ProtoMappingNet: Interpretable Hierarchical Prototypes through Relational Prototype Mappings

Jaehun Park, Jongmin Lim, Soobin Cha, and Kwangsu Kim*

Sungkyunkwan University, Suwon, South Korea
{pk9403, jm.lim, soobin99, kim.kwangsu}@skku.edu

Abstract. Interpretable machine learning aims to develop models whose internal reasoning is transparent and aligned with human perception. While prototype-based networks provide intuitive part-based interpretations, most existing methods rely on flat or independently learned prototype sets, failing to capture the hierarchical and relational organization of visual concepts. Inspired by theories of structured part-whole perception, we introduce **ProtoMappingNet**, a prototype-based architecture that learns hierarchically structured visual concepts through learnable cross-layer prototype mappings. These mappings align representations across texture, part, and object levels, enabling coherent parent-child relations between prototypes. To quantitatively evaluate the learned hierarchy, we propose a relational evaluation framework consisting of Hierarchical Relational Consistency (HRC) and Hierarchical Spatial Stability (HSS), which measure cross-level relational consistency and spatial containment without requiring part annotations. Experiments show that ProtoMappingNet produces coherent multi-level explanations and substantially improves structural consistency through bidirectional relational modeling while maintaining competitive classification accuracy. Our code is available at `https://github.com/pk9403/ProtoMappingNet`

Keywords: prototype-based networks · interpretable deep learning · hierarchical prototypes · part-whole reasoning

1 Introduction

Deep learning has achieved remarkable success across many fields, yet its decision-making process often remains opaque, making it difficult to interpret and trust in practical applications. This limitation has motivated growing interest in model interpretability, which seeks not merely to expose internal model mechanisms but to provide explanations that are understandable and meaningful to humans. Among various approaches, prototype-based neural networks have emerged as a compelling family of inherently interpretable models [32]. By grounding decisions in comparisons with learned prototypical image regions, they provide intuitive explanations that resemble example-based human reasoning.

* Corresponding author.

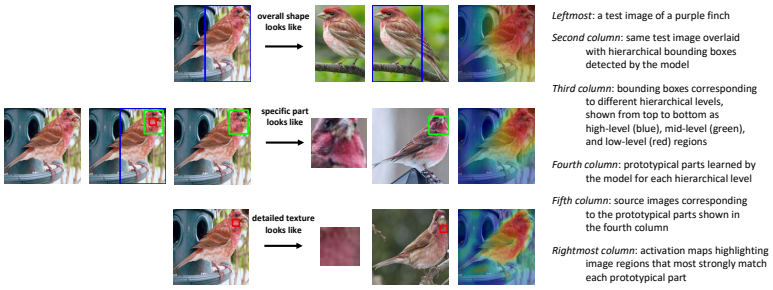


Fig. 1: Hierarchical prototype reasoning in ProtoMappingNet. The model explains its prediction by matching image regions with prototypical parts learned at different abstraction levels (object, part, and texture).

Despite their intuitive appeal, most existing prototype-based models do not explicitly capture the hierarchical and relational structures that play an important role in human visual perception. Cognitive theories suggest that human visual perception relies on structured part-whole representations [2, 12, 30]. Building on this perspective, recent work has argued that explicit part-whole representations are important for enabling more human-like visual understanding in neural networks [13]. In contrast, most prototype-based approaches operate at a single abstraction level or learn prototypes at multiple levels without explicitly connecting them. ProtoPNet and related models [3,25,26,29,33,42] operate primarily on high-level feature maps, localizing prototypical regions without modeling interactions across abstraction levels. Although recent extensions introduce multi-level or hierarchical prototype structures [39, 41], they do not explicitly learn compositional prototype-to-prototype dependencies across abstraction levels. Furthermore, although post-hoc methods such as VCC [19] recover cross-layer concept relationships, these relations are not explicitly involved in the model's predictive computation, which can limit their faithfulness to the underlying decision process [31]. Consequently, existing prototype-based models generally lack the explicit cross-level interactions needed to support cognitively motivated hierarchical reasoning, leaving an important gap in inherently interpretable visual recognition.

To bridge this cognitive and structural gap, we propose **ProtoMappingNet**, a hierarchical architecture that explicitly models relationships between visual concepts through learnable cross-layer prototype mappings and explains predictions through texture-, part-, and object-level prototypes, as illustrated in Fig. 1. These mappings encode parent-child dependencies by bidirectionally aligning prototype representations between adjacent abstraction levels. This bidirectional alignment encourages high-level prototypes to reflect coherent compositions of their lower-level counterparts, yielding an inherently interpretable hierarchical structure that promotes structural integrity. To quantitatively evaluate this cross-level alignment, we move beyond existing part-grounding metrics [16], which rely on ground-truth part annotations and evaluate prototypes

within a single layer. We instead introduce two annotation-free metrics: Hierarchical Relational Consistency (HRC), which measures agreement between learned cross-level mappings and inference-time prototype activations, and Hierarchical Spatial Stability (HSS), which evaluates spatial containment between mapped prototype activation regions.

The main contributions of this paper are summarized as follows:

- **Hierarchical Prototype Architecture**: We propose ProtoMappingNet, a prototype-based model that learns hierarchical visual concepts through bidirectional mappings between adjacent prototype levels, explicitly modeling parent-child relationships across abstraction levels.
- **Relational Evaluation Framework**: We introduce two complementary metrics—Hierarchical Relational Consistency (HRC) and Hierarchical Spatial Stability (HSS)—to evaluate the consistency of learned prototype mappings and the spatial containment of mapped activation regions without requiring ground-truth part annotations.
- **Hierarchical Interpretability Analysis**: We validate ProtoMappingNet on CUB-200-2011, Stanford Cars, and PartImageNet, demonstrating competitive classification performance and coherent relationships among texture-level, part-level, and object-level prototypes.

