

# LUCIDPPN: UNAMBIGUOUS PROTOTYPICAL PARTS NETWORK FOR USER-CENTRIC INTERPRETABLE COMPUTER VISION

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

Prototypical parts networks combine the power of deep learning with the explainability of case-based reasoning to make accurate, interpretable decisions. They follow the this looks like that reasoning, representing each prototypical part with patches from training images. However, a single image patch comprises multiple visual features, such as color, shape, and texture, making it difficult for users to identify which feature is important to the model. To reduce this ambiguity, we introduce the Lucid Prototypical Parts Network (LucidPPN), a novel prototypical parts network that separates color prototypes from other visual features. Our method employs two reasoning branches: one for non-color visual features, processing grayscale images, and another focusing solely on color information. This separation allows us to clarify whether the model’s decisions are based on color, shape, or texture. Additionally, LucidPPN identifies prototypical parts corresponding to semantic parts of classified objects, making comparisons between data classes more intuitive, e.g., when two bird species might differ primarily in belly color. Our experiments demonstrate that the two branches are complementary and together achieve results comparable to baseline methods. More importantly, LucidPPN generates less ambiguous prototypical parts, enhancing user understanding.

## 1 INTRODUCTION

Increased adoption of deep neural networks across critical fields, such as healthcare (Rymarczyk et al., 2022b), and autonomous driving (Wu et al., 2017), shows the need to develop models in which decisions are interpretable, ensuring accountability and transparency in decision-making processes (Rudin, 2019; Rudin et al., 2022). One promising approach is based on prototypical parts (Chen et al., 2019; Nauta et al., 2023; Rymarczyk et al., 2021; 2022d), which integrate the power of deep learning with interpretability, particularly in fine-grained image classification tasks. During training, these models learn visual concepts characteristic for each class, called Prototypical Parts (PPs). In inference, predictions are made by identifying the PPs of distinct classes within an image. This way, the user is provided with explanations in the form of “This looks like that”.

The primary benefit of PPs-based methods over post hoc approaches is their ability to incorporate explanations into the prediction process (Chen et al., 2019) directly. Nevertheless, a significant challenge with these methods lies in the ambiguity of prototypical parts, visualized using five to ten nearest patches. Each patch encodes a range of visual features, including color<sup>1</sup>, shape, texture, and contrast (Nauta et al., 2021a), making it difficult for users to identify which of them are relevant. This issue is compounded by the fact that neural networks are generally biased towards texture (Geirhos et al., 2019) and color (Hosseini et al., 2018), whereas humans are typically biased towards shape (Landau et al., 1988).

Therefore, recent works have attempted to solve this problem using various strategies. Some works propose to reduce the ambiguity of prototypical parts by visualizing them through

<sup>1</sup>We follow the color definition from the research of (Berga et al., 2020; Khan et al., 2012)

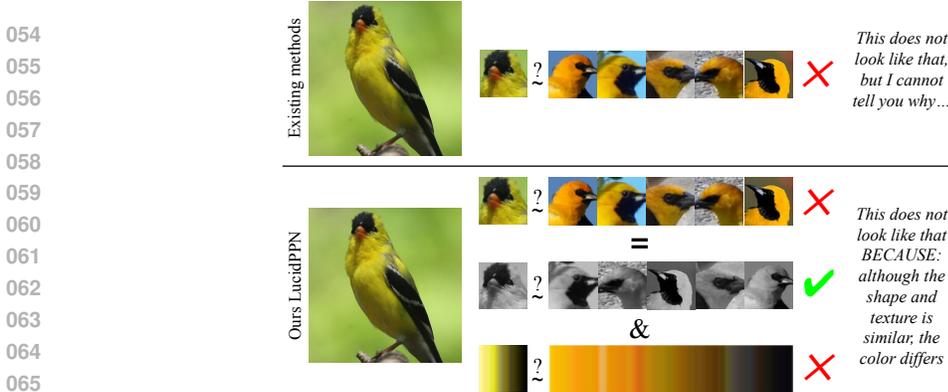


Figure 1: Our novel prototypical parts-based model, LucidPPN, enables the disentangling of color information from the prototypical parts. This capability allows us to examine more closely the differences between an image patch and patches representing a prototypical part. As shown in the image, our model can visualize that the head of a bird, compared to the prototypical part of a bird’s head from different classes, shows a high resemblance in shape and texture but differs in color. Such detailed analysis was not possible with previous prototypical parts-based approaches.

a larger number of patches (Ma et al., 2024; Nauta et al., 2023). However, it does not solve the problem with various visual features encoded in each patch. Other approaches tend to solve this problem by quantifying the appearance of specific visual features (Nauta et al., 2021a) or concepts (Wan et al., 2024) on prototypical parts. However, they generate ambiguous statements such as “color is important”, leading to further questions (e.g. about which color) that complicate understanding (Ma et al., 2024; Xu-Darme et al., 2023).

Motivated by the challenge of decoding the crucial visual attributes of prototypical parts, we introduce the Lucid Prototypical Parts Network (LucidPPN). It uniquely divides the model into two branches: the first focuses on identifying visual features of texture and shape corresponding to specific object parts (e.g. heads, tails, wings for birds), while the second is dedicated solely to color. It allows us to disentangle color features from the prototypical parts and present pairs of a simplified gray prototypical part and corresponding color (see Figure 1). The second advantage of LucidPPN is that the successive prototypes in each class correspond to the same object parts (e.g., the first prototypes are heads, the second prototypes are legs, etc.). Altogether, it enabled us to introduce a novel type of visualization presented in Figure 2, more intuitive and less ambiguous according to our user studies.

Extensive experiments demonstrate that LucidPPN achieves results competitive with current PPs-based models while successfully disentangling and fusing color information. Additionally, using LucidPPN, we can identify tasks where color information is an unimportant feature, as demonstrated on the Stanford Cars dataset (Krause et al., 2013). Finally, a user study showed that participants, guided by LucidPPN explanations, more accurately identified the ground truth compared to those using PIP-Net.

Our contributions can be summarized as follows:

- We introduce LucidPPN, a novel architecture based on PPs, which disentangles color features from the PPs in inference. Consequently, thanks to LucidPPN we know the relevance of the color and shape with texture in the decision process<sup>2</sup>.
- We propose a mechanism that ensures successive prototypes within each class consistently correspond to the same object parts.
- We introduce a more intuitive type of visualization incorporating the assumption about the fine-grained classification.
- We conduct a comprehensive examination demonstrating the usability and limitations of LucidPPN. Specifically, we highlight scenarios where color information may not be pivotal or even confuses the model in fine-grained image classification.

<sup>2</sup>See the discussion in paragraph Color Impact in Section 5.

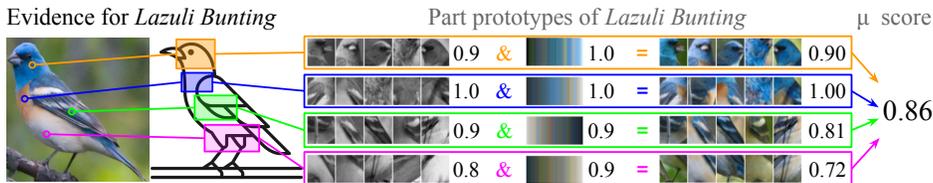


Figure 2: Our novel type of visualization utilizes the fact that the successive prototypes in each class of LucidPPN correspond to the same object parts. That is why we use a schematic drawing of a bird to show the location of the specific prototypical parts. Moreover, LucidPPN disentangles color features from the prototypical parts to present pairs of a simplified gray prototypical part and a corresponding color. The aggregated resemblance is obtained by multiplying the resemblance to the prototypical part and the resemblance to the corresponding color.

## 2 RELATED WORKS

**Ante-hoc methods for XAI.** Self-explainable models (ante-hoc) aim to make the decision process more transparent by providing the explanation together with the prediction, and they have attracted significant attention (Alvarez Melis & Jaakkola, 2018; Böhle et al., 2022; Brendel & Bethge, 2019). Much of this attention has focused on enhancing the concept of prototypical parts introduced in ProtoPNet (Chen et al., 2019) to represent the activation patterns of networks. Several extensions have been proposed, including TesNet (Wang et al., 2021) and Deformable ProtoPNet (Donnelly et al., 2022), which exploit orthogonality in prototype construction. ProtoPShare (Rymarczyk et al., 2021), ProtoTree (Nauta et al., 2021b), ProtKNN (Ukai et al., 2022), and ProtoPool (Rymarczyk et al., 2022d) reduce the number of prototypes used in classification. Other methods consider hierarchical classification with prototypes (Hase et al., 2019), prototypical part transformation (Li et al., 2018), and knowledge distillation techniques from prototypes (Keswani et al., 2022). Prototype-based solutions have been widely adopted in various applications such as medical imaging (Afnan et al., 2021; Barnett et al., 2021; Kim et al., 2021; Rymarczyk et al., 2022b), time-series analysis (Gee et al., 2019), graph classification (Rymarczyk et al., 2023a; Zhang et al., 2022), semantic segmentation (Sacha et al., 2023), and class incremental learning (Rymarczyk et al., 2023b).

However, prototypical parts still need to be improved, especially regarding the understandability and clarity of the underlying features responsible for the prediction (Kim et al., 2022). Issues such as spatial misalignment of prototypical parts (Carmichael et al., 2024; Sacha et al., 2024) and imprecise visualization techniques (Gautam et al., 2023; Xu-Darme et al., 2023) have been identified. There are also post-hoc explainers analyzing visual features such as color, shape, and textures (Nauta et al., 2021a), and approaches using multiple image patches to visualize the prototypical parts (Ma et al., 2024; Nauta et al., 2023). In this work, we address the ambiguity of prototypical parts by presenting a dedicated architecture, LucidPPN, that detects separate sets of prototypes for shapes with textures and another set for colors. This approach aims to enhance the interpretability and clarity of the interpretations.

**Usage of low-level vision features for image classification.** Multiple approaches to extracting features based on texture (Armi & Fekri-Ershad, 2019; Haralick et al., 1973), shape (Khan et al., 2012; Mingqiang et al., 2008), and color (Chen et al., 2010; Kobayashi & Otsu, 2009) have been proposed before the deep learning era, based on the knowledge about human perception (Fan et al., 2017). Similar features are trained by shallow layers of deep networks, which can be visualized with methods such as (Zeiler, 2014; Springenberg et al., 2014). However, while these methods effectively illustrate low-level features, they struggle with deeper layers, where the visualized concepts entangle multiple visual features and lead to ambiguous explanations. Similar behavior can be observed in recent eXplainable AI (XAI) methods (Basaj et al., 2021; Laina et al., 2022; Rymarczyk et al., 2022a; Zieliński & Górszczak, 2021; Nauta et al., 2021a). By using two branches, one for color and one for remaining visual features, our method explicitly disentangles these visual features, reducing ambiguity of explanations based on high-level concepts.

## 3 METHOD

### 3.1 PRELIMINARIES

**Problem formulation.** Our objective is to train a fine-grained classification model based on prototypical parts, which accurately predicts one of  $M$  subtly differentiating classes. We use  $N$  image-label pairs  $\{(x_0, y_0), \dots, (x_N, y_N)\} \subset I \times \{1, \dots, M\}$  as a training set to obtain the model returning highly accurate predictions and lucid explanations. For this, we separate color from other visual features at the input and process them through two network branches with separate sets of PPs.

**PDiscoNet.** PDiscoNet (van der Klis et al., 2023) generates segmentation masks of object parts, used in training of LucidPPN to align  $K$  successive prototypical parts of each class with  $K$  successive object parts. We decided to use it instead of human annotators because it is more efficient and cost-effective. However, it can be replaced with any method of object part segmentation due to the modularity of our approach.

PDiscoNet model  $f_{Disco}$  utilizes a convolutional neural network (CNN) to generate a feature map  $Z_{Disco} = [z_{ij}]_{i,j} \in (\mathbb{R}^{D_{Disco}})^{H_{Disco} \times W_{Disco}}$  from a given image  $x$ . Each of  $H_{Disco} \times W_{Disco}$  vectors from such feature map is then compared to trainable vectors  $q^k \in \mathbb{R}^{D_{Disco}}$  representing  $K$  object parts and background, using similarity based on Euclidean distance

$$t_{ij}^k = \frac{\exp(-\|z_{ij} - q^k\|^2)}{\sum_{k'=1}^{K+1} \exp(-\|z_{ij} - q^{k'}\|^2)}, \quad (1)$$

for  $i = 1, \dots, W_{Disco}$  and  $j = 1, \dots, H_{Disco}$ , and  $k \in \{1, \dots, K + 1\}$ . This way, we obtain an attention map  $T^k = [t_{ij}^k]_{i,j} \in \mathbb{R}^{H_{Disco} \times W_{Disco}}$  for each object part and background. Such attention maps are multiplied by feature map  $Z_{Disco}$  and averaged to obtain one vector per object part. Those vectors are passed to the classification part of PDiscoNet, which involves learnable modulation vectors and a linear classifier.

A vital observation is that the maps  $T^k$  continuously split the image into regions corresponding to discovered object parts thanks to a well-conceived set of loss functions added to the usual cross-entropy. They assure the distinctiveness, consistency, and presence of the semantic regions. Yet, the only annotations used in training are the class labels.

In the subsequent sections, we ignore the PDiscoNet predictions  $P_{Disco}$ , using only the attention maps  $T^k$ , which we will call *segmentation masks* from now on.

### 3.2 LUCIDPPN

#### 3.2.1 ARCHITECTURE

LucidPPN is a deep architecture, presented in Figure 3, consisting of two branches: one for revealing information about shape and texture (*ShapeTexNet*), and the second dedicated to color (*ColorNet*). That is why *ShapeTexNet* operates on grayscaled input, while *ColorNet* uses aggregated information about the color.

**ShapeTexNet.** A grayscaled version of image  $x$  is obtained by converting its channels  $x = (r, g, b)$  to  $x_S = (w, w, w)$ , where  $w = 0.299r + 0.587g + 0.114b$ . This formula approximates human perception of brightness (Pratt, 2013) and is a default grayscale method used in computational libraries, such as PyTorch (Paszke et al., 2019).

Grayscaled image  $x_S$  is fed to a convolutional neural network backbone  $f_{S_b}$ . For this purpose, we adapt the ConvNeXt-tiny (Liu et al., 2022) without classification head and with increased stride at the two last convolutional layers to increase the resolution of the feature map, like in PIP-Net (Nauta et al., 2023). As an output of  $f_{S_b}$ , we obtain a matrix of dimension  $(D \times H \times W)$ , which is projected to dimension  $KM \times H \times W$  using  $1 \times 1$  convolution layer  $f_{S_{cl}}$  (where  $K$  is the number of object parts and  $M$  is the number of classes) so that each prototype has its channel. Then, it is reshaped to the size of  $K \times M \times H \times W$  on which we apply the sigmoid. As a result, we obtain *ShapeTexNet feature map* defined as

$$Z_S = [z_S^{km}]_{k,m} = \sigma(f_{S_{cl}}(f_{S_b}(x_S))) = f_S(x_S) \in (\mathbb{R}^{H \times W})^{K \times M} \quad (2)$$

216  
217  
218  
219  
220  
221  
222  
223  
224  
225  
226  
227  
228  
229  
230  
231  
232  
233  
234  
235  
236  
237  
238  
239  
240  
241  
242  
243  
244  
245  
246  
247  
248  
249  
250  
251  
252  
253  
254  
255  
256  
257  
258  
259  
260  
261  
262  
263  
264  
265  
266  
267  
268  
269

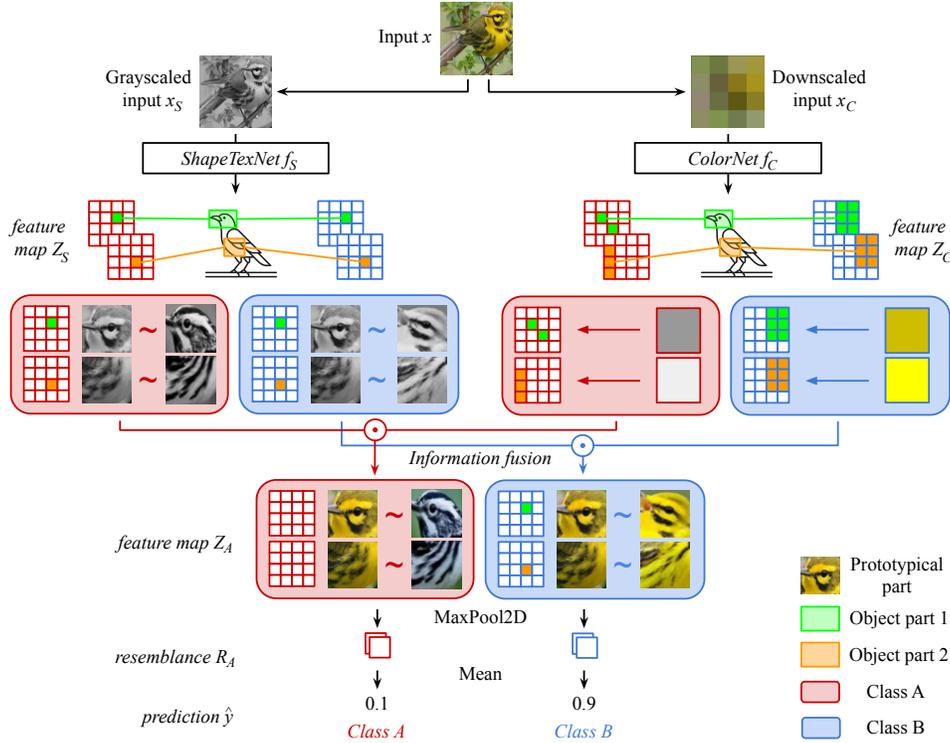


Figure 3: LucidPPN architecture consists of two branches: *ColorNet* and *ShapeTexNet* that encode color and shape with texture in feature maps  $Z_C$  and  $Z_S$ , respectively. Thanks to a special type of training each channel of a feature map corresponds to similarity to a specific object part of a given class. In this image, green and orange correspond to two object parts: head and belly, and red and blue correspond to classes A and B. Therefore, each feature map consists of four channels for head of A, belly of A, head of B, and belly of B. Corresponding channels from both branches are multiplied to obtain feature map  $Z_A$ , which is then pooled with maximum to obtain the resemblance of prototypical parts fusion and aggregated through mean to obtain final logits.

Thus, we link each map  $z_S^{km}$  to a unique *prototypical part* of an object part  $k$  for class  $m$ , from which we compute *ShapeTexNet* resemblance  $R_S = [r_S^{km}]_{k,m} \in [0, 1]^{K \times M}$ , where

$$r_S^{km} = \text{MaxPool2D}(z_S^{km}) \quad (3)$$

Finally, we obtain *ShapeTexNet* predictions  $P_S = [p_S^m]_m \in [0, 1]^M$  by taking the mean over the resemblance of all parts of a specific class

$$p_S^m = \frac{1}{K} \sum_{k=1}^K r_S^{km} \quad (4)$$

**ColorNet.** To obtain aggregated information about color, as an input of *ColorNet*, image  $x$  is downsampled through bilinear interpolation to  $H \times W$  resolution, marked as  $x_C$ . Then,  $x_C$  is passed to convolutional neural network  $f_C$ , composed of six  $1 \times 1$  convolutional layers with ReLU activations, except the last layer after which we apply sigmoid. This way, we process each input pixel of  $x_C$  separately, taking into account only its color. As a result, we obtain *ColorNet* feature map

$$Z_C = [z_C^{km}]_{k,m} = f_C(x_C) \in (\mathbb{R}^{H \times W})^{K \times M}. \quad (5)$$

Analogously to *ShapeTexNet*, each dimension in the feature map is related to a unique *prototypical part* of an object part  $k$  in class  $m$ . Hence, as before, we calculate *ColorNet* resemblance  $R_C = [r_C^{km}]_{k,m} \in [0, 1]^{K \times M}$ , where

$$r_C^{km} = \text{MaxPool2D}(z_C^{km}). \quad (6)$$

**Information fusion and prediction.** To obtain aggregated feature map  $Z_A = [z_A^{km}]_{k,m} \in (\mathbb{R}^{H \times W})^{K \times M}$  from both branches, we multiply the *ShapeTexNet* feature map with *ColorNet* feature map element-wise

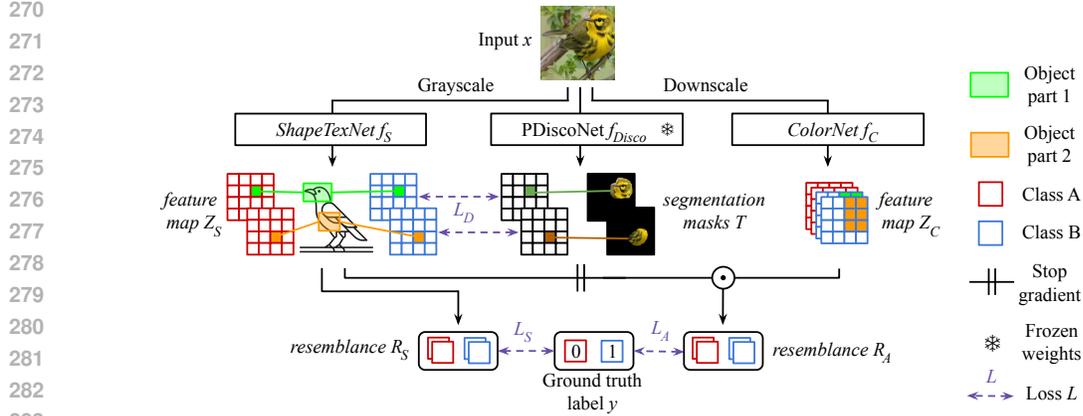


Figure 4: LucidPPN training schema. We use segmentation masks from PDiscoNet to align the activation of prototypical parts with object parts. Additionally, we enforce the *ShapeTexNet* to encode as much predictive information as possible through the usage of  $L_S$ . Lastly, we learn how to classify images through  $L_A$  which is a binary cross-entropy loss.

$$z_A^{km} = z_S^{km} \odot z_C^{km}, \quad (7)$$

and define *aggregated resemblance*  $R_A = [r_A^{km}]_{k,m} \in [0, 1]^{K \times M}$  as

$$r_A^{km} = \text{MaxPool2D}(z_A^{km}). \quad (8)$$

The final predictions  $\hat{y} = [\hat{y}^m]_m \in [0, 1]^M$  for all classes are obtained by averaging  $r_A^{km}$  over class-related parts

$$\hat{y}^m = \frac{1}{K} \sum_{k=1}^K r_A^{km}. \quad (9)$$

### 3.2.2 TRAINING

As a result of LucidPPN training (presented in Figure 4), we aim to achieve three primary goals: 1) obtaining a high-accuracy model, 2) ensuring the correspondence of prototypical parts to object parts, 3) and disentangling color information from other visual features. To accomplish these goals, we design three loss functions: 1) prototypical-object part correspondence loss  $L_D$ , 2) loss disentangling color from shape with texture  $L_S$ , 3) and classification loss  $L_A$  that contribute to the final loss

$$L = \alpha_D L_D + \alpha_S L_S + \alpha_A L_A, \quad (10)$$

where  $\alpha_D, \alpha_S, \alpha_A$  are weighting factors whose values are found through hyperparameter search. The definition of each loss component is presented in the following paragraphs. Please note that we assume that PDiscoNet was already trained, and we denote  $\bar{y} \in \mathbb{B}^M$  as a one-hot encoding of  $y$ .

**Correspondence of prototypical parts to object parts.** To ensure that each prototypical part assigned to a given class corresponds to distinct object parts, we define the prototypical-object part correspondence loss  $L_D$ . This function leverages *segmentation masks*  $T^k$  from PDiscoNet to align the activations of prototypical parts, represented by the *ShapeTexNet* feature map  $Z_S$  with the locations of object parts. Hence, the activations from the *aggregated feature map*  $Z_A$  will be aligned with these object parts. It is defined as

$$L_D = \frac{1}{K} \sum_{k=1}^K \text{MBCE}(Z_S^{ky}, T^k), \quad (11)$$

where  $\text{MBCE}(u, v)$  is defined as the mean binary cross-entropy loss between two maps  $u, v \in [0, 1]^{H \times W}$ .

$$\text{MBCE}(u, v) = \frac{1}{HW} \sum_{i=1}^H \sum_{j=1}^W \text{BCE}(u_{ij}, v_{ij}), \quad (12)$$

and  $y$  is the ground truth class. We align only the maps corresponding to  $y$  because the prototypical parts assigned to other classes should not be highly activated.

