This Looks Distinctly Like That: Grounding Interpretable Recognition in Stiefel Geometry against Neural Collapse

Junhao Jia[1,3†,‡], Jiaqi Wang[2†], Yunyou Liu[3], Haodong Jing[4], Yueyi Wu[3], Xian Wu[2(✉)], and Yefeng Zheng[1(✉)]

[1] Medical Artificial Intelligence Lab, Westlake University, Hangzhou, China
[2] Tencent Jarvis Lab, Shenzhen, China
[3] Hangzhou Dianzi University, Hangzhou, China
[4] Xi'an Jiaotong University, Xi'an, China
junhaojia530@gmail.com, kevinxwu@tencent.com, zhengyefeng@westlake.edu.cn

Abstract. Prototype networks provide an intrinsic case-based explanation mechanism, but their interpretability is often undermined by prototype collapse, where multiple prototypes degenerate to highly redundant evidence. We attribute this failure mode to the terminal dynamics of Neural Collapse, where cross-entropy optimization suppresses intra-class variance and drives class-conditional features toward a low-dimensional limit. To mitigate this, we propose Adaptive Manifold Prototypes (AMP), a framework that leverages Riemannian optimization on the Stiefel manifold to represent class prototypes as orthonormal bases and make rank-one prototype collapse infeasible by construction. AMP further learns class-specific effective rank via a proximal gradient update on a non-negative capacity vector, and introduces spatial regularizers that reduce rotational ambiguity and encourage localized, non-overlapping part evidence. Extensive experiments on fine-grained benchmarks demonstrate that AMP achieves state-of-the-art classification accuracy while significantly improving causal faithfulness over prior interpretable models.

Keywords: Explainable AI · Stiefel Manifold · Neural Collapse

1 Introduction

Deep visual recognition systems are increasingly deployed in high-stakes settings where authoritative decisions demand intrinsic transparency and verifiable reasoning [57]. This need is especially salient in medical and remote-sensing scenarios, where robust diagnosis and target detection must remain trustworthy under domain-specific variation [17, 63]. Within expert domains, this cognitive process hinges on structural deconstruction: practitioners verify a concept by explicitly isolating distinct anatomical parts and evaluating their resemblance to

† Equal contribution; (✉) Corresponding author. ‡ J. Jia contributed to this work as a visiting student at Westlake University.

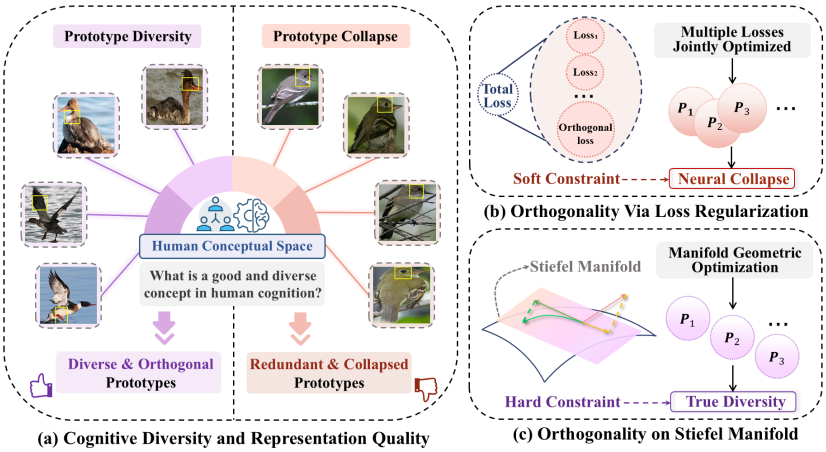

Fig. 1: Motivation of Adaptive Manifold Prototypes (AMP). (a) Contrast between diverse and redundant prototypes. (b) Loss regularizations yield soft constraints. (c) AMP imposes a geometric constraint on the Stiefel manifold, ensuring part diversity.

established archetypal evidence [7]. To emulate this human-like reasoning, prototype networks [10] operationalize this intuitive framework by learning a diverse set of representative visual exemplars, which are subsequently matched against local spatial features extracted from the input image. Theoretically, this architecture builds a robust semantic bridge between abstract convolutional representations and human-understandable spatial evidence, decisively transforming opaque logit aggregations into a visually verifiable similarity-matching paradigm.

Despite their conceptual promise, existing prototype architectures systematically suffer from severe functional degradation in fine-grained visual recognition tasks, where categorical distinctions rely on subtle morphological cues [24]. As shown in Fig. 1(a), empirical observations reveal a pervasive homogenization phenomenon known as prototype collapse [53]. Ideally, a model should exhibit representational diversity, explicitly capturing a comprehensive array of distinct anatomical components [81]. However, these freely optimized vectors frequently lose their intended structural diversity, collapsing to focus on the exact same highly discriminative spatial region. This yields redundant prototypes that entirely ignore the broader physical context. Ultimately, this localized overfitting structurally undermines the foundational premise of compositional interpretability, creating a gap between the model's representations and human cognition [28].

We argue that pervasive prototype homogenization is not merely an architectural flaw, but a geometric inevitability. It stems directly from the algorithmic friction between interpretable representation learning and standard cross-entropy optimization. Recent theoretical analyses of deep neural networks formalize this dynamic as Neural Collapse [51, 77]. During the terminal phase of training, the optimization landscape actively penalizes intra-class variance to maximize the inter-class decision margin. Driven by this overwhelming dis-

criminative pressure, spatial features of a given category are aggressively compressed, converging toward a singular, highly symmetric mean vector in the latent space [86]. This exposes a profound structural paradox: while compositional reasoning relies on internal feature diversity to isolate distinct physical parts, the overarching classification objective relentlessly destroys this variance, collapsing the learned manifold into a zero-dimensional terminal state [9, 48].

To mitigate prototype redundancy, prior works commonly add auxiliary loss terms that softly penalize similarity among Euclidean prototypes (Fig. 1(b)). While such penalties can encourage diversity early in training, they do not provide a feasibility guarantee: under strong cross-entropy gradients and the terminal dynamics associated with Neural Collapse, Euclidean prototypes may still become nearly collinear and concentrate on the same discriminative region.

In this paper, we propose Adaptive Manifold Prototypes (AMP), which replaces soft orthogonality penalties with a hard geometric constraint, as depicted in Fig. 1(c). Specifically, AMP parameterizes class prototypes as an orthonormal basis on the Stiefel manifold, making the rank-one prototype configuration infeasible by construction. In addition, AMP introduces two spatial gauge-fixing objectives (spatial entropy minimization and overlap suppression) whose purpose is not to enforce orthogonality, but to select a semantically stable and localized basis within the rotation-invariant subspace optimized by the classifier. Overall, our contributions can be summarized as follows:

- We theoretically bridge prototype collapse with the terminal dynamics of Neural Collapse, revealing how standard cross-entropy optimization geometrically drives unconstrained prototypes toward low-rank degeneracy.
- We propose Adaptive Manifold Prototypes (AMP), which formulates prototypes as orthonormal bases on the Stiefel manifold and employs dynamic rank calibration alongside spatial regularizers to structurally guarantee diverse, localized part discovery.
- Extensive experiments on fine-grained benchmarks demonstrate that AMP achieves state-of-the-art classification accuracy while setting new standards for causal faithfulness and stability among intrinsically interpretable models.