2 Related Works

Prototype-based Models Prototype-based explanation methods provide intuitive interpretability by grounding predictions in visually meaningful prototypical parts [18]. ProtoPNet [3] established this paradigm by learning class-specific prototype patches from the final feature maps and classifying inputs based on their similarity to these prototypes. TesNet [42] constructs a transparent Grassmann embedding space in which visual patches are linked to output categories through disentangled orthogonal concepts. ProtoPShare [34], ProtoPool [33], and SIDE [7] reduce prototype redundancy through class sharing or sparsity-inducing mechanisms. PIP-Net [26] learns prototypes using self-supervised objectives, whereas InfoDisent [37] derives interpretable concept representations post-hoc from pretrained models. ProtoConcepts [25] improves the semantic interpretation of individual prototypes by explaining each prototype through multiple representative visualizations. LucidPPN [29] separates prototypical reasoning into color and non-color branches to clarify whether decisions rely on color, shape, or texture. However, most existing prototype-based models operate primarily at a single abstraction level and therefore do not capture the multi-level relational structure associated with part-whole visual reasoning.

Multi-Level Prototype-based Models To overcome the limitations of single-level prototypes, several extensions introduce multi-level or structured variants. ProtoTree [27] organizes prototypes in a tree-structured decision hierarchy, yet its prototypes remain anchored to a single feature representation rather than

forming cross-level compositional relations. HPnet [9] aligns prototypes with hierarchical class taxonomies, but relies on a predefined semantic class taxonomy rather than learning compositional dependencies among prototypes across feature levels. Deformable ProtoPNet [6] and ProtoViT [24] improve the spatial flexibility of prototype matching, but primarily model within-level geometric variation rather than cross-level compositional dependencies. ProtoArgNet [1] combines prototypical-part activations into class-level super-prototypes using MLPs, emphasizing bottom-up aggregation rather than explicit prototype-to-prototype relations across feature levels. FPN-IAIA-BL [43] and MCPNet [41] learn prototypes from multiple feature scales, but they do not explicitly model prototype-to-prototype relations across abstraction levels. Meanwhile, VCC [19] explores cross-layer concept connections, and HVC [17] decomposes complex medical concepts into multi-level visual representations while modeling their relations. However, both operate in a post-hoc manner on pretrained networks, and thus do not enforce hierarchical relational consistency during learning nor guarantee alignment with the model's internal reasoning process. In contrast, our approach jointly learns bidirectional relational mappings between adjacent prototype levels.

Some studies [8, 21, 39] leverage hyperbolic geometry to embed features in a space naturally suited for representing hierarchical structures [28]. These approaches encode hierarchical organization through geometric relations in a latent space, rather than through explicit prototype-to-prototype mappings between visual abstraction levels. They also do not explicitly evaluate whether hierarchical prototype relations correspond to spatial containment across feature levels. In contrast, our method learns bidirectional mappings between prototypes at different visual abstraction levels and evaluates whether these relations are reflected in their spatial activation patterns.

Hierarchical and Part-Whole Representations Decades of cognitive science research suggest that human visual perception relies on structured part-whole representations [2,12,30]. Related theories of predictive processing emphasize bidirectional interactions between bottom-up sensory evidence and top-down predictions [4]. Motivated by similar principles, Hinton [13] proposes GLOM as a framework for representing dynamic part-whole hierarchies in neural networks, arguing that such representations are important for more human-like visual understanding. However, integrating such compositional structures into inherently interpretable deep models remains challenging. Compositional frameworks such as Capsule Networks [14, 35] model part-whole dependencies through routing between capsule layers, but their explanations remain encoded in latent representations rather than grounded in exemplar-based visual evidence. Conversely, prototype-based networks provide visually grounded, patch-based explanations, but generally do not connect these explanations through explicit part-whole relations across feature levels. To the best of our knowledge, no prior prototype-based model jointly learns bidirectional cross-level prototype mappings and evaluates their relational and spatial consistency without part annotations.

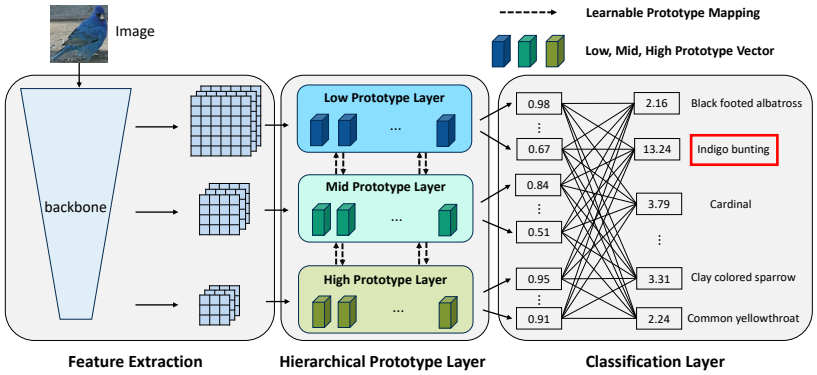


Fig. 2: Architecture of ProtoMappingNet.

3 Method

3.1 Overall Architecture

To operationalize the hierarchical part-whole structure discussed earlier, we propose **ProtoMappingNet**, a hierarchical prototype-based architecture organized across three levels of abstraction: low, mid, and high. Low-level prototypes tend to capture localized textures and primitive edges, mid-level prototypes correspond to semantically meaningful object parts, and high-level prototypes represent holistic object concepts. While each level learns prototypical representations at a different granularity, learnable cross-level mappings explicitly model compositional dependencies between adjacent levels. These mappings encourage higher-level prototypes to emerge as structured compositions of lower-level prototypes, enabling the network to maintain coherent hierarchical reasoning.