**Disentangling color from other visual information.** To maximize the usage of information about shape and texture during the classification with prototypical parts, we maximize the accuracy of the *ShapeTexNet* through the usage of binary cross-entropy as classification loss function on *ShapeTexNet* resemblances values

$$L_S = \frac{1}{KM} \sum_{m=1}^M \sum_{k=1}^K \text{BCE}(r_S^{km}, \bar{y}_m). \quad (13)$$

**Classification loss.** Lastly, to ensure the high accuracy of the model and to combine information from *ColorNet* and *ShapeTexNet*, we employ binary cross-entropy on *aggregated resemblances* as our classification loss

$$L_A = \frac{1}{KM} \sum_{m=1}^M \sum_{k=1}^K \text{BCE}(r_A^{km}, \bar{y}_m). \quad (14)$$

### 3.3 PREDICTION INTERPRETATION

LucidPPN adopts the definition of prototypical parts from PIP-Net (Nauta et al., 2023), where each prototypical part is represented by ten patches, typically activated by ten colored images from the training set. However, in LucidPPN, the visualization must demonstrate how each prototypical part is disentangled into color and shape with texture features. That is why we propose a method to present the disentangled visual features of a prototypical part by combining five grayscale patches, a color bar, and five colored patches. The grayscale and colored patches are selected from the training images with the highest *ShapeTexNet* resemblance and *aggregated resemblance*, respectively. The color bar is created by sampling RGB color values from the ten colored patches with the highest *aggregated resemblance* and projecting them using t-SNE (Van der Maaten & Hinton, 2008). Moreover, in contrast to PIP-Net, LucidPPN creates prototypical parts corresponding to the same object parts in all classes. Therefore, we can use the information about the specific object part location to enrich the explainability.

**Local (prediction) interpretation.** Figure 2 demonstrates how LucidPPN classifies a specific sample  $x$  into class  $\hat{y}$  by examining the prototypical parts assigned to  $\hat{y}$  that are disentangled into color and other visual features. The views are enhanced with pointers to regions of highest *aggregated resemblance*, clearly associated with the object parts.

**Comparison explanation.** Users may wish to inspect and compare local explanations for multiple classes. LucidPPN facilitates this comparison by allowing users to compare prototypical parts of corresponding object parts, making the process intuitive, as shown in supplementary Figure 8.

**Class (global) characteristic.** Disentangled prototypical parts corresponding to object parts reveal the patterns the model uses to classify a given class. This enables the identification of texture and shape features, as well as colors (see Sup. Figure 21), that describe a class without the need to analyze the final-layer connections, unlike other prototypical part-based approaches (Chen et al., 2019; Donnelly et al., 2022; Ma et al., 2024; Nauta et al., 2023; Rymarczyk et al., 2021; 2022c).

## 4 EXPERIMENTAL SETUP

**Datasets.** We train and evaluate our model on four datasets: CUB-200-2011 (CUB) with 200 bird species (Wah et al., 2011), Stanford Cars (CARS) with 196 car models (Krause et al., 2013), Stanford Dogs (DOGS) with 120 breeds of dogs (Khosla et al., 2011), and Oxford 102 Flower (FLOWER) with 102 kinds of flowers (Nilsback & Zisserman, 2008). More details on image preprocessing are in the Supplement.

**Implementation details.** Trainings are repeated 3 times. We made the code public. The size of *ShapeTexNet* feature map is  $768 \times 28 \times 28$ . The channel number of *ColorNet*'s convolutional layers is 20, 50, 150, 200, 600,  $K \cdot M$ . The values of loss weights are found through hyperparameter search ( $\alpha_D = 1.4, \alpha_S = 1.0, \alpha_A = 1.0$ ). Details are in the Supplement.

**Metrics** During the evaluation, we report the top-1 accuracy classification score. Additionally, we measure the quality of *prototypical parts* alignment with object parts by calculating intersection-over-union (IoU). In PDiscoNet, the segmentation map is the size of

Table 1: Comparison of accuracy of PPs-based models on 4 datasets. LucidPPN achieves competitive results to all methods, and SOTA on 2 datasets. Note that, LucidPPN is trained with  $K = 12$ , and “-“ means that the model did not converge during training when using the code provided by the authors.

	CUB	CARS	DOGS	FLOWER
ProtoPNet (Chen et al., 2019)	79.2	86.1	<b>77.4±0.2</b>	<b>92.1±0.3</b>
ProtoTree (Nauta et al., 2021b)	82.2 ± 0.7	86.6 ± 0.2	-	-
ProtoPShare (Rymarczyk et al., 2021)	74.7	86.4	<b>74.1±0.3</b>	<b>90.3±0.2</b>
ProtoPool (Rymarczyk et al., 2022c)	<b>85.5 ± 0.1</b>	88.9 ± 0.1	<b>71.7±0.2</b>	<b>92.7±0.1</b>
PIP-Net (Nauta et al., 2023)	84.3 ± 0.2	88.2 ± 0.5	<b>80.8 ± 0.4</b>	91.8 ± 0.5
LucidPPN	81.5 ± 0.4	<b>91.6 ± 0.2</b>	79.4 ± 0.4	<b>95.0 ± 0.3</b>

the input image, while our activation map is the size of the latent space. Hence, we are downsizing the segmentation map to  $26 \times 26$  resolution to match its dimensions with the activation map before calculating the IoU between the corresponding patches of both maps.

**User studies.** Using ClickWorker System<sup>3</sup>, we run two user studies to compare the quality of patch-based prototypes and the influence of disentangled resemblance scores provided by LucidPPN. For the first study, we collect the testing examples from CUB which are correctly classified by PIP-Net, *single branch CNN*<sup>4</sup> and LucidPPN. These are joined with information about the two most probable classes per model and associated prototypical parts. Ninety workers (30 per method) answer the survey, which consists of 10 questions. They are asked to predict the model’s decision based on the evidence for the top two output classes without the numerical scores. This approach mimics the user study presented in HIVE (Kim et al., 2022) and is also inspired by the study performed in (Ma et al., 2024). In the second study, we also collect images from CUB. This time we join them with prototypical parts of the correctly predicted class and one other class. Each of the forty workers answers 10 questions in which he/she rates from 1 (least) to 5 (most) to assess the influence of the color features on the model’s prediction. The users give ratings based on LucidPPN prototypical parts visualization, with or without included numerical resemblance scores. More details and the survey templates are in the Supplement.

## 5 RESULTS

In this section, we show the effectiveness of LucidPPN, the influence of the color disentanglement in the processing on the model’s performance, and the results related to the interpretability of learned prototypical parts based on the user study.

### Comparison to other PPs-based models.

In Table 1 we compare the classification quality of LucidPPN and other PPs-based methods. We present the mean accuracy and standard deviation. We report best performing LucidPPN, which in the case of all datasets was trained with fixed  $K = 12$ . Our LucidPPN achieves the highest accuracy for CARS and FLOWER datasets, and competitive results on CUB and DOGS.

**Color impact.** The influence of *ColorNet* on LucidPPN predictions is shown in Table 2. We compare the accuracy of *ShapeTexNet* with the LucidPPN predictions. The *information fusion* enhances the results on the CUB, DOGS, and FLOWER. However, it does not affect performance on the CARS. This can be attributed to the characteristics of the CARS dataset, where vehicles of the same model can differ in colors, indicating that color is not

Table 2: Comparison of the accuracy of *ShapeTexNet* to LucidPPN. Integrating color with other visual features proves advantageous for datasets containing objects found in nature. However, for the CARS dataset, adding color information does not enhance the model’s performance. This is because color is not a significant feature when classifying vehicles, as the same car model can appear in various colors.

	CUB	CARS	DOGS	FLOWER
<i>ShapeTexNet</i>	80.4	<b>91.7</b>	78.6	93.6
LucidPPN	<b>81.8</b>	<b>91.7</b>	<b>78.9</b>	<b>95.3</b>

<sup>3</sup><https://www.clickworker.com/>

<sup>4</sup>For ablation analysis we also report results of a *single branch CNN* which has the same architecture as *ShapeTexNet* and receives colored images as input. Its local interpretation is visualized similarly to LucidPPN, but without the gray patches and color bar as presented in Sup. Figure 13

critical for this task. This contrasts with the fine-grained classification of natural objects, such as birds and flowers, where color plays a significant role.

In Table 3 we show the results of experiments aiming to analyze how the model is susceptible to the change of the color on the image. We report the accuracy of PIP-Net, *ShapeTexNet*, and LucidPPN on original and hue-perturbed images from the CUB dataset. One can notice that PIP-Net is highly dependent on color information and its score drops by over 37% after perturbation. At the same time *ShapeTexNet* is immune to this transformation, while LucidPPN loses approximately 12.5% accuracy because of it. To alter hue we randomly rotate hue values in the HSV color space. After rotation, we adjust the luminosity of each pixel by proportionately scaling its RGB channels. This step is key to modifying the hue without changing the brightness perceived by humans.

**User studies.** Statistics from the user study assessing the lucidity of explanations generated by LucidPPN, *single branch*, and PIP-Net are in Figure 5 and Supplementary Table 5. We report the mean user accuracy with a standard deviation and  $p$ -values. Users basing their responses on LucidPPN explanations score significantly better than both PIP-Net and random guess baselines. Additionally, we conclude that most of the accuracy in this user study can be attributed to the PDiscoNet part supervision as *single branch* scores similarly to

LucidPPN, without a statistically significant difference. While both of our explanation variants with prototypical parts corresponding to the same object parts prove to be more intuitive for users, we also want to highlight the advantage of using full LucidPPN over *single branch*. To this end, in Supplementary Figure 7 and Supplementary Table 6 we show the outcomes of the study evaluating the user’s ability to recognize the importance of color features in LucidPPN’s decisions. Users without information about resemblance values struggle in this task achieving the same performance as if they answered at random. In contrast, users provided with the resemblances in LucidPPN visualizations score 23% better. Note that neither *single branch* nor PIP-Net gives the disentangled resemblance values. In both studies, we perform a one-sided  $t$ -test and one-sample  $t$ -test to compare methods against each other and 50% accuracy, respectively. More details can be found in the Supplement.

## 6 ABLATION AND ANALYSIS

In this section, we examine how LucidPPN’s performance is impacted by object part supervision and the weights of the loss function components.

**Influence of part supervision on the performance of LucidPPN.** One of the features of LucidPPN is object part supervision based on PDiscoNet. To check its influence on the PPs-based model without disentanglement, in Table 4 we compare the accuracy of a *single branch* to LucidPPN and PIP-Net. The *single branch* scores better than both models. The disentanglement in LucidPPN causes a small (<6%) or negligible drop in accuracy while offering more insights from the model.

**Loss weighting.** In Figure 6 we investigate the impact of the loss weight  $\alpha_D$ , which is responsible for prototypical-object parts alignment, on training outcomes. In this analysis, the weights of the other losses are fixed at  $\alpha_S = \alpha_C = 1$ . We evaluate the accuracy and intersection-over-union (IoU) between the highest activated *ShapeTexNet feature map* and PDiscoNet’s *segmentation masks* for each object part. The results show that increasing  $\alpha_D$  enhances the IoU, but after a certain point, it gradually

Table 3: Robustness of the model to changes in image color. When the hue value is perturbed, the accuracy of PIP-Net drops significantly. In contrast, the accuracy drop for LucidPPN is only half as much, and for *ShapeTexNet* none.

	Original	Hue-perturbed
PIP-Net	83.9	53.0
<i>ShapeTexNet</i>	80.3	80.3
LucidPPN	81.9	71.7

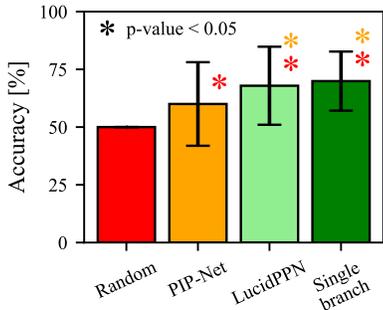


Figure 5: User study results show that users based on LucidPPN explanations outperform those with explanations from PIP-Net to a statistically significant degree.

reduces accuracy. Notably, omitting the loss  $\alpha_D$  (part supervision) significantly diminishes the network’s classification performance and makes the learned prototypical parts collapse into a single, most descriptive one as presented in Supplementary Figure 10. While other works address this issue by adding novel regularization losses (Nauta et al., 2023; Wang et al., 2021), these solutions fail to ensure consistency of the considered parts across different classes.

We provide additional results in the Supplementary Materials. They include ablations on the LucidPPN’s backbone, the size of a *ColorNet*, type of input color representation, number of object parts  $K$ , and different training schedules for LucidPPN’s branches. Also, we show examples of PIPNet failures mitigated by LucidPPN, and we discuss the reasons for introducing  $L_S$ .

## 7 CONCLUSIONS

In this work, we propose LucidPPN, an inherently interpretable model that uses prototypical parts to disentangle color from other visual features in its explanations. Our extensive results demonstrate the effectiveness of our method, and user studies confirm that our explanations are less ambiguous than those from PIP-Net. In future research, we aim to further refine the model architecture to separately process shape and texture features, as well as analyze different visualization strategies of disentanglement and their recognition by the users. Additionally, we plan to explore the human perception system in greater depth to inform the design of the next generation of interpretable neural network architectures.

**Limitations.** Our work faces a significant constraint: while our designed mechanism adeptly disentangles color information from input images, it cannot currently extract other crucial visual features such as texture, shape, and contrast. This highlights a broader challenge within the field: the absence of a universal mechanism capable of encompassing diverse visual attributes. Furthermore, our approach inherits limitations from other PPs-based architectures, including issues such as spatial misalignment (Sacha et al., 2024), the non-obvious interpretation of PPs (Ma et al., 2024) and those of PIP-Net (Nauta et al., 2023). The latter could be addressed with textual descriptions of concepts discovered by PPs. Lastly, LucidPPN increases the transparency of the decision made by the deep neural networks however it still has a performance gap to black-box models, or even to those offering some insights into the model reasoning process such as PDiscoNet (van der Klis et al., 2023). This shall be under further investigation to fill this performance gap if possible.

**Broader Impact.** Our work advances the field of interpretability, a crucial component for trustworthy AI systems, where users have the right to understand the decisions made by these systems (Kaminski, 2021; Tabassi, 2023). LucidPPN enhances the quality of explanations derived from PPs-based neural networks, which are among the most promising techniques for ante-hoc interpretability methods. Consequently, it can facilitate the derivation of scientific insights and the creation of better human-AI interfaces for complex, high-stakes applications.

Additionally, LucidPPN provides visual characteristics for PPs, which are especially beneficial in domains lacking standardized semantic textual descriptions of concepts. This is particularly useful in fields such as medicine, where it aids in analyzing radiology and histopathology images.

Table 4: Accuracy of PIP-Net, LucidPPN, and a *single branch* CNN supervised by PDiscoNet.

	CUB	CARS	DOGS	FLOWER
PIP-Net	84.3	88.2	80.8	91.8
LucidPPN	81.5	91.6	79.4	95.0
<i>single branch</i>	<b>86.6</b>	<b>91.9</b>	<b>82.7</b>	<b>95.6</b>

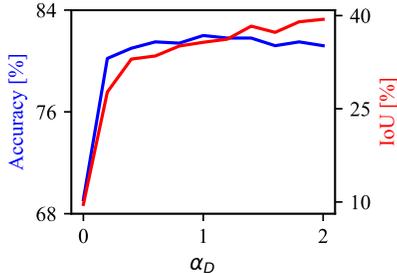


Figure 6: Influence of the weight of prototypical-object part correspondence loss on accuracy and Intersection-over-Union (IoU). An increase of  $\alpha_D$  improves IoU but at a certain point gradually reduces accuracy.

540 REPRODUCIBILITY STATEMENT

541  
542 We ensured that our experiments are reproducible by thoroughly describing them in Sec-  
543 tion 4 and the Supplement. Additionally, the Supplementary Materials include the code used  
544 to perform the experiments, along with a README.md file providing further instructions.  
545

546 REFERENCES

- 547  
548 Michael Anis Mihdi Afnan, Yanhe Liu, Vincent Conitzer, Cynthia Rudin, Abhishek Mishra,  
549 Julian Savulescu, and Masoud Afnan. Interpretable, not black-box, artificial intelligence  
550 should be used for embryo selection. *Human Reproduction Open*, 2021.  
551  
552 David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-  
553 explaining neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman,  
554 N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Sys-*  
555 *tems*, volume 31. Curran Associates, Inc., 2018.  
556  
557 Laleh Armi and Shervan Fekri-Ershad. Texture image analysis and texture classification  
558 methods-a review. *arXiv preprint arXiv:1904.06554*, 2019.  
559  
560 Alina Jade Barnett, Fides Regina Schwartz, Chaofan Tao, Chaofan Chen, Yinhao Ren,  
561 Joseph Y Lo, and Cynthia Rudin. A case-based interpretable deep learning model for  
562 classification of mass lesions in digital mammography. *Nature Machine Intelligence*, 3  
(12):1061–1070, 2021.  
563  
564 Dominika Basaj, Witold Oleszkiewicz, Igor Sieradzki, Michał Górszczak, B Rychalska,  
565 T Trzcinski, and B Zielinski. Explaining self-supervised image representations with visual  
566 probing. In *International Joint Conference on Artificial Intelligence*, 2021.  
567  
568 David Berga, Marc Masana, and Joost Van de Weijer. Disentanglement of color and shape  
569 representations for continual learning. *arXiv preprint arXiv:2007.06356*, 2020.  
570  
571 Moritz Böhle, Mario Fritz, and Bernt Schiele. B-cos networks: Alignment is all we need for  
572 interpretability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and*  
573 *Pattern Recognition*, pp. 10329–10338, 2022.  
574  
575 Wieland Brendel and Matthias Bethge. Approximating CNNs with bag-of-local-features  
576 models works surprisingly well on imagenet. In *International Conference on Learning*  
577 *Representations*, 2019. URL <https://openreview.net/forum?id=SkfMWhAqYQ>.  
578  
579 Zachariah Carmichael, Suhas Lohit, Anoop Cherian, Michael J Jones, and Walter J Scheirer.  
Pixel-grounded prototypical part networks. In *Proceedings of the IEEE/CVF Winter*  
*Conference on Applications of Computer Vision*, pp. 4768–4779, 2024.  
580  
581 Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su.  
582 This looks like that: deep learning for interpretable image recognition. *Advances in neural*  
*information processing systems*, 32, 2019.  
583  
584 Wei-Ta Chen, Wei-Chuan Liu, and Ming-Syan Chen. Adaptive color feature extraction based  
585 on image color distributions. *IEEE Transactions on image processing*, 19(8):2005–2016,  
586 2010.  
587  
588 Jon Donnelly, Alina Jade Barnett, and Chaofan Chen. Deformable protopnet: An inter-  
589 pretable image classifier using deformable prototypes. In *Proceedings of the IEEE/CVF*  
*conference on computer vision and pattern recognition*, pp. 10265–10275, 2022.  
590  
591 Shaojing Fan, Tian-Tsong Ng, Bryan Lee Koenig, Jonathan Samuel Herberg, Ming Jiang,  
592 Zhiqi Shen, and Qi Zhao. Image visual realism: From human perception to machine  
593 computation. *IEEE transactions on pattern analysis and machine intelligence*, 40(9):  
2180–2193, 2017.

- 594 Srishti Gautam, Marina M-C Höhne, Stine Hansen, Robert Jenssen, and Michael  
595 Kampffmeyer. This looks more like that: Enhancing self-explaining models by proto-  
596 typical relevance propagation. *Pattern Recognition*, 136:109172, 2023.
- 597
- 598 Alan H Gee, Diego Garcia-Olano, Joydeep Ghosh, and David Paydarfar. Explaining deep  
599 classification of time-series data with learned prototypes. In *CEUR workshop proceedings*,  
600 volume 2429, pp. 15. NIH Public Access, 2019.
- 601 Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann,  
602 and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape  
603 bias improves accuracy and robustness. *7th International Conference on Learning Repre-  
604 sentations, ICLR*, 2019.
- 605
- 606 Robert M Haralick, Karthikeyan Shanmugam, and Its' Hak Dinstein. Textural features for  
607 image classification. *IEEE Transactions on systems, man, and cybernetics*, (6):610–621,  
608 1973.
- 609 Peter Hase, Chaofan Chen, Oscar Li, and Cynthia Rudin. Interpretable image recogni-  
610 tion with hierarchical prototypes. In *Proceedings of the AAAI Conference on Human  
611 Computation and Crowdsourcing*, volume 7, pp. 32–40, 2019.
- 612
- 613 Ju He, Shuo Yang, Shaokang Yang, Adam Kortylewski, Xiaoding Yuan, Jie-Neng Chen,  
614 Shuai Liu, Cheng Yang, Qihang Yu, and Alan Yuille. Partimagenet: A large, high-quality  
615 dataset of parts. In *European Conference on Computer Vision*, pp. 128–145. Springer,  
616 2022.
- 617 Hossein Hosseini, Baicen Xiao, Mayoore Jaiswal, and Radha Poovendran. Assessing shape  
618 bias property of convolutional neural networks. In *Proceedings of the IEEE Conference  
619 on Computer Vision and Pattern Recognition Workshops*, pp. 1923–1931, 2018.
- 620
- 621 Qihan Huang, Mengqi Xue, Wenqi Huang, Haofei Zhang, Jie Song, Yongcheng Jing, and  
622 Mingli Song. Evaluation and improvement of interpretability for self-explainable part-  
623 prototype networks. In *Proceedings of the IEEE/CVF International Conference on Com-  
624 puter Vision*, pp. 2011–2020, 2023.
- 625 Margot E Kaminski. The right to explanation, explained. In *Research Handbook on Infor-  
626 mation Law and Governance*, pp. 278–299. Edward Elgar Publishing, 2021.
- 627
- 628 Monish Keswani, Sriranjani Ramakrishnan, Nishant Reddy, and Vineeth N Balasubrama-  
629 nian. Proto2proto: Can you recognize the car, the way i do? In *Proceedings of the  
630 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10233–10243,  
631 2022.
- 632
- 633 Fahad Shahbaz Khan, Joost Van de Weijer, and Maria Vanrell. Modulating shape features  
634 by color attention for object recognition. *International Journal of Computer Vision*, 98:  
635 49–64, 2012.
- 636
- 637 Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset  
638 for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-  
639 grained visual categorization (FGVC)*, volume 2. Citeseer, 2011.
- 640
- 641 Eunji Kim, Siwon Kim, Minji Seo, and Sungroh Yoon. Xprotonet: Diagnosis in chest radio-  
642 graphy with global and local explanations. In *Proceedings of the IEEE/CVF Conference  
643 on Computer Vision and Pattern Recognition*, pp. 15719–15728, 2021.
- 644
- 645 Sunnie SY Kim, Nicole Meister, Vikram V Ramaswamy, Ruth Fong, and Olga Russakovsky.  
646 Hive: Evaluating the human interpretability of visual explanations. In *European Confer-  
647 ence on Computer Vision*, pp. 280–298. Springer, 2022.
- 648
- 649 Takumi Kobayashi and Nobuyuki Otsu. Color image feature extraction using color index  
650 local auto-correlations. In *2009 IEEE International Conference on Acoustics, Speech and  
651 Signal Processing*, pp. 1057–1060. IEEE, 2009.