2 Related Work

2.1 Post-hoc Interpretability Methods

Post-hoc explanation techniques have long served as the standard approach for understanding deep neural networks, attempting to reverse-engineer pre-trained black-box models without architectural modifications. Among these, widely used gradient-based methods like Grad-CAM aim to highlight discriminative spatial regions [59], yet extensive sanity checks reveal that their saliency maps are often weakly tied to learned parameters, behaving more like generic edge detectors than faithful representations of internal reasoning [2, 39]. Benchmark-style evaluations further suggest that popular attribution methods are often no better

than simple baselines at identifying truly important features [30, 79]. Explanations can also be fragile or manipulated by imperceptible perturbations that preserve model outputs [15, 26]. Perturbation-based surrogates like LIME and SHAP [47, 55] remain vulnerable to adversarial scaffolding and rationalization, enabling biased models to output deceptive explanations [3, 62]. Recent multimodal systems further attach generated textual descriptions or semantic cues to predictions, offering auxiliary post-hoc evidence for human inspection [11, 71]. Restoration, inpainting, and quality-assessment modules can provide additional perceptual evidence for degraded visual inputs [70, 72, 73]. However, such auxiliary narratives or maps remain decoupled from the decision mechanism itself. Absent direct regularization of the training dynamics, post-hoc methods cannot stop the network from learning shortcut rules or non-robust features [25, 34]. Nor do they guarantee disentangled and non-redundant internal representations [4, 46]. Training-time acceleration and distillation strategies can shape model behavior, but they likewise do not impose faithful explanatory structure by construction [41, 42, 82]. This structural limitation prompts a necessary shift toward inherently interpretable architectures and concept-based reasoning pipelines [57].

2.2 Prototype-based Interpretability Methods

To address the inherent unfaithfulness of post-hoc techniques, prototype networks explicitly embed a transparent, case-based reasoning process into the architecture [4]. Originating with ProtoPNet [10], these models classify images by matching local patches to learned prototypical vectors. Owing to its intuitive transparency, this case-based paradigm has seen extensive applications beyond standard visual recognition. In remote sensing, related models address hyperspectral super-resolution, pansharpening, thin-cloud removal, and infrared small target detection [17, 19–21]. Medical diagnosis has also benefited from concept-based and prototype-style pipelines [36, 63, 68]. Pathology-aware prototype evolution further shows that LLM-driven semantic disambiguation can refine prototype-based multicenter diabetic retinopathy diagnosis [84]. Parallel medical vision-language post-training studies explore dynamic visual focus, adaptive causal reasoning, and latent-memory evolution for trustworthy progressive diagnosis [45, 83, 85]. NLP and time-series analysis have adopted prototype trajectories, textual rationales, or hierarchical prototypes [29, 54, 74]. Recent text classification and 3D visual decoding further demonstrate the broader value of structured evidence across modalities [37, 38]. Unconstrained prototypes naturally gravitate toward the single most discriminative region, prompting numerous extensions to enforce structural diversity via hierarchical trees [49], spatial matching [16], and nearest-neighbor alignments [64]. Concurrently, many works rely on heuristic loss regularizers, such as soft orthogonality penalties or spatial separation margins, to encourage diverse anatomical part discovery [5, 33, 67, 69]. However, lacking absolute geometric boundaries, these soft constraints are systematically overpowered by the primary cross-entropy objective [24, 28], inevitably driving multiple prototypes to homogenize and redundantly attend to identical spatial regions [53]. This severe prototype collapse fatally undermines

compositional interpretability, demonstrating that robust part discovery requires a fundamental geometric restructuring of the representation space rather than mere auxiliary penalties.

2.3 Optimization on Orthogonal Manifolds

Optimization under orthogonality constraints has a rich history on Grassmann and Stiefel manifolds, progressing from early Newton and conjugate gradient algorithms [22] to modern unified frameworks utilizing retractions and vector transports [1]. This geometric foundation has been further adapted for large-scale applications via constraint-preserving feasible methods [75] and stochastic Riemannian algorithms for data-driven learning [8, 80].

In deep learning, orthogonality frequently serves as a structural prior to stabilize training and enhance generalization by preserving gradient norms and mitigating optimization instabilities. Applications traditionally impose these constraints directly on the model's parameters. This ranges from orthogonal weight normalization [32] and convolutional regularizers [6] to geometry-aware optimizers [56] that maintain weight matrices on the Stiefel manifold, sometimes employing efficient approximations like the Cayley transform [44]. Similarly, in the latent space, representation learning leverages orthogonal and decorrelated objectives to promote disentanglement while reducing feature redundancy [13, 50]. Geometry-aware regularization has also been used to encode ordered disease progression in medical grading, as in directed ordinal diffusion regularization [12].

Unlike prior works that apply orthogonality to network weights or latent factors, AMP leverages these geometric principles to enforce a strict Stiefel constraint on class-specific prototype bases to counteract Neural Collapse-induced prototype collapse, coupling this with adaptive rank pruning and spatial regularization to translate orthogonality into localized, compositional explanations.

3 Preliminaries on Prototypical Networks

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N} \subset \mathcal{X} \times \mathcal{Y}$ be the training dataset, where $\mathcal{X}$ is the image space and $\mathcal{Y} = \{1, 2, \ldots, C\}$ denotes the label space. A prototype network employs a parameterized convolutional backbone $f_\theta : \mathcal{X} \to \mathbb{R}^{D \times H \times W}$ to project an input x_i into a spatial feature tensor $F^{(i)} = f_\theta(x_i)$. We denote the local feature vector at spatial location $(h, w) \in \mathcal{H} \times \mathcal{W}$ as $F^{(i)}_{h,w} \in \mathbb{R}^D$, where $\mathcal{H} = \{1, \ldots, H\}$ and $\mathcal{W} = \{1, \ldots, W\}$. The model maintains a learnable prototype set $\mathcal{P} = \{P_1, \ldots, P_C\}$, where $P_c = [p_{c,1}, \ldots, p_{c,K}] \in \mathbb{R}^{D \times K}$ aggregates K latent basis vectors assigned to class c.

During forward propagation, the network evaluates the spatial correspondence between a prototype $p_{c,k}$ and the input x_i via global spatial max-pooling over a localized similarity metric $g : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$:

$$S_{c,k}(x_i) = \max_{(h,w) \in \mathcal{H} \times \mathcal{W}} g\left(F^{(i)}_{h,w}, p_{c,k}\right) \tag{1}$$

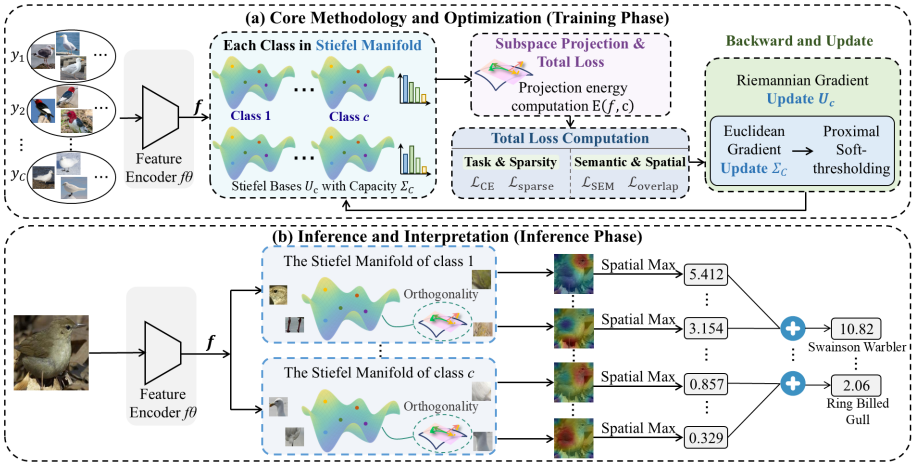

Fig. 2: The overview of our proposed AMP framework. (a) Training: learn Stiefel bases and capacity matrices via Riemannian gradients. (b) Inference: project features onto class manifolds, then aggregate activations to class scores for interpretable predictions.