Figure 2 illustrates the overall architecture of ProtoMappingNet, a hierarchical prototype-based model designed to learn a consistent part-whole concept space. The network consists of three main components: a convolutional backbone, a hierarchical prototype layer, and a classification layer. Given an input image x, the backbone extracts multi-level feature maps $f_l(x) \in \mathbb{R}^{H_l \times W_l \times D_l}$, where $l \in \{\text{low}, \text{mid}, \text{high}\}$ denotes the abstraction level, and H_l, W_l, D_l represent the spatial and channel dimensions. The hierarchical prototype layer maintains sets of learnable prototype vectors $\mathcal{P} = \{\mathcal{P}_{\text{low}}, \mathcal{P}_{\text{mid}}, \mathcal{P}_{\text{high}}\}$, together with learnable prototype mappings M. At each level l, the model learns $m_l = C\, k_l$ prototypes, allocating k_l class-specific prototypes per class. Each prototype $p \in \mathcal{P}_l$ has spatial dimensions $p \in \mathbb{R}^{H_l^{(p)} \times W_l^{(p)} \times D_l}$, with $H_l^{(p)} \leq H_l$ and $W_l^{(p)} \leq W_l$, representing localized conceptual patches at different granularities. To establish a structured part–whole hierarchy, the prototype mappings M operate as cross-layer semantic constraints. Specifically, bottom-up mappings $(M_{\text{low}\to\text{mid}}, M_{\text{mid}\to\text{high}})$ and top-down mappings $(M_{\text{high}\to\text{mid}}, M_{\text{mid}\to\text{low}})$ facilitate predictive alignment between adjacent abstraction levels.

Building upon the extracted feature representations $f_l(x)$ for each level $l \in \{\text{low}, \text{mid}, \text{high}\}$, the hierarchical prototype layer computes dense similarity maps between these features and their corresponding prototype sets $\mathcal{P}_l$. Specifically, for each prototype $p_j^l \in \mathcal{P}_l$, the cosine similarity between p_j^l and all local patches of $f_l(x)$ sharing the same channel dimension is computed as

$$s_j^l(u,v) = \frac{\langle f_l(x)_{u,v},\, p_j^l \rangle}{\|f_l(x)_{u,v}\|_2 \, \|p_j^l\|_2}, \tag{1}$$

where $f_l(x)_{u,v}$ denotes the local feature patch starting at spatial location (u, v) with the same spatial and channel dimensions as p_j^l, and the inner product and norms are computed over all spatial and channel dimensions. This operation produces spatial similarity maps $S_l = \{s_j^l\}_{j=1}^{m_l}$, where $s_j^l \in \mathbb{R}^{\bar{H}_l \times \bar{W}_l}$, $\bar{H}_l = H_l - H_l^{(p)} + 1$, and $\bar{W}_l = W_l - W_l^{(p)} + 1$.

To aggregate these localized responses, ProtoMappingNet applies global max pooling over the spatial dimensions of each similarity map $s_j^l(u, v)$, extracting a single activation per prototype:

$$a_j^l = \max_{(u,v)} s_j^l(u,v), \tag{2}$$

where a_j^l denotes the activation of prototype p_j^l at level l. The activations of all prototypes at level l are then concatenated into an activation vector $a_l \in \mathbb{R}^{m_l}$, $a_l = [a_1^l, a_2^l, \ldots, a_{m_l}^l]^\top$.

To preserve prototype interpretability, each level maintains an independent, bias-free classification head. These heads are parameterized by weight matrices $W_l \in \mathbb{R}^{C \times m_l}$, which are initialized with a binary class-identity matrix to allocate specific prototypes to their designated classes. The class probabilities at level l are computed as:

$$\hat{y}_l = \text{softmax}(\gamma_l W_l a_l), \tag{3}$$

where γ_l is a fixed scaling factor controlling the sharpness of the class distribution. During inference, the level-specific logits are aggregated using the same normalized level weights employed for the classification losses.

3.2 Intra-layer Prototype Learning

ProtoMappingNet learns class-discriminative prototypes at each level of the hierarchy. Unlike flat prototype models, our formulation explicitly organizes prototypes across multiple abstraction levels, enabling fine-to-coarse concept representations. Following ProtoPNet [3], we use a combination of cluster loss ($\mathcal{L}_{\text{clst}}^l$) and separation loss ($\mathcal{L}_{\text{sep}}^l$) at each level. In addition, we incorporate the orthogonality loss ($\mathcal{L}_{\text{ortho}}^l$) from TesNet [42] to enhance intra-class diversity among prototypes. Unlike ProtoPNet's Euclidean formulation, our cosine similarity-based objectives emphasize directional alignment between input features and their corresponding prototypes, providing a scale-invariant similarity measure well suited for multi-level feature spaces. Let $P_c^l \in \mathbb{R}^{k_l \times d_l^{(p)}}$ denote the matrix whose rows

are the vectorized and ℓ_2-normalized prototypes assigned to class c at level l, where $d_l^{(p)} = H_l^{(p)} W_l^{(p)} D_l$. Given a training set of n input images $\{x_i, y_i\}_{i=1}^n$, the intra-layer objectives are defined as follows:

$$\mathcal{L}_{\text{clst}}^l = 1 - \frac{1}{n} \sum_{i=1}^{n} \max_{p_j^l \in \mathcal{P}_l^{y_i}} a_j^l(x_i), \tag{4}$$

$$\mathcal{L}_{\text{sep}}^l = \frac{1}{n} \sum_{i=1}^{n} \max_{p_j^l \notin \mathcal{P}_l^{y_i}} \max\big(a_j^l(x_i),\, 0\big), \tag{5}$$

$$\mathcal{L}_{\text{ortho}}^l = \frac{1}{C} \sum_{c=1}^{C} \left\| P_c^l (P_c^l)^\top - I \right\|_F^2, \tag{6}$$

Overall intra-layer objective. The overall intra-layer objective is a weighted combination of the losses described above:

$$\mathcal{L}_{\text{intra}}^l = \lambda_{\text{CE}}^l \mathcal{L}_{\text{CE}}^l + \lambda_{\text{clst}}^l \mathcal{L}_{\text{clst}}^l + \lambda_{\text{sep}}^l \mathcal{L}_{\text{sep}}^l + \lambda_{\text{ortho}}^l \mathcal{L}_{\text{ortho}}^l. \tag{7}$$

This objective enables each hierarchical level to independently learn discriminative and semantically meaningful prototypes, providing a robust foundation for the cross-layer relational constraints introduced in the next section.