- 648 Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-  
649 grained categorization. In *Proceedings of the IEEE international conference on computer*  
650 *vision workshops*, pp. 554–561, 2013.
- 651 Iro Laina, Yuki M Asano, and Andrea Vedaldi. Measuring the interpretability of unsuper-  
652 vised representations via quantized reverse probing. *ICLR*, 2022.
- 653 Barbara Landau, Linda B Smith, and Susan S Jones. The importance of shape in early  
654 lexical learning. *Cognitive development*, 3(3):299–321, 1988.
- 655 Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. Deep learning for case-based reason-  
656 ing through prototypes: A neural network that explains its predictions. In *Proceedings of*  
657 *the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- 658 Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and  
659 Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on*  
660 *computer vision and pattern recognition*, pp. 11976–11986, 2022.
- 661 Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint*  
662 *arXiv:1711.05101*, 2017.
- 663 Chiyu Ma, Brandon Zhao, Chaofan Chen, and Cynthia Rudin. This looks like those: Illu-  
664 minating prototypical concepts using multiple visualizations. *Advances in Neural Informa-*  
665 *tion Processing Systems*, 36, 2024.
- 666 Yang Mingqiang, Kpalma Kidiyo, Ronsin Joseph, et al. A survey of shape feature extraction  
667 techniques. *Pattern recognition*, 15(7):43–90, 2008.
- 668 Meike Nauta, Annemarie Jutte, Jesper Provoost, and Christin Seifert. This looks like that,  
669 because... explaining prototypes for interpretable image recognition. In *Joint European*  
670 *Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 441–456.  
671 Springer, 2021a.
- 672 Meike Nauta, Ron Van Bree, and Christin Seifert. Neural prototype trees for interpretable  
673 fine-grained image recognition. In *Proceedings of the IEEE/CVF conference on computer*  
674 *vision and pattern recognition*, pp. 14933–14943, 2021b.
- 675 Meike Nauta, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. Pip-net: Patch-  
676 based intuitive prototypes for interpretable image classification. In *Proceedings of the*  
677 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2744–2753,  
678 2023.
- 679 Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large  
680 number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image*  
681 *processing*, pp. 722–729. IEEE, 2008.
- 682 Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan,  
683 Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An im-  
684 perative style, high-performance deep learning library. *Advances in neural information*  
685 *processing systems*, 32, 2019.
- 686 William K Pratt. *Introduction to digital image processing*. CRC press, 2013.
- 687 Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions  
688 and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.
- 689 Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi  
690 Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges.  
691 *Statistic Surveys*, 16:1–85, 2022.
- 692 Dawid Rymarczyk, Łukasz Struski, Jacek Tabor, and Bartosz Zielinski. Protopshare: Pro-  
693 totypical parts sharing for similarity discovery in interpretable image classification. In  
694 *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Min-*  
695 *ing*, pp. 1420–1430, 2021.

- 702 Dawid Rymarczyk, Adriana Borowa, Anna Bracha, Maurycy Chronowski, Wojciech Ozimek,  
703 and Bartosz Zieliński. Comparison of supervised and self-supervised deep representations  
704 trained on histological images. In *MEDINFO 2021: One World, One Health–Global  
705 Partnership for Digital Innovation*, pp. 1052–1053. IOS Press, 2022a.
- 706 Dawid Rymarczyk, Adam Pardyl, Jarosław Kraus, Aneta Kaczyńska, Marek Skomorowski,  
707 and Bartosz Zieliński. Protomil: multiple instance learning with prototypical parts for  
708 whole-slide image classification. In *Joint European Conference on Machine Learning and  
709 Knowledge Discovery in Databases*, pp. 421–436. Springer, 2022b.
- 710 Dawid Rymarczyk, Łukasz Struski, Michał Górszczak, Koryna Lewandowska, Jacek Tabor,  
711 and Bartosz Zieliński. Interpretable image classification with differentiable prototypes  
712 assignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*,  
713 2022c.
- 714 Dawid Rymarczyk, Łukasz Struski, Michał Górszczak, Koryna Lewandowska, Jacek Tabor,  
715 and Bartosz Zieliński. Interpretable image classification with differentiable prototypes  
716 assignment. In *European Conference on Computer Vision*, pp. 351–368. Springer, 2022d.
- 717 Dawid Rymarczyk, Daniel Dobrowolski, and Tomasz Danel. ProgreSt: Prototypical graph  
718 regression soft trees for molecular property prediction. *SIAM International Conference  
719 on Data Mining*, 2023a.
- 720 Dawid Rymarczyk, Joost van de Weijer, Bartosz Zieliński, and Bartłomiej Twardowski.  
721 Icycle: Interpretable class incremental continual learning. In *Proceedings of the IEEE/CVF  
722 International Conference on Computer Vision*, pp. 1887–1898, 2023b.
- 723 Mikolaj Sacha, Bartosz Jura, Dawid Rymarczyk, Łukasz Struski, Jacek Tabor, and Bartosz  
724 Zieliński. Interpretability benchmark for evaluating spatial misalignment of prototypical  
725 parts explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*,  
726 volume 38, pp. 21563–21573, 2024.
- 727 Mikolaj Sacha, Dawid Rymarczyk, Lukasz Struski, Jacek Tabor, and Bartosz Zieliński. Pro-  
728 toseg: Interpretable semantic segmentation with prototypical parts. In *Winter Conference  
729 on Applications of Computer Vision (WACV)*, 2023.
- 730 Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striv-  
731 ing for simplicity: The all convolutional net. In *International Conference on Learning  
732 Representations (ICLR)*, 2014.
- 733 Elham Tabassi. Artificial intelligence risk management framework (ai rmf 1.0). 2023.
- 734 Yuki Ukai, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. This looks like  
735 it rather than that: Protoknn for similarity-based classifiers. In *The Eleventh International  
736 Conference on Learning Representations*, 2022.
- 737 Robert van der Klis, Stephan Alaniz, Massimiliano Mancini, Cassio F Dantas, Dino Ienco,  
738 Zeynep Akata, and Diego Marcos. Pdisconet: Semantically consistent part discovery for  
739 fine-grained recognition. In *Proceedings of the IEEE/CVF International Conference on  
740 Computer Vision*, pp. 1866–1876, 2023.
- 741 Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of  
742 machine learning research*, 9(11), 2008.
- 743 Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The  
744 caltech-ucsd birds-200-2011 dataset. 2011.
- 745 Qiyang Wan, Ruiping Wang, and Xilin Chen. Interpretable object recognition by semantic  
746 prototype analysis. In *Proceedings of the IEEE/CVF Winter Conference on Applications  
747 of Computer Vision*, pp. 800–809, 2024.
- 748 Jiaqi Wang et al. Interpretable image recognition by constructing transparent embedding  
749 space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*,  
750 pp. 895–904, 2021.

756 Bichen Wu, Forrest Iandola, Peter H Jin, and Kurt Keutzer. Squeezedet: Unified, small, low  
757 power fully convolutional neural networks for real-time object detection for autonomous  
758 driving. In *Proceedings of the IEEE conference on computer vision and pattern recognition*  
759 *workshops*, pp. 129–137, 2017.

760 Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J Geras. Characterizing  
761 and overcoming the greedy nature of learning in multi-modal deep neural networks. In  
762 *International Conference on Machine Learning*, pp. 24043–24055. PMLR, 2022.

763 Romain Xu-Darme, Georges Quénot, Zakaria Chihani, and Marie-Christine Rousset. Sanity  
764 checks for patch visualisation in prototype-based image classification. In *Proceedings of*  
765 *the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3690–3695,  
766 2023.

767 MD Zeiler. Visualizing and understanding convolutional networks. In *European conference*  
768 *on computer vision/arXiv*, volume 1311, 2014.

769 Zaixi Zhang, Qi Liu, Hao Wang, Chengqiang Lu, and Cheekong Lee. Protgmn: Towards  
770 self-explaining graph neural networks. 2022.

771 Bartosz Zieliński and Michał Górszczak. What pushes self-supervised image representations  
772 away? In *International Conference on Neural Information Processing*, pp. 514–521.  
773 Springer, 2021.

774  
775  
776  
777  
778  
779  
780  
781  
782  
783  
784  
785  
786  
787  
788  
789  
790  
791  
792  
793  
794  
795  
796  
797  
798  
799  
800  
801  
802  
803  
804  
805  
806  
807  
808  
809

810 SUPPLEMENT FOR LUCIDPPN: UNAMBIGUOUS PROTOTYPICAL PARTS  
 811 NETWORK FOR USER-CENTRIC INTERPRETABLE COMPUTER VISION  
 812

813 MORE DETAILS ON DATA PREPROCESSING  
 814

815 In training, we apply transformations as follows: `Resize(size=224+8)`, `TAWideNoColor()`,  
 816 `RandomHorizontalFlip()`, `RandomResizedCrop(size=(224, 224), scale=(0.95, 1.))`,  
 817 where `TAWideNoColor()` is the same variation of `TrivialAugment` augmentation as  
 818 in PIP-Net. Additionally, the image entering the *ShapeTexNet* is normalized with  
 819 `Normalize(mean=0.445, std=0.269)` after being converted to grayscale. At test time and  
 820 when finding the prototypical parts patches, we only apply `Resize(size=224)` followed by  
 821 grayscaling and normalization in case of *ShapeTexNet* input. The CUB images used for  
 822 training and evaluation are first cropped to the bounding boxes similarly to other PP-based  
 823 methods.

824 We do not modify any parameters in PDiscoNet. CUB settings are used for datasets not  
 825 trained in the PDiscoNet paper. For efficiency, we generated and saved the segmentation  
 826 masks to avoid inferencing PDiscoNet during LucidPPN’ training.

827 MORE DETAILS ON EXPERIMENTAL SETUP  
 828

829 The networks (*ShapeTexNet* and *ColorNet*) are optimized together in minibatches of size  
 830 64 for 40 epochs using AdamW (Loshchilov & Hutter, 2017) optimizer with beta values of  
 831 0.9 and 0.999, epsilon of  $10^{-8}$ , and weight decay of 0. The learning rate of *ShapeTexNet*  
 832 parameters is initialized to 0.002 and lowered to 0.0002 after 15 epochs. The learning rate  
 833 of the *ColorNet* is fixed at 0.002. We freeze the weights of *ShapeTexNet* backbone for the  
 834 first 15 epochs as a warm-up stage similar to other PPs-based approaches (Chen et al., 2019;  
 835 Nauta et al., 2023; Rymarczyk et al., 2022c).

836 MORE DETAILS ON COMPUTING RESOURCES  
 837

838 We ran our experiments on an internal cluster and a local cloud provider, a single GPU, it  
 839 was either NVIDIA A100 40GB or NVIDIA H100 80GB. The node we ran the experiments  
 840 on has 40GB of RAM and an 8-core CPU. The model on average trains for 3 hours.  
 841

842 MORE DETAILS ON USER STUDIES WITH EXEMPLARY SURVEYS.  
 843

844 Each worker answering a short 10-question survey was paid 1.50 euros. Questions between  
 845 users may differ as they are randomly composed. Participants are gender-balanced and have  
 846 ages from 18 to 60.  
 847

848 **User study on quality of prototypical parts.** For PIP-Net, we randomly select sam-  
 849 ples with  $K' = 4, 3, 2, 1$  in the proportion of  $5 : 3 : 2 : 1$  based on the frequency of occurrence  
 850 as PIP-Net doesn’t have the same number of prototypical parts assigned to data classes. The  
 851 LucidPPN pieces of evidence for classes in the same samples always show four prototypical  
 852 parts as we use a model trained with  $K = 4$  here.

853 Example surveys for LucidPPN, PIP-Net, and *single branch* are presented in Figures 25 to  
 854 37, 38 to 50, and 51 to 63, respectively.  
 855

856 **User study on the importance of disentangled visual features.** Because we focus  
 857 on the influence of the color features, we use visualizations with a random single object  
 858 and prototypical part to let the user focus on the influence of the color. When gathering  
 859 samples for the survey, we make sure that for nearly half of them color was important for the  
 860 correct prediction, and for half of them, it was not. We define that the color was important,  
 861 when LucidPPN was correct, but *ShapeTexNet* was wrong. And, we define that color was  
 862 unimportant if both outputs were correct.

863 Example surveys for LucidPPN with color feature scores and without them are presented  
 in Figures 64 to 76, and 77 to 89, respectively.

DETAILED RESULTS OF THE USER STUDY

In Tables 5 and 6, we present detailed results of the user studies. We also visualize the results of the user study on the importance of scores in the Figure 7.

Table 5: User study results indicate that users based on LucidPPN explanations outperform those with explanations from PIP-Net to a statistically significant degree.

	Mean Acc. [%] ± Std.	random	<i>p</i> -value PIP-Net	LucidPPN
PIP-Net	60.0 ± 18.1	0.002	–	–
LucidPPN	67.9 ± 16.9	$2.13 \cdot 10^{-6}$	0.044	–
<i>single branch</i>	69.9 ± 12.8	$1.11 \cdot 10^{-9}$	0.008	0.299

Table 6: Details of the user study about assessing the importance of color.

	Mean Acc. [%] ± Std.	random	<i>p</i> -value without resemblances
without resemblances	49.50 ± 11.3	0.577	–
with resemblances (LucidPPN)	60.87 ± 19.9	0.012	0.016

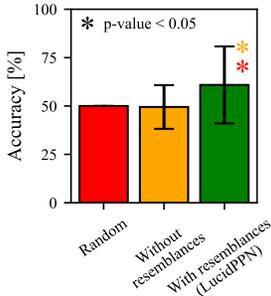


Figure 7: User study shows that disentangled resemblance scores enable users to better understand the relevance of color in model’s decisions.

COMPARISON EXPLANATION EXAMPLE

We show how our model can generate explanations by comparison of two potential classes in Figure 8.

COLOR REPRESENTATION

We have performed an ablation study to evaluate how different color representations of the input  $x_c$  influence the model’s results. Instead of directly downsizing the RGB image, we first transformed it into HSV space, replaced the S and V values with the Hue value, and then downsized the image. In other words, the input to the network was an image composed of the Hue channel repeated three times. The results demonstrate that LucidPPN with this input still outperforms *ShapeTexNet*, with performance similar to the basic LucidPPN as presented in the Table 7.

COLORNET SIZE

Since the architecture of *ColorNet* may significantly impact LucidPPN’s performance, we conducted an ablation study on the architecture’s size. Table 8 presents the accuracy comparison for CUB. All layers are  $1 \times 1$  convolutions followed by ReLU, except the last layer, which is followed by a sigmoid activation. The results indicate that using at least two layers

918  
919  
920  
921  
922  
923  
924  
925  
926  
927  
928  
929  
930  
931  
932  
933  
934  
935  
936  
937  
938  
939  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
950  
951  
952  
953  
954  
955  
956  
957  
958  
959  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
970  
971

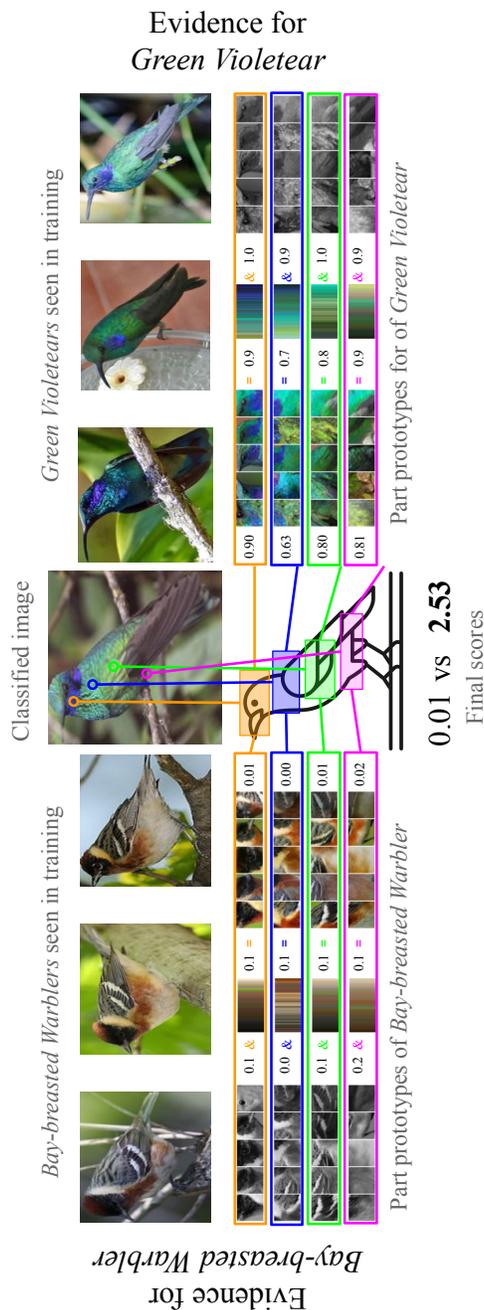


Figure 8: Comparison explanation example. Best viewed in landscape orientation.

to introduce non-linearity is beneficial. Additional layers have a smaller impact but can be added to ensure sufficient expressiveness of the network.

QUALITATIVE EXAMPLES OF FAILURE CASES WITHOUT DISENTANGLEMENT THAT ARE IMPROVED THROUGH LUCIDPPN

The main goal of the disentanglement is not to improve the accuracy but to provide a better understanding of the model’s reasoning based on color and shape with texture information. The explanations containing *ShapeTexNet*, *ColorNet*, and *aggregated resemblances* intro-

Table 7: Accuracy of LucidPPN when ColorNet receives RGB values vs. only hue value. *ShapeTexNet* added for comparison.

	ColorNet input	CUB	CARS	DOGS	FLOWER
<i>ShapeTexNet</i>	-	80.1 ± 0.2	91.7 ± 0.1	79.0 ± 0.3	93.6 ± 0.3
LucidPPN	RGB	81.5 ± 0.4	91.6 ± 0.2	79.4 ± 0.4	95.0 ± 0.3
	Only Hue	81.1 ± 0.4	91.6 ± 0.2	79.5 ± 0.2	94.1 ± 0.5

Table 8: LucidPPN’s accuracy for CUB vs the size of *ColorNet*.

Number of layers	Hidden dimensions	Accuracy [%]
1	-	80.3
2	(600)	81.4
4	(50, 200, 600)	81.2
6	(20, 50, 150, 200, 600)	<b>81.5</b>

duced in this work offer this additional information. Such an insight was missing in the previous prototypical parts models (Chen et al., 2019; Nauta et al., 2023).

Nevertheless, disentanglement can enhance accuracy in scenarios where shape and texture are the primary decision factors, with color serving to refine decisions that are difficult to make based on other features alone. Figure 9 illustrates examples from CARS and CUB where LucidPPN adheres to this principle, whereas the *single branch* CNN does not. Table 9 shows how often *single branch* CNN with colored input misclassifies color-altered images compared to the LucidPPN, and vice versa, for two specific data classes scenarios:

1. For test images of the typically red *Lamborghini Aventador*, which were converted to green and yellow via hue rotation, the *single branch* CNN mistakenly classified these altered images as either the typically green *Lamborghini Gallardo* or the usually yellow *Lamborghini Diablo* 24 times, despite the noticeable differences in shape (e.g., headlights and bumpers). In contrast, LucidPPN correctly classified these altered images. LucidPPN only made such a mistake 3 times, while the *single branch* CNN did not.
2. For test images of the red *Cardinal*, which were similarly converted to yellow and indigo, the *single branch* CNN misclassified the *Cardinal* as any other bird 34 times. LucidPPN made this mistake only once, while the *single branch* CNN was correct in all other instances.

Table 9: Number of found examples for which LucidPPN and *single branch* outperformed each other when asked to predict class in the color altered images.

	<i>Lamborghini Aventador</i>	<i>Cardinal</i>
LucidPPN correct but <i>single branch</i> wrong	24	34
LucidPPN wrong but <i>single branch</i> correct	3	1

PROTOTYPICAL PARTS EXAMPLES TRAINED WITHOUT PART SUPERVISION

are presented in Figure 10.

1026  
1027  
1028  
1029  
1030  
1031  
1032  
1033  
1034  
1035  
1036  
1037  
1038  
1039  
1040  
1041  
1042  
1043  
1044  
1045  
1046  
1047  
1048  
1049  
1050  
1051  
1052  
1053  
1054  
1055  
1056  
1057  
1058  
1059  
1060  
1061  
1062  
1063  
1064  
1065  
1066  
1067  
1068  
1069  
1070  
1071  
1072  
1073  
1074  
1075  
1076  
1077  
1078  
1079

Original image	Altered images	Predictions	Wrong class guessed by single branch
 <b>Correct label:</b> <i>Lamborghini Aventador</i>		<i>single branch:</i> <i>Lamborghini Gallardo</i> LucidPPN: <i>Lamborghini Aventador</i> <i>single branch:</i> <i>Lamborghini Diablo</i> LucidPPN: <i>Lamborghini Aventador</i>	
 <b>Correct label:</b> <i>Cardinal</i>		<i>single branch:</i> <i>Blue-winged Warbler</i> LucidPPN: <i>Cardinal</i> <i>single branch:</i> <i>Indigo Bunting</i> LucidPPN: <i>Cardinal</i>	

Figure 9: Examples of images with altered colors that change the prediction of a single branch CNN include a Lamborghini Aventador (top) and a Cardinal (bottom). Both are incorrectly classified by the single branch CNN, while LucidPPN with color disentangling classifies them correctly.

NUMBER OF PARTS

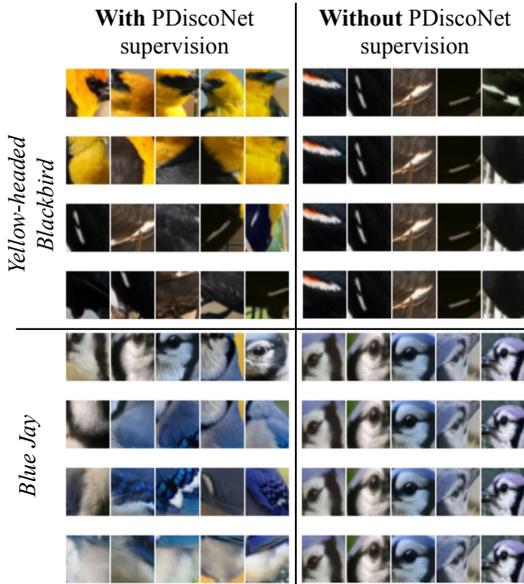
In Figure 11, we show the impact of choosing a different number of parts  $K$ . LucidPPN achieves high results for all tested  $K$ , however it is noticeable that increasing  $K$  improves classification. Especially on CARS, our method seems to strongly benefit from choosing  $K \geq 4$ . The reasonably high scores for all  $K$  allow for a choice between sparse explanations and higher accuracy.

NEED FOR  $L_S$

Many prototypical-parts-based models, such as ProtoPNet (Chen et al., 2019), ProtoPool (Rymarczyk et al., 2022d), and PIP-Net (Nauta et al., 2023), involve complex training schemes with warm-up and pretraining phases. Initially, we believed that *ShapeTexNet* should be pretrained before training *ColorNet*, given that *ShapeTexNet* processes more complex data. However, the ablation study presented in Figure 12 shows that warming up *ShapeTexNet* (or delaying the training of *ColorNet*) is either unnecessary or may even negatively influence color-based explanations.

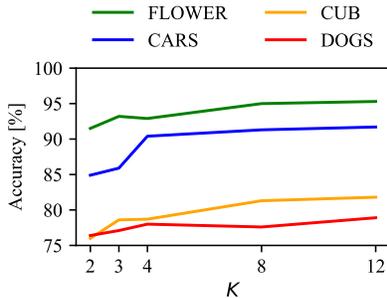
During the initial development of LucidPPN, we used  $L_S$  to guide the learning of *ShapeTexNet* during its warm-up phase. Once we observed that *ShapeTexNet* did not require a warm-up, we switched to jointly using  $L_S$  and  $L_A$  in training. We found that removing  $L_S$  negatively impacted LucidPPN’s performance which is presented in Figure 6 in the manuscript. Consequently, we retained  $L_S$ , as it provides essential guidance for *ShapeTexNet* to effectively extract important features from its more complex input.

1080  
1081  
1082  
1083  
1084  
1085  
1086  
1087  
1088  
1089  
1090  
1091  
1092  
1093  
1094  
1095  
1096  
1097  
1098  
1099



1100 Figure 10: Examples from two classes demonstrate that prototypes learned without PDiscoNet  
1101 supervision focus on a single object part, in contrast to those more diverse learned  
1102 by LucidPPN.

1103  
1104  
1105  
1106  
1107  
1108  
1109  
1110  
1111  
1112  
1113  
1114



1115 Figure 11: Influence of the number of object parts  $K$  on LucidPPN accuracy. Increasing  
1116 the number of parts improves the accuracy of the model. Note that each dataset is plotted  
1117 in a unique color.

1118  
1119  
1120  
1121  
1122  
1123  
1124  
1125  
1126  
1127  
1128

This need for stronger guidance aligns with observations in multimodal learning, where separate learning of different modality branches maximizes the information extracted from each modality (Wu et al., 2022). Here, we can think of each branch as different modalities. Alternatively, using a weighted average could yield similar accuracy, but it would complicate the final prediction. This approach would necessitate analyzing the contribution of each logit vector separately and understanding their aggregation, with potentially different weights for each dataset, making the process less transparent for the user.

