Prototype-based geometry of local object patches in the pigeon Mesopallium Ventrolaterale

Xiaoteng Zhang[1,#], Shan Lu[2,#], Yongxiang Lian[1], Pingge Hu[3], Shaoju Zeng[2], Li Shi[1,4*]

1 Department of Automation, Tsinghua University, Beijing 100084, China

2 Beijing Key Laboratory of Gene Resource and Molecular Development, College of Life Sciences, Beijing Normal University, Beijing 100875, China

3 China Academy of Information and Communications Technology, Beijing 100191, China

4 Institute for Brain and Cognitive Sciences, Tsinghua University, Beijing 100084, China

Li Shi*: shilits@tsinghua.edu.cn; 15010185566

ABSTRACT

In natural vision, objects are frequently partially occluded and thus perceived through local fragments, yet how such fragments are organized in neural representational space remains poorly understood. Using pigeons as a model, we combined behavior from a whole–part image categorization paradigm with population electrophysiological recordings from the mesopallium ventrolaterale (MVL) during passive viewing. We used CEBRA-time to embed neural time series into a low-dimensional space, defined category neural prototypes from population responses to whole images, and projected local patches into the same space to quantify their distances to the prototypes. This framework allowed systematic tests of how prototype distance relates to behavioral accuracy and patch semantics. In MVL, animal categories formed compact, prototype-centered clusters. Across patches, greater prototype distance predicted lower behavioral accuracy, and patches closer to the prototype were enriched for diagnostic semantic parts (e.g., heads and faces). By contrast, across multiple deep-network baselines, prototype distance neither reliably predicted behavior nor consistently reflected the semantic distribution of patches. Building on these baselines, we then introduced a local prototype learning mechanism in which category matching is governed by the proximity between local features and learned prototypes. The resulting network reproduced several MVL-consistent geometric signatures, including prototype-centered clustering for animal categories and a negative relationship between prototype distance and behavior. Together, these results support the hypothesis that incomplete-object representations in pigeon MVL are organized by a diagnostic-part-related prototype geometry, and suggest that this geometry provides

actionable neural constraints for brain-inspired vision models.

Keywords: Prototype representation; Local recognition; MVL population code; CEBRA-time embedding; Neural constrained network.

INTRODUCTION

How the brain constructs category representations that are both stable and decision-relevant from high-dimensional, noise-corrupted and often incomplete sensory inputs is a central problem in visual recognition. Classic prototype theories propose that categories are represented by "prototypes" in a psychological or neural representational space: through experience, observers form a central tendency for each category, and the similarity or distance between a novel stimulus and category prototypes determines both its classification and the difficulty of that judgment (Posner & Keele, 1968; Rosch, 1975; Nosofsky, 1992). In vision, a prototype can be viewed as the center of a category in representational space, estimated from responses to multiple exemplars; the distance from an individual image to the prototype naturally indexes its typicality or deviation. Converging evidence from neuroimaging and electrophysiology further suggests that high-level visual cortices encode a prototype–typicality geometry: across both face space and object space, response magnitude and multivariate patterns often vary systematically with distance to a prototype or an average face, consistent with a representational organization anchored around central category statistics (Leopold et al., 2001; Kahn et al., 2010; Iordan et al., 2016).

Although prior work has shown that prototype structure and its semantic composition can be characterized in detail under fully visible viewing conditions, real-world objects are often partially occluded and may appear only as local fragments. Under such locally visible conditions, visual recognition typically requires integrating part relations across space or engaging additional computations to support pattern completion and robust decisions (Orlov et

al., 2018; Tang et al., 2018). Consistent with this view, computational modeling of human neural and behavioral data suggests that a single feedforward sweep is often insufficient to account for recognition under occlusion, and that recurrent and feedback processes play a critical role (Rajaei et al., 2019; Kar et al., 2019). From a prototype-representation perspective, however, a more specific question that remains largely untested is: when only local patches are available, how is this local evidence organized in representational space relative to the prototype? Put differently, to what extent is the distance from an individual patch to the category prototype—defined from whole-object responses—driven by low-level visual statistics versus the semantic cues carried by the patch (e.g., whether it contains diagnostic parts such as the head/face), and how does this "patch-level prototype distance" map onto actual categorization behavior?