Crucially, to bridge the latent geometry with human-understandable visual concepts, the network relies on a strict prototype projection mechanism. The optimal coordinate $(h^*, w^*) = \arg\max_{(h,w)} g(F^{(i)}_{h,w}, p_{c,k})$ precisely localizes the semantic part in the input. Concurrently, to ensure intrinsic transparency, each continuous prototype vector is explicitly projected onto the nearest spatial feature patch from the training subset of class c:

$$p_{c,k} \leftarrow \arg \min_{\substack{j \in \{1,\dots,N\} \\ s.t.\ y_j = c}} \min_{(h,w) \in \mathcal{H} \times \mathcal{W}} \left\| p_{c,k} - F^{(j)}_{h,w} \right\|_2 \tag{2}$$

This projection grounds the abstract basis vectors into discrete, visually verifiable exemplars. These semantic activation scores form a representation vector $\mathbf{S}(x_i) \in \mathbb{R}^{CK}$, which is projected into the categorical probability simplex Δ^{C-1} via a linear fully connected layer to output the final prediction.

4 Methodology

As illustrated in Fig. 2, we propose the AMP framework, which replaces unconstrained Euclidean prototypes with Stiefel-constrained orthonormal bases for prototype matching, enabling diverse and collapse-free interpretable inference.

4.1 Orthogonal Subspaces on the Stiefel Manifold

Traditional prototype networks parameterize class visual features as unconstrained matrices $P_c \in \mathbb{R}^{D \times K}$. However, this formulation is structurally vulnerable to prototype collapse under the dynamics of Neural Collapse (NC) [51].

Let $\mu_c \in \mathbb{R}^D$ denote the global mean vector of class c, and Σ_W represent the intra-class covariance matrix of the spatial features. During the terminal phase of cross-entropy optimization, NC dictates that $\lim_{t\to\infty} \Sigma_W^{(t)} = \mathbf{0}$, aggressively compressing all spatial features $\{F_{h,w}^{(i)} \mid y_i = c\}$ into the singular state μ_c, while the class means self-organize into a Simplex Equiangular Tight Frame (ETF) [86].

To maximize the objective against this completely homogenized feature space, the unconstrained prototype matrix P_c is forced to collapse, aligning all K column vectors along the exact same mean coordinate:

$$\lim_{t\to\infty} p_{c,k}^{(t)} = \mu_c \quad \forall k \in \{1, \dots, K\} \tag{3}$$

Geometrically, the effective rank of the class-specific prototype matrix fatally degenerates: $\lim_{t\to\infty} \operatorname{rank}(P_c^{(t)}) = 1$. This zero-dimensional point collapse annihilates the local representational diversity required for compositional part discovery. To preclude this collapse, we abandon the unconstrained parameterization P_c and strictly redefine the prototype set as an orthogonal basis matrix $U_c \in \mathbb{R}^{D\times K}$. We restrict U_c to the Stiefel manifold $St(D, K)$ [22]:

$$St(D, K) = \{U \in \mathbb{R}^{D\times K} \mid U^\top U = I_K\} \tag{4}$$

By anchoring the prototypes to $St(D, K)$, we structurally endow the K latent dimensions with absolute orthogonal separation. Strictly bound by the intrinsic curvature of the manifold, discriminative gradients are fundamentally prevented from compressing these basis vectors into a singular coordinate.

Under this rigid constraint, we replace the traditional Euclidean similarity metric with the projection energy onto the manifold's orthogonal subspace. Given a generic local feature vector $f \in \mathbb{R}^D$ (which represents any spatial feature $F_{h,w}^{(i)}$ from the input tensor), the mapping matrix projecting f onto the subspace spanned by U_c is $U_c U_c^\top$. The corresponding projection energy $E(f, c)$ is defined as the squared ℓ_2 norm of this projection:

$$E(f, c) = \|U_c^\top f\|_2^2 = f^\top U_c U_c^\top f \tag{5}$$

This paradigm shift fundamentally transforms the catastrophic point collapse (where rank degrades to 1) into a controlled subspace representation (where rank is strictly preserved at K), geometrically guaranteeing the retention of representational capacity.

4.2 Dynamic Rank Calibration via Proximal Gradients

Real-world visual categories exhibit extreme asymmetry in their intrinsic semantic complexity. Imposing a uniform, fixed-rank subspace across all classes inevitably induces representational redundancy and overfitting. Motivated by the variable rank topology of nested subspaces [78], we explicitly inject a learnable, non-negative diagonal capacity matrix $\Sigma_c = \operatorname{diag}(\sigma_{c,1}, \dots, \sigma_{c,K})$ into the

orthogonal basis. This operation transforms the rigid constraints of $St(D, K)$ into a dynamically weighted manifold framework defined by $F_c = U_c \Sigma_c$.

Within this dynamic scheme, the projection energy $E(f, c)$ mapping $f \in \mathbb{R}^D$ to the target class c expands into an adaptively weighted quadratic form:

$$E(f, c) = f^\top U_c \Sigma_c U_c^\top f = \sum_{k=1}^{K} \sigma_{c,k} (U_{c,k}^\top f)^2 \tag{6}$$

To enforce the dimensional collapse of extraneous bases, we penalize the capacity matrix with a strict ℓ_1 sparsity regularization. Given the non-negativity constraint on $\sigma_{c,k}$, this penalty simplifies to a direct linear summation:

$$\mathcal{L}_{sparse} = \lambda \sum_{c=1}^{C} \sum_{k=1}^{K} \sigma_{c,k} \tag{7}$$

Standard stochastic gradient descent (SGD) is notoriously incapable of producing exact zero values for continuous parameters, leading to persistent, low-magnitude noise rather than true structural sparsity [43,60]. To enforce absolute dimensional collapse, we decouple the optimization and execute a proximal gradient descent step. Specifically, we apply a Euclidean soft-thresholding operator to the capacity weights at each iteration t:

$$\sigma_{c,k}^{(t+1)} = \max \left(\sigma_{c,k}^{(t)} - \eta \frac{\partial \mathcal{L}_{CE}}{\partial \sigma_{c,k}} - \eta\lambda, 0 \right) \tag{8}$$

This exact truncation mechanism permits the effective subspace rank, denoted as $\text{rank}(F_c) = \|\text{diag}(\Sigma_c)\|_0$, to dynamically collapse to an optimal low-rank state. Such rigorous calibration ensures the active representational capacity perfectly mirrors the intrinsic physical complexity of each distinct visual category, completely suppressing the absorption of high-frequency noise.

4.3 Semantic Gauge Fixing and Unsupervised Part Discovery

Although the Stiefel constraint guarantees orthonormal columns for each class basis $U_c \in St(D, K)$, the classification score primarily depends on subspace geometry. Because an isotropic projector $U_c U_c^\top$ is invariant to right multiplication by any orthogonal matrix $Q \in O(K)$, and a learned diagonal capacity matrix Σ_c only partially reduces this symmetry, rotational ambiguities persist. Consequently, the classification objective alone cannot guarantee a semantically stable basis. To mitigate this, we introduce two spatial regularizers that break rotational invariance, acting as an inductive bias to encourage individual basis directions to yield spatially localized and minimally overlapping visual evidence.

Given a spatial feature tensor $F \in \mathbb{R}^{D \times H \times W}$ for an input image, we compute a dense energy response map $M_{c,k,h,w} = \left(U_{c,k}^\top F_{h,w} \right)^2$ for each class c and basis direction k, where $F_{h,w} \in \mathbb{R}^D$ is the local feature at spatial location (h, w)

and $U_{c,k}$ is the kth column of U_c. We then normalize this map into a spatial probability distribution via a spatial softmax:

$$P_{c,k,h,w} = \frac{\exp(M_{c,k,h,w})}{\sum_{h'=1}^{H}\sum_{w'=1}^{W}\exp(M_{c,k,h',w'})} \tag{9}$$

Rank calibration yields an active set of basis indices:

$$\mathcal{K}_c = \{k \in \{1,\dots,K\} \mid \sigma_{c,k} > 0\}, \qquad R_c = |\mathcal{K}_c|. \tag{10}$$

The spatial regularizers are applied exclusively to this active set, which is treated as a discrete selection without backpropagating through its membership.