START OF COLOR NETWORK TRAINING

1129  
1130  
1131  
1132  
1133

It is natural to ask whether delaying the start of *ColorNet* optimization could improve LucidPPN. In Figure 12, we report the accuracy and color sparsity after delaying the training of *ColorNet*. The change in classification quality is negligible. However, we observe a drop in color sparsity, indicating that *ColorNet* is less focused on relevant colors. It is important to note that despite the delay, the number of training epochs for *ColorNet* remains constant for comparability.

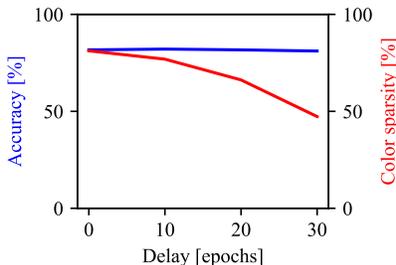


Figure 12: Influence of a delay when *ColorNet* starts to train on LucidPPN’s accuracy and color sparsity. While this delay does not negatively affect accuracy, it results in lower color sparsity. This means that the network is not concentrating on a single color when processing the PP.

#### LUCIDPPN WITH DIFFERENT BACKBONES

We evaluate LucidPPN on an additional ResNet50 backbone and compare the results to other models and backbones in Table 10. LucidPPN with ResNet50 backbone performs worse than the one with ConvNeXt-tiny, which is similar to PIP-Net. When comparing the results, note that iNaturalist-pretrained backbones have an advantage over ImageNet resulting in a few points higher accuracy.

Table 10: Accuracy for different PP-based methods and backbones. The asterisk means that the used backbone was pretrained on the iNaturalist dataset instead of ImageNet.

		CUB	CARS	DOGS	FLOWER
ProtoPNet	ResNet34	79.2	86.1	-	-
ProtoPShare	ResNet34	74.7	86.4	-	-
ProtoTree	ResNet50	82.2 ± 0.7*	86.6 ± 0.2	-	-
ProtoPool	ResNet50	<b>85.5 ± 0.1*</b>	88.9 ± 0.1	-	-
PIP-Net	ResNet50	82.0 ± 0.3*	86.5 ± 0.3	-	-
	ConvNeXt-tiny	84.3 ± 0.2	88.2 ± 0.5	<b>80.8 ± 0.4</b>	91.8 ± 0.5
LucidPPN	ResNet50	75.5 ± 1.1	89.0 ± 0.3	70.8 ± 0.2	89.5 ± 0.4
	ConvNeXt-tiny	81.5 ± 0.4	<b>91.6 ± 0.2</b>	79.4 ± 0.4	<b>95.0 ± 0.3</b>

#### DISCUSSION ON PATCHES OF THE COLOR INFO AND THE PATCH OF GRAY-SCALED INPUT ALIGNED IN THE LATENT SPACE

Using a convolutional backbone, we assume a spatial correspondence between the latent map from *ShapeTexNet* and the input, similar to the approach in ProtoPNet (Chen et al., 2019). As we downsize the colorful image to match the height and width of the activation map, the input dimensions for *ColorNet* are maintained consistently. *ColorNet* employs  $1 \times 1$  convolutions to encode color information, ensuring the latent map has the same dimensions as both the downsized input and the *ShapeTexNet* activation map.

Given the use of  $1 \times 1$  convolutions on the downsized image and a convolutional backbone for the full-resolution image, we can assume that the  $(i, j)$  position on one map corresponds to the  $(i, j)$  position on the other. Finally, we extract color information from the latent representation at the same location where the prototypical part is most active, ensuring alignment between the color and shape features.

#### LOCAL INTERPRETATION OF SINGLE BRANCH

An example of prediction interpretation of *single branch* is presented in Figure 13.

1188  
 1189  
 1190  
 1191  
 1192  
 1193  
 1194  
 1195  
 1196  
 1197  
 1198  
 1199  
 1200  
 1201  
 1202  
 1203  
 1204  
 1205  
 1206  
 1207  
 1208  
 1209  
 1210  
 1211  
 1212  
 1213  
 1214  
 1215  
 1216  
 1217  
 1218  
 1219  
 1220  
 1221  
 1222  
 1223  
 1224  
 1225  
 1226  
 1227  
 1228  
 1229  
 1230  
 1231  
 1232  
 1233  
 1234  
 1235  
 1236  
 1237  
 1238  
 1239  
 1240  
 1241

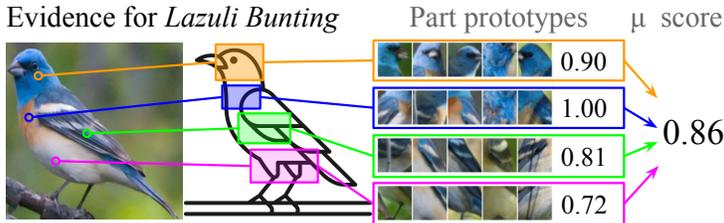


Figure 13: An example local interpretation of *single branch* for *Lazuli Bunting*.

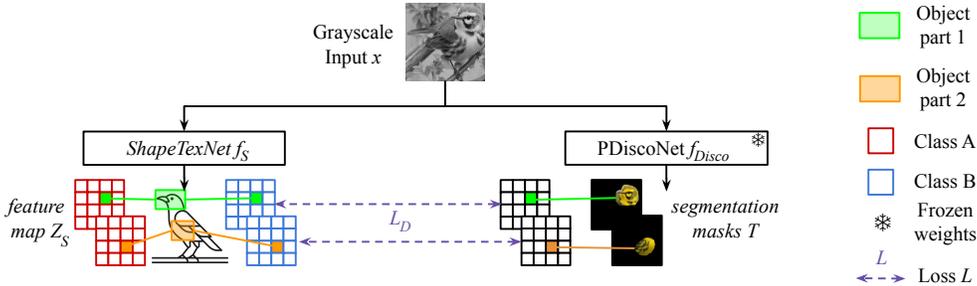


Figure 14: The image illustrates the object-part correspondence loss, which is applied solely to the outputs of *ShapeTexNet* and *PDiscoNet*. First, we identify object parts through *PDiscoNet* (e.g., first on the head, second on the wing). Next, we align the corresponding prototypical parts with the object parts identified by the segmentation results through  $L_D$  loss.

OBJECT-PART CORRESPONDENCE

On Figure 14 we present how object-part correspondence  $L_D$  loss works.

CONSISTENCY AND STABILITY OF PROTOTYPICAL PARTS

One way to evaluate the quality of prototypical parts is to measure their consistency and stability Huang et al. (2023). In Table 11. we present the results of those metrics. The results show that LucidPPN achieves state-of-the-art results on stability without any additional loss components, and is comparable to other metrics when it comes to stability. This improvement is likely due to the enhanced object-part correspondence enabled by its prototypical parts.

COMPUTATIONAL COSTS

In Table 12, we provide computational costs of training LucidPPN when compared to other prototypical-parts-based architectures.

GENERALIZATION TO NOT FINE-GRAINED DATASET

To assess whether LucidPPN generalizes to broader classification tasks (beyond fine-grained datasets), we present results on PartImageNet He et al. (2022). On this dataset, LucidPPN achieves an accuracy of 84.1%, outperforming PIPNet, which achieves 82.8%.

COMPARISON OF EXPLANATION VISUALIZATIONS

In Figures 15, 16, 17, 18, and 19 we compare the decision explanations generated by different methods.

Table 11: Results of LucidPPN on consistency and stability metrics from the work of Huang et al. (2023). The results indicate that LucidPPN is more robust than other prototypical-parts-based approaches and achieves state-of-the-art results for Consistency while still remaining competitive in Stability.

Method	Consistency	Stability
ProtoPNet	28.3	56.7
ProtoTree	16.4	23.2
ProtoPool	35.7	58.4
TesNet	48.6	60.0
Deformable ProtoPNet	44.2	53.5
Huang et al. (2023)	70.6	<b>72.1</b>
LucidPPN (our)	<b>71.2</b>	66.3

Table 12: Computational costs of prototypical-parts-based methods. One can observe that training of LucidPPN requires fewer hours and less RAM memory than PIP-Net, but more GFLOPs. Generally, LucidPPN and PIPNet require more RAM memory than ProtoPNet and ProtoPool, however they converge faster.

Method	Training time	GFLOPs for 1 batch of data	Avg. Training Memory Usage
ProtoPNet	3h	586	4.9GB
ProtoPool	18h	658	14.4GB
PIP-Net	3h	354	41.5GB
LucidPPN (our)	2h	475	22.9GB

#### FAITHFULLNESS OF PATCH VISUALIZATIONS

LucidPPN introduces a key difference in the definition of prototypical parts compared to PIPNet. While PIPNet employs Softmax across channels in the latent feature map, LucidPPN uses the sigmoid activation function. The sigmoid function allows each channel’s activation to be learned independently, not influenced by the relative activations of other channels. At the same time, Softmax normalization can distort activations by emphasizing values that are only relatively high compared to others, even if they are low in absolute terms. Therefore, using the sigmoid function instead of Softmax, one can easily verify if the image patches selected for visualization are faithful because such patches should have a resemblance score close to 1. In Figure 20, we provide a distribution of the sigmoid function values obtained for patches used in prototype visualization. For LucidPPN trained on the CUB dataset (blue curve), 61.04% of those patches have values above 0.9, which indicates that prototype visualizations are relatively faithful. Moreover, higher faithfulness can be obtained when training with an additional loss component  $L_C$  that punishes the model if the sigmoid function value for a given prototype is smaller than 1 for all samples in the batch:

$$L_C = \frac{1}{KM} \sum_{k=1}^K \sum_{m=1}^M \max_{b \in B} (1 - r_{A,b}^{km}),$$

where  $B$  is the number of samples in a batch and  $r_{A,b}^{km}$  is the value of  $r_A^{km}$  for sample  $b$  in the batch. As we observe in Figure 20 (green and yellow curves), the distribution of sigmoid function values moves right with increasing weight  $\alpha_C$  of  $L_C$ . However, it also comes with a small decrease in accuracy.

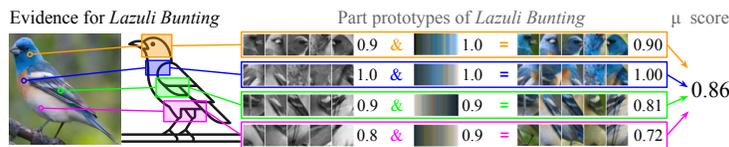
#### PRUNING PROTOTYPES WITH LESS FAITHFUL VISUALIZATIONS

To increase the faithfulness of LucidPPN, we analyze the effects of pruning the prototypes with less faithful representation (those with resemblance scores  $< 0.9$ ). As shown in Table 13, LucidPPN accuracy after pruning drops only by around 2% (from 81.6% to 79.3%). However, interestingly, the accuracy stays the same for  $L_C = 0.05$ . It suggests that combi-

1296 nation of using loss  $L_C$  and applying the pruning allows to enforce high resemblance scores  
 1297 ( $>0.9$ ) of visualized patches without sacrificing on the accuracy.  
 1298

1299 REASON BEHIND USING THE BINARY CROSS ENTROPY WITH SIGMOID INSTEAD OF THE  
 1300 CROSS ENTROPY WITH SOFTMAX  
 1301

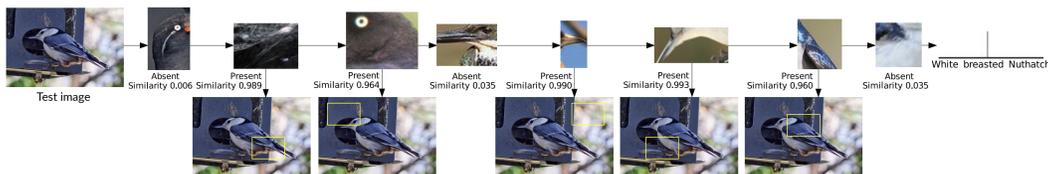
1302 The intuition behind Binary Cross Entropy (BCE) usage is rooted from multilabel classification.  
 1303 To some degree, ShapeTexNet operates in a multilabel setting from the prototypical parts  
 1304 perspective, as they may match multiple classes. Hence, to enable multiple classes  
 1305 having high similarity to the same prototypical parts, we use sigmoid instead of softmax  
 1306 when computing the feature maps. This necessitates a shift from Cross-Entropy (CE) to  
 1307 Binary Cross-Entropy (BCE) because CE would then solely maximize the activation of the  
 1308 correct class while ignoring crucial signals from negative classes. Another reason behind our  
 1309 choice is to make it easier to verify the faithfulness of visualizations, as the sigmoid function  
 1310 allows each channel’s activation to be learned independently, not influenced by the relative  
 1311 activations of other channels. While, Softmax normalization used with CE can distort acti-  
 1312 vations by emphasizing values that are only relatively high compared to others, even if they  
 1313 are low in absolute terms.



1314 Figure 15: Local interpretation visualization in LucidPPN



1322 Figure 16: Local interpretation visualization in PIP-Net



1337 Figure 17: Local interpretation visualization in ProtoTree

1340 GLOBAL CHARACTERISTICS EXAMPLES  
 1341

1342 We present global characteristics for different datasets in Figures 21, 22, 23, 24, 25.  
 1343  
 1344  
 1345  
 1346  
 1347  
 1348  
 1349

1350 Why is that *Ford Freestar Minivan 2007*?

1351	Image	Prototypical part	Activation map	Similarity and weigh
1352				$4.08 \cdot 0.89 = 3.63$
1353				$3.87 \cdot 0.83 = 3.21$
1354				
1355				
1356				
1357				
1358				
1359				
1360				$3.82 \cdot 0.81 = 3.09$
1361				<u>SUM: 21.64</u>

1362 Figure 18: Local interpretation visualization in ProtoPool

1366 Why is this bird classified as a red-bellied woodpecker?

1367

1368 Evidence for this bird being a red-bellied woodpecker:

1370	Original image (box showing part that looks like prototype)	Prototype	Training image where prototype comes from	Activation map	Similarity score	Class connection	Points contributed
1371					6.499	1.180	$7.669$
1372							
1373							
1374							
1375							
1376							
1377					4.392	1.127	$4.950$
1378							
1379							
1380							
1381					3.890	1.108	$4.310$
1382							
1383							
1384							
1385							
1386							
1387							
1388							
1389							
1390							
1391							
1392							
1393							
1394							
1395							
1396							
1397							
1398							
1399							
1400							
1401							
1402							
1403							

Total points to red-bellied woodpecker: 32.736

Figure 19: Local interpretation visualization in ProtoPNet

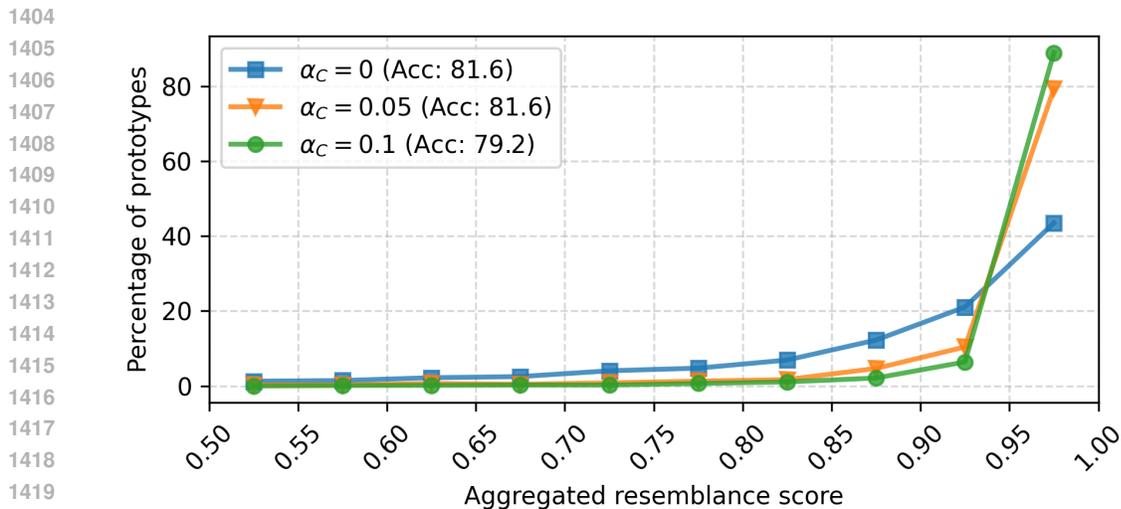


Figure 20: Distribution of the aggregated resemblance scores for different weights of cluster loss  $L_C$ . For LucidPPN ( $\alpha_C = 0$ ) trained on CUB dataset, 61.04% of patches representing prototypes have aggregated resemblance score above 0.9, which indicates that prototype visualizations are relatively faithful. Moreover, when training with additional loss function  $L_C$ , we obtain 95.33% patches with values above 0.9, with only a small drop in accuracy.

Table 13: Accuracy before and after pruning the prototypes with less faithful visualizations. The results show that the combination of training with loss  $L_C$  and pruning can enforce faithfulness of visualizations without the loss in performance.

$\alpha_C$	Before pruning	After pruning
0	81.6	79.3
0.05	81.6	81.6
0.1	79.2	79.2

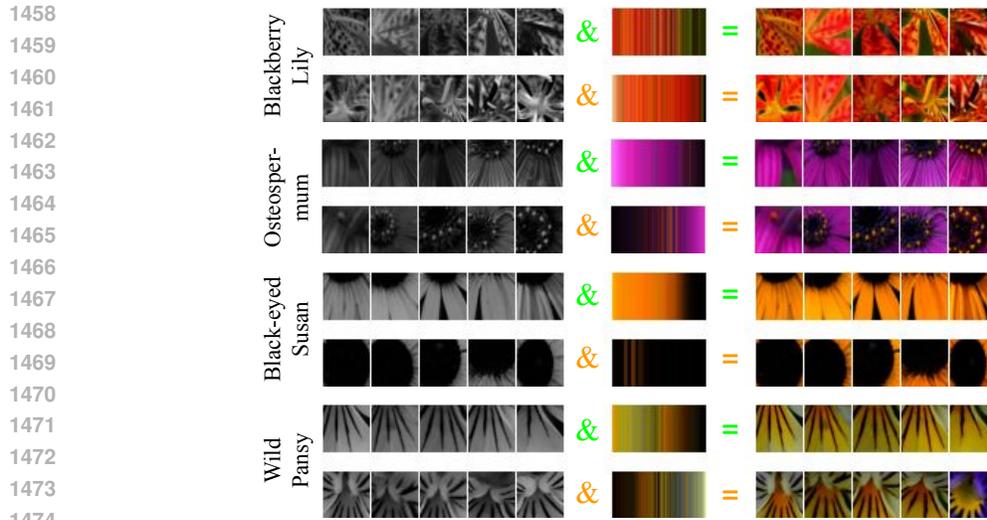
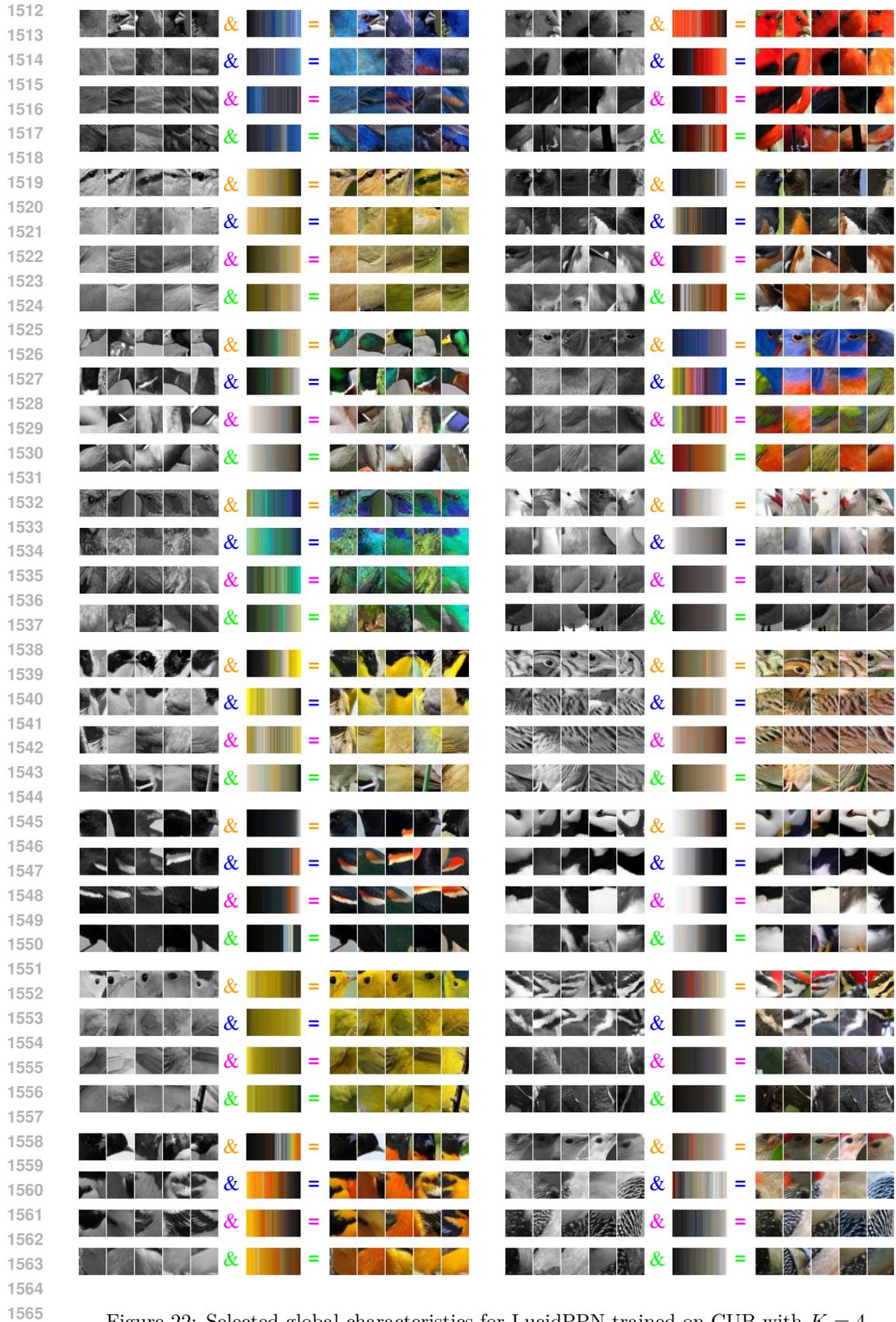
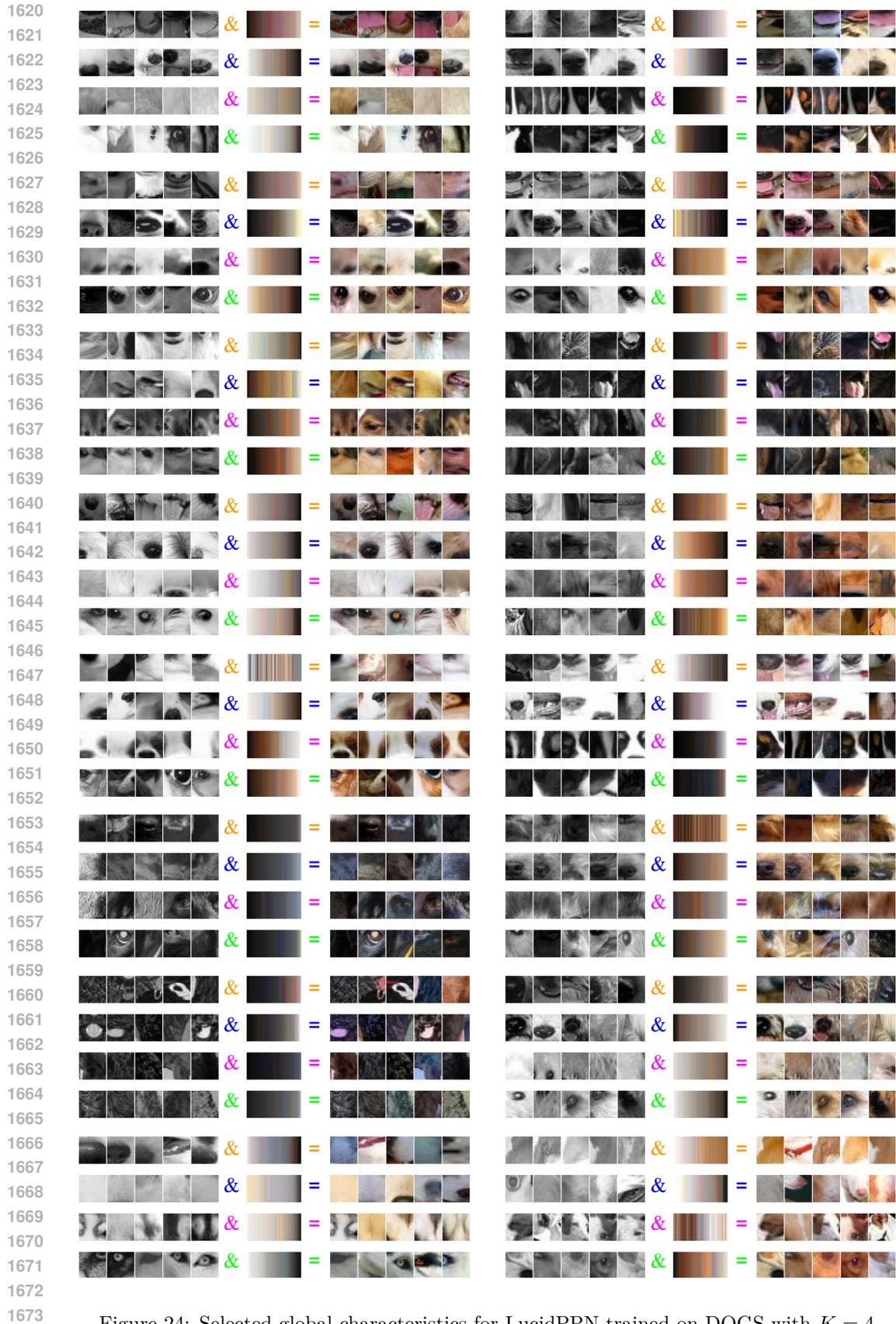


Figure 21: An example showcasing global characteristics of four classes in the FLOWER dataset, using prototypical parts from LucidPPN trained with  $K = 2$ . This visualization demonstrates the ability to detect differences between data classes. For instance, the *osteospermum* and *black-eyed susan* exhibit more variation in color, while the *blackberry lilly* and *wild pansay* classes differ in texture and shape.





