac{\exp(\ell_y)}{\sum_{c=1}^{C} \exp(\ell_c)}, \tag{5}$$

where ℓ_y denotes the logit of the ground-truth class y. Gradients are backpropagated through the full model, while the anchors remain fixed. This leads to a single-stage pipeline without auxiliary clustering, projection steps, or additional regularization losses.

At inference time, OPAL produces visual explanations as a direct byproduct of the forward pass. The 1×1 projection yields $K = C \cdot m$ output channels that are in one-to-one correspondence with the fixed prototype coordinates defined in Sec. 3.2. Since each class anchor a_c has support only on its block B_c, a prediction $\hat{y}$ naturally selects the subset of channels $k \in B_{\hat{y}}$ as the prototype slots used to form class evidence. For each selected channel k, global max pooling identifies a unique maximizer $(u_k^*, v_k^*) \in \Omega$ that determines the pooled response s_k, ensuring that the contribution of z_k is sourced from a single spatial location. Finally, the maximizer coordinates are mapped back to the input resolution by linearly rescaling from the feature-map grid $H \times W$ to $H_0 \times W_0$, yielding the localized evidence patches that constitute the explanation.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate OPAL on five standard fine-grained benchmarks: CUB-200-2011 (200 bird species) [40], Stanford Cars (196 car models) [15], Oxford-IIIT Pet (37 breeds of cats and dogs) [30], Stanford Dogs (120 dog breeds) [12], and Oxford Flowers-102 (102 flower categories) [26]. We follow the official train/test splits for all datasets. Additional dataset details and preprocessing steps are provided in Sec. B.1 of the supplementary material.

Implementation Details. We evaluate OPAL across three convolutional backbone families at three model scales each: ResNet-50/101/152 [9], EfficientNet-V2 S/M/L [37], and ConvNeXt Tiny/Small/Base [19]. Images are resized to 224×224 and augmented with TrivialAugment [23], as in [24]. We fix the number of prototypes per class to $m = 5$ across all datasets and backbones, with sensitivity to m analyzed in Sec. 4.4. Across all settings, models are trained for 30 epochs using Adam with a learning rate of 1×10^{-4} and a batch size of 64. All experiments were run on an NVIDIA A100 (40GB) GPU and repeated with three different random seeds, reporting results averaged across runs. In addition to mean performance, we report the absolute percentage-point difference relative to the corresponding CNN backbone trained under the same protocol. Additional implementation details are provided in Sec. B.2 of the supplementary material.

4.2 Quantitative Evaluation

Comparison with Inherently Interpretable Models. Tab. 2 reports top-1 accuracy on five fine-grained benchmarks and compares OPAL against eight

Table 2: Top-1 accuracy (%) comparison on five fine-grained benchmarks between OPAL and representative inherently interpretable models. ↑ and ↓ denote the absolute percentage point difference with respect to the ConvNeXt-Tiny baseline. ProtoTree exhibits low performance on DOGS and FLOWER, a behavior also observed on these datasets in [28] and reported on another fine-grained benchmark in [41].

Method	Accuracy (↑)				
	CUB	**CARS**	**PETS**	**DOGS**	**FLOWER**
Baseline (ConvNext-Tiny)	84.2	87.8	91.0	85.3	95.6
ProtoPNet NeurIPS '19 [3]	79.2↓5.0	86.1↓1.7	80.9↓10.1	77.4↓7.9	92.1↓3.5
ProtoTree CVPR '21 [25]	82.2↓2.0	86.6↓1.2	75.0↓16.0	58.9↓26.4	17.5↓78.1
ProtoPShare KDD '21 [34]	74.7↓9.5	86.4↓1.4	81.2↓9.8	74.1↓11.2	90.3↓5.3
TesNet ICCV '21 [43]	84.6↑0.4	**92.6**↑4.8	88.7↓2.3	80.7↓4.6	83.4↓12.2
ProtoPool ECCV '22 [33]	85.5↑1.3	88.9↑1.1	80.8↓10.2	71.7↓13.6	92.7↓2.9
PIP-Net CVPR '23 [24]	84.3↑0.1	88.2↑0.4	92.0↑1.0	80.8↓4.5	91.8↓3.8
MCPNet CVPR '24 [41]	83.5↓0.7	83.2↓4.6	91.3↑0.3	76.5↓8.8	94.7↓0.9
LucidPPN ICLR '25 [28]	81.5↓2.7	91.6 ↑3.8	92.1↑1.1	79.5↓5.8	95.0↓0.6
OPAL *(Ours)*	**85.6**↑1.4	90.3↑2.5	**92.8**↑1.8	**86.9**↑1.6	**96.1**↑0.5

interpretable methods, together with a non-interpretable counterpart. We use ConvNeXt-Tiny as the baseline backbone, consistent with recent prototype-based approaches that report results on this architecture [24, 28, 41], enabling direct backbone-aligned comparisons. Overall, this setting provides a broad side-by-side evaluation across datasets and representative interpretable baselines.