Spatial Entropy Minimization. To encourage focal and localized evidence for each active basis direction, we minimize the spatial entropy of $P_{c,k}$:

$$\mathcal{L}_{SEM} = -\frac{1}{R_c}\sum_{k\in\mathcal{K}_c}\sum_{h=1}^{H}\sum_{w=1}^{W} P_{c,k,h,w}\log(P_{c,k,h,w}). \tag{11}$$

Lower entropy corresponds to more concentrated maps, which empirically improves the localization quality of part evidence.

Spatial Overlap Penalty. Focal attention alone does not prevent multiple basis directions from attending to the same spatial region. To encourage distinct evidence among active directions, we penalize pairwise overlap using cosine similarity between heatmaps:

$$\mathcal{L}_{overlap} = \frac{1}{R_c(R_c-1)}\sum_{k\in\mathcal{K}_c}\sum_{\substack{j\in\mathcal{K}_c\\ j\neq k}} \frac{\langle P_{c,k}, P_{c,j}\rangle}{\|P_{c,k}\|_F\|P_{c,j}\|_F}, \tag{12}$$

where $\langle\cdot,\cdot\rangle$ and $\|\cdot\|_F$ are the Frobenius inner product and Frobenius norm over spatial dimensions. When $R_c < 2$, we set $\mathcal{L}_{overlap}$ to zero.

Together, these spatial objectives encourage the optimization to converge to a Stiefel basis whose columns correspond to focal and distinct spatial evidence. This reduces rotational ambiguity in practice and improves semantic stability of part-based explanations without requiring part annotations.

4.4 Decoupled Optimization and Interpretable Inference

We optimize AMP with a composite objective that combines classification, spatial regularization, and sparsity-induced rank calibration:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \gamma_1\mathcal{L}_{SEM} + \gamma_2\mathcal{L}_{overlap} + \lambda\sum_{c=1}^{C}\sum_{k=1}^{K}\sigma_{c,k}, \tag{13}$$

subject to the constraints $U_c^\top U_c = I_K$ for all classes and $\sigma_{c,k} \geq 0$. The parameters live on different spaces, so we use a decoupled update rule: Euclidean

updates for the backbone parameters θ, Riemannian updates for the Stiefel bases U_c, and a proximal gradient update for the capacity weights $\sigma_{c,k}$.

AMP employs a decoupled optimization strategy to accommodate parameters residing in fundamentally different geometric spaces. While the feature backbone is updated using standard Euclidean optimizers, the class-specific orthogonal bases U_c are optimized via Riemannian gradient descent, utilizing tangent space projections and QR-based retractions to strictly preserve the Stiefel manifold constraints at every iteration [1,22]. Concurrently, the capacity matrix Σ_c is updated using a Euclidean gradient step followed by a proximal soft-thresholding operation [52]. This proximal step enforces exact ℓ_1 sparsity on the continuous variables, physically driving the discrete dynamic rank calibration. The complete unified training algorithm, including exact mathematical derivations for the Riemannian retractions and proximal steps, is detailed in the Appendix.

Interpretable inference. Given a feature tensor $F^{(i)}$ for input x_i, we define the class logit as the maximum projection energy across spatial locations:

$$z_c(x_i) = \sum_{k=1}^{K} \sigma_{c,k} \left(\max_{(h,w)\in\mathcal{H}\times\mathcal{W}} (U_{c,k}^{\top} F_{h,w}^{(i)})^2 \right) \tag{14}$$

The prediction is $\hat{c} = \arg\max_c z_c(x)$. For explanation, we consider only the active directions $k \in \mathcal{K}_{\hat{c}}$, computing an energy heatmap:

$$\widetilde{M}_{\hat{c},k,i,j} = \sigma_{\hat{c},k} \left(U_{\hat{c},k}^{\top} F_{i,j} \right)^2, \tag{15}$$

and localize the most contributing region by $(h^*, w^*) = \arg\max_{i,j} \widetilde{M}_{\hat{c},k,i,j}$. Finally, we retrieve a nearest training patch from the same class by searching over cached training features and selecting the patch that maximizes the corresponding energy response to ground each basis direction to a discrete exemplar:

$$F_{source} = \arg \max_{j\in\{1,\ldots,N_{\hat{c}}\},h,w} \left(U_{\hat{c},k}^{\top} F_{h,w}^{(j)} \right)^2 \tag{16}$$

5 Experiments

5.1 Experimental Setup

Datasets and Metrics. We conduct experiments on CUB-200-2011 [66] and Stanford Cars [40], two widely used benchmarks for fine-grained visual categorization. CUB-200-2011 contains 11,788 images from 200 bird species, while Stanford Cars contains 16,185 images from 196 car models. Following the established evaluation protocol [33,35,67], we assess model performance by reporting Accuracy (Acc), Consistency (Cons.), Stability (Stab.), OIRR and DAUC.

Table 1: Predictive performance comparison (top-1 accuracy, %) on CUB-200-2011 and Stanford Cars. Best results among *intrinsically interpretable models* are highlighted as **first**, second and third. † denotes ResNet50 pre-trained on iNaturalist for CUB.

Method	CUB-200-2011				Stanford Cars			
	V16	R34	R50†	D161	V16	R34	R50	D161
Baseline	76.5 ± 0.8	80.0 ± 0.7	85.1 ± 0.5	84.5 ± 0.4	85.1 ± 0.4	86.7 ± 0.4	88.0 ± 0.3	89.0 ± 0.3
Black-box Classifiers								
RA-CNN	83.1 ± 0.6	84.9 ± 0.5	88.7 ± 0.4	88.0 ± 0.3	91.4 ± 0.3	92.2 ± 0.3	92.9 ± 0.2	93.4 ± 0.2
NTS-Net	82.4 ± 0.6	84.1 ± 0.5	87.9 ± 0.4	87.2 ± 0.3	90.8 ± 0.3	91.7 ± 0.3	92.4 ± 0.2	92.9 ± 0.2
PMG	84.7 ± 0.5	85.8 ± 0.5	89.2 ± 0.3	88.5 ± 0.3	92.0 ± 0.3	92.9 ± 0.2	93.6 ± 0.2	94.0 ± 0.2
Intrinsically Interpretable Models								
ProtoPNet	79.0 ± 0.7	80.5 ± 0.7	84.0 ± 0.5	83.3 ± 0.5	87.3 ± 0.4	88.0 ± 0.4	88.7 ± 0.3	89.5 ± 0.3
TesNet	80.8 ± 0.6	82.1 ± 0.6	86.2 ± 0.4	85.3 ± 0.4	88.3 ± 0.4	89.2 ± 0.3	89.6 ± 0.3	90.0 ± 0.3
ProtoKNN	80.3 ± 0.6	82.7 ± 0.6	84.9 ± 0.4	84.2 ± 0.4	88.4 ± 0.3	90.0 ± 0.3	89.9 ± 0.3	90.6 ± 0.2
SDFA-SA	81.3 ± 0.6	83.0 ± 0.6	86.3 ± 0.6	85.1 ± 0.4	87.9 ± 0.4	88.7 ± 0.3	89.3 ± 0.3	89.9 ± 0.3
MGProto	82.1 ± 0.6	83.6 ± 0.5	86.6 ± 0.4	85.2 ± 0.4	88.6 ± 0.3	89.7 ± 0.3	90.5 ± 0.2	90.5 ± 0.2
CBC	82.4 ± 0.5	83.8 ± 0.5	86.0 ± 0.5	85.4 ± 0.4	88.2 ± 0.3	89.8 ± 0.3	90.0 ± 0.2	91.0 ± 0.2
ProtoArgNet	82.0 ± 0.6	84.0 ± 0.5	85.8 ± 0.4	85.7 ± 0.4	88.3 ± 0.3	89.5 ± 0.3	90.3 ± 0.2	90.3 ± 0.2
AMP (Ours)	**84.2** ± 0.5	**85.5** ± 0.5	**88.4** ± 0.3	**87.8** ± 0.3	**90.4** ± 0.3	**91.3** ± 0.2	**92.0** ± 0.2	**92.5** ± 0.2

Baselines and Compared Methods. As foundational convolutional backbones, we utilize VGG16, ResNet34, ResNet50, and DenseNet161 [27, 31, 61]. All are initialized with ImageNet [14] weights, except for the ResNet50 model on CUB, which employs iNaturalist [65] pre-training. We benchmark our approach against representative black-box fine-grained classifiers, including RA-CNN, NTS-Net, and PMG [18, 23, 76]. We further compare against intrinsically interpretable baselines, including ProtoPNet, TesNet, ProtoKNN, and SDFA-SA [10,33,64,69]. Finally, we include MGProto, CBC, and ProtoArgNet [5,58,67].