Figure 24: Selected global characteristics for LucidPPN trained on DOGS with  $K = 4$

1674  
 1675  
 1676  
 1677  
 1678  
 1679  
 1680  
 1681  
 1682  
 1683  
 1684  
 1685  
 1686  
 1687  
 1688  
 1689  
 1690  
 1691  
 1692  
 1693  
 1694  
 1695  
 1696  
 1697  
 1698  
 1699  
 1700  
 1701  
 1702  
 1703  
 1704  
 1705  
 1706  
 1707  
 1708  
 1709  
 1710  
 1711  
 1712  
 1713  
 1714  
 1715  
 1716  
 1717  
 1718  
 1719  
 1720  
 1721  
 1722  
 1723  
 1724  
 1725  
 1726  
 1727

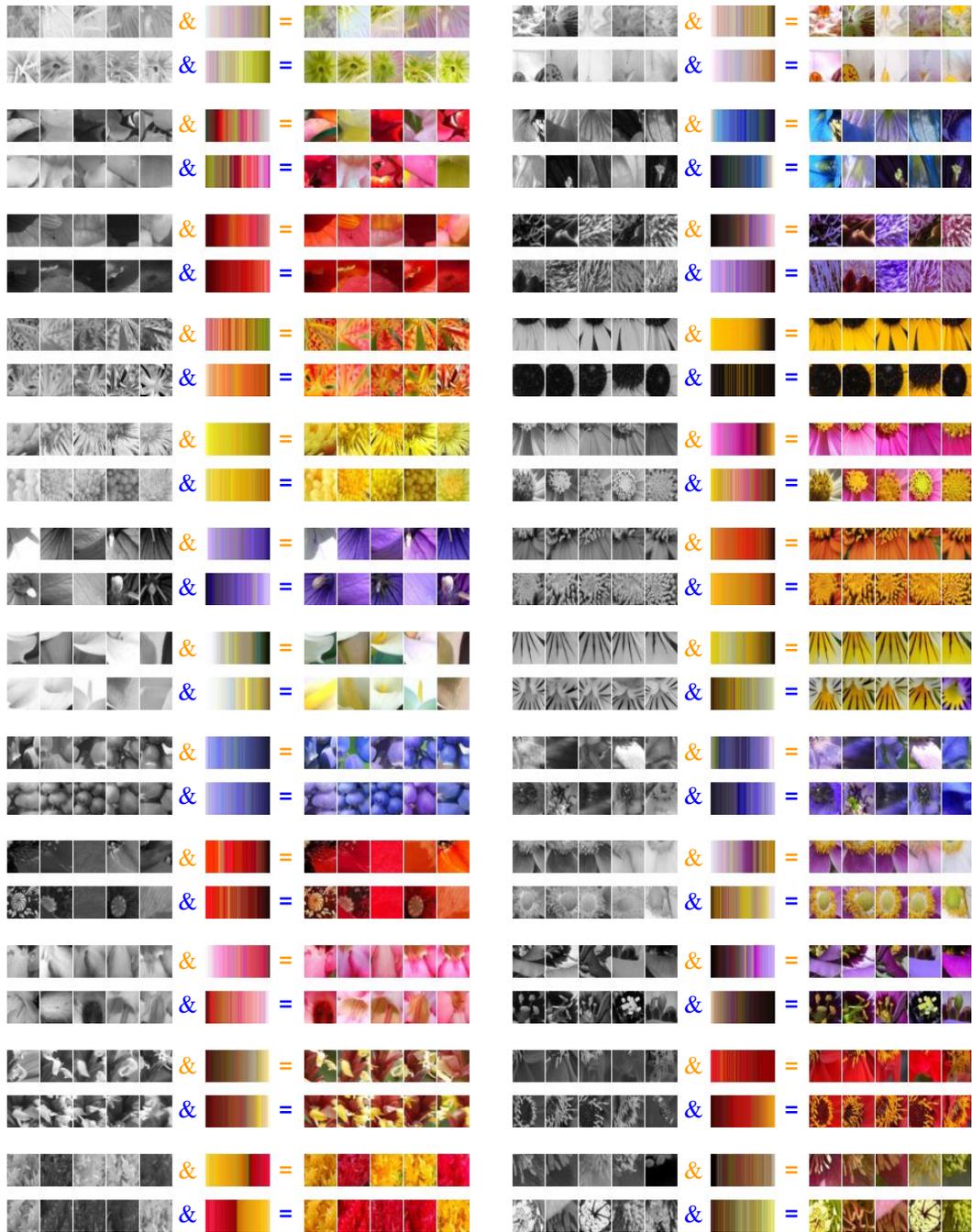


Figure 25: Selected global characteristics for LucidPPN trained on FLOWER with  $K = 2$

1728  
1729  
1730  
1731  
1732  
1733  
1734  
1735  
1736  
1737  
1738  
1739  
1740  
1741  
1742  
1743  
1744  
1745  
1746  
1747  
1748  
1749  
1750  
1751  
1752  
1753  
1754  
1755  
1756  
1757  
1758  
1759  
1760  
1761  
1762  
1763  
1764  
1765  
1766  
1767  
1768  
1769  
1770  
1771  
1772  
1773  
1774  
1775  
1776  
1777  
1778  
1779  
1780  
1781

Given a bird photo a model predicts the species based on prototypes it has learned from previously seen photos. Specifically for each prototype, the model identifies a region in the photo that looks the most similar to the prototype and rates their similarity. Greater similarity determines the classification of the bird into a specific species.

**For a given photo we show evidence for 2 bird species found by the model. Your task is to choose the species you think is predicted by the model based on that evidence.**

Random guessing will get you 50% accuracy. You will receive a reward only if your performance is reasonably beyond the random chance.

**Interface explanation**

An image being classified by the model

Prototypes' location

Each row shows a prototype of a bird part composed of two components:  
- shape with texture  
- color

Next

Figure 26: Page 1 of survey for LucidPPN

1782  
1783  
1784  
1785  
1786  
1787  
1788  
1789  
1790  
1791  
1792  
1793  
1794  
1795  
1796  
1797  
1798  
1799  
1800  
1801  
1802  
1803  
1804  
1805  
1806  
1807  
1808  
1809  
1810  
1811  
1812  
1813  
1814  
1815  
1816  
1817  
1818  
1819  
1820  
1821  
1822  
1823  
1824  
1825  
1826  
1827  
1828  
1829  
1830  
1831  
1832  
1833  
1834  
1835

Example 1. \*

Evidence for *species A*

Part prototypes of *species A*

Evidence for *species B*

Part prototypes of *species B*

Based on given evidence, the model would say the photo shows...

Correct answer: **species A**

Explanation: Evidence for species A is more similar to the image

species A

species B

Previous Next

Figure 27: Page 2 of survey for LucidPPN

1836  
1837  
1838  
1839  
1840  
1841  
1842  
1843  
1844  
1845  
1846  
1847  
1848  
1849  
1850  
1851  
1852  
1853  
1854  
1855  
1856  
1857  
1858  
1859  
1860  
1861  
1862  
1863  
1864  
1865  
1866  
1867  
1868  
1869  
1870  
1871  
1872  
1873  
1874  
1875  
1876  
1877  
1878  
1879  
1880  
1881  
1882  
1883  
1884  
1885  
1886  
1887  
1888  
1889

Example 2. \*

Evidence for *species A*

Part prototypes of *species A*

Evidence for *species B*

Part prototypes of *species B*

Based on given evidence, the model would say the photo shows...

Correct answer: **species B**

Explanation: Evidence for species B is more similar to the image

species A

species B

Previous Next

Figure 28: Page 3 of survey for LucidPPN

1890  
1891  
1892  
1893  
1894  
1895  
1896  
1897  
1898  
1899  
1900  
1901  
1902  
1903  
1904  
1905  
1906  
1907  
1908  
1909  
1910  
1911  
1912  
1913  
1914  
1915  
1916  
1917  
1918  
1919  
1920  
1921  
1922  
1923  
1924  
1925  
1926  
1927  
1928  
1929  
1930  
1931  
1932  
1933  
1934  
1935  
1936  
1937  
1938  
1939  
1940  
1941  
1942  
1943

Question 0. \*

Evidence for *species A*

Part prototypes of *species A*

Evidence for *species B*

Part prototypes of *species B*

Based on given evidence, the model would say the photo shows...

species A

species B

Previous Next

Figure 29: Page 4 of survey for LucidPPN

1944  
1945  
1946  
1947  
1948  
1949  
1950  
1951  
1952  
1953  
1954  
1955  
1956  
1957  
1958  
1959  
1960  
1961  
1962  
1963  
1964  
1965  
1966  
1967  
1968  
1969  
1970  
1971  
1972  
1973  
1974  
1975  
1976  
1977  
1978  
1979  
1980  
1981  
1982  
1983  
1984  
1985  
1986  
1987  
1988  
1989  
1990  
1991  
1992  
1993  
1994  
1995  
1996  
1997

Question 1. \*

Evidence for *species A*

Part prototypes of *species A*

Evidence for *species B*

Part prototypes of *species B*

Based on given evidence, the model would say the photo shows...

species A

species B

Previous Next

Figure 30: Page 5 of survey for LucidPPN

1998  
1999  
2000  
2001  
2002  
2003  
2004  
2005  
2006  
2007  
2008  
2009  
2010  
2011  
2012  
2013  
2014  
2015  
2016  
2017  
2018  
2019  
2020  
2021  
2022  
2023  
2024  
2025  
2026  
2027  
2028  
2029  
2030  
2031  
2032  
2033  
2034  
2035  
2036  
2037  
2038  
2039  
2040  
2041  
2042  
2043  
2044  
2045  
2046  
2047  
2048  
2049  
2050  
2051

Question 2. \*

Evidence for *species A*

Part prototypes of *species A*

Evidence for *species B*

Part prototypes of *species B*

Based on given evidence, the model would say the photo shows...

species A

species B

Previous Next

Figure 31: Page 6 of survey for LucidPPN

2052  
2053  
2054  
2055  
2056  
2057  
2058  
2059  
2060  
2061  
2062  
2063  
2064  
2065  
2066  
2067  
2068  
2069  
2070  
2071  
2072  
2073  
2074  
2075  
2076  
2077  
2078  
2079  
2080  
2081  
2082  
2083  
2084  
2085  
2086  
2087  
2088  
2089  
2090  
2091  
2092  
2093  
2094  
2095  
2096  
2097  
2098  
2099  
2100  
2101  
2102  
2103  
2104  
2105

Question 3. \*

Evidence for *species A*

Part prototypes of *species A*

Evidence for *species B*

Part prototypes of *species B*

Based on given evidence, the model would say the photo shows...

species A

species B

Previous Next

Figure 32: Page 7 of survey for LucidPPN

2106  
2107  
2108  
2109  
2110  
2111  
2112  
2113  
2114  
2115  
2116  
2117  
2118  
2119  
2120  
2121  
2122  
2123  
2124  
2125  
2126  
2127  
2128  
2129  
2130  
2131  
2132  
2133  
2134  
2135  
2136  
2137  
2138  
2139  
2140  
2141  
2142  
2143  
2144  
2145  
2146  
2147  
2148  
2149  
2150  
2151  
2152  
2153  
2154  
2155  
2156  
2157  
2158  
2159

Question 4. \*

Evidence for *species A*

Part prototypes of *species A*

Evidence for *species B*

Part prototypes of *species B*

Based on given evidence, the model would say the photo shows...

species A

species B

Previous Next

The figure displays a survey question interface. At the top, it says 'Question 4. \*'. Below this, there are two main sections: 'Evidence for species A' and 'Evidence for species B'. Each section contains a photograph of a bird on a branch. The bird in the 'Evidence for species A' section has four colored bounding boxes: orange on the head, blue on the neck, green on the back, and pink on the tail. To the right of each photo is a set of 'Part prototypes' for that species, consisting of four rows. Each row shows a sequence of images: a small crop of the part from the photo, followed by an ampersand (&), a vertical gradient bar, an equals sign (=), and a larger crop of the part. The colors of the bounding boxes and the corresponding part prototype rows match. Below these two sections, there is a text prompt: 'Based on given evidence, the model would say the photo shows...'. Underneath this prompt are two radio button options: 'species A' and 'species B'. At the bottom of the interface are two buttons: 'Previous' (light blue) and 'Next' (dark blue).

Figure 33: Page 8 of survey for LucidPPN

2160  
2161  
2162  
2163  
2164  
2165  
2166  
2167  
2168  
2169  
2170  
2171  
2172  
2173  
2174  
2175  
2176  
2177  
2178  
2179  
2180  
2181  
2182  
2183  
2184  
2185  
2186  
2187  
2188  
2189  
2190  
2191  
2192  
2193  
2194  
2195  
2196  
2197  
2198  
2199  
2200  
2201  
2202  
2203  
2204  
2205  
2206  
2207  
2208  
2209  
2210  
2211  
2212  
2213

Question 5. \*

Evidence for *species A*

Part prototypes of *species A*

Evidence for *species B*

Part prototypes of *species B*

Based on given evidence, the model would say the photo shows...

species A

species B

Previous Next

Figure 34: Page 9 of survey for LucidPPN

2214  
2215  
2216  
2217  
2218  
2219  
2220  
2221  
2222  
2223  
2224  
2225  
2226  
2227  
2228  
2229  
2230  
2231  
2232  
2233  
2234  
2235  
2236  
2237  
2238  
2239  
2240  
2241  
2242  
2243  
2244  
2245  
2246  
2247  
2248  
2249  
2250  
2251  
2252  
2253  
2254  
2255  
2256  
2257  
2258  
2259  
2260  
2261  
2262  
2263  
2264  
2265  
2266  
2267

Question 6. \*

Evidence for *species A*

Part prototypes of *species A*

Evidence for *species B*

Part prototypes of *species B*

Based on given evidence, the model would say the photo shows...

species A

species B

Previous Next

Figure 35: Page 10 of survey for LucidPPN

2268  
2269  
2270  
2271  
2272  
2273  
2274  
2275  
2276  
2277  
2278  
2279  
2280  
2281  
2282  
2283  
2284  
2285  
2286  
2287  
2288  
2289  
2290  
2291  
2292  
2293  
2294  
2295  
2296  
2297  
2298  
2299  
2300  
2301  
2302  
2303  
2304  
2305  
2306  
2307  
2308  
2309  
2310  
2311  
2312  
2313  
2314  
2315  
2316  
2317  
2318  
2319  
2320  
2321

Question 7. \*

Evidence for *species A*

Part prototypes of *species A*

Evidence for *species B*

Part prototypes of *species B*

Based on given evidence, the model would say the photo shows...

species A

species B

Previous Next

Figure 36: Page 11 of survey for LucidPPN

2322  
2323  
2324  
2325  
2326  
2327  
2328  
2329  
2330  
2331  
2332  
2333  
2334  
2335  
2336  
2337  
2338  
2339  
2340  
2341  
2342  
2343  
2344  
2345  
2346  
2347  
2348  
2349  
2350  
2351  
2352  
2353  
2354  
2355  
2356  
2357  
2358  
2359  
2360  
2361  
2362  
2363  
2364  
2365  
2366  
2367  
2368  
2369  
2370  
2371  
2372  
2373  
2374  
2375

Question 8. \*

Evidence for *species A*

Part prototypes of *species A*

Evidence for *species B*

Part prototypes of *species B*

Based on given evidence, the model would say the photo shows...

species A

species B

Previous Next

Figure 37: Page 12 of survey for LucidPPN

2376  
2377  
2378  
2379  
2380  
2381  
2382  
2383  
2384  
2385  
2386  
2387  
2388  
2389  
2390  
2391  
2392  
2393  
2394  
2395  
2396  
2397  
2398  
2399  
2400  
2401  
2402  
2403  
2404  
2405  
2406  
2407  
2408  
2409  
2410  
2411  
2412  
2413  
2414  
2415  
2416  
2417  
2418  
2419  
2420  
2421  
2422  
2423  
2424  
2425  
2426  
2427  
2428  
2429

Question 9. \*

Evidence for *species A*

Part prototypes of *species A*

Evidence for *species B*

Part prototypes of *species B*

Based on given evidence, the model would say the photo shows...

species A

species B

Previous Send Job

Figure 38: Page 13 of survey for LucidPPN

2430  
2431  
2432  
2433  
2434  
2435  
2436  
2437  
2438  
2439  
2440  
2441  
2442  
2443  
2444  
2445  
2446  
2447  
2448  
2449  
2450  
2451  
2452  
2453  
2454  
2455  
2456  
2457  
2458  
2459  
2460  
2461  
2462  
2463  
2464  
2465  
2466  
2467  
2468  
2469  
2470  
2471  
2472  
2473  
2474  
2475  
2476  
2477  
2478  
2479  
2480  
2481  
2482  
2483

Given a bird photo a model predicts the species based on prototypes it has learned from previously seen photos. Specifically for each prototype, the model identifies a region in the photo that looks the most similar to the prototype and rates their similarity. Greater similarity determines the classification of the bird into a specific species.

**For a given photo we show evidence for 2 bird species found by the model. Your task is to choose the species you think is predicted by the model based on that evidence.**

Random guessing will get you 50% accuracy. You will receive a reward only if your performance is reasonably beyond the random chance.

**Interface explanation**

An image being classified by the model

Each row shows a prototype of a bird part

The interface displays two panels, one for 'Evidence for species A' and one for 'Evidence for species B'. Each panel shows a main image of a bird with colored bounding boxes (yellow, blue, green, purple) around different parts. To the right of each main image is a grid of 'Learned Part Prototypes' corresponding to the colored regions. Red arrows point from the text boxes to the main images and the prototype grids.

Next

Figure 39: Page 1 of survey for PIP-Net

2484  
2485  
2486  
2487  
2488  
2489  
2490  
2491  
2492  
2493  
2494  
2495  
2496  
2497  
2498  
2499  
2500  
2501  
2502  
2503  
2504  
2505  
2506  
2507  
2508  
2509  
2510  
2511  
2512  
2513  
2514  
2515  
2516  
2517  
2518  
2519  
2520  
2521  
2522  
2523  
2524  
2525  
2526  
2527  
2528  
2529  
2530  
2531  
2532  
2533  
2534  
2535  
2536  
2537

Example 1. \*

**Evidence for *species A***

*Learned Part Prototypes*

**Evidence for *species B***

*Learned Part Prototypes*

Based on given evidence, the model would say the photo shows...

Correct answer: **species A**

Explanation: Evidence for species A is more similar to the image

species A

species B

Previous Next

Figure 40: Page 2 of survey for PIP-Net

2538  
2539  
2540  
2541  
2542  
2543  
2544  
2545  
2546  
2547  
2548  
2549  
2550  
2551  
2552  
2553  
2554  
2555  
2556  
2557  
2558  
2559  
2560  
2561  
2562  
2563  
2564  
2565  
2566  
2567  
2568  
2569  
2570  
2571  
2572  
2573  
2574  
2575  
2576  
2577  
2578  
2579  
2580  
2581  
2582  
2583  
2584  
2585  
2586  
2587  
2588  
2589  
2590  
2591

Example 2. \*

Evidence for *species A*

Learned Part Prototypes

Evidence for *species B*

Learned Part Prototypes

Based on given evidence, the model would say the photo shows...