3.3 Inter-layer Relational Learning

While each level independently learns class-discriminative prototypes, ProtoMappingNet additionally models cross-level relations through learnable prototype mappings. These mappings connect adjacent levels in both bottom-up and top-down directions. The bottom-up mappings project lower-level prototypes into the representation space of higher-level prototypes, encouraging class-consistent alignment across levels. Conversely, the top-down mappings project higher-level prototypes back to lower levels, promoting mutual consistency between prototype representations. Together, this bidirectional interaction encourages hierarchical consistency across prototype levels.

Cross-layer Alignment losses To align prototypes across hierarchical levels, we introduce cross-layer cluster and separation losses. For any two adjacent levels l and l', the prototypes from the source level $\mathcal{P}_l$ are projected into the prototype space of the target level $\mathcal{P}_{l'}$ through a learnable projection head $M_{l \to l'}$:

$$\tilde{p}_i^{l \to l'} = M_{l \to l'}(p_i^l), \tag{8}$$

where $p_i^l \in \mathcal{P}_l$, and the projection head $M_{l \to l'}$ consists of a small convolutional projection head. The similarity is computed as:

$$S_{l \to l'}(i, j) = \frac{\langle \tilde{p}_i^{l \to l'}, p_j^{l'} \rangle}{\|\tilde{p}_i^{l \to l'}\|_2 \, \|p_j^{l'}\|_2}, \tag{9}$$

where $S_{l\to l'}(i,j) \in [-1,1]$ denotes the similarity between the i-th projected prototype from level l and the j-th prototype at level l'.

Let c_i^l and $c_j^{l'}$ denote the class labels associated with prototypes p_i^l and $p_j^{l'}$, respectively. The cross-layer cluster loss aligns each projected prototype with the most similar same-class prototype in the adjacent level, encouraging high similarity between semantically corresponding prototypes. Conversely, the cross-layer separation loss penalizes positive similarity with different-class prototypes in the target level. The losses are formulated as:

$$\mathcal{L}_{\text{clst}}^{l\to l'} = 1 - \frac{1}{m_l}\sum_{i=1}^{m_l} \max_{j:c_j^{l'}=c_i^l} \max\big(S_{l\to l'}(i,j),\,0\big). \tag{10}$$

$$\mathcal{L}_{\text{sep}}^{l\to l'} = \frac{1}{m_l}\sum_{i=1}^{m_l} \max_{j:c_j^{l'}\neq c_i^l} \max\big(S_{l\to l'}(i,j),\,0\big). \tag{11}$$

Overall Cross-layer Objective The cross-layer alignment losses are applied to all bottom-up and top-down projection directions. By optimizing these mappings jointly, the network is encouraged to maintain bidirectional consistency across adjacent abstraction levels. The overall inter-layer objective is defined as:

$$\mathcal{L}_{\text{inter}} = \sum_{(l,l')\in\mathcal{E}} \big(\lambda_{\text{clst}}^{l\to l'}\mathcal{L}_{\text{clst}}^{l\to l'} + \lambda_{\text{sep}}^{l\to l'}\mathcal{L}_{\text{sep}}^{l\to l'}\big), \tag{12}$$

where $\mathcal{E} = \{(\text{low},\text{mid}),(\text{mid},\text{high}),(\text{high},\text{mid}),(\text{mid},\text{low})\}$ denotes the set of interacting adjacent level pairs. This relational learning complements the intra-layer objectives by encouraging consistent relations between prototypes across hierarchical levels.

3.4 Training Objective and Optimization

To prevent prototype collapse and ensure semantically meaningful representations, ProtoMappingNet follows the established multi-stage training paradigm of ProtoPNet [3], extending it with cross-layer relational objectives for hierarchical concept learning. During the warm-up stage, the backbone and prototype mapping modules are frozen, while the add-on layers and prototype vectors are optimized using the level-specific intra-layer losses $\mathcal{L}_{\text{intra}}^l$. In the joint stage, the backbone, prototype layers, and mapping modules are optimized together using both intra-layer and cross-layer objectives. In the push stage, each prototype vector is replaced with its nearest feature patch in the training set to preserve interpretability. After the push step, the backbone and prototype vectors are frozen, and only the classifier parameters and mapping weights are optimized using: $\mathcal{L}_{\text{fine}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{inter}}\mathcal{L}_{\text{inter}} + \lambda_1\|W\|_1$. Further details regarding the multi-stage training schedules and hyperparameter configurations are provided in the supplementary material.

4 Experiment

4.1 Datasets and Implementation Details

Datasets. We evaluate ProtoMappingNet on three benchmarks: CUB-200-2011 [40], Stanford Cars [20], and PartImageNet [10]. For all datasets, the model is trained using only image-level class labels without part annotations, allowing hierarchical prototypes to be learned under weak supervision.

Implementation Details. We implement ProtoMappingNet on multiple backbone architectures, including ResNet [11], DenseNet [15], and ConvNeXt [23]. For ResNet-50, we use a model pretrained on the iNaturalist2017 [38], while all other backbones are initialized with ImageNet-pretrained weights [5]. Following our hierarchical design, feature maps are extracted from three depths of the backbone network. Specifically, we select layers with spatial resolutions of approximately 1/4 (Low), 1/16 (Mid), and 1/32 (High) of the input image. These resolutions roughly correspond to low-level textures, mid-level parts, and high-level object structures. The spatial dimensions of the prototype kernels are set to 5×5, 3×3, and 5×5 for the low-, mid-, and high-level prototypes, respectively. We allocate five prototypes per class at each hierarchical level.

4.2 Evaluation Metrics

While Huang et al. [16] and Sacha et al. [36] proposed metrics for prototype evaluation, they primarily focus on single-layer part grounding or spatial misalignment. Such measures are limited for hierarchical models, as they do not directly evaluate relationships across abstraction levels. To address this limitation, we introduce annotation-free metrics that assess the intrinsic structural consistency of the learned hierarchy. Specifically, **Hierarchical Relational Consistency (HRC)** quantifies the agreement between learned cross-level prototype mappings and inference-time activation patterns, while **Hierarchical Spatial Stability (HSS)** evaluates the spatial containment between mapped prototype activation regions. For robustness evaluation, we additionally adopt the IoU-based stability score introduced in the appendix of [16].