OPAL achieves the best performance on 4 out of 5 datasets and ranks third on CARS, behind TesNet [43] and LucidPPN [28]. Notably, OPAL is the only method that consistently surpasses the corresponding CNN backbone across all benchmarks, with an average improvement of +1.56 absolute percentage points over the baseline. These results suggest that imposing a fixed, highly discriminative latent space enables a single-stage training pipeline without auxiliary losses, thereby improving not only interpretability but also performance.

Backbone Generalization and Scalability. Tab. 3 evaluates OPAL as a plug-in module across three convolutional backbone families and three model scales per family, always comparing against the corresponding CNN backbone trained under the same protocol. This setting isolates the effect of OPAL from backbone-specific design choices and tests whether the approach transfers across architectural families and capacities.

Overall, OPAL improves or matches the non-interpretable counterpart in 39 out of 45 backbone–dataset experiments, with an average gain of +2.94 absolute percentage points in top-1 accuracy across all datasets and backbones. In the few cases where OPAL does not outperform the baseline, the largest drop is 1.0 percentage points, indicating that the method is broadly robust and does not trade performance for interpretability when integrated into diverse convolutional architectures.

Table 3: Backbone generalization and scalability of OPAL across three convolutional backbone families and three model scales per family. Each backbone is evaluated in its standard form and with OPAL integrated (shaded rows). ↑ and ↓ denote the absolute percentage-point difference with respect to the corresponding CNN backbone.

Backbone	Accuracy (↑)				
	CUB	CARS	PETS	DOGS	FLOWER
ResNet-50	72.6	71.9	89.5	81.3	92.8
w/ OPAL	80.8↑8.2	84.3↑12.4	89.8↑0.3	81.0↓0.3	92.7↓0.1
ResNet-101	74.9	73.8	88.5	83.5	91.6
w/ OPAL	81.4↑6.5	83.9↑10.1	91.0↑2.5	83.0↓0.5	93.0↑1.4
ResNet-152	76.0	73.7	90.3	83.1	92.4
w/ OPAL	80.7↑4.7	85.1↑11.4	90.1↓0.2	84.1↑1.0	93.8↑1.4
EfficientNet-V2 S	78.3	76.7	89.6	84.5	95.3
w/ OPAL	83.5↑5.2	87.2↑10.5	90.4↑0.8	85.8↑1.3	95.3- 0.0
EfficientNet-V2 M	79.4	80.3	89.0	83.3	93.9
w/ OPAL	82.6↑3.2	82.7↑2.4	90.6↑1.6	85.7↑2.4	92.9↓1.0
EfficientNet-V2 L	84.1	85.9	90.0	81.4	94.9
w/ OPAL	86.9↑2.8	89.4↑3.5	90.0- 0.0	83.1↑1.7	96.6↑1.7
ConvNeXt-Tiny	84.2	87.8	91.0	85.3	95.6
w/ OPAL	85.6↑1.4	90.3↑2.5	92.8↑1.8	86.9↑1.6	96.1↑0.5
ConvNeXt-Small	81.7	83.9	91.8	86.5	95.3
w/ OPAL	85.6↑3.9	90.6↑6.7	93.4↑1.6	87.7↑1.2	95.5↑0.2
ConvNeXt-Base	82.1	82.1	92.0	88.9	97.2
w/ OPAL	86.3↑4.2	91.4↑9.3	93.9↑1.9	90.2↑1.3	96.6↓0.6

4.3 Qualitative Interpretability Analysis

Interpretable predictions are essential to understand which visual evidence drives a model decision. Fig. 3 illustrates OPAL explanations across all five datasets. Across benchmarks, OPAL yields non-collapsed prototypes that activate on multiple, distinct regions of the object within a single image. In each example, the $m = 5$ prototypes provide a compact set of complementary, localized evidence spanning different discriminative parts of the object. This is especially important in fine-grained recognition, where class separation is often driven by subtle part-level attributes rather than a single salient region.