In artificial vision, deep neural networks (DNNs) achieve high classification accuracy under fully visible conditions. However, extensive work has shown systematic discrepancies between their robustness and that of humans under occlusion, partial visibility, and cluttered backgrounds, and that they tend to rely more on texture-like local statistics than on compositional shape or part structure (Geirhos et al., 2019). This has motivated two closely related lines of research in recent years: first, developing models with stronger part-based inductive biases to improve occlusion robustness (Kortylewski et al., 2020); and second, evaluating and constraining models using human behavioral and neural data, with particular emphasis on recurrent computations that may be required under occlusion to explain the dynamics of human recognition in challenging conditions (Yamins & DiCarlo, 2016; Schrimpf

et al., 2020; Maniquet et al., 2025). Together, these efforts provide important context for understanding local-to-global integration and how part evidence supports category decisions, yet mechanistic tests that link these questions to interpretable geometric quantities in avian visual systems remain scarce.

Relatedly, a separate line of machine-vision research has shown that prototype-, metric-, and matching-based learning frameworks can improve class discrimination by emphasizing class-aware anchors, structured similarity, and informative local features (Tang et al., 2020, 2022, 2023, 2025) and also highlight the broader usefulness of prototype-based representations in visual learning and interpretation (Wang et al., 2023; Zhou & Wang, 2024). Although these studies are not aimed at neural interpretability, they provide useful computational background for the prototype-based modeling strategy examined here.

Birds—pigeons in particular—offer a powerful model system for addressing these questions. Extensive behavioral work shows that pigeons can learn and maintain stable category judgments even when objects are fragmented, occluded, or presented as isolated patches, and that they tend to rely on diagnostic local cues rather than complete global contours (Wasserman et al., 2024; Clark & Colombo, 2022; Qadri & Cook, 2015). This behavioral competence makes pigeons well suited for probing how local fragments are encoded relative to category prototypes. At the neural level, the avian mesopallium ventrolaterale (MVL) is a key node in the tectofugal visual pathway and exhibits selectivity for object categories and complex shapes; population activity in MVL can support category discrimination (Azizi et al., 2019; Anderson et al., 2020; Clark et al., 2022). Nevertheless, prior studies have largely focused on single-neuron tuning

and decoding with whole-object stimuli, leaving open whether MVL category "centers" (i.e., neural prototypes) are systematically linked to the semantic content of local patches and to behavioral performance.

Here we address this gap by integrating behavioral experiments, population electrophysiology from pigeon MVL, and deep-network modeling within a unified representational framework. We first established pigeons' robust category recognition in an active object categorization task, and then recorded MVL population responses during passive viewing to the same set of whole images and their corresponding local patches. Using population responses to whole images, we defined an MVL prototype for each category and projected local patches into the same representational space. We then tested, along the axis of prototype distance, how patch-to-prototype relationships predict behavioral accuracy and relate to semantic part annotations and low-level visual features. Our results show that MVL prototype distance differentiates patches carrying diagnostic semantic evidence from nondiagnostic fragments. We next examined internal representations of a canonical deep network on the same stimulus set and found that its prototype distance fails to jointly account for behavioral performance and the semantic distribution of patches. Building on this analysis, we constructed a bird-like network by introducing local prototype learning mechanism, demonstrating that a semantics-sensitive prototype geometry can be abstracted into an explicit computational principle and transferred to artificial models. Together, this framework both characterizes the prototype-based organization of category representations in avian MVL under fragmentary viewing and provides concrete neural constraints for developing more biologically

121 grounded vision models.

122

MATERIALS AND METHODS

Subjects

We performed chronic electrophysiological recordings in 10 adult pigeons (*Columba livia*; both sexes; ~1 year old; 300–400 g) prepared for electrophysiological experiments. Before experiments, birds were housed individually under a natural light–dark cycle, with mixed grain diet and ad libitum water. Food intake was controlled to maintain body weight at 80–90% of free-feeding weight. Among these birds, 6 adult pigeons were also used for behavioral training under the same housing conditions; mild food restriction was used such that birds obtained their daily food through training. All procedures were approved by the Life Science Ethics Review Committee of Zhengzhou University and complied with institutional animal care and use guidelines.