Hierarchical Relational Consistency (HRC). HRC quantifies the agreement between learned cross-level prototype mappings $\mathcal{M}$ and inference-time activation patterns. Let $\mathcal{P}_s$ and $\mathcal{P}_t$ denote the source and target prototype sets, respectively. For each source prototype p_i^s, we define the learned Top-k static mapping $\mathcal{M}_{s\to t}^k(p_i^s) \subseteq \mathcal{P}_t$ by selecting the k target prototypes with the highest projected cosine similarities $S_{s\to t}(i,j)$ defined in Eq. (9). For test images x belonging to the class associated with p_i^s, we define the confident activation set $\mathcal{X}_i = \{x \mid a_i^s(x) \geq \tau_{\text{act}}\}$. We set $\tau_{\text{act}} = 0.8$. Let $\mathcal{P}_s^* = \{p_i^s \mid |\mathcal{X}_i| > 0\}$ denote the set of prototypes that activate on at least one test image. The hierarchical relational consistency score is then defined as:

$$\text{HRC}_{s\to t}@k = \frac{1}{|\mathcal{P}_s^*|} \sum_{p_i^s \in \mathcal{P}_s^*} \frac{1}{|\mathcal{X}_i|} \sum_{x \in \mathcal{X}_i} \mathbf{1}\left[\arg\max_{p_j^t \in \mathcal{P}_t} a_j^t(x) \in \mathcal{M}_{s\to t}^k(p_i^s)\right], \tag{13}$$

Table 1: Classification accuracy (%) on CUB-200-2011. * denotes models utilizing iNaturalist pretraining.

Method	ResNet34	ResNet50	DenseNet161	ConvNeXt-Tiny
ProtoPNet	79.2 ± 0.1	-	80.1 ± 0.3	-
ProtoTree	-	82.2 ± 0.7 *	-	-
ProtoPool	80.3 ± 0.2	85.5 ± 0.1 *	-	-
Deformable	76.8	86.4 *	81.2	-
TesNet	82.8 ± 0.1	-	84.6 ± 0.3	-
EvalProtoPNet	84.0	87.1 *	86.5	-
PIPNet	-	82.0 ± 0.3 *	-	84.3 ± 0.2
MCPNet	-	80.2	-	83.5
HAPPI (ProtoPNet)	-	82.4	-	-
ProtoConcepts	80.4 ± 0.1	85.2 ± 0.2 *	81.5 ± 0.6	-
LucidPPN	-	75.5 ± 1.1	-	81.5 ± 0.4
ProtoArgNet	-	85.4 ± 0.2	-	-
ProtoMappingNet (Ours)	83.1 ± 0.2	86.5 ± 0.2 *	84.5 ± 0.3	86.1 ± 0.3

By conditioning the evaluation on the confident activation set $\mathcal{X}_i$, the metric focuses on reliable prototype activations rather than random background responses.

Hierarchical Spatial Stability (HSS). HSS quantifies the spatial containment relationship between the activation regions of mapped prototypes. For each mapped pair (p_i^s, p_j^t) with $p_j^t \in \mathcal{M}_{s\to t}^k(p_i^s)$, we consider the subset of images $\mathcal{X}_{i,j} = \{x \in \mathcal{X}_i \mid |r_j^t(x)| > 0\}$, where $r_i^l(x)$ denotes the binary activation region obtained by thresholding the min-max normalized similarity map of prototype p_i^l (upsampled to the input resolution) with a layer-specific threshold τ_l. We set τ_l to 0.95, 0.80, and 0.50 for the low-, mid-, and high-level prototypes, respectively, to account for the increasing receptive field size across hierarchical feature layers. A sensitivity analysis demonstrating the robustness of HSS to different activation thresholds is provided in the supplementary material. Let $\mathcal{P}_{s\to t}^* = \{(p_i^s, p_j^t) \mid |\mathcal{X}_{i,j}| > 0\}$. The hierarchical spatial stability score is then defined as:

$$\mathrm{HSS}_{s\to t}@k = \frac{1}{|\mathcal{P}_{s\to t}^*|} \sum_{(p_i^s,p_j^t)\in\mathcal{P}_{s\to t}^*} \frac{1}{|\mathcal{X}_{i,j}|} \sum_{x\in\mathcal{X}_{i,j}} \frac{|r_i^s(x) \cap r_j^t(x)|}{\min(|r_i^s(x)|, |r_j^t(x)|)}. \tag{14}$$

Unlike IoU, the IoMin formulation measures whether the smaller activation region is contained within the larger one, making it suitable for evaluating spatial nesting between mapped prototypes at different abstraction levels.

4.3 Quantitative Results

In this section, we empirically evaluate ProtoMappingNet from two perspectives: (i) predictive performance, verifying that the proposed hierarchical architecture maintains competitive accuracy, and (ii) structural evaluation, assessing whether the model learns more coherent hierarchical relations than baseline architectures.

Table 2: Classification accuracy (%) on Stanford Cars.

Method	ResNet34	ResNet50	DenseNet121	ConvNeXt-Tiny
ProtoPNet	86.1 ± 0.1	-	86.8 ± 0.1	-
ProtoTree	-	86.6 ± 0.2	-	-
ProtoPool	89.3 ± 0.1	88.9 ± 0.1	-	-
TesNet	90.9 ± 0.2	-	91.9 ± 0.3	-
EvalProtoPNet	92.0	-	92.4	-
PIPNet	-	86.5 ± 0.3	-	88.2 ± 0.5
ProtoConcepts	89.4 ± 0.5	-	87.1 ± 0.4	-
LucidPPN	-	89.0 ± 0.3	-	91.6 ± 0.2
ProtoArgNet	-	89.3 ± 0.3	-	-
ProtoMappingNet (Ours)	88.5 ± 0.2	89.9 ± 0.2	88.9 ± 0.1	92.4 ± 0.1

Classification Accuracy. We evaluate classification accuracy across multiple datasets and backbone architectures. As shown in Tables 1 and 2, ProtoMappingNet achieves competitive results compared with a broad range of representative prototype-based models, indicating that cross-layer relational modeling can be introduced without degrading predictive performance. Unless otherwise specified, ProtoMappingNet results are reported as the mean and standard deviation over five random seeds.