Beyond diversity, the retrieved evidence is also consistent within class across images, suggesting that individual prototypes acquire stable semantic roles. The same prototype repeatedly activates on visually and semantically related parts for different instances, indicating that the fixed class-wise orthonormal structure encourages different prototype directions to align with recurring attributes (e.g., ears for Abyssinian cats in PETS or eyes for Maltese dogs in DOGS). Further qualitative results are provided in the supplementary material, including additional prototype-based explanations (Sec. C.1), the qualitative effect of the number of prototypes per class (Sec. C.2), and an analysis of prototype collapse (Sec. D.1). As a complementary quantitative assessment, and since part-level an-

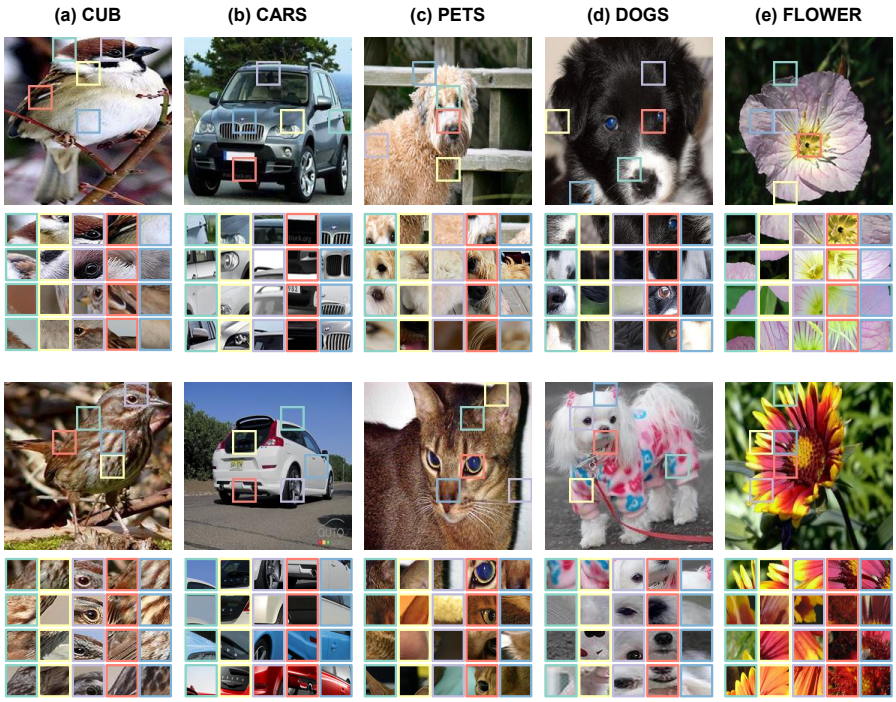

Fig. 3: Qualitative prototype-based explanations produced by OPAL on five fine-grained benchmarks ($m = 5$). For each dataset, we show two test images and the corresponding prototype-wise evidence. Colored boxes indicate the spatial maximizers selected for each prototype. The patch grids visualize the evidence associated with each prototype, with the first row extracted from the shown image and subsequent rows taken from other images of the same predicted class.

notations are available only for CUB-200-2011, we report prototype consistency and stability on this dataset in Sec. E.3 of the supplementary material, where OPAL attains the second-best results among prototype-based methods.

4.4 Ablation Studies

Architectural Component Analysis. Tab. 4 analyzes the contribution of OPAL components by selectively enabling the projection, competitive normalization, pooling, and orthonormal anchoring mechanisms. Since the orthonormal anchors are defined in a $K = C \cdot m$ space, they require the 1×1 projection to match the channel dimensionality, so configurations with Orth. necessarily include Proj.

Two observations are consistent across datasets. First, the orthonormal anchors play a central role: enabling Orth. yields a clear and consistent improvement over the corresponding non-orthogonal variants (+4.07 percentage points

Table 4: Architectural ablation of OPAL components on five fine-grained benchmarks, reported as top-1 accuracy (%). Columns correspond to the 1×1 convolutional projection (Proj.), global max pooling (GMP), channel-wise softmax (CWS), and orthonormal anchors (Orth.). ✓ and ✗ indicate whether a component is enabled or not. When GMP is disabled, global average pooling is used. The all-crosses configuration (✗✗✗✗) corresponds to the convolutional baseline, while the all-checks configuration (✓✓✓✓) corresponds to full OPAL. ↑ and ↓ denote absolute percentage-point differences with respect to the baseline.