Implementation Details. We implement AMP in PyTorch on a single NVIDIA RTX 4090 GPU by resizing images to 448×448 pixels and cropping via bounding boxes. Backbones follow a decoupled optimization strategy where backbone parameters use SGD while Stiefel bases employ Riemannian SGD with QR retraction [1]. Initial prototype count K is 10 for each class with sparsity λ at 0.0001 and spatial weights γ_1 and γ_2 both at 0.01. The model undergoes training for 100 epochs with a batch size of 32 using a cosine annealing scheduler to decay the learning rate from 0.001 to 0.00001.

5.2 Quantitative Results

Table 1 reports the predictive comparison on CUB-200-2011 and Stanford Cars where AMP consistently achieves the highest top-1 accuracy among intrinsically interpretable models across all backbones. On CUB-200-2011 with ResNet50, AMP reaches 88.4% accuracy to surpass the strongest interpretable baseline MGProto at 86.6% while remaining competitive with the black-box model PMG at 89.2%. Stanford Cars exhibits a similar pattern with AMP achieving 92.0% to exceed MGProto at 90.5% and approach PMG at 93.6%. Furthermore, AMP maintains stable state-of-the-art interpretable performance across VGG16, ResNet34,

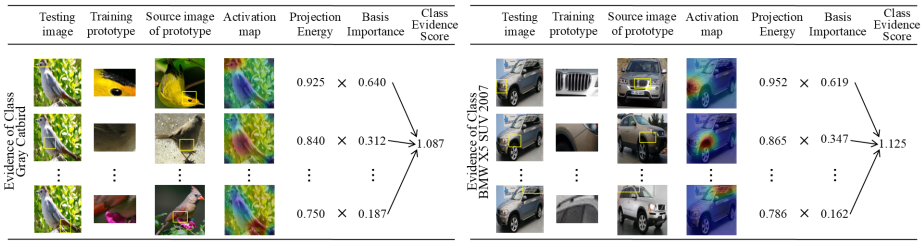

Fig. 3: Qualitative visualization of the AMP reasoning process on the CUB-200-2011 (left) and Stanford Cars (right) datasets. For a given test image, AMP explicitly decomposes the final Class Evidence Score into a sparse sum of localized visual evidence.

and DenseNet161 on both datasets. These results indicate that enforcing geometric diversity avoids an accuracy penalty and instead improves discrimination by preventing redundant prototype usage.

Table 2 evaluates interpretability through Consistency, Stability, OIRR, and DAUC. AMP achieves the best scores across all four metrics on both datasets. On CUB-200-2011, AMP reaches 76.80 Consistency and 49.20 Stability to surpass the previous bests of MGProto and CBC, while attaining 28.10 OIRR and 3.45 DAUC to outperform ProtoArgNet and ProtoKNN. AMP similarly sets state-of-the-art interpretability records on Stanford Cars by achieving 50.20 Consistency, 76.40 Stability, 33.50 OIRR, and 5.65 DAUC. Overall, this quantitative evidence demonstrates that the Stiefel formulation effectively prevents prototype collapse and yields explanations that are stable across images and causally aligned with model decisions.

Table 2: Interpretability evaluation on CUB-200-2011 and Stanford Cars. Best results are highlighted as **first**, second and third.

Method	CUB-200-2011				Stanford Cars			
	Cons.	**Stab.**	**OIRR**	**DAUC**	**Cons.**	**Stab.**	**OIRR**	**DAUC**
Black-box Classifiers								
RA-CNN	$23.40_{\pm 1.90}$	$36.70_{\pm 1.60}$	$51.60_{\pm 1.40}$	$8.05_{\pm 0.58}$	$15.30_{\pm 1.95}$	$59.60_{\pm 1.35}$	$56.20_{\pm 1.30}$	$10.35_{\pm 0.60}$
NTS-Net	$24.60_{\pm 2.05}$	$37.20_{\pm 1.90}$	$50.40_{\pm 1.70}$	$7.82_{\pm 0.61}$	$16.50_{\pm 2.10}$	$60.90_{\pm 1.55}$	$54.90_{\pm 1.55}$	$9.95_{\pm 0.58}$
PMG	$26.00_{\pm 2.25}$	$38.60_{\pm 2.15}$	$47.90_{\pm 2.00}$	$7.35_{\pm 0.60}$	$18.10_{\pm 2.30}$	$62.30_{\pm 1.90}$	$53.20_{\pm 1.95}$	$9.38_{\pm 0.56}$
Intrinsically Interpretable Models								
ProtoPNet	$56.00_{\pm 1.70}$	$42.00_{\pm 1.60}$	$39.00_{\pm 1.70}$	$6.20_{\pm 0.50}$	$26.50_{\pm 2.10}$	$69.60_{\pm 1.30}$	$49.50_{\pm 1.40}$	$8.70_{\pm 0.51}$
TesNet	$61.50_{\pm 1.60}$	$41.80_{\pm 1.70}$	$37.00_{\pm 1.70}$	$5.45_{\pm 0.47}$	$31.00_{\pm 2.00}$	$70.70_{\pm 1.40}$	$45.80_{\pm 1.60}$	$7.80_{\pm 0.48}$
ProtoKNN	$63.00_{\pm 1.80}$	$42.60_{\pm 1.90}$	$35.20_{\pm 1.90}$	$3.80_{\pm 0.45}$	$34.50_{\pm 2.20}$	$71.40_{\pm 1.60}$	$42.60_{\pm 1.80}$	$6.15_{\pm 0.46}$
MGProto	$71.40_{\pm 1.90}$	$45.80_{\pm 2.30}$	$30.80_{\pm 2.10}$	$4.02_{\pm 0.44}$	$45.00_{\pm 2.40}$	$74.20_{\pm 2.00}$	$36.80_{\pm 2.10}$	$6.05_{\pm 0.43}$
CBC	$69.00_{\pm 2.20}$	$46.80_{\pm 2.20}$	$31.20_{\pm 2.20}$	$3.92_{\pm 0.45}$	$42.00_{\pm 2.50}$	$75.00_{\pm 2.10}$	$35.90_{\pm 2.20}$	$6.18_{\pm 0.46}$
ProtoArgNet	$67.50_{\pm 2.20}$	$46.20_{\pm 2.30}$	$30.60_{\pm 2.10}$	$3.88_{\pm 0.46}$	$41.00_{\pm 2.40}$	$73.80_{\pm 2.20}$	$36.20_{\pm 2.20}$	$6.25_{\pm 0.47}$
AMP (Ours)	**76.80**$_{\pm 1.80}$	**49.20**$_{\pm 2.10}$	**28.10**$_{\pm 2.00}$	**3.45**$_{\pm 0.41}$	**50.20**$_{\pm 2.20}$	**76.40**$_{\pm 1.90}$	**33.50**$_{\pm 2.00}$	**5.65**$_{\pm 0.40}$