Surgery and electrode implantation

Pigeons were anesthetized with 20% ethyl carbamate at 0.3 mL/100 g body weight. After anesthesia, feathers over the head and around the ears were removed, and the animal was secured in a custom stereotaxic apparatus (Shenzhen RWD Life Science Co., Ltd., RWD). The scalp was incised to expose the skull. Using the anterior–posterior (AP) axis as reference (AP zero), the reference position was determined by measuring the distance between the bregma point and the ear bars. Target coordinates for MVL followed the pigeon brain atlas (MVL: AP 11, ML 4.5, DV 2.0) (Karten & Hodos, 1967). Recordings were performed with a 16-channel tungsten microelectrode array (Suzhou Kedou Brain–Computer Technology Co., Ltd.; 4×4 array; 250 μm inter-electrode spacing; 3 mm electrode length; ~800 kΩ impedance). DV was

144 set to 0 when electrodes contacted the brain surface. A small craniotomy was drilled and the
145 dura was carefully opened. Electrodes were lowered at ~1 mm per 10 min to minimize tissue
146 deformation. The reference wire was placed subcutaneously as the ground electrode
147 (Supplementary Figure S1A). The electrode array was fixed with dental resin. After recovery,
148 birds were returned to their home cages and neural recordings were performed after 5 days of
149 recovery.

150 **Active behavioral paradigm**

151 To probe prototype organization in MVL, the stimulus set comprised whole and local natural
152 images from four object categories: human, pigeon, ship, and tree. Human and pigeon were
153 grouped as "animal," whereas ship and tree were grouped as "non-animal." Whole-category
154 images ("whole images") were 1080×1080 px. Local patches were generated by sampling
155 random centers and cropping 270×270 px patches to cover different object parts (Figure 1A;
156 Supplementary Figure S2).

157 Six pigeons were trained in a delayed matching-to-sample task using a behavioral training
158 chamber (Supplementary Figure S1C). Before formal training, birds learned to peck a key to
159 obtain food reward. In each trial (Figure 1C), a white dot appeared to prompt trial initiation; a
160 peck started the trial. After a 1.5 s blank period, a whole image was presented for 1 s, followed
161 by a 0.5 s delay. Two choice images then appeared: one from the correct category and one from
162 a randomly selected incorrect category. The bird indicated its decision by pecking the
163 corresponding key. Training was conducted daily and performance was tracked; pigeons that
164 reached ≥80% accuracy advanced to the test phase. Five pigeons met this criterion and

proceeded to testing (Supplementary Figure S3). The testing procedure was identical except that sample stimuli were local patches rather than whole images.

Passive viewing paradigm

To examine the bottom-up, perception-driven organization of neural prototypes in MVL, we performed passive-viewing recordings to collect neural signals from trained pigeons. The paasssive task used the same stimulus set as the aactive task.

Stimuli were presented using MATLAB Psychophysics Toolbox (Psychtoolbox, PTB). Images were displayed on a 50-inch, 16:9 monitor positioned ~20–40 cm from the pigeon's head (3840×2160 resolution; 100 Hz refresh rate; 36.486 pixels/cm), covering a maximum visual angle of ~138°×112°. Because pigeons show a right-eye dominance (Güntürkün & Hahmann, 1994; Güntürkün et al., 2000), recordings were performed in the left MVL while the left eye was occluded, ensuring stimulation through the right eye only. The display was color-calibrated prior to experiments and monitor placement was adjusted to match the bird's typical viewing geometry (Wang et al., 2025; Supplementary Figure S1B). Before each recording session, receptive fields were coarsely mapped and monitor position was fine-tuned so that the receptive-field center aligned with the center of the display area. In the passive task, stimulus duration was 1 s, followed by a randomly jittered inter-stimulus interval of 10–20 s.

Neural data acquisition and spike sorting

Neural signals were acquired using a Cerebus multichannel recording system (Blackrock Microsystems, Salt Lake City, UT, USA), amplified, filtered, and digitized at 30 kHz. For spike extraction, signals were bandpass-filtered between 300 Hz and 5 kHz, and threshold crossings

were detected using a threshold of −4 to −6 standard deviations of the signal. Task events (e.g., stimulus onset/offset) were synchronized and recorded for subsequent analyses.