To better understand how predictive performance evolves across the hierarchy, we analyze the layer-wise classification accuracy on the CUB-200-2011 dataset using a ResNet-50 backbone. The individual accuracies are $53.3 \pm 1.1\%$ for the low level, $82.2 \pm 0.3\%$ for the mid level, and $86.1 \pm 0.2\%$ for the high level. Combining predictions across all three levels further improves the final accuracy to $86.5 \pm 0.2\%$. This trend suggests that lower-level visual cues provide complementary information to high-level object representations, supporting the effectiveness of multi-level information fusion. Similar trends were observed across other datasets and backbones (see Supplementary Material).

Evaluation of Hierarchical Relational Integrity. Although several prototype models incorporate hierarchical structures, existing methods [9, 27, 41, 43] do not explicitly model cross-level relational constraints during training. Consequently, directly applying HRC/HSS to these architectures would not provide a meaningful assessment of cross-level relational consistency, as the resulting scores would primarily reflect incidental spatial overlap among independently learned concepts rather than learned prototype relationships. To isolate the effect of relational modeling while avoiding these architectural limitations, we construct a strictly controlled *ProtoHierarchy* baseline that retains the same hierarchical capacity but removes all cross-level mappings. We compare this baseline with three relational variants: *Bottom-up only*, *Top-down only*, and our full *Bidirectional* model. For models without explicit mapping modules, source prototypes cannot be projected into the target prototype space. In such cases, prototype vectors are summarized using global average pooling over spatial dimensions to obtain comparable prototype embeddings.

Table 3: Evaluation of hierarchical relational integrity. Comparison of our ProtoMappingNet with relation-aware variants (Bottom-up, Top-down) and the ProtoHierarchy baseline on CUB-200-2011 (ResNet-50). Mean ± std. over five random runs.

Model		ProtoHier.	Bottom-up	Top-down	Ours (Bi)
Accuracy (%)		86.8 ± 0.2	86.8 ± 0.3	86.8 ± 0.1	86.5 ± 0.2
Direction	**Mapping**	***HRC@1 (%)*** ↑			
Bottom-up	$Low \rightarrow Mid$	0.05 ± 0.03	25.13 ± 1.68	0.05 ± 0.09	**25.28** ± 1.74
Top-down	$Mid \rightarrow Low$	0.23 ± 0.33	0.04 ± 0.02	**3.63** ± 0.60	3.20 ± 0.31
Bottom-up	$Mid \rightarrow High$	0.35 ± 0.30	26.46 ± 0.20	0.25 ± 0.18	**26.61** ± 0.61
Top-down	$High \rightarrow Mid$	0.05 ± 0.08	0.06 ± 0.14	17.19 ± 1.33	**17.98** ± 1.52
Overall HRC		0.17 ± 0.17	12.92 ± 0.43	5.28 ± 0.40	**18.27** ± 0.49
Direction	**Mapping**	***HSS@1 (%)*** ↑			
Bottom-up	$Low \rightarrow Mid$	14.64 ± 3.64	**49.52** ± 1.97	15.41 ± 1.80	48.88 ± 1.72
Top-down	$Mid \rightarrow Low$	12.73 ± 2.16	11.10 ± 1.66	**35.91** ± 1.54	32.16 ± 2.35
Bottom-up	$Mid \rightarrow High$	48.04 ± 3.82	93.36 ± 0.56	47.26 ± 2.72	**94.29** ± 0.47
Top-down	$High \rightarrow Mid$	63.49 ± 2.17	63.66 ± 1.12	**86.39** ± 1.91	86.11 ± 1.86
Overall HSS		34.72 ± 1.35	54.41 ± 0.57	46.24 ± 0.95	**65.36** ± 1.16
Layer		***Stability (%)*** ↑			
low		14.99 ± 1.47	14.18 ± 0.86	14.05 ± 0.62	14.22 ± 0.60
mid		60.76 ± 1.57	60.47 ± 0.81	59.56 ± 0.78	60.23 ± 1.03
high		88.70 ± 0.97	86.84 ± 1.16	87.40 ± 0.73	87.54 ± 1.27
Overall Stability		54.81 ± 1.03	53.83 ± 0.68	53.67 ± 0.28	54.00 ± 0.67

As shown in Table 3, ProtoMappingNet consistently improves cross-level relational consistency over the ProtoHierarchy baseline and single-directional variants, achieving the highest HRC in three of four mapping directions and the best overall HRC. Interestingly, the $Mid \rightarrow Low$ mapping consistently exhibits lower HRC scores across all models. This behavior is expected because low-level prototypes typically capture distributed texture patterns that may appear across multiple object parts. HSS results further confirm stronger spatial containment between hierarchical prototypes, with particularly strong improvements observed in the $Mid \rightarrow High$ mapping. Importantly, these structural gains are achieved while maintaining comparable classification accuracy. In addition, the Overall Stability scores remain comparable across all variants, indicating that hierarchical relational constraints improve structural consistency without materially reducing prototype robustness under input perturbations.