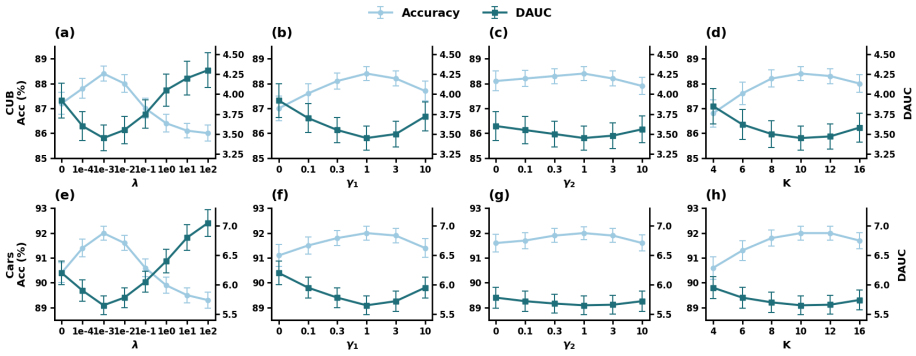

Fig. 4: Hyperparameter sensitivity on CUB-200-2011 and Stanford Cars.

5.3 Qualitative Results

Fig. 3 visually demonstrates that AMP avoids prototype collapse and yields faithful compositional explanations on the CUB-200-2011 and Stanford Cars datasets. The model leverages geometric constraints to successfully discover diverse semantic parts such as the head and wing of a bird or the grille and wheel of a car. Visualizing each active Stiefel basis direction alongside its nearest training exemplar and spatial activation map confirms the efficacy of dynamic rank calibration. The exact contribution of each localized part is strictly quantified and cumulatively summed to derive the final class evidence score.

5.4 Sensitivity Analysis

We analyze the sensitivity of AMP to sparsity strength λ, spatial entropy weight γ_1, overlap suppression weight γ_2, and subspace size K as illustrated in Fig. 4. Across both datasets, a moderate λ yields a U-shaped trend in faithfulness metrics to optimally prune redundancy without collapsing discriminative capacity. Similarly, intermediate values for γ_1 and γ_2 resolve rotational ambiguity by encouraging localized and distinct evidence while avoiding the diffuse attention or artificial disjointness caused by extreme settings. Finally, an intermediate subspace size K ensures sufficient representational capacity without the unconstrained redundancy of overly large bases. Overall, the broad optimal regions confirm that AMP is practically robust and consistently delivers causally faithful explanations without fragile tuning.

5.5 Ablation Study

Table 3 ablates the key components of AMP on CUB-200-2011 and Stanford Cars. The results demonstrate that removing the Stiefel manifold constraint

Table 3: Ablation study on CUB-200-2011 and Stanford Cars. Best results are highlighted as **first**, second and third.

Method	CUB-200-2011					Stanford Cars				
	Acc	**Cons.**	**Stab.**	**OIRR**	**DAUC**	**Acc**	**Cons.**	**Stab.**	**OIRR**	**DAUC**
w/o Stiefel	85.2	65.0	44.2	36.5	4.95	90.2	41.2	72.3	40.8	6.95
w/o Σ_c	87.9	75.8	47.6	28.9	3.55	91.8	48.7	76.0	35.4	5.85
w/o $\mathcal{L}_{SEM}$	87.0	73.2	47.8	30.4	3.92	91.1	46.2	74.5	35.0	6.20
w/o $\mathcal{L}_{overlap}$	88.1	74.6	48.6	29.1	3.60	91.6	49.0	75.7	34.0	5.78
AMP (Full)	**88.4**	**76.8**	**49.2**	**28.1**	**3.45**	**92.0**	**50.2**	**76.4**	**33.5**	**5.65**

causes the most severe performance degradation by dropping CUB-200-2011 accuracy from 88.4 to 85.2, decreasing Consistency from 76.8 to 65.0, and raising DAUC to 4.95, confirming that hard orthogonality is essential to preserve representational capacity and prevent collapse. Similarly, omitting the dynamic capacity matrix Σ_c slightly reduces accuracy and worsens interpretability by increasing OIRR on Stanford Cars from 33.5 to 35.4, underscoring the necessity of rank calibration for pruning redundant dimensions without sacrificing discriminative power. Furthermore, discarding the spatial gauge-fixing terms degrades both performance and faithfulness. Specifically, removing Spatial Entropy Minimization diffuses attention to drop CUB accuracy to 87.0 and raise DAUC to 3.92, while eliminating the Spatial Overlap Penalty fails to separate part evidence and negatively impacts OIRR. Ultimately, retaining all components enables AMP to achieve the best overall balance, delivering the highest accuracy and the most reliable explanations across both datasets.

5.6 Human Evaluation

To assess the qualitative interpretability of AMP across three key dimensions and validate our dynamic rank design, we conducted a human evaluation study with 50 participants. As outlined in Fig. 5(a), we defined Part Diversity to measure region distinctness, Evidence Sufficiency to evaluate recognition confidence, and Explanation Parsimony to verify noise pruning. The results (Fig. 5(b)) demonstrate that AMP significantly outperforms both ProtoPNet and TesNet on the evaluated datasets, indicating that the Stiefel manifold effectively prevents redundancy while dynamic calibration ensures concise explanations. Furthermore, these human assessments align with the distribution of active prototypes (Fig. 5(c)). The data reveals that AMP adaptively adjusts its complexity according to the dataset: most CUB images activate three prototypes, whereas Stanford Cars typically require four. This adaptive behavior confirms that our dynamic rank mechanism successfully identifies the optimal number of semantic parts necessary for different categories, ultimately driving the high sufficiency and parsimony scores reported by the evaluators.

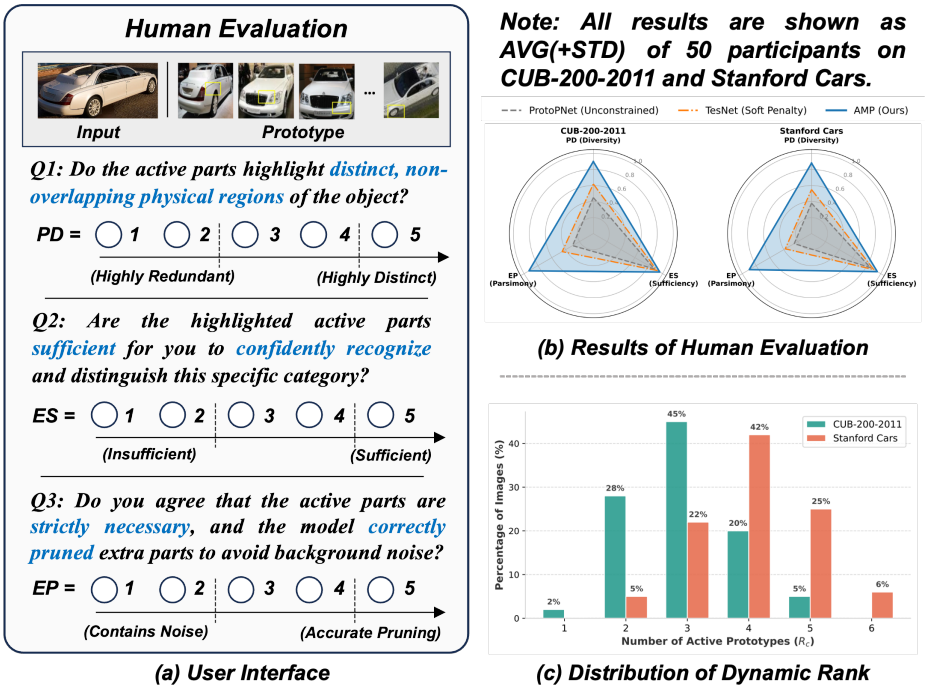

Fig. 5: Human evaluation results and dynamic rank distribution.