Correct answer: **species B**

Explanation: Evidence for species B is more similar to the image

species A

species B

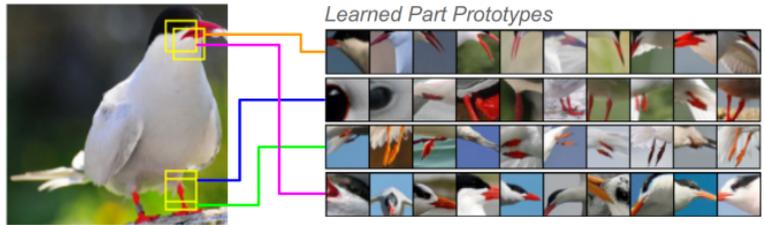
Previous Next

Figure 41: Page 3 of survey for PIP-Net

2592  
2593  
2594  
2595  
2596  
2597  
2598  
2599  
2600  
2601  
2602  
2603  
2604  
2605  
2606  
2607  
2608  
2609  
2610  
2611  
2612  
2613  
2614  
2615  
2616  
2617  
2618  
2619  
2620  
2621  
2622  
2623  
2624  
2625  
2626  
2627  
2628  
2629  
2630  
2631  
2632  
2633  
2634  
2635  
2636  
2637  
2638  
2639  
2640  
2641  
2642  
2643  
2644  
2645

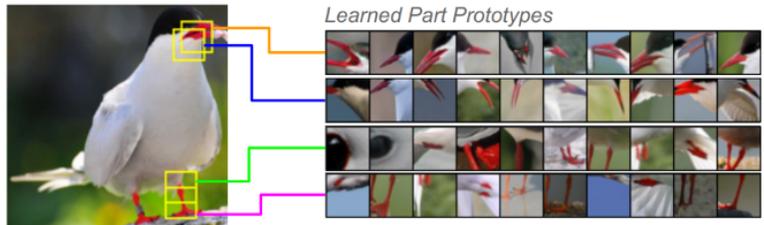
Question 0. \*

Evidence for *species A*



Learned Part Prototypes

Evidence for *species B*



Learned Part Prototypes

Based on given evidence, the model would say the photo shows...

species A

species B

Previous Next

Figure 42: Page 4 of survey for PIP-Net

2646  
2647  
2648  
2649  
2650  
2651  
2652  
2653  
2654  
2655  
2656  
2657  
2658  
2659  
2660  
2661  
2662  
2663  
2664  
2665  
2666  
2667  
2668  
2669  
2670  
2671  
2672  
2673  
2674  
2675  
2676  
2677  
2678  
2679  
2680  
2681  
2682  
2683  
2684  
2685  
2686  
2687  
2688  
2689  
2690  
2691  
2692  
2693  
2694  
2695  
2696  
2697  
2698  
2699

Question 1. \*

Evidence for *species A*

Learned Part Prototypes

Evidence for *species B*

Learned Part Prototypes

Based on given evidence, the model would say the photo shows...

species A

species B

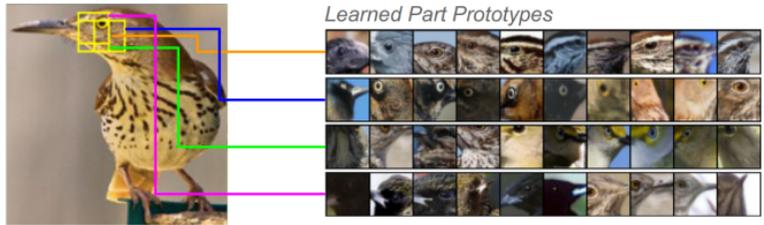
Previous Next

Figure 43: Page 5 of survey for PIP-Net

2700  
2701  
2702  
2703  
2704  
2705  
2706  
2707  
2708  
2709  
2710  
2711  
2712  
2713  
2714  
2715  
2716  
2717  
2718  
2719  
2720  
2721  
2722  
2723  
2724  
2725  
2726  
2727  
2728  
2729  
2730  
2731  
2732  
2733  
2734  
2735  
2736  
2737  
2738  
2739  
2740  
2741  
2742  
2743  
2744  
2745  
2746  
2747  
2748  
2749  
2750  
2751  
2752  
2753

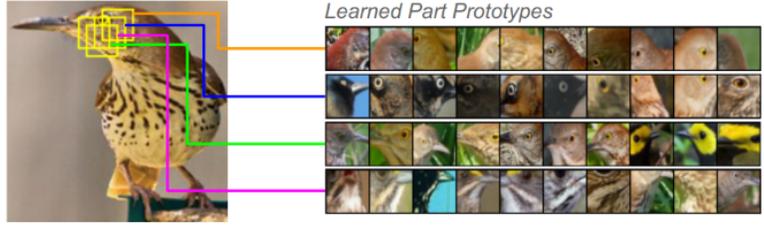
Question 2. \*

Evidence for *species A*



Learned Part Prototypes

Evidence for *species B*



Learned Part Prototypes

Based on given evidence, the model would say the photo shows...

species A

species B

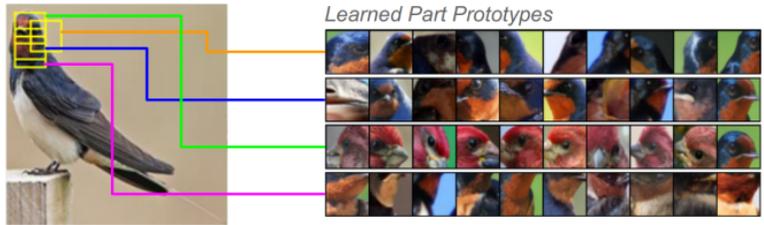
Previous Next

Figure 44: Page 6 of survey for PIP-Net

2754  
2755  
2756  
2757  
2758  
2759  
2760  
2761  
2762  
2763  
2764  
2765  
2766  
2767  
2768  
2769  
2770  
2771  
2772  
2773  
2774  
2775  
2776  
2777  
2778  
2779  
2780  
2781  
2782  
2783  
2784  
2785  
2786  
2787  
2788  
2789  
2790  
2791  
2792  
2793  
2794  
2795  
2796  
2797  
2798  
2799  
2800  
2801  
2802  
2803  
2804  
2805  
2806  
2807

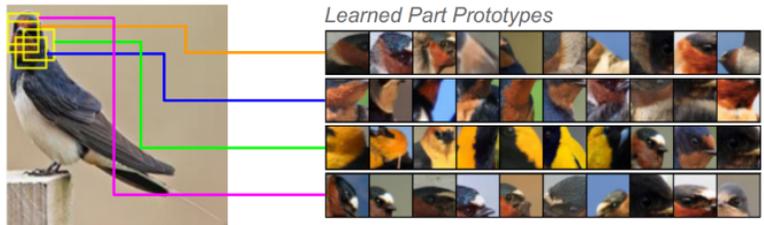
Question 3. \*

Evidence for *species A*



Learned Part Prototypes

Evidence for *species B*



Learned Part Prototypes

Based on given evidence, the model would say the photo shows...

species A

species B

Previous Next

Figure 45: Page 7 of survey for PIP-Net

2808  
2809  
2810  
2811  
2812  
2813  
2814  
2815  
2816  
2817  
2818  
2819  
2820  
2821  
2822  
2823  
2824  
2825  
2826  
2827  
2828  
2829  
2830  
2831  
2832  
2833  
2834  
2835  
2836  
2837  
2838  
2839  
2840  
2841  
2842  
2843  
2844  
2845  
2846  
2847  
2848  
2849  
2850  
2851  
2852  
2853  
2854  
2855  
2856  
2857  
2858  
2859  
2860  
2861

Question 4. \*

Evidence for *species A*

Evidence for *species B*

Based on given evidence, the model would say the photo shows...

species A

species B

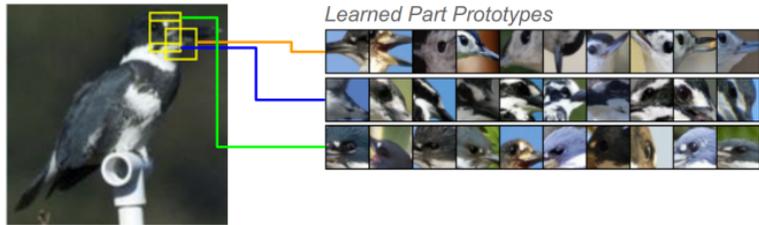
Previous Next

Figure 46: Page 8 of survey for PIP-Net

2862  
2863  
2864  
2865  
2866  
2867  
2868  
2869  
2870  
2871  
2872  
2873  
2874  
2875  
2876  
2877  
2878  
2879  
2880  
2881  
2882  
2883  
2884  
2885  
2886  
2887  
2888  
2889  
2890  
2891  
2892  
2893  
2894  
2895  
2896  
2897  
2898  
2899  
2900  
2901  
2902  
2903  
2904  
2905  
2906  
2907  
2908  
2909  
2910  
2911  
2912  
2913  
2914  
2915

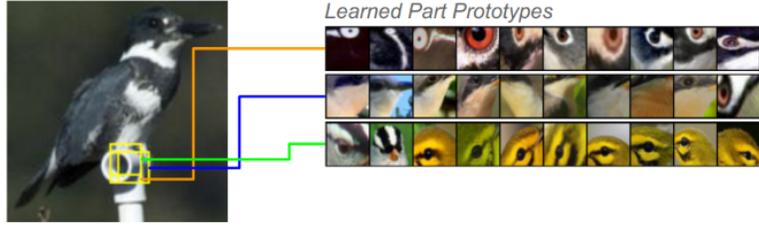
Question 5. \*

Evidence for *species A*



Learned Part Prototypes

Evidence for *species B*



Learned Part Prototypes

Based on given evidence, the model would say the photo shows...

species A

species B

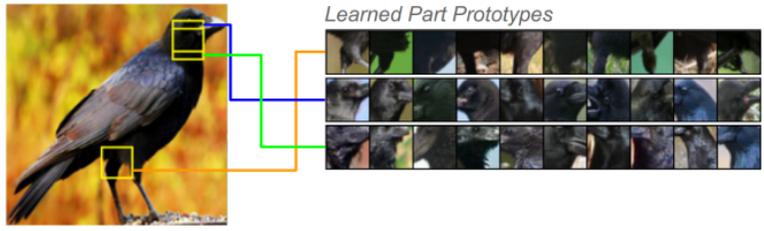
Previous Next

Figure 47: Page 9 of survey for PIP-Net

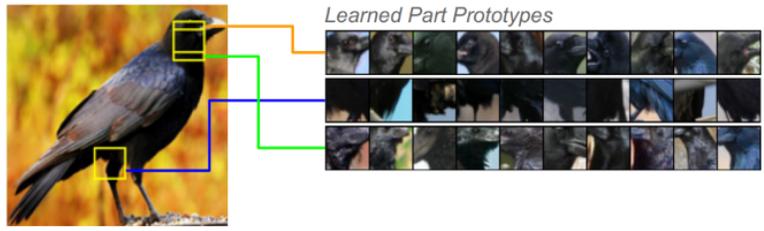
2916  
2917  
2918  
2919  
2920  
2921  
2922  
2923  
2924  
2925  
2926  
2927  
2928  
2929  
2930  
2931  
2932  
2933  
2934  
2935  
2936  
2937  
2938  
2939  
2940  
2941  
2942  
2943  
2944  
2945  
2946  
2947  
2948  
2949  
2950  
2951  
2952  
2953  
2954  
2955  
2956  
2957  
2958  
2959  
2960  
2961  
2962  
2963  
2964  
2965  
2966  
2967  
2968  
2969

Question 6. \*

Evidence for *species A*



Evidence for *species B*



Based on given evidence, the model would say the photo shows...

species A

species B

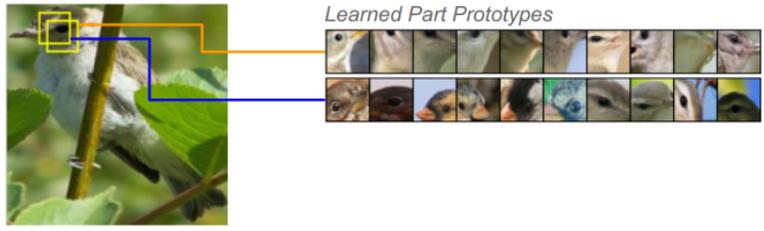
Previous Next

Figure 48: Page 10 of survey for PIP-Net

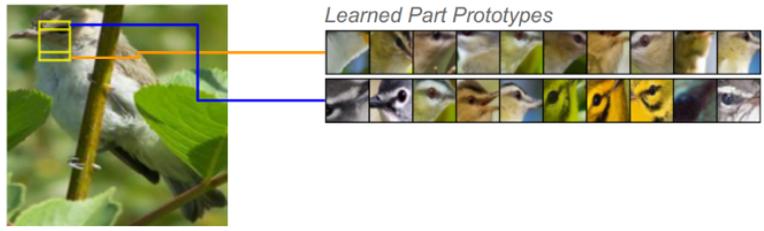
2970  
2971  
2972  
2973  
2974  
2975  
2976  
2977  
2978  
2979  
2980  
2981  
2982  
2983  
2984  
2985  
2986  
2987  
2988  
2989  
2990  
2991  
2992  
2993  
2994  
2995  
2996  
2997  
2998  
2999  
3000  
3001  
3002  
3003  
3004  
3005  
3006  
3007  
3008  
3009  
3010  
3011  
3012  
3013  
3014  
3015  
3016  
3017  
3018  
3019  
3020  
3021  
3022  
3023

Question 7. \*

Evidence for *species A*



Evidence for *species B*



Based on given evidence, the model would say the photo shows...

species A

species B

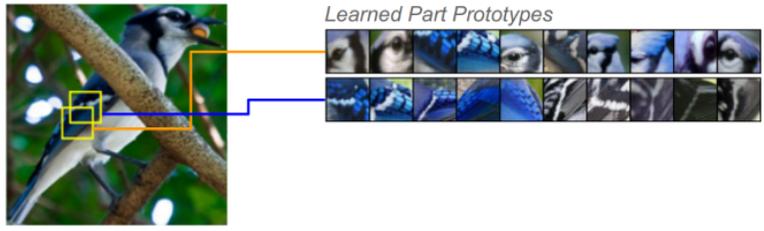
Previous Next

Figure 49: Page 11 of survey for PIP-Net

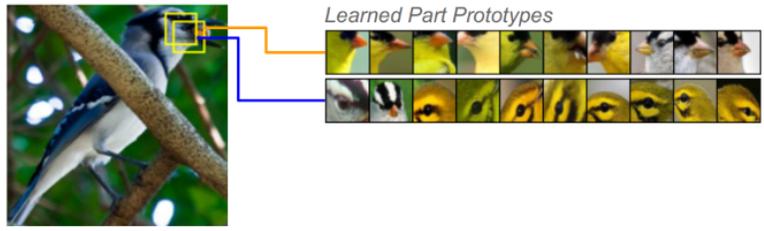
3024  
3025  
3026  
3027  
3028  
3029  
3030  
3031  
3032  
3033  
3034  
3035  
3036  
3037  
3038  
3039  
3040  
3041  
3042  
3043  
3044  
3045  
3046  
3047  
3048  
3049  
3050  
3051  
3052  
3053  
3054  
3055  
3056  
3057  
3058  
3059  
3060  
3061  
3062  
3063  
3064  
3065  
3066  
3067  
3068  
3069  
3070  
3071  
3072  
3073  
3074  
3075  
3076  
3077

Question 8. \*

Evidence for *species A*



Evidence for *species B*



Based on given evidence, the model would say the photo shows...

species A

species B

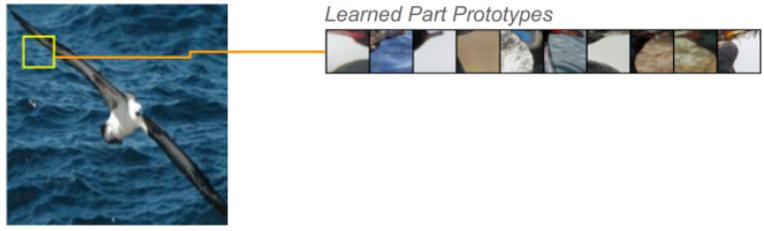
Previous Next

Figure 50: Page 12 of survey for PIP-Net

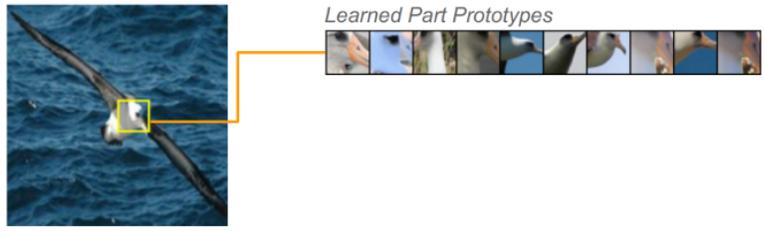
3078  
3079  
3080  
3081  
3082  
3083  
3084  
3085  
3086  
3087  
3088  
3089  
3090  
3091  
3092  
3093  
3094  
3095  
3096  
3097  
3098  
3099  
3100  
3101  
3102  
3103  
3104  
3105  
3106  
3107  
3108  
3109  
3110  
3111  
3112  
3113  
3114  
3115  
3116  
3117  
3118  
3119  
3120  
3121  
3122  
3123  
3124  
3125  
3126  
3127  
3128  
3129  
3130  
3131

Question 9. \*

Evidence for *species A*



Evidence for *species B*



Based on given evidence, the model would say the photo shows...

species A

species B

Previous Send Job

Figure 51: Page 13 of survey for PIP-Net

3132  
3133  
3134  
3135  
3136  
3137  
3138  
3139  
3140  
3141  
3142  
3143  
3144  
3145  
3146  
3147  
3148  
3149  
3150  
3151  
3152  
3153  
3154  
3155  
3156  
3157  
3158  
3159  
3160  
3161  
3162  
3163  
3164  
3165  
3166  
3167  
3168  
3169  
3170  
3171  
3172  
3173  
3174  
3175  
3176  
3177  
3178  
3179  
3180  
3181  
3182  
3183  
3184  
3185

Given a bird photo a model predicts the species based on prototypes it has learned from previously seen photos. Specifically for each prototype, the model identifies a region in the photo that looks the most similar to the prototype and rates their similarity. Greater similarity determines the classification of the bird into a specific species.

**For a given photo we show evidence for 2 bird species found by the model. Your task is to choose the species you think is predicted by the model based on that evidence.**

Random guessing will get you 50% accuracy. You will receive a reward only if your performance is reasonably beyond the random chance.

**Interface explanation**

An image being classified by the model

Prototypes' location

Each row shows a prototype of a bird

Evidence for *species A*

Part prototypes of *species A*

Evidence for *species B*

Part prototypes of *species B*

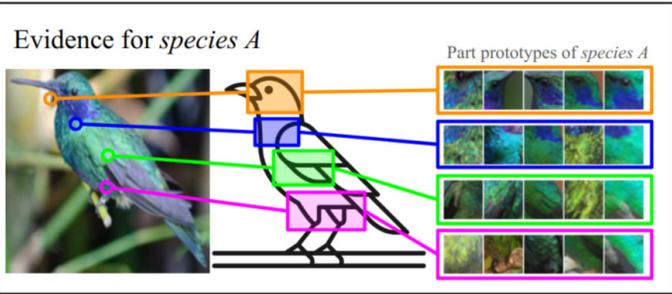
Next

Figure 52: Page 1 of survey for *single branch*

3186  
3187  
3188  
3189  
3190  
3191  
3192  
3193  
3194  
3195  
3196  
3197  
3198  
3199  
3200  
3201  
3202  
3203  
3204  
3205  
3206  
3207  
3208  
3209  
3210  
3211  
3212  
3213  
3214  
3215  
3216  
3217  
3218  
3219  
3220  
3221  
3222  
3223  
3224  
3225  
3226  
3227  
3228  
3229  
3230  
3231  
3232  
3233  
3234  
3235  
3236  
3237  
3238  
3239

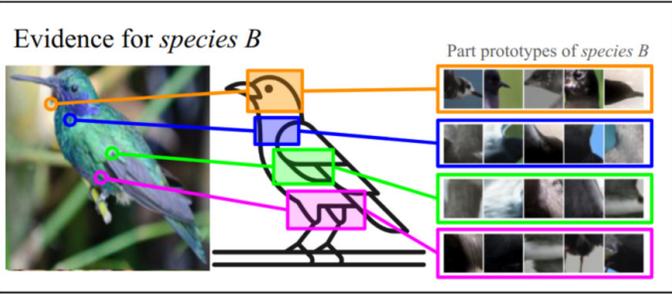
Example 1. \*

Evidence for *species A*



Part prototypes of *species A*

Evidence for *species B*



Part prototypes of *species B*

Based on given evidence, the model would say the photo shows...

Correct answer: **species A**  
Explanation: Evidence for species A is more similar to the image

species A  
 species B

Previous Next

Figure 53: Page 2 of survey for *single branch*

3240  
3241  
3242  
3243  
3244  
3245  
3246  
3247  
3248  
3249  
3250  
3251  
3252  
3253  
3254  
3255  
3256  
3257  
3258  
3259  
3260  
3261  
3262  
3263  
3264  
3265  
3266  
3267  
3268  
3269  
3270  
3271  
3272  
3273  
3274  
3275  
3276  
3277  
3278  
3279  
3280  
3281  
3282  
3283  
3284  
3285  
3286  
3287  
3288  
3289  
3290  
3291  
3292  
3293

Example 2. \*

Evidence for *species A*

Part prototypes of *species A*

Evidence for *species B*

Part prototypes of *species B*

Based on given evidence, the model would say the photo shows...

Correct answer: **species B**  
Explanation: Evidence for species B is more similar to the image

species A  
 species B

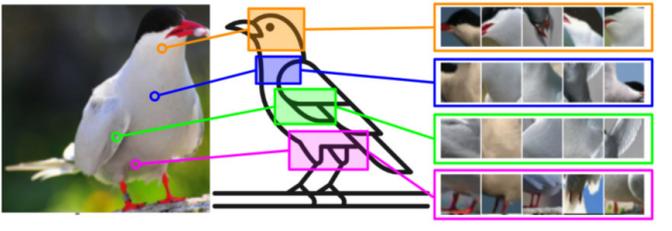
[Previous](#) [Next](#)

Figure 54: Page 3 of survey for *single branch*

3294  
3295  
3296  
3297  
3298  
3299  
3300  
3301  
3302  
3303  
3304  
3305  
3306  
3307  
3308  
3309  
3310  
3311  
3312  
3313  
3314  
3315  
3316  
3317  
3318  
3319  
3320  
3321  
3322  
3323  
3324  
3325  
3326  
3327  
3328  
3329  
3330  
3331  
3332  
3333  
3334  
3335  
3336  
3337  
3338  
3339  
3340  
3341  
3342  
3343  
3344  
3345  
3346  
3347

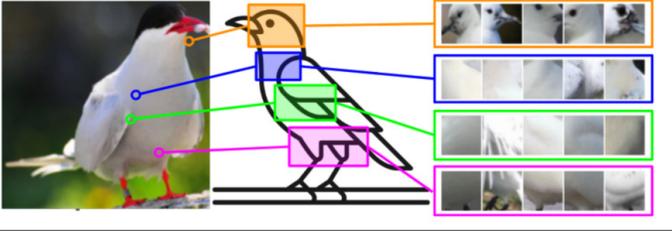
Question 0. \*

Evidence for *species A*



Part prototypes of *species A*

Evidence for *species B*



Part prototypes of *species B*

Based on given evidence, the model would say the photo shows...

species A

species B

Previous Next

Figure 55: Page 4 of survey for *single branch*

3348  
3349  
3350  
3351  
3352  
3353  
3354  
3355  
3356  
3357  
3358  
3359  
3360  
3361  
3362  
3363  
3364  
3365  
3366  
3367  
3368  
3369  
3370  
3371  
3372  
3373  
3374  
3375  
3376  
3377  
3378  
3379  
3380  
3381  
3382  
3383  
3384  
3385  
3386  
3387  
3388  
3389  
3390  
3391  
3392  
3393  
3394  
3395  
3396  
3397  
3398  
3399  
3400  
3401

Question 1. \*

**Evidence for *species A***

Part prototypes of *species A*

**Evidence for *species B***

Part prototypes of *species B*

Based on given evidence, the model would say the photo shows...

species A

species B

Previous Next

Figure 56: Page 5 of survey for *single branch*

3402  
3403  
3404  
3405  
3406  
3407  
3408  
3409  
3410  
3411  
3412  
3413  
3414  
3415  
3416  
3417  
3418  
3419  
3420  
3421  
3422  
3423  
3424  
3425  
3426  
3427  
3428  
3429  
3430  
3431  
3432  
3433  
3434  
3435  
3436  
3437  
3438  
3439  
3440  
3441  
3442  
3443  
3444  
3445  
3446  
3447  
3448  
3449  
3450  
3451  
3452  
3453  
3454  
3455

Question 2. \*

Evidence for *species A*

Part prototypes of *species A*

Evidence for *species B*

Part prototypes of *species B*

Based on given evidence, the model would say the photo shows...

species A

species B

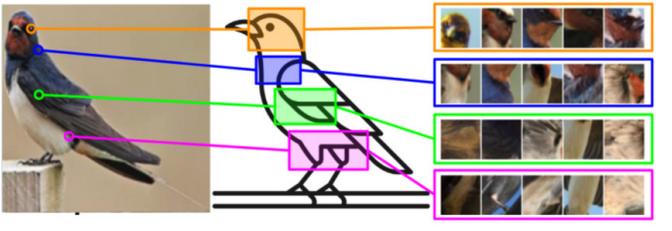
Previous Next

Figure 57: Page 6 of survey for *single branch*

3456  
3457  
3458  
3459  
3460  
3461  
3462  
3463  
3464  
3465  
3466  
3467  
3468  
3469  
3470  
3471  
3472  
3473  
3474  
3475  
3476  
3477  
3478  
3479  
3480  
3481  
3482  
3483  
3484  
3485  
3486  
3487  
3488  
3489  
3490  
3491  
3492  
3493  
3494  
3495  
3496  
3497  
3498  
3499  
3500  
3501  
3502  
3503  
3504  
3505  
3506  
3507  
3508  
3509

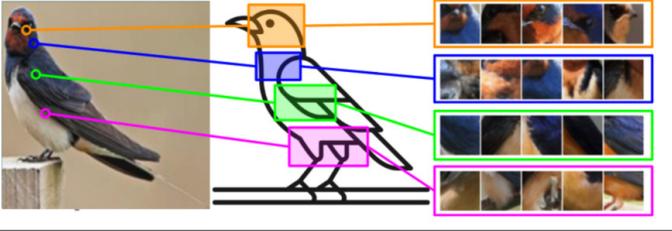
Question 3. \*

Evidence for *species A*



Part prototypes of *species A*

Evidence for *species B*



Part prototypes of *species B*

Based on given evidence, the model would say the photo shows...