Generalization Beyond Fine-Grained Datasets. To evaluate whether ProtoMappingNet generalizes beyond fine-grained recognition settings, we conduct additional experiments on the PartImageNet dataset using a ConvNeXt-Tiny backbone. Across three random seeds, ProtoMappingNet achieves an accuracy of $86.2 \pm 0.4\%$, outperforming PIPNet (85.0%) and LucidPPN (84.1%). In addition to classification performance, we analyze the relational metrics HRC and HSS

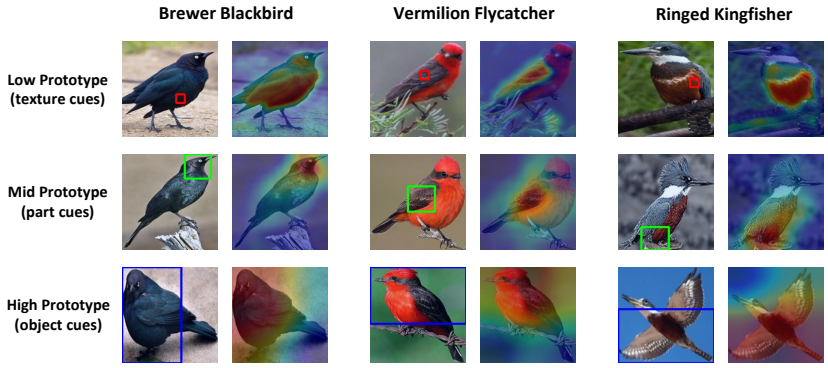


Fig. 3: Hierarchical prototype visualization in ProtoMappingNet. Rows correspond to abstraction levels. Columns are arranged in pairs, where each pair shows a prototype patch (odd column) and its activation heatmap on the input image (even column) for the same class.

Table 4: Performance and parameter comparison on Transformer-based extensions.

Method	Backbone (# Params)	CUB-200-2011	Stanford Cars
ProtoViT	DeiT-S (22M)	85.37 ± 0.13	**91.84** ± 0.30
Ours	Swin-T (28M)	**87.19** ± 0.14	90.97 ± 0.08

under different structural configurations (ProtoHierarchy, Bottom-up, and Top-down). The results show trends consistent with those observed on fine-grained datasets, suggesting that the relational structure can extend beyond fine-grained recognition settings. Detailed experimental results are provided in the supplementary material.

Extension to Transformer Architectures. To demonstrate that ProtoMappingNet is not restricted to CNNs, we extend it to a hierarchical vision Transformer using a Swin-Transformer (Swin-T) backbone [22]. As summarized in Table 4, our framework achieves competitive performance on fine-grained benchmarks. These results demonstrate that our cross-layer relational modeling is compatible with Transformer architectures and suggest its potential extension to vision-language models with hierarchical visual representations.

4.4 Qualitative Results

In this section, we visually demonstrate that ProtoMappingNet not only learns semantically meaningful features across different granularities but also maintains consistent hierarchical spatial relationships during inference. Additional qualitative examples and visualizations are provided in the supplementary material.

Hierarchical Prototype Visualization. We visualize the learned hierarchical prototypes in Figure 3, showing representative image regions that strongly

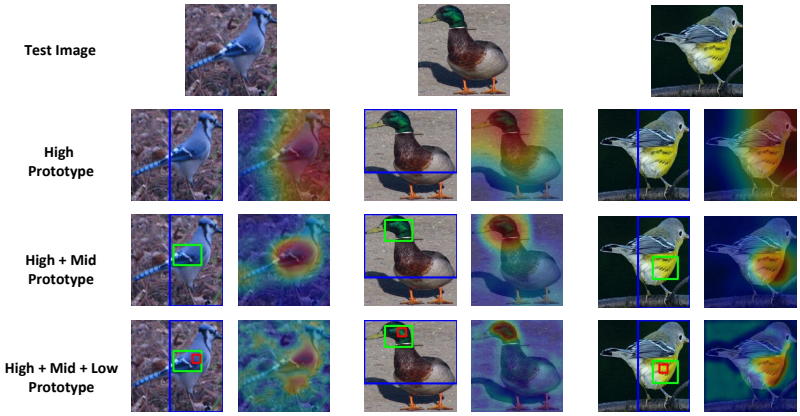
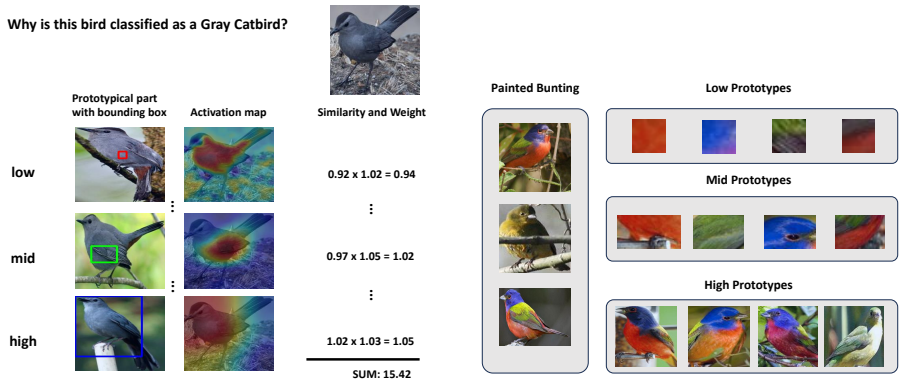


Fig. 4: Hierarchical spatial containment in ProtoMappingNet. Low-level prototype activations (red) are spatially contained within mid-level activations (green), which in turn lie within high-level activations (blue). Activation maps illustrate the prototype responses at each hierarchical level.



(a) Local explanation **(b)** Global explanation

Fig. 5: Examples of local and global explanations.

activate each prototype together with their activation heatmaps. Specifically, low-level prototypes capture fine-grained textures from 20×20 patches around the maximal activation location, mid-level prototypes highlight semantic parts (e.g., heads or wings) using the top 5% activations, and high-level prototypes capture object-level structures (e.g., pose and shape) from the top 50% activation regions. Visualization thresholds are architectural choices grounded in the network's spatial design; further justification is provided in the supplementary material. This progression from texture- to part- to object-level cues illustrates that ProtoMappingNet organizes visual concepts across abstraction levels in an interpretable hierarchy.

Hierarchical Spatial Containment. ProtoMappingNet captures spatial part-whole relations between prototypes across hierarchical levels. Figure 4 visualizes this spatial containment structure on representative test images. The high-level prototype establishes a global semantic anchor (blue box) over the most discriminative object region. Within this context, the mid-level prototype (green box) localizes semantically meaningful parts, while the low-level prototype (red box) focuses on fine-grained texture cues. These regions form a consistent nested structure (Blue $\supset$ Green $\supset$ Red), indicating a spatial hierarchy across abstraction levels.