6 Conclusion

In this paper, we identified that the pervasive issue of prototype collapse in interpretable recognition is not merely an architectural flaw, but a geometric inevitability driven by the Neural Collapse under standard cross-entropy optimization. To resolve this, we introduce Adaptive Manifold Prototypes (AMP), which formulates class prototypes as orthonormal bases on the Stiefel manifold to mathematically preclude rank-one degeneracy. By incorporating proximal rank calibration and spatial gauge-fixing regularizers, AMP adaptively prunes redundancy while ensuring localized and semantically distinct part discovery. Extensive evaluations on fine-grained benchmarks demonstrate that AMP achieves state-of-the-art predictive performance while delivering exceptionally stable and causally faithful explanations. Ultimately, our work underscores a necessary paradigm shift for inherently interpretable AI: robust compositional reasoning demands strict geometric boundaries rather than heuristic soft penalties.

Acknowledgements

This work was supported by Zhejiang Leading Innovative and Entrepreneur Team Introduction Program (2024R01007).

References

1. Absil, P.A., Mahony, R., Sepulchre, R.: Optimization algorithms on matrix manifolds. Princeton University Press (2008)
2. 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)
3. Aïvodji, U., Arai, H., Fortineau, O., Gambs, S., Hara, S., Tapp, A.: Fairwashing: the risk of rationalization. In: International Conference on Machine Learning. pp. 161–170. PMLR (2019)
4. Alvarez Melis, D., Jaakkola, T.: Towards robust interpretability with self-explaining neural networks. Advances in Neural Information Processing Systems **31** (2018)
5. Ayoobi, H., Potyka, N., Toni, F.: Protoargnet: Interpretable image classification with super-prototypes and argumentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 1791–1799 (2025)
6. Bansal, N., Chen, X., Wang, Z.: Can we gain more from orthogonality regularizations in training deep networks? Advances in Neural Information Processing Systems **31** (2018)
7. Biederman, I.: Recognition-by-components: a theory of human image understanding. Psychological Review **94**(2), 115 (1987)
8. Bonnabel, S.: Stochastic gradient descent on riemannian manifolds. IEEE Transactions on Automatic Control **58**(9), 2217–2229 (2013)
9. Branson, S., Van Horn, G., Perona, P., Belongie, S.: Improved bird species recognition using pose normalized deep convolutional nets. In: Proceedings of the British Machine Vision Conference (2014)
10. 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)
11. Chen, H., Chen, Y., Yan, Z., Ding, M., Li, C., Qin, F., Zhang, D., Zhu, Z.: MMLNB: Multi-Modal Learning for Neuroblastoma subtyping classification assisted with textual description generation. Biomedical Signal Processing and Control **126**, 110798 (2026)
12. Chen, H., Jia, J., Li, R., Yang, C., Li, W., Pang, X., Chen, Y., Wang, H., Bu, J., Wu, L.: Directed ordinal diffusion regularization for progression-aware diabetic retinopathy grading. arXiv preprint arXiv:2602.21942 (2026)
13. Cogswell, M., Ahmed, F., Girshick, R., Zitnick, L., Batra, D.: Reducing overfitting in deep networks by decorrelating representations. In: International Conference on Learning Representations (2016)
14. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009)
15. Dombrowski, A.K., Alber, M., Anders, C., Ackermann, M., Müller, K.R., Kessel, P.: Explanations can be manipulated and geometry is to blame. Advances in Neural Information Processing Systems **32** (2019)
16. Donnelly, J., Barnett, A.J., Chen, C.: Deformable protopnet: An interpretable image classifier using deformable prototypes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10265–10275 (2022)
17. Du, N., Wei, S., Tong, Y., Du, S., Liu, Y.: Fame-net: Frequency-adaptive mixture-of-experts network for robust infrared small target detection. IEEE Transactions on Geoscience and Remote Sensing (2026)

18. Du, R., Chang, D., Bhunia, A.K., Xie, J., Ma, Z., Song, Y.Z., Guo, J.: Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In: European Conference on Computer Vision. pp. 153–168 (2020)
19. Du, S., Bai, Y., Ma, J., Liu, M., Li, Y.: Frequency-decoupled learning for joint thin-cloud removal and pansharpening. In: ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 9282–9286. IEEE (2026)
20. Du, S., Zou, Y., Li, J., Liu, M., Li, Y., Shang, C., Shen, Q.: Pansharpening for thin-cloud contaminated remote sensing images: a unified framework and benchmark dataset. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 3696–3704 (2026)
21. Du, S., Zou, Y., Wang, Z., Li, X., Li, Y., Shang, C., Shen, Q.: Unsupervised hyperspectral image super-resolution via self-supervised modality decoupling. International Journal of Computer Vision **134**(4), 152 (2026)
22. Edelman, A., Arias, T.A., Smith, S.T.: The geometry of algorithms with orthogonality constraints. SIAM Journal on Matrix Analysis and Applications **20**(2), 303–353 (1998)
23. Fu, J., Zheng, H., Mei, T.: Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4438–4446 (2017)
24. Gautam, S., Boubekki, A., Hansen, S., Salahuddin, S., Jenssen, R., Höhne, M., Kampffmeyer, M.: Protovae: A trustworthy self-explainable prototypical variational model. Advances in Neural Information Processing Systems **35**, 17940–17952 (2022)
25. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (2020)
26. Ghorbani, A., Abid, A., Zou, J.: Interpretation of neural networks is fragile. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 3681–3688 (2019)
27. 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)
28. Hoffmann, A., Fanconi, C., Rade, R., Kohler, J.: This looks like that... does it? shortcomings of latent space prototype interpretability in deep networks. In: ICML 2021 Workshop on Theoretic Foundation, Criticism, and Application Trend of Explainable AI (2021)
29. Hong, D., Wang, T., Baek, S.: Protorynet-interpretable text classification via prototype trajectories. Journal of Machine Learning Research **24**(264), 1–39 (2023)
30. Hooker, S., Erhan, D., Kindermans, P.J., Kim, B.: A benchmark for interpretability methods in deep neural networks. Advances in Neural Information Processing Systems **32** (2019)
31. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4700–4708 (2017)
32. Huang, L., Liu, X., Lang, B., Yu, A., Wang, Y., Li, B.: Orthogonal weight normalization: Solution to optimization over multiple dependent stiefel manifolds in deep neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)