Architectural Components				Accuracy (↑)				
Proj.	GMP	CWS	Orth.	CUB	CARS	PETS	DOGS	FLOWER
✗	✗	✗	✗	84.2	87.8	91.0	85.3	95.6
✓	✗	✗	✗	82.3↓1.9	86.3↓1.5	92.0↑1.0	86.2↑0.9	96.3↑0.7
✗	✓	✗	✗	79.6↓4.6	75.7↓12.1	88.9↓2.1	84.1↓1.2	93.4↓2.2
✗	✗	✓	✗	57.4↓26.8	48.9↓38.9	88.1↓2.9	74.6↓10.7	91.5↓4.1
✓	✓	✗	✗	77.1↓7.1	73.5↓14.3	91.6↑0.6	86.5↑1.2	94.2↓1.4
✗	✓	✓	✗	54.6↓29.6	31.9↓55.9	86.2↓4.8	80.2↓5.1	86.3↓9.3
✓	✗	✓	✗	73.7↓10.5	70.5↓17.3	91.1↑0.1	84.7↓0.6	95.0↓0.6
✓	✗	✗	✓	85.0↑0.8	90.6↑2.8	92.6↑1.6	85.8↑0.5	96.3↑0.7
✓	✓	✗	✓	82.0↓2.2	85.5↓2.3	93.2↑2.2	87.2↑1.9	94.0↓1.6
✓	✗	✓	✓	84.0↓0.2	90.4↑2.6	93.4↑2.4	86.8↑1.5	95.2↓0.4
✓	✓	✓	✓	85.6↑1.4	90.3↑2.5	92.8↑1.8	86.9↑1.6	96.1↑0.5

in top-1 accuracy on average, across matched settings). Second, pooling-based localization alone is not sufficient. Configurations that rely on GMP without Orth. (rows 3/5/6) incur a substantial average drop (-9.86 percentage points in top-1 accuracy) relative to the baseline, despite producing localized evidence. Consequently, the ablation indicates that OPAL benefits arise from coupling competitive, localized aggregation with the fixed class-structured geometry induced by orthonormal anchoring.

Although certain ablated variants outperform full OPAL on isolated datasets (e.g., Proj+GMP+Orth without CWS in PETS/DOGS, or Proj+CWS+Orth without GMP in CARS/PETS), these configurations compromise key properties that OPAL is designed to provide. In particular, removing GMP breaks the one-maximizer-per-channel mechanism, weakening the direct traceability between each prototype activation and a unique localized image region, which is central to faithful, part-based explanations. Conversely, removing CWS eliminates explicit spatial competition, which empirically increases prototype redundancy and makes it harder to obtain diverse prototypes with distinct semantic roles. Both quantitative and qualitative evidence of the resulting prototype collapse is provided in Sec. D.1 of the supplementary material. Overall, these findings indicate that strong performance together with reliable prototype diversity and localized interpretability requires combining orthonormal anchoring with spatial competition and max-based aggregation.

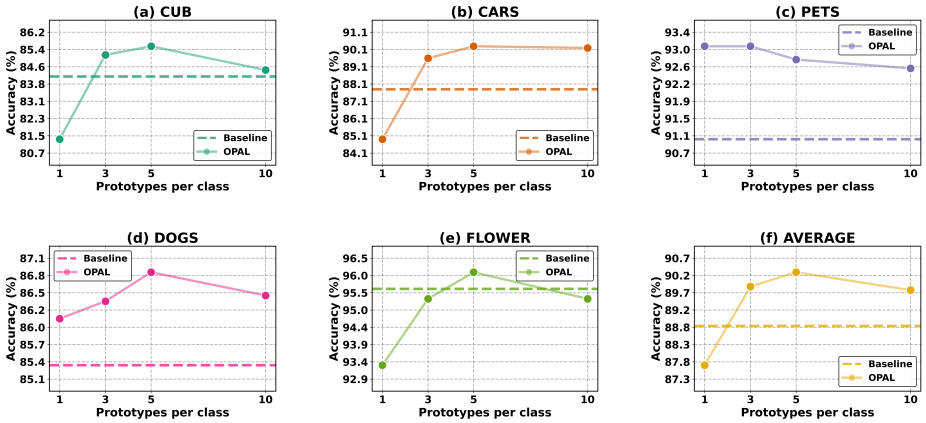

Fig. 4: Sensitivity to the number of prototypes per class m. OPAL is evaluated with $m \in \{1, 3, 5, 10\}$ on (a) CUB, (b) CARS, (c) PETS, (d) DOGS, and (e) FLOWER, with (f) reporting the average across datasets. The dashed line denotes the ConvNeXt-Tiny baseline, and points correspond to OPAL under the same training protocol.