Local and Global Interpretations. Local explanations illustrate how the hierarchical prototype evidence contributes to individual predictions, as shown in Figure 5a. For a given test image, we show the most activated prototypes at each hierarchical level together with their activation maps and contribution scores. The contributions are computed by combining the prototype similarity with its classification weight. This visualization highlights which texture-, part-, and object-level cues support the model's decision. Global explanations summarize the class-level concept dictionary learned by ProtoMappingNet, as shown in Figure 5b. For each class, we visualize representative prototypes at different hierarchical levels. Low-level prototypes capture color and texture patterns, mid-level prototypes represent discriminative object parts, and high-level prototypes correspond to holistic object configurations. Together, these prototypes provide an interpretable summary of the visual concepts representing each class.

5 Conclusion

We introduced ProtoMappingNet, a prototype-based architecture that explicitly models part-whole relationships through a hierarchical prototype structure. By learning cross-layer prototype mappings, the model connects texture-, part-, and object-level concepts, enabling coherent hierarchical explanations. Experiments demonstrate that ProtoMappingNet improves cross-level relational and spatial consistency while maintaining competitive classification performance. Future work will explore sharing lower-level prototypes across classes to improve efficiency and scalability.

Acknowledgements

This work was partly supported by Korea Internet & Security Agency (KISA) grant funded by the Korea government (PIPC) (No. RS-2026-25530348, Development of Evaluation Technologies for the Safety and Utility of Unstructured Synthetic Data) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2019-II190421, Artificial Intelligence Graduate School Program (Sungkyunkwan University)).

References

1. 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)
2. Biederman, I.: Recognition-by-components: a theory of human image understanding. Psychological review **94**(2), 115 (1987)
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. Clark, A.: Whatever next? predictive brains, situated agents, and the future of cognitive science. Behavioral and brain sciences **36**(3), 181–204 (2013)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
6. 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)
7. Dubovik, V., Struski, Ł., Tabor, J., Rymarczyk, D.: Side: Sparse information disentanglement for explainable artificial intelligence. arXiv preprint arXiv:2507.19321 (2025)
8. Gulshad, S., Long, T., van Noord, N.: Hierarchical explanations for video action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 3703–3708 (June 2023)
9. Hase, P., Chen, C., Li, O., Rudin, C.: Interpretable image recognition with hierarchical prototypes. In: Proceedings of the AAAI conference on human computation and crowdsourcing. vol. 7, pp. 32–40 (2019)
10. He, J., Yang, S., Yang, S., Kortylewski, A., Yuan, X., Chen, J.N., Liu, S., Yang, C., Yu, Q., Yuille, A.: Partimagenet: A large, high-quality dataset of parts. In: European Conference on Computer Vision. pp. 128–145. Springer (2022)
11. 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)
12. Hinton, G.: Some demonstrations of the effects of structural descriptions in mental imagery. Cognitive Science **3**(3), 231–250 (1979)
13. Hinton, G.: How to represent part-whole hierarchies in a neural network. Neural Computation **35**(3), 413–452 (2023)
14. Hinton, G.E., Sabour, S., Frosst, N.: Matrix capsules with em routing. In: International conference on learning representations (2018)
15. 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)
16. 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)
17. Khaleel, M., Tavanapong, W., Wong, J., Oh, J., De Groen, P.: Hierarchical visual concept interpretation for medical image classification. In: 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS). pp. 25–30. IEEE (2021)

18. Kim, S.S., Watkins, E.A., Russakovsky, O., Fong, R., Monroy-Hernández, A.: "Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI). pp. 1–17 (2023)
19. Kowal, M., Wildes, R.P., Derpanis, K.G.: Visual concept connectome (vcc): Open world concept discovery and their interlayer connections in deep models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10895–10905 (June 2024)
20. 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)
21. Lebedeva, I., Bah, M.J., Li, T.: Interpretable image recognition in hyperbolic space. In: 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). pp. 643–650. IEEE (2023)
22. 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)
23. 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)
24. Ma, C., Donnelly, J., Liu, W., Vosoughi, S., Rudin, C., Chen, C.: Interpretable image classification with adaptive prototype-based vision transformers. Advances in Neural Information Processing Systems **37**, 41447–41493 (2024)
25. Ma, C., Zhao, B., Chen, C., Rudin, C.: This looks like those: Illuminating prototypical concepts using multiple visualizations. Advances in Neural Information Processing Systems **36**, 39212–39235 (2023)
26. 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)
27. 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)
28. Nickel, M., Kiela, D.: Poincaré embeddings for learning hierarchical representations. Advances in neural information processing systems **30** (2017)
29. 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, accessed: 2026-06-30
30. Palmer, S.E.: Hierarchical structure in perceptual representation. Cognitive psychology **9**(4), 441–474 (1977)
31. 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)
32. Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., Zhong, C.: Interpretable machine learning: Fundamental principles and 10 grand challenges. Statistics Surveys **16**(none), 1 – 85 (2022). https://doi.org/10.1214/21-SS133

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. Sabour, S., Frosst, N., Hinton, G.E.: Dynamic routing between capsules. Advances in neural information processing systems **30** (2017)
36. Sacha, M., Jura, B., Rymarczyk, D., Struski, Ł., Tabor, J., Zieliński, B.: Interpretability benchmark for evaluating spatial misalignment of prototypical parts explanations. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 21563–21573 (2024)
37. Struski, Ł., Rymarczyk, D., Tabor, J.: Infodisent: Explainability of image classification models by information disentanglement. arXiv preprint arXiv:2409.10329 (2024)
38. 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)
39. Vaseli, H., Wu, V., Kondori, N., To, N.N.M., Fung, A., Gu, A.N., Abolmaesumi, P.: Happi: Hyperbolic hierarchical part prototypes for image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 696–705 (October 2025)
40. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011)
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, 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)
43. Yang, J., Barnett, A.J., Donnelly, J., Kishore, S., Fang, J., Schwartz, F.R., Chen, C., Lo, J.Y., Rudin, C.: Fpn-iaia-bl: A multi-scale interpretable deep learning model for classification of mass margins in digital mammography. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 5003–5009 (June 2024)