33. Huang, Q., Xue, M., Huang, W., Zhang, H., Song, J., Jing, Y., Song, M.: Evaluation and improvement of interpretability for self-explainable part-prototype networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2011–2020 (2023)
34. Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., Madry, A.: Adversarial examples are not bugs, they are features. Advances in Neural Information Processing Systems **32** (2019)
35. Jia, J., Liu, Y., Sun, Y., Chen, H., Qin, F., Wang, C., Peng, Y.: Geodesic prototype matching via diffusion maps for interpretable fine-grained recognition. In: ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1786–1790. IEEE (2026)
36. Jia, J., Wu, Y., Chen, H., Jing, H., Wang, H., Bu, J., Wu, L.: Unsupervised causal prototypical networks for de-biased interpretable dermoscopy diagnosis. arXiv preprint arXiv:2602.23752 (2026)
37. Jia, J., Zheng, H., Wu, Y., Chen, H., Wang, H., Bu, J., Wu, L.: SCOUT: Selective coupling via optimal unbalanced transport for interpretable text classification. In: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 6413–6431. Association for Computational Linguistics (2026)
38. Jing, H., Jiang, D., Wang, J., Jia, J., Li, Y., Ma, Y., Zheng, N.: More natural, more real: Object-aware gaussian splatting for 3d visual decoding from human brain. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19033–19044 (2026)
39. Kindermans, P.J., Hooker, S., Adebayo, J., Alber, M., Schütt, K.T., Dähne, S., Erhan, D., Kim, B.: The (un) reliability of saliency methods. In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, pp. 267–280. Springer (2019)
40. 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)
41. Lan, Q., Tian, Q.: Acam-kd: Adaptive and cooperative attention masking for knowledge distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3957–3966 (2025)
42. Lan, Q., Tian, Q.: Clockdistill: Consistent location and context aware knowledge distillation for detrs. In: Proceedings of the IEEE/CVF Winter Conference on Computer Vision. pp. 7188–7197 (2026)
43. Langford, J., Li, L., Zhang, T.: Sparse online learning via truncated gradient. Advances in Neural Information Processing Systems **21** (2008)
44. Li, J., Li, F., Todorovic, S.: Efficient riemannian optimization on the stiefel manifold via the cayley transform. In: International Conference on Learning Representations (2020)
45. Lin, J., Zhu, C., Kneuertz, P.J., Bai, Y., Xue, Y.: When Models Learn to Ask Why: Adaptive causal reasoning for trustworthy medical Vision-Language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5556–5568 (2026)
46. Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., Bachem, O.: Challenging common assumptions in the unsupervised learning of disentangled representations. In: International Conference on Machine Learning. pp. 4114–4124. PMLR (2019)
47. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems **30** (2017)

48. Markou, E., Ajanthan, T., Gould, S.: Guiding neural collapse: Optimising towards the nearest simplex equiangular tight frame. Advances in Neural Information Processing Systems **37**, 35544–35573 (2024)
49. 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)
50. Pandey, A., Fanuel, M., Schreurs, J., Suykens, J.A.: Disentangled representation learning and generation with manifold optimization. Neural Computation **34**(10), 2009–2036 (2022)
51. 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)
52. Parikh, N., Boyd, S.: Proximal algorithms. Foundations and Trends in Optimization **1**(3), 127–239 (2014)
53. Parise, A., Namee, B.M.: This looks like that and that and that: Multi-objective optimization for diverse prototype learning. In: International Conference on Innovative Techniques and Applications of Artificial Intelligence. pp. 49–62 (2025)
54. Peng, Z., Ren, S., Gu, X., Yang, L., Wang, X., Sun, L.: ProtoTS: Learning hierarchical prototypes for explainable time series forecasting. In: International Conference on Learning Representations (2026)
55. 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)
56. Roy, S.K., Mhammedi, Z., Harandi, M.: Geometry aware constrained optimization techniques for deep learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4460–4469 (2018)
57. 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)
58. Saralajew, S., Rana, A., Villmann, T., Shaker, A.: A robust prototype-based network with interpretable rbf classifier foundations. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 20273–20282 (2025)
59. 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)
60. Shalev-Shwartz, S., Tewari, A.: Stochastic methods for l 1-regularized loss minimization. Journal of Machine Learning Research **12**, 1865–1892 (2011)
61. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)
62. Slack, D., Hilgard, S., Jia, E., Singh, S., Lakkaraju, H.: Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. pp. 180–186 (2020)
63. Sun, Y., He, Y., Jia, J., Wang, J., Ge, R., Wang, C., Xu, H.: Wdt-md: Wavelet diffusion transformers for microaneurysm detection in fundus images. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 9242–9250 (2026)
64. Ukai, Y., Hirakawa, T., Yamashita, T., Fujiyoshi, H.: This looks like it rather than that: ProtoKNN for similarity-based classifiers. In: International Conference on Learning Representations (2023)

65. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8769–8778 (2018)
66. 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)
67. Wang, C., Chen, Y., Liu, F., Liu, Y., McCarthy, D.J., Frazer, H., Carneiro, G.: Mixture of gaussian-distributed prototypes with generative modelling for interpretable and trustworthy image recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
68. Wang, C., Liu, F., Chen, Y., Frazer, H., Carneiro, G.: Cross-and intra-image prototypical learning for multi-label disease diagnosis and interpretation. IEEE Transactions on Medical Imaging (2025)
69. 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)
70. Wang, J., Chen, Z., Yuan, C., Li, B., Ma, W., Hu, W.: Hierarchical curriculum learning for no-reference image quality assessment. International Journal of Computer Vision **131**, 3074–3093 (2023)
71. Wang, J., Duan, Y., Tao, X., Xu, M., Lu, J.: Semantic perceptual image compression with a laplacian pyramid of convolutional networks. IEEE Transactions on Image Processing **30**, 4225–4237 (2021)
72. Wang, J., Yuan, C., Li, B., Deng, Y., Hu, W., Maybank, S.: Self-prior guided pixel adversarial networks for blind image inpainting. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(10), 12377–12393 (2023)
73. Wang, J., Zhang, K., Guo, J., Li, B., Pan, W., Hu, W.: Fine-grained and multi-instruction image restoration with attribute-aware attention learning. In: ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 9017–9021. IEEE (2026)
74. Wei, B., Zhu, Z.: ProtoLens: Advancing prototype learning for fine-grained interpretability in text classification. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 4503–4523 (2025)
75. Wen, Z., Yin, W.: A feasible method for optimization with orthogonality constraints. Mathematical Programming **142**(1), 397–434 (2013)
76. Yang, Z., Luo, T., Wang, D., Hu, Z., Gao, J., Wang, L.: Learning to navigate for fine-grained classification. In: Proceedings of the European Conference on Computer Vision. pp. 420–435 (2018)
77. Yaras, C., Wang, P., Zhu, Z., Balzano, L., Qu, Q.: Neural collapse with normalized features: A geometric analysis over the riemannian manifold. Advances in Neural Information Processing Systems **35**, 11547–11560 (2022)
78. Ye, K., Wong, K.S.W., Lim, L.H.: Optimization on flag manifolds. Mathematical Programming **194**(1), 621–660 (2022)
79. Yeh, C.K., Hsieh, C.Y., Suggala, A., Inouye, D.I., Ravikumar, P.K.: On the (in) fidelity and sensitivity of explanations. Advances in Neural Information Processing Systems **32** (2019)
80. Zhang, H., J Reddi, S., Sra, S.: Riemannian svrg: Fast stochastic optimization on riemannian manifolds. Advances in Neural Information Processing Systems **29** (2016)

81. Zhang, N., Donahue, J., Girshick, R., Darrell, T.: Part-based r-cnns for fine-grained category detection. In: European Conference on Computer Vision. pp. 834–849 (2014)
82. Zhao, Q., Li, Y., Sun, Q., Yan, Z.: Resilphase: Plug-and-play phase mapping and noise-resilient macro-trajectory extrapolation for diffusion acceleration. arXiv preprint arXiv:2606.26769 (2026)
83. Zhu, C., Lin, Y., Chen, S., Wang, Y., Lin, J.: MedEyes: Learning dynamic visual focus for medical progressive diagnosis. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 13916–13924 (2026)
84. Zhu, C., Lin, Y., Shao, J., Lin, J., Wang, Y.: Pathology-Aware Prototype Evolution via LLM-Driven Semantic Disambiguation for Multicenter Diabetic Retinopathy Diagnosis. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 9196–9205 (2025)
85. Zhu, C., Zeng, J., Jiang, J., Lin, J., Wang, Y.: Medsynapse-v: Bridging visual perception and clinical intuition via latent memory evolution. arXiv preprint arXiv:2604.26283 (2026)
86. Zhu, Z., Ding, T., Zhou, J., Li, X., You, C., Sulam, J., Qu, Q.: A geometric analysis of neural collapse with unconstrained features. Advances in Neural Information Processing Systems **34**, 29820–29834 (2021)