Sensitivity to the Number of Prototypes We analyze the sensitivity of OPAL to the number of prototypes per class m by varying $m \in \{1, 3, 5, 10\}$ while keeping the remaining training protocol fixed. Fig. 4 reports results across datasets and their average. Using a single prototype per class ($m = 1$) is insufficient and underperforms the ConvNeXt-Tiny baseline by -1.04 percentage points on average. In contrast, selecting $m \in \{3, 5, 10\}$ yields consistent gains, with an average improvement of $+1.25$ percentage points over the baseline when averaging the three configurations.

A clear trend emerges in most benchmarks: allocating too few prototypes limits representational capacity, while an excessive number of prototypes can dilute discriminative evidence, leading to slightly lower accuracy than the best setting. These results indicate that a moderate number of prototypes per class ($m = 5$ in our main experiments) provides a robust operating point, enabling OPAL to deliver interpretable predictions while improving performance over the non-interpretable counterpart. Additional qualitative results for $m \in \{1, 3, 5, 10\}$ are presented in Sec. C.2 of the supplementary material.

5 Conclusion

In this work, we introduced OPAL, an inherently interpretable framework that shifts prototype-based classification from learning latent prototypes to aligning representations within a fixed, class-structured geometry. OPAL anchors each class in a predefined orthonormal subspace and couples this structure with competitive feature aggregation, yielding localized evidence directly tied to the model computation. This design enables a single-stage, end-to-end training pipeline

with a standard classification objective, while preserving interpretable, part-based explanations without auxiliary regularizers or multi-stage optimization schedules.

OPAL also has limitations that motivate future research directions. First, the current formulation leverages the inductive bias of convolutional backbones, where locality and hierarchical feature composition naturally ground prototype evidence in the image plane. Extending OPAL to vision transformers will require designing token-level competition and aggregation mechanisms that preserve the benefits of fixed, discriminative latent geometries while remaining competitive and interpretable. Second, OPAL currently uses a fixed number of prototypes per class and does not share prototypes across classes, which can be restrictive when class complexity varies or when attributes recur across categories. Future work will explore adaptive prototype allocation and controlled prototype sharing strategies to better reflect class complexity and cross-class regularities while preserving interpretability and performance guarantees.

Acknowledgements

This work has been supported by the Generalitat Valenciana (GVA) through the project CIPROM/2022/20 (PROMETEO). The contribution of GJ. Angulo was supported by his AI Chair PR[AI]RIE–PSAI (ANR-23-IACL-0008, France 2030).

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. Advances in neural information processing systems **31** (2018)
2. Borycki, P., Trędowicz, M., Janusz, S., Tabor, J., Spurek, P., Lewicki, A., Struski, Ł.: Epic: Explanation of pretrained image classification networks via prototypes. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 17366–17373 (2026)
3. Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J.K.: This looks like that: deep learning for interpretable image recognition. Advances in neural information processing systems **32** (2019)
4. Choi, H., Jin, S., Han, K.: Icev2: Interpretability, comprehensiveness, and explainability in vision transformer: H. choi et al. International Journal of Computer Vision **133**(5), 2487–2504 (2025)
5. Codevilla, F., Müller, M., López, A., Koltun, V., Dosovitskiy, A.: End-to-end driving via conditional imitation learning. In: 2018 IEEE international conference on robotics and automation (ICRA). pp. 4693–4700. IEEE (2018)
6. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

8. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. nature **542**(7639), 115–118 (2017)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
10. Hoffer, E., Hubara, I., Soudry, D.: Fix your classifier: the marginal value of training the last weight layer. arXiv preprint arXiv:1801.04540 (2018)
11. Kazmierczak, R., Berthier, E., Frehse, G., Franchi, G.: Explainability and vision foundation models: A survey. Information Fusion **122**, 103184 (2025)
12. Khosla, A., Jayadevaprakash, N., Yao, B., Li, F.F.: Novel dataset for fine-grained image categorization: Stanford dogs. In: Proc. CVPR workshop on fine-grained visual categorization (FGVC). vol. 2 (2011)
13. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in neural information processing systems **33**, 18661–18673 (2020)
14. Kim, S., Kim, D., Cho, M., Kwak, S.: Proxy anchor loss for deep metric learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3238–3247 (2020)
15. 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)
16. Li, Z., Bao, W., Zheng, J., Xu, C.: Deep grouping model for unified perceptual parsing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4053–4063 (2020)
17. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 212–220 (2017)
18. 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. pp. 10012–10022 (2021)
19. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)
20. Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Del Ser, J., Guidotti, R., Hayashi, Y., Herrera, F., Holzinger, A., et al.: Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions. Information Fusion **106**, 102301 (2024)
21. Mettes, P., Van der Pol, E., Snoek, C.: Hyperspherical prototype networks. Advances in neural information processing systems **32** (2019)
22. Movshovitz-Attias, Y., Toshev, A., Leung, T.K., Ioffe, S., Singh, S.: No fuss distance metric learning using proxies. In: Proceedings of the IEEE international conference on computer vision. pp. 360–368 (2017)
23. Müller, S.G., Hutter, F.: Trivialaugment: Tuning-free yet state-of-the-art data augmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 774–782 (2021)
24. Nauta, M., Schlötterer, J., Van Keulen, M., Seifert, C.: Pip-net: Patch-based intuitive prototypes for interpretable image classification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2744–2753 (2023)

25. 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. pp. 14933–14943 (2021)
26. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: 2008 Sixth Indian conference on computer vision, graphics & image processing. pp. 722–729. IEEE (2008)
27. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Is object localization for free?-weakly-supervised learning with convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 685–694 (2015)
28. Pach, M., Lewandowska, K., Tabor, J., Zieliński, B.M., Rymarczyk, D.D.: LucidPPN: Unambiguous prototypical parts network for user-centric interpretable computer vision. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=BM9qfolt6p
29. Papyan, V., Han, X., Donoho, D.L.: Prevalence of neural collapse during the terminal phase of deep learning training. Proceedings of the National Academy of Sciences **117**(40), 24652–24663 (2020)
30. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3498–3505. IEEE (2012)
31. Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)
32. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence **1**(5), 206–215 (2019)
33. Rymarczyk, D., Struski, Ł., Górszczak, M., Lewandowska, K., Tabor, J., Zieliński, B.: Interpretable image classification with differentiable prototypes assignment. In: European Conference on Computer Vision. pp. 351–368. Springer (2022)
34. Rymarczyk, D., Struski, Ł., Tabor, J., Zieliński, B.: Protopshare: Prototypical parts sharing for similarity discovery in interpretable image classification. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. pp. 1420–1430 (2021)
35. 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. pp. 618–626 (2017)
36. Tan, A., Zhou, F., Chen, H.: Post-hoc part-prototype networks. Proceedings of the 41st International Conference on Machine Learning (2024)
37. Tan, M., Le, Q.: Efficientnetv2: Smaller models and faster training. In: International conference on machine learning. pp. 10096–10106. PMLR (2021)
38. Teh, E.W., DeVries, T., Taylor, G.W.: Proxynca++: Revisiting and revitalizing proxy neighborhood component analysis. In: European conference on computer vision. pp. 448–464. Springer (2020)
39. Van Der Klis, R., Alaniz, S., Mancini, M., Dantas, C.F., Ienco, D., Akata, Z., Marcos, D.: Pdisconet: Semantically consistent part discovery for fine-grained recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1866–1876 (2023)
40. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S., et al.: The caltech-ucsd birds-200-2011 dataset. Tech. rep., Technical Report CNS-TR-2011-001, California Institute of Technology (2011)

41. Wang, B.S., Wang, C.Y., Chiu, W.C.: Mcpnet: An interpretable classifier via multi-level concept prototypes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10885–10894 (2024)
42. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5265–5274 (2018)
43. 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. pp. 895–904 (2021)
44. Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: International conference on machine learning. pp. 9929–9939. PMLR (2020)
45. Zheng, H., Fu, J., Mei, T., Luo, J.: Learning multi-attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE international conference on computer vision. pp. 5209–5217 (2017)
46. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)