species A

species B

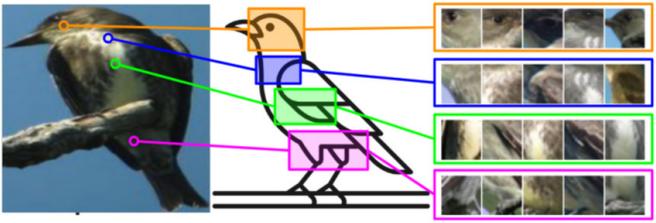
Previous Next

Figure 58: Page 7 of survey for *single branch*

3510  
3511  
3512  
3513  
3514  
3515  
3516  
3517  
3518  
3519  
3520  
3521  
3522  
3523  
3524  
3525  
3526  
3527  
3528  
3529  
3530  
3531  
3532  
3533  
3534  
3535  
3536  
3537  
3538  
3539  
3540  
3541  
3542  
3543  
3544  
3545  
3546  
3547  
3548  
3549  
3550  
3551  
3552  
3553  
3554  
3555  
3556  
3557  
3558  
3559  
3560  
3561  
3562  
3563

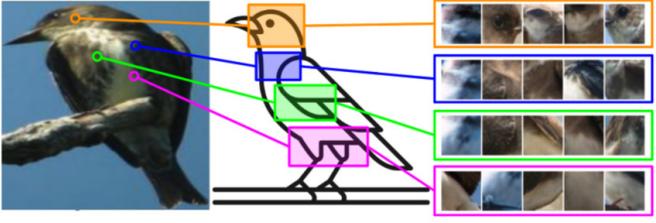
Question 4. \*

Evidence for *species A*



Part prototypes of *species A*

Evidence for *species B*



Part prototypes of *species B*

Based on given evidence, the model would say the photo shows...

species A

species B

Previous Next

Figure 59: Page 8 of survey for *single branch*

3564  
3565  
3566  
3567  
3568  
3569  
3570  
3571  
3572  
3573  
3574  
3575  
3576  
3577  
3578  
3579  
3580  
3581  
3582  
3583  
3584  
3585  
3586  
3587  
3588  
3589  
3590  
3591  
3592  
3593  
3594  
3595  
3596  
3597  
3598  
3599  
3600  
3601  
3602  
3603  
3604  
3605  
3606  
3607  
3608  
3609  
3610  
3611  
3612  
3613  
3614  
3615  
3616  
3617

Question 5. \*

**Evidence for species A**

Part prototypes of species A

**Evidence for species B**

Part prototypes of species B

Based on given evidence, the model would say the photo shows...

species A

species B

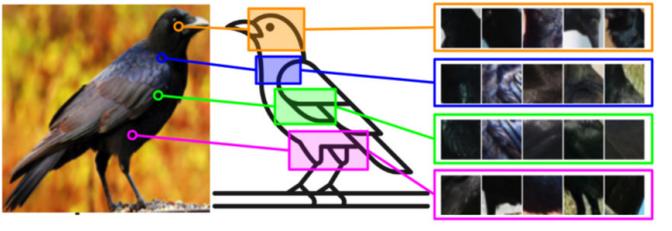
Previous Next

Figure 60: Page 9 of survey for *single branch*

3618  
3619  
3620  
3621  
3622  
3623  
3624  
3625  
3626  
3627  
3628  
3629  
3630  
3631  
3632  
3633  
3634  
3635  
3636  
3637  
3638  
3639  
3640  
3641  
3642  
3643  
3644  
3645  
3646  
3647  
3648  
3649  
3650  
3651  
3652  
3653  
3654  
3655  
3656  
3657  
3658  
3659  
3660  
3661  
3662  
3663  
3664  
3665  
3666  
3667  
3668  
3669  
3670  
3671

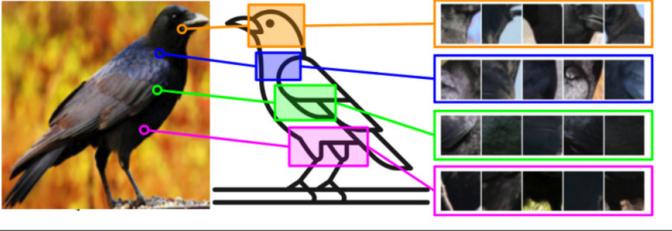
Question 6. \*

Evidence for *species A*



Part prototypes of *species A*

Evidence for *species B*



Part prototypes of *species B*

Based on given evidence, the model would say the photo shows...

species A

species B

Previous Next

Figure 61: Page 10 of survey for *single branch*

3672  
3673  
3674  
3675  
3676  
3677  
3678  
3679  
3680  
3681  
3682  
3683  
3684  
3685  
3686  
3687  
3688  
3689  
3690  
3691  
3692  
3693  
3694  
3695  
3696  
3697  
3698  
3699  
3700  
3701  
3702  
3703  
3704  
3705  
3706  
3707  
3708  
3709  
3710  
3711  
3712  
3713  
3714  
3715  
3716  
3717  
3718  
3719  
3720  
3721  
3722  
3723  
3724  
3725

Question 7. \*

Evidence for *species A*

Part prototypes of *species A*

Evidence for *species B*

Part prototypes of *species B*

Based on given evidence, the model would say the photo shows...

species A

species B

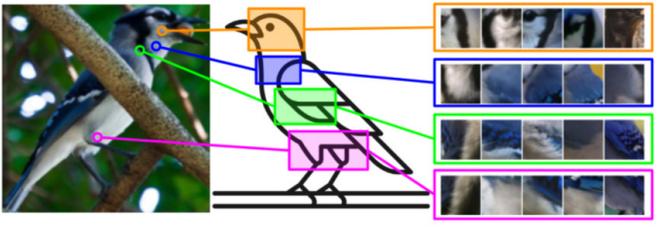
Previous Next

Figure 62: Page 11 of survey for *single branch*

3726  
3727  
3728  
3729  
3730  
3731  
3732  
3733  
3734  
3735  
3736  
3737  
3738  
3739  
3740  
3741  
3742  
3743  
3744  
3745  
3746  
3747  
3748  
3749  
3750  
3751  
3752  
3753  
3754  
3755  
3756  
3757  
3758  
3759  
3760  
3761  
3762  
3763  
3764  
3765  
3766  
3767  
3768  
3769  
3770  
3771  
3772  
3773  
3774  
3775  
3776  
3777  
3778  
3779

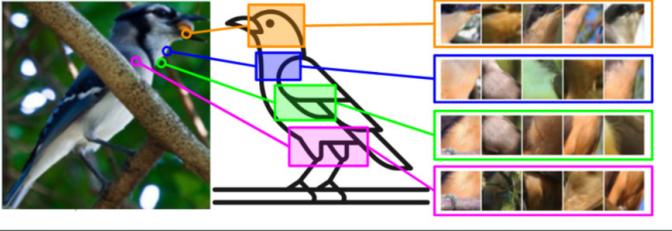
Question 8. \*

Evidence for *species A*



Part prototypes of *species A*

Evidence for *species B*



Part prototypes of *species B*

Based on given evidence, the model would say the photo shows...

species A

species B

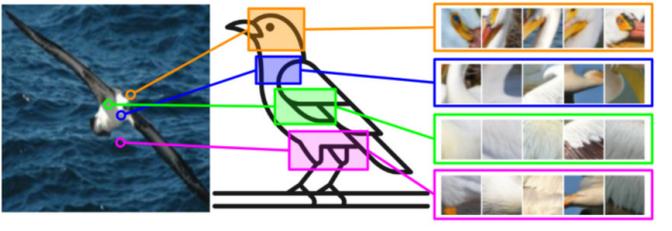
Previous Next

Figure 63: Page 12 of survey for *single branch*

3780  
3781  
3782  
3783  
3784  
3785  
3786  
3787  
3788  
3789  
3790  
3791  
3792  
3793  
3794  
3795  
3796  
3797  
3798  
3799  
3800  
3801  
3802  
3803  
3804  
3805  
3806  
3807  
3808  
3809  
3810  
3811  
3812  
3813  
3814  
3815  
3816  
3817  
3818  
3819  
3820  
3821  
3822  
3823  
3824  
3825  
3826  
3827  
3828  
3829  
3830  
3831  
3832  
3833

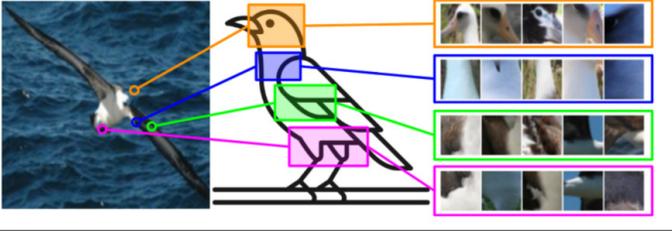
Question 9. \*

Evidence for *species A*



Part prototypes of *species A*

Evidence for *species B*



Part prototypes of *species B*

Based on given evidence, the model would say the photo shows...

species A

species B

[Previous](#) [Send job](#)

Figure 64: Page 13 of survey for *single branch*

3834  
3835  
3836  
3837  
3838  
3839  
3840  
3841  
3842  
3843  
3844  
3845  
3846  
3847  
3848  
3849  
3850  
3851  
3852  
3853  
3854  
3855  
3856  
3857  
3858  
3859  
3860  
3861  
3862  
3863  
3864  
3865  
3866  
3867  
3868  
3869  
3870  
3871  
3872  
3873  
3874  
3875  
3876  
3877  
3878  
3879  
3880  
3881  
3882  
3883  
3884  
3885  
3886  
3887

Given a bird photo a model predicts the species based on prototypes it has learned from previously seen photos. Specifically for each prototype, the model identifies a region in the photo that looks the most similar to the prototype and rates their similarity. Greater similarity determines the classification of the bird into a specific species. Model **first** tries to decide based on the **shape with texture** (gray prototypes), and **then** it looks at **color** which may correct the prediction.

The AI model made a decision to classify an image of a bird as belonging to a specific bird species. In the following questions, you will see the two most probable bird species present in the image, according to the model. Based on the explanation provided by the model, try to answer the following question: "In your opinion, what is the influence of color on the model's decision process?" (1 - None, 2 - Weak, 3 - Don't know, 4 - Moderate, 5 - Substantial)

Interface explanation

An image being classified by the model

Both rows show a prototype of a bird part composed of two components:  
- shape with texture  
- color

**Evidence for *correct species***

Part prototypes of *correct species*

0.90 & 0.90 = 0.81

**Evidence for *wrong species***

Part prototypes of *wrong species*

0.20 & 0.20 = 0.04

Similarity of prototype components to the image, the higher the more similar

[Next](#)

Figure 65: Page 1 of survey for full LucidPPN (with scores)

3888  
3889  
3890  
3891  
3892  
3893  
3894  
3895  
3896  
3897  
3898  
3899  
3900  
3901  
3902  
3903  
3904  
3905  
3906  
3907  
3908  
3909  
3910  
3911  
3912  
3913  
3914  
3915  
3916  
3917  
3918  
3919  
3920  
3921  
3922  
3923  
3924  
3925  
3926  
3927  
3928  
3929  
3930  
3931  
3932  
3933  
3934  
3935  
3936  
3937  
3938  
3939  
3940  
3941

Example 1. \*

**Evidence for *correct species***

Part prototypes of *correct species*

0.90      0.90      0.81

Shape with texture similarity to image is HIGHER for the *correct species*. Color was NOT needed.

**Evidence for *wrong species***

Part prototypes of *wrong species*

0.20      0.20      0.04

In your opinion, what is the influence of color on the model's decision process?

Correct answer: 1 - None  
Explanation: Shape with texture similarity of image to prototypes was enough to predict the correct class.

1 - None  
 2 - Weak  
 3 - Don't know  
 4 - Moderate  
 5 - Substantial

Previous
Next

Figure 66: Page 2 of survey for full LucidPPN (with scores)

3942  
3943  
3944  
3945  
3946  
3947  
3948  
3949  
3950  
3951  
3952  
3953  
3954  
3955  
3956  
3957  
3958  
3959  
3960  
3961  
3962  
3963  
3964  
3965  
3966  
3967  
3968  
3969  
3970  
3971  
3972  
3973  
3974  
3975  
3976  
3977  
3978  
3979  
3980  
3981  
3982  
3983  
3984  
3985  
3986  
3987  
3988  
3989  
3990  
3991  
3992  
3993  
3994  
3995

Example 2. \*

**Evidence for *correct species***

Part prototypes of *correct species*

0.15      1.00      0.15

Shape with texture similarity to image is LOWER for the *correct species*. Color WAS needed.

**Evidence for *wrong species***

Part prototypes of *wrong species*

0.29      0.01      0.00

In your opinion, what is the influence of color on the model's decision process?

Correct answer: 5 - Substantial  
Explanation: Shape with texture similarity of image to prototypes would lead to a mistake and it was avoided by considering color.

1 - None  
 2 - Weak  
 3 - Don't know  
 4 - Moderate  
 5 - Substantial

Previous
Next

Figure 67: Page 3 of survey for full LucidPPN (with scores)

3996  
3997  
3998  
3999  
4000  
4001  
4002  
4003  
4004  
4005  
4006  
4007  
4008  
4009  
4010  
4011  
4012  
4013  
4014  
4015  
4016  
4017  
4018  
4019  
4020  
4021  
4022  
4023  
4024  
4025  
4026  
4027  
4028  
4029  
4030  
4031  
4032  
4033  
4034  
4035  
4036  
4037  
4038  
4039  
4040  
4041  
4042  
4043  
4044  
4045  
4046  
4047  
4048  
4049

Question 0. \*

Evidence for *correct species* Part prototypes of *correct species*

0.38      0.99      0.38

Evidence for *wrong species* Part prototypes of *wrong species*

0.42      0.48      0.20

In your opinion, what is the influence of color on the model's decision process?

- 1 - None
- 2 - Weak
- 3 - Don't know
- 4 - Moderate
- 5 - Substantial

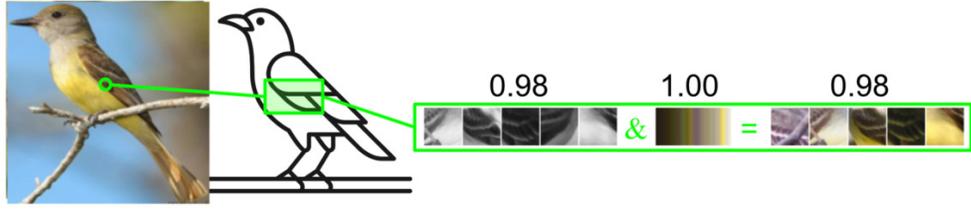
[Previous](#)   [Next](#)

Figure 68: Page 4 of survey for full LucidPPN (with scores)

4050  
4051  
4052  
4053  
4054  
4055  
4056  
4057  
4058  
4059  
4060  
4061  
4062  
4063  
4064  
4065  
4066  
4067  
4068  
4069  
4070  
4071  
4072  
4073  
4074  
4075  
4076  
4077  
4078  
4079  
4080  
4081  
4082  
4083  
4084  
4085  
4086  
4087  
4088  
4089  
4090  
4091  
4092  
4093  
4094  
4095  
4096  
4097  
4098  
4099  
4100  
4101  
4102  
4103

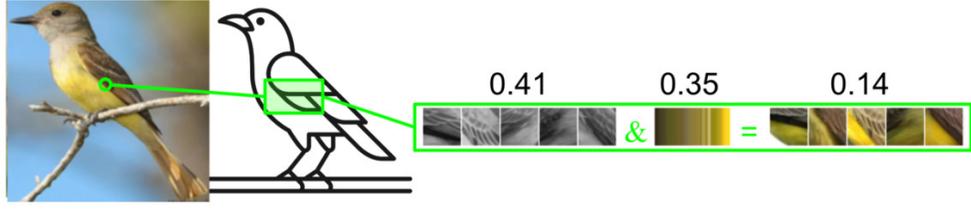
Question 1. \*

Evidence for *correct species* Part prototypes of *correct species*



0.98      1.00      0.98

Evidence for *wrong species* Part prototypes of *wrong species*



0.41      0.35      0.14

In your opinion, what is the influence of color on the model's decision process?

- 1 - None
- 2 - Weak
- 3 - Don't know
- 4 - Moderate
- 5 - Substantial

[Previous](#)   [Next](#)

Figure 69: Page 5 of survey for full LucidPPN (with scores)

4104  
4105  
4106  
4107  
4108  
4109  
4110  
4111  
4112  
4113  
4114  
4115  
4116  
4117  
4118  
4119  
4120  
4121  
4122  
4123  
4124  
4125  
4126  
4127  
4128  
4129  
4130  
4131  
4132  
4133  
4134  
4135  
4136  
4137  
4138  
4139  
4140  
4141  
4142  
4143  
4144  
4145  
4146  
4147  
4148  
4149  
4150  
4151  
4152  
4153  
4154  
4155  
4156  
4157

Question 2. \*

Evidence for *correct species* Part prototypes of *correct species*

0.89      0.79      0.70

Evidence for *wrong species* Part prototypes of *wrong species*

0.35      0.82      0.29

In your opinion, what is the influence of color on the model's decision process?

- 1 - None
- 2 - Weak
- 3 - Don't know
- 4 - Moderate
- 5 - Substantial

[Previous](#)   [Next](#)

Figure 70: Page 6 of survey for full LucidPPN (with scores)

4158  
4159  
4160  
4161  
4162  
4163  
4164  
4165  
4166  
4167  
4168  
4169  
4170  
4171  
4172  
4173  
4174  
4175  
4176  
4177  
4178  
4179  
4180  
4181  
4182  
4183  
4184  
4185  
4186  
4187  
4188  
4189  
4190  
4191  
4192  
4193  
4194  
4195  
4196  
4197  
4198  
4199  
4200  
4201  
4202  
4203  
4204  
4205  
4206  
4207  
4208  
4209  
4210  
4211

Question 3. \*

Evidence for *correct species*

Part prototypes of *correct species*

0.37      1.00      0.37

Evidence for *wrong species*

Part prototypes of *wrong species*

0.82      0.02      0.02

In your opinion, what is the influence of color on the model's decision process?

- 1 - None
- 2 - Weak
- 3 - Don't know
- 4 - Moderate
- 5 - Substantial

Previous    Next

Figure 71: Page 7 of survey for full LucidPPN (with scores)

4212  
4213  
4214  
4215  
4216  
4217  
4218  
4219  
4220  
4221  
4222  
4223  
4224  
4225  
4226  
4227  
4228  
4229  
4230  
4231  
4232  
4233  
4234  
4235  
4236  
4237  
4238  
4239  
4240  
4241  
4242  
4243  
4244  
4245  
4246  
4247  
4248  
4249  
4250  
4251  
4252  
4253  
4254  
4255  
4256  
4257  
4258  
4259  
4260  
4261  
4262  
4263  
4264  
4265

Question 4. \*

**Evidence for *correct species***

Part prototypes of *correct species*

0.31      0.85      0.26

---

**Evidence for *wrong species***

Part prototypes of *wrong species*

0.62      0.16      0.10

In your opinion, what is the influence of color on the model's decision process?

- 1 - None
- 2 - Weak
- 3 - Don't know
- 4 - Moderate
- 5 - Substantial

Previous
Next

Figure 72: Page 8 of survey for full LucidPPN (with scores)

4266  
4267  
4268  
4269  
4270  
4271  
4272  
4273  
4274  
4275  
4276  
4277  
4278  
4279  
4280  
4281  
4282  
4283  
4284  
4285  
4286  
4287  
4288  
4289  
4290  
4291  
4292  
4293  
4294  
4295  
4296  
4297  
4298  
4299  
4300  
4301  
4302  
4303  
4304  
4305  
4306  
4307  
4308  
4309  
4310  
4311  
4312  
4313  
4314  
4315  
4316  
4317  
4318  
4319

Question 5. \*

Evidence for *correct species*

Part prototypes of *correct species*

0.96      1.00      0.96

Evidence for *wrong species*

Part prototypes of *wrong species*

0.12      1.00      0.12

In your opinion, what is the influence of color on the model's decision process?

- 1 - None
- 2 - Weak
- 3 - Don't know
- 4 - Moderate
- 5 - Substantial

Previous    Next

Figure 73: Page 9 of survey for full LucidPPN (with scores)

4320  
4321  
4322  
4323  
4324  
4325  
4326  
4327  
4328  
4329  
4330  
4331  
4332  
4333  
4334  
4335  
4336  
4337  
4338  
4339  
4340  
4341  
4342  
4343  
4344  
4345  
4346  
4347  
4348  
4349  
4350  
4351  
4352  
4353  
4354  
4355  
4356  
4357  
4358  
4359  
4360  
4361  
4362  
4363  
4364  
4365  
4366  
4367  
4368  
4369  
4370  
4371  
4372  
4373

Question 6. \*

Evidence for *correct species* Part prototypes of *correct species*

Evidence for *wrong species* Part prototypes of *wrong species*

In your opinion, what is the influence of color on the model's decision process?

- 1 - None
- 2 - Weak
- 3 - Don't know
- 4 - Moderate
- 5 - Substantial

[Previous](#) [Next](#)

Figure 74: Page 10 of survey for full LucidPPN (with scores)

4374  
4375  
4376  
4377  
4378  
4379  
4380  
4381  
4382  
4383  
4384  
4385  
4386  
4387  
4388  
4389  
4390  
4391  
4392  
4393  
4394  
4395  
4396  
4397  
4398  
4399  
4400  
4401  
4402  
4403  
4404  
4405  
4406  
4407  
4408  
4409  
4410  
4411  
4412  
4413  
4414  
4415  
4416  
4417  
4418  
4419  
4420  
4421  
4422  
4423  
4424  
4425  
4426  
4427

Question 7. \*

Evidence for *correct species* Part prototypes of *correct species*

0.95      0.74      0.70

Evidence for *wrong species* Part prototypes of *wrong species*

0.29      0.92      0.27

In your opinion, what is the influence of color on the model's decision process?

- 1 - None
- 2 - Weak
- 3 - Don't know
- 4 - Moderate
- 5 - Substantial

[Previous](#)   [Next](#)

Figure 75: Page 11 of survey for full LucidPPN (with scores)

4428  
4429  
4430  
4431  
4432  
4433  
4434  
4435  
4436  
4437  
4438  
4439  
4440  
4441  
4442  
4443  
4444  
4445  
4446  
4447  
4448  
4449  
4450  
4451  
4452  
4453  
4454  
4455  
4456  
4457  
4458  
4459  
4460  
4461  
4462  
4463  
4464  
4465  
4466  
4467  
4468  
4469  
4470  
4471  
4472  
4473  
4474  
4475  
4476  
4477  
4478  
4479  
4480  
4481

Question 8. \*

Evidence for *correct species* Part prototypes of *correct species*

0.23      0.46      0.10

Evidence for *wrong species* Part prototypes of *wrong species*

0.18      0.41      0.07

In your opinion, what is the influence of color on the model's decision process?

- 1 - None
- 2 - Weak
- 3 - Don't know
- 4 - Moderate
- 5 - Substantial

[Previous](#)   [Next](#)

Figure 76: Page 12 of survey for full LucidPPN (with scores)

4482  
4483  
4484  
4485  
4486  
4487  
4488  
4489  
4490  
4491  
4492  
4493  
4494  
4495  
4496  
4497  
4498  
4499  
4500  
4501  
4502  
4503  
4504  
4505  
4506  
4507  
4508  
4509  
4510  
4511  
4512  
4513  
4514  
4515  
4516  
4517  
4518  
4519  
4520  
4521  
4522  
4523  
4524  
4525  
4526  
4527  
4528  
4529  
4530  
4531  
4532  
4533  
4534  
4535

Question 9. \*

Evidence for *correct species* Part prototypes of *correct species*

0.01      0.99      0.01

Evidence for *wrong species* Part prototypes of *wrong species*

0.11      0.02      0.00

In your opinion, what is the influence of color on the model's decision process?

- 1 - None
- 2 - Weak
- 3 - Don't know
- 4 - Moderate
- 5 - Substantial

[Previous](#) [Send job](#)

Figure 77: Page 13 of survey for full LucidPPN (with scores)

4536  
4537  
4538  
4539  
4540  
4541  
4542  
4543  
4544  
4545  
4546  
4547  
4548  
4549  
4550  
4551  
4552  
4553  
4554  
4555  
4556  
4557  
4558  
4559  
4560  
4561  
4562  
4563  
4564  
4565  
4566  
4567  
4568  
4569  
4570  
4571  
4572  
4573  
4574  
4575  
4576  
4577  
4578  
4579  
4580  
4581  
4582  
4583  
4584  
4585  
4586  
4587  
4588  
4589

Given a bird photo a model predicts the species based on prototypes it has learned from previously seen photos. Specifically for each prototype, the model identifies a region in the photo that looks the most similar to the prototype and rates their similarity. Greater similarity determines the classification of the bird into a specific species. Model **first** tries to decide based on the **shape with texture** (gray prototypes), and **then** it looks at **color** which may correct the prediction.

The AI model made a decision to classify an image of a bird as belonging to a specific bird species. In the following questions, you will see the two most probable bird species present in the image, according to the model. Based on the explanation provided by the model, try to answer the following question: "In your opinion, what is the influence of color on the model's decision process?" (1 - None, 2 - Weak, 3 - Don't know, 4 - Moderate, 5 - Substantial)

Interface explanation

An image being classified by the model

Both rows show a prototype of a bird part composed of two components:  
- shape with texture  
- color

**Evidence for *correct species*** Part prototypes of *correct species*

**Evidence for *wrong species*** Part prototypes of *wrong species*

[Next](#)

Figure 78: Page 1 of survey for LucidPPN with *no scores*

4590  
4591  
4592  
4593  
4594  
4595  
4596  
4597  
4598  
4599  
4600  
4601  
4602  
4603  
4604  
4605  
4606  
4607  
4608  
4609  
4610  
4611  
4612  
4613  
4614  
4615  
4616  
4617  
4618  
4619  
4620  
4621  
4622  
4623  
4624  
4625  
4626  
4627  
4628  
4629  
4630  
4631  
4632  
4633  
4634  
4635  
4636  
4637  
4638  
4639  
4640  
4641  
4642  
4643

Example 1. \*

**Evidence for *correct species***

Part prototypes of *correct species*

Shape with texture similarity to image is HIGHER for the *correct species*.  
Color was NOT needed.

**Evidence for *wrong species***

Part prototypes of *wrong species*

In your opinion, what is the influence of color on the model's decision process?

Correct answer: 1 - None  
Explanation: Shape with texture similarity of image to prototypes was enough to predict the correct class.

- 1 - None
- 2 - Weak
- 3 - Don't know
- 4 - Moderate
- 5 - Substantial

Previous
Next

Figure 79: Page 2 of survey for LucidPPN with *no scores*

4644  
4645  
4646  
4647  
4648  
4649  
4650  
4651  
4652  
4653  
4654  
4655  
4656  
4657  
4658  
4659  
4660  
4661  
4662  
4663  
4664  
4665  
4666  
4667  
4668  
4669  
4670  
4671  
4672  
4673  
4674  
4675  
4676  
4677  
4678  
4679  
4680  
4681  
4682  
4683  
4684  
4685  
4686  
4687  
4688  
4689  
4690  
4691  
4692  
4693  
4694  
4695  
4696  
4697

Example 2. \*

**Evidence for *correct species***

Part prototypes of *correct species*

Shape with texture similarity to image is LOWER for the *correct species*.  
Color WAS needed.

**Evidence for *wrong species***

Part prototypes of *wrong species*

In your opinion, what is the influence of color on the model's decision process?

Correct answer: 5 - Substantial

Explanation: Shape with texture similarity of image to prototypes would lead to a mistake and it was avoided by considering color.

- 1 - None
- 2 - Weak
- 3 - Don't know
- 4 - Moderate
- 5 - Substantial

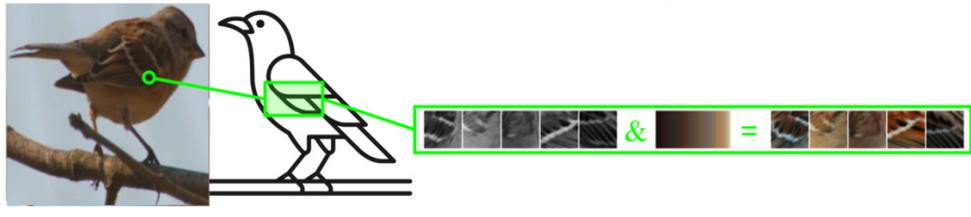
Previous
Next

Figure 80: Page 3 of survey for LucidPPN with *no scores*

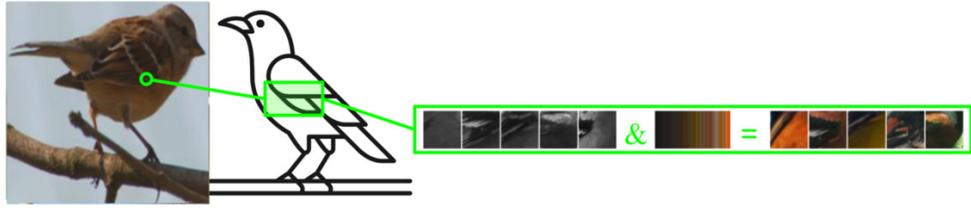
4698  
4699  
4700  
4701  
4702  
4703  
4704  
4705  
4706  
4707  
4708  
4709  
4710  
4711  
4712  
4713  
4714  
4715  
4716  
4717  
4718  
4719  
4720  
4721  
4722  
4723  
4724  
4725  
4726  
4727  
4728  
4729  
4730  
4731  
4732  
4733  
4734  
4735  
4736  
4737  
4738  
4739  
4740  
4741  
4742  
4743  
4744  
4745  
4746  
4747  
4748  
4749  
4750  
4751

Question 0. \*

Evidence for *correct species* Part prototypes of *correct species*



Evidence for *wrong species* Part prototypes of *wrong species*



In your opinion, what is the influence of color on the model's decision process?

- 1 - None
- 2 - Weak
- 3 - Don't know
- 4 - Moderate
- 5 - Substantial

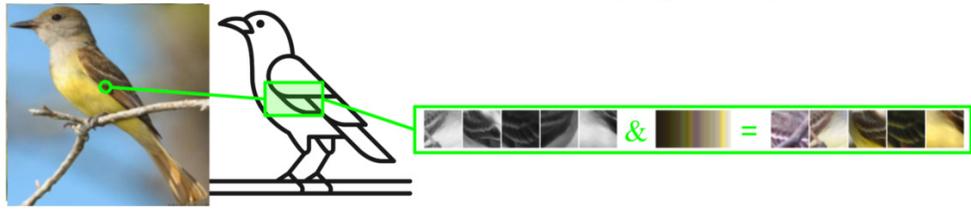
[Previous](#) [Next](#)

Figure 81: Page 4 of survey for LucidPPN with *no scores*

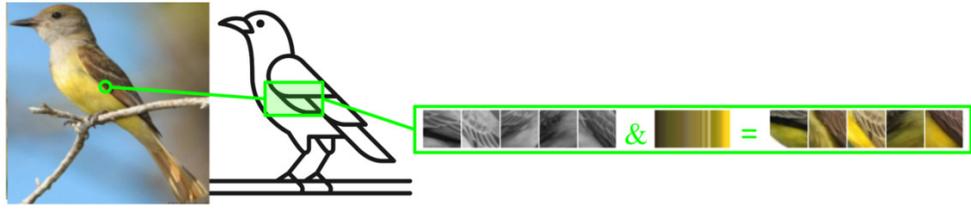
4752  
4753  
4754  
4755  
4756  
4757  
4758  
4759  
4760  
4761  
4762  
4763  
4764  
4765  
4766  
4767  
4768  
4769  
4770  
4771  
4772  
4773  
4774  
4775  
4776  
4777  
4778  
4779  
4780  
4781  
4782  
4783  
4784  
4785  
4786  
4787  
4788  
4789  
4790  
4791  
4792  
4793  
4794  
4795  
4796  
4797  
4798  
4799  
4800  
4801  
4802  
4803  
4804  
4805

Question 1. \*

Evidence for *correct species* Part prototypes of *correct species*



Evidence for *wrong species* Part prototypes of *wrong species*



In your opinion, what is the influence of color on the model's decision process?

- 1 - None
- 2 - Weak
- 3 - Don't know
- 4 - Moderate
- 5 - Substantial

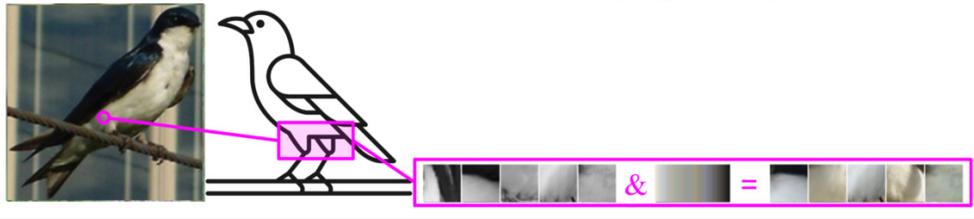
[Previous](#) [Next](#)

Figure 82: Page 5 of survey for LucidPPN with *no scores*

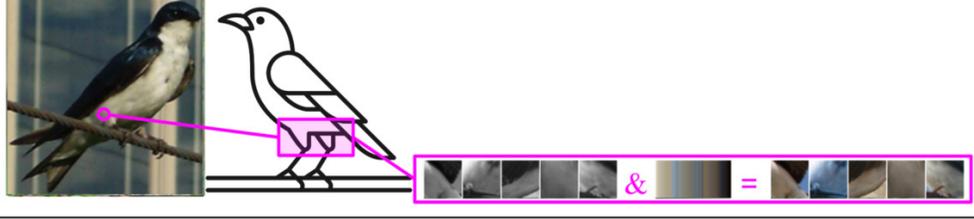
4806  
4807  
4808  
4809  
4810  
4811  
4812  
4813  
4814  
4815  
4816  
4817  
4818  
4819  
4820  
4821  
4822  
4823  
4824  
4825  
4826  
4827  
4828  
4829  
4830  
4831  
4832  
4833  
4834  
4835  
4836  
4837  
4838  
4839  
4840  
4841  
4842  
4843  
4844  
4845  
4846  
4847  
4848  
4849  
4850  
4851  
4852  
4853  
4854  
4855  
4856  
4857  
4858  
4859

Question 2. \*

Evidence for *correct species* Part prototypes of *correct species*



Evidence for *wrong species* Part prototypes of *wrong species*



In your opinion, what is the influence of color on the model's decision process?

- 1 - None
- 2 - Weak
- 3 - Don't know
- 4 - Moderate
- 5 - Substantial

[Previous](#) [Next](#)

Figure 83: Page 6 of survey for LucidPPN with *no scores*

4860  
4861  
4862  
4863  
4864  
4865  
4866  
4867  
4868  
4869  
4870  
4871  
4872  
4873  
4874  
4875  
4876  
4877  
4878  
4879  
4880  
4881  
4882  
4883  
4884  
4885  
4886  
4887  
4888  
4889  
4890  
4891  
4892  
4893  
4894  
4895  
4896  
4897  
4898  
4899  
4900  
4901  
4902  
4903  
4904  
4905  
4906  
4907  
4908  
4909  
4910  
4911  
4912  
4913

Question 3. \*

Evidence for *correct species*

Part prototypes of *correct species*

Evidence for *wrong species*

Part prototypes of *wrong species*

In your opinion, what is the influence of color on the model's decision process?

- 1 - None
- 2 - Weak
- 3 - Don't know
- 4 - Moderate
- 5 - Substantial

Previous Next

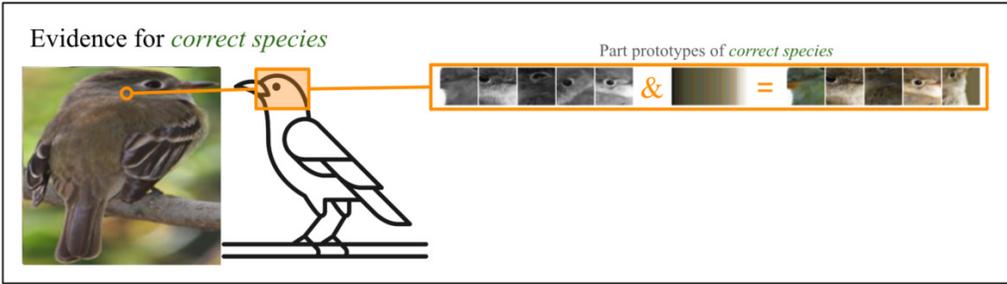
Figure 84: Page 7 of survey for LucidPPN with *no scores*

4914  
4915  
4916  
4917  
4918  
4919  
4920  
4921  
4922  
4923  
4924  
4925  
4926  
4927  
4928  
4929  
4930  
4931  
4932  
4933  
4934  
4935  
4936  
4937  
4938  
4939  
4940  
4941  
4942  
4943  
4944  
4945  
4946  
4947  
4948  
4949  
4950  
4951  
4952  
4953  
4954  
4955  
4956  
4957  
4958  
4959  
4960  
4961  
4962  
4963  
4964  
4965  
4966  
4967

Question 4. \*

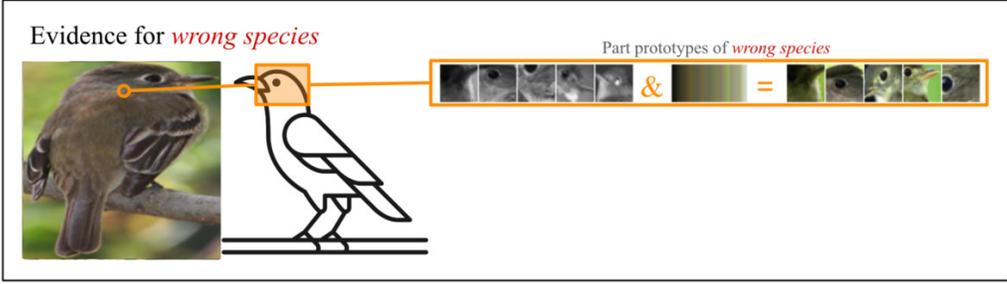
Evidence for *correct species*

Part prototypes of *correct species*



Evidence for *wrong species*

Part prototypes of *wrong species*



In your opinion, what is the influence of color on the model's decision process?

- 1 - None
- 2 - Weak
- 3 - Don't know
- 4 - Moderate
- 5 - Substantial

Previous Next

Figure 85: Page 8 of survey for LucidPPN with *no scores*

4968  
4969  
4970  
4971  
4972  
4973  
4974  
4975  
4976  
4977  
4978  
4979  
4980  
4981  
4982  
4983  
4984  
4985  
4986  
4987  
4988  
4989  
4990  
4991  
4992  
4993  
4994  
4995  
4996  
4997  
4998  
4999  
5000  
5001  
5002  
5003  
5004  
5005  
5006  
5007  
5008  
5009  
5010  
5011  
5012  
5013  
5014  
5015  
5016  
5017  
5018  
5019  
5020  
5021

Question 5. \*

Evidence for *correct species*

Part prototypes of *correct species*

Evidence for *wrong species*

Part prototypes of *wrong species*

In your opinion, what is the influence of color on the model's decision process?

1 - None

2 - Weak

3 - Don't know

4 - Moderate

5 - Substantial

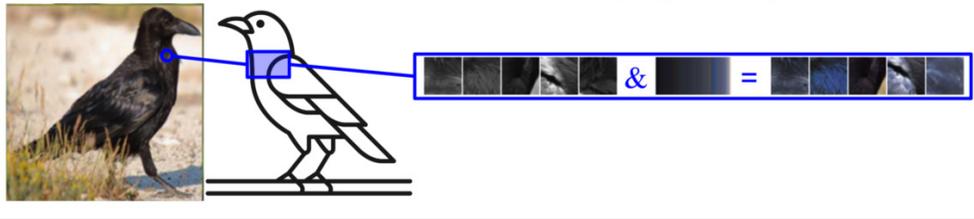
Previous Next

Figure 86: Page 9 of survey for LucidPPN with *no scores*

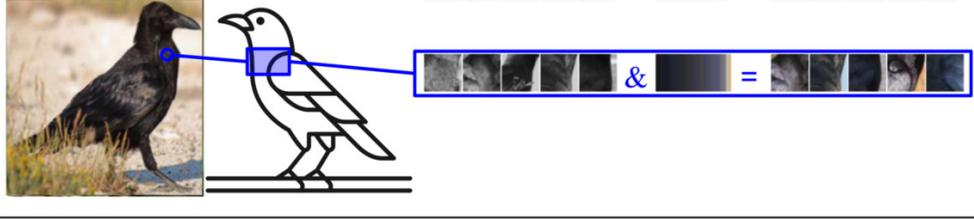
5022  
5023  
5024  
5025  
5026  
5027  
5028  
5029  
5030  
5031  
5032  
5033  
5034  
5035  
5036  
5037  
5038  
5039  
5040  
5041  
5042  
5043  
5044  
5045  
5046  
5047  
5048  
5049  
5050  
5051  
5052  
5053  
5054  
5055  
5056  
5057  
5058  
5059  
5060  
5061  
5062  
5063  
5064  
5065  
5066  
5067  
5068  
5069  
5070  
5071  
5072  
5073  
5074  
5075

Question 6. \*

Evidence for *correct species* Part prototypes of *correct species*



Evidence for *wrong species* Part prototypes of *wrong species*



In your opinion, what is the influence of color on the model's decision process?

- 1 - None
- 2 - Weak
- 3 - Don't know
- 4 - Moderate
- 5 - Substantial

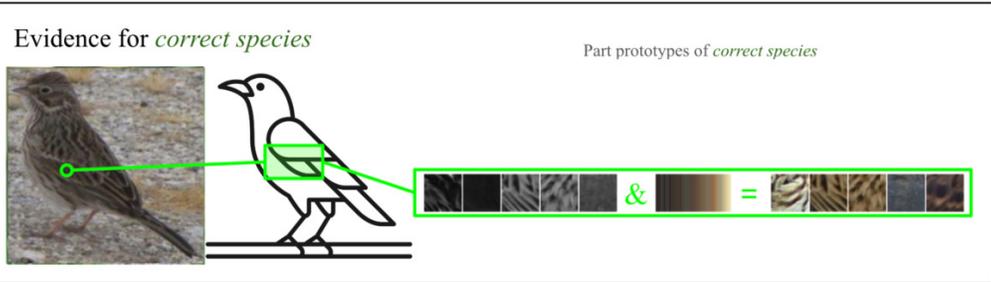
[Previous](#) [Next](#)

Figure 87: Page 10 of survey for LucidPPN with *no scores*

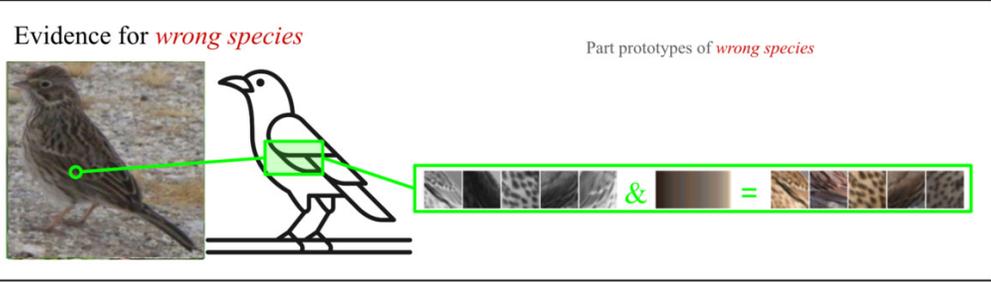
5076  
5077  
5078  
5079  
5080  
5081  
5082  
5083  
5084  
5085  
5086  
5087  
5088  
5089  
5090  
5091  
5092  
5093  
5094  
5095  
5096  
5097  
5098  
5099  
5100  
5101  
5102  
5103  
5104  
5105  
5106  
5107  
5108  
5109  
5110  
5111  
5112  
5113  
5114  
5115  
5116  
5117  
5118  
5119  
5120  
5121  
5122  
5123  
5124  
5125  
5126  
5127  
5128  
5129

Question 7. \*

Evidence for *correct species* Part prototypes of *correct species*



Evidence for *wrong species* Part prototypes of *wrong species*



In your opinion, what is the influence of color on the model's decision process?

- 1 - None
- 2 - Weak
- 3 - Don't know
- 4 - Moderate
- 5 - Substantial

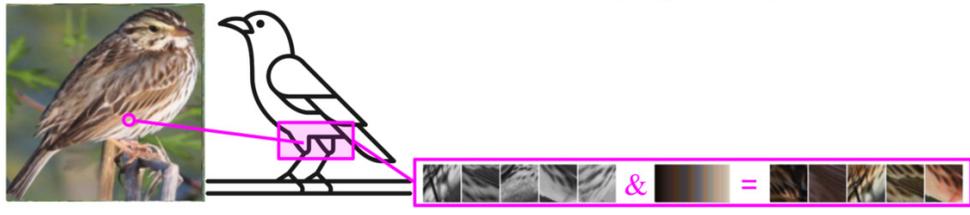
[Previous](#) [Next](#)

Figure 88: Page 11 of survey for LucidPPN with *no scores*

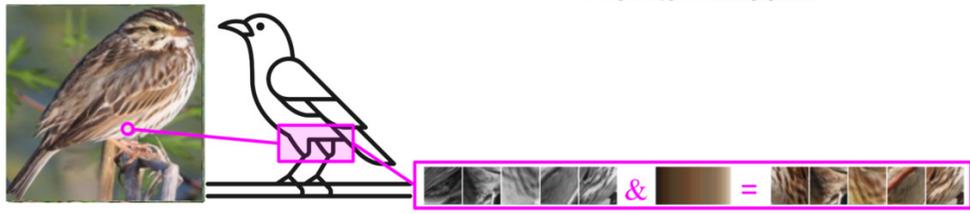
5130  
5131  
5132  
5133  
5134  
5135  
5136  
5137  
5138  
5139  
5140  
5141  
5142  
5143  
5144  
5145  
5146  
5147  
5148  
5149  
5150  
5151  
5152  
5153  
5154  
5155  
5156  
5157  
5158  
5159  
5160  
5161  
5162  
5163  
5164  
5165  
5166  
5167  
5168  
5169  
5170  
5171  
5172  
5173  
5174  
5175  
5176  
5177  
5178  
5179  
5180  
5181  
5182  
5183

Question 8. \*

Evidence for *correct species* Part prototypes of *correct species*



Evidence for *wrong species* Part prototypes of *wrong species*



In your opinion, what is the influence of color on the model's decision process?

- 1 - None
- 2 - Weak
- 3 - Don't know
- 4 - Moderate
- 5 - Substantial

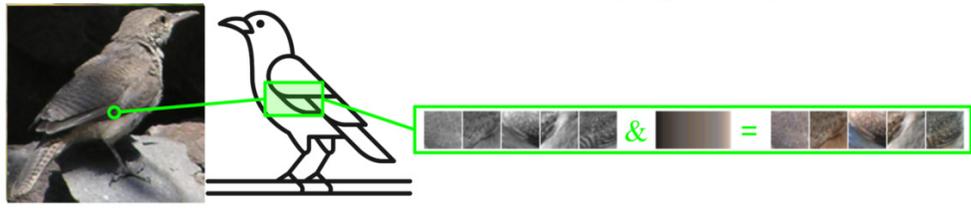
[Previous](#) [Next](#)

Figure 89: Page 12 of survey for LucidPPN with *no scores*

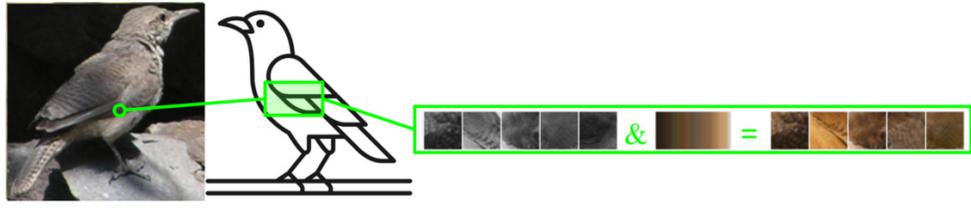
5184  
5185  
5186  
5187  
5188  
5189  
5190  
5191  
5192  
5193  
5194  
5195  
5196  
5197  
5198  
5199  
5200  
5201  
5202  
5203  
5204  
5205  
5206  
5207  
5208  
5209  
5210  
5211  
5212  
5213  
5214  
5215  
5216  
5217  
5218  
5219  
5220  
5221  
5222  
5223  
5224  
5225  
5226  
5227  
5228  
5229  
5230  
5231  
5232  
5233  
5234  
5235  
5236  
5237

Question 9. \*

Evidence for *correct species* Part prototypes of *correct species*



Evidence for *wrong species* Part prototypes of *wrong species*



In your opinion, what is the influence of color on the model's decision process?

- 1 - None
- 2 - Weak
- 3 - Don't know
- 4 - Moderate
- 5 - Substantial

[Previous](#) [Send job](#)

Figure 90: Page 13 of survey for LucidPPN with *no scores*