Spike sorting was performed using Kilosort (Pachitariu et al., 2016), followed by manual curation and quality control in Phy2 (Rossant et al., 2016). Only well-isolated units with statistically reliable separability were retained. Neurons were classified into broad- versus narrow-spiking groups based on waveform features (e.g., spike width/repolarization-related metrics) and baseline firing statistics (Ditz et al., 2022; Trainito et al., 2019) (Supplementary Figure S4). Because no significant differences in response properties were observed between the two classes (Supplementary Figure S4), neurons were pooled for all main analyses.

Basic response quantification

Visual responses were quantified using baseline-subtracted firing rates. For each unit, baseline firing rate was computed in a pre-stimulus window (−3000 to 0 ms) and subtracted from the mean firing rate during the response window (50–300 ms after stimulus onset). For each image, responses were averaged across repeated trials (n=3) to yield a single response value. Peri-stimulus firing-rate time courses were computed using a sliding window (100 ms window length; 1 ms step).

Decoding analysis

To assess category separability in population activity for whole images and local patches, we performed both 4-way (human/pigeon/ship/tree) and 2-way (animal vs non-animal) decoding. Trial labels were assigned based on stimulus category. Decoding performance was evaluated with five-fold cross-validation. To examine how decoding depended on population

size, we randomly subsampled increasing numbers of neurons from the full set recorded in each pigeon, repeated this procedure 50 times, and reported mean accuracy ± standard deviation across repetitions. A linear support vector machine (linear SVM) classifier was used (Majaj et al., 2015), treating each neuron's response as one feature dimension and optimizing a linear decision boundary.

Neural prototype geometry in MVL

Following classic prototype theory, we defined the category prototype as the category center in MVL population representational space: the prototype for category c was the mean population response to the whole images of that category (Posner & Keele, 1968; Leopold et al., 2001; Iordan et al., 2016). To construct time-resolved population vectors while controlling dimensionality, firing-rate time courses were smoothed and downsampled. For each neuron, we extracted the firing-rate trace from 50–300 ms post-stimulus, yielding 250 time points, and applied a moving average with a 50 ms window and 10 ms step to compress to $T = 25$time bins. This produced a single-trial time-series vector $f_n(t) \in \mathbb{R}^T$per neuron. For each trial, vectors were concatenated across neurons to form a population time series of shape $N_{\text{neurons}} \times T$, which served as input to CEBRA-time embedding.

CEBRA-time embedding and trial-level representation: To characterize population-level representational geometry, we embedded the population time series using CEBRA-time (Schneider et al., 2023), a contrastive learning method that leverages temporal proximity: temporally adjacent segments within the same trial are treated as similar samples, whereas temporally distant segments or segments from other trials are treated as dissimilar. The model

learns smooth low-dimensional representations without using category labels.

In implementation, each trial's $N_{\text{neurons}} \times T$population time series was fed into CEBRA-time. We used the offset10-model architecture described for time-series data in the original work, with batch size 512, learning rate 3×10^{-4}, and 4000 maximum iterations. The embedding dimension was set to 3 for visualization and downstream geometric analyses. After training, each time bin was mapped to a 3D embedding $z_{c,i}(t) \in \mathbb{R}^3$, corresponding to category c, image i, and time t. Because our focus was prototype geometry rather than temporal dynamics, we summarized each trial trajectory by its temporal mean, yielding a single trial-level point:

$$\mathbf{z}_{c,i} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{z}_{c,i}\,(t).$$

Category prototypes were defined as the mean of whole-image embeddings within each category:

$$\mathbf{p}_c = \frac{1}{n_c}\sum_{i=1}^{n_c} \mathbf{z}_{c,i}\,,$$

where $n_c$is the number of whole images in category c. For each local patch, we computed its embedding $\mathbf{z}_{\text{patch}}$using the same preprocessing and the trained CEBRA mapping, and defined prototype distance as the Euclidean distance to the corresponding category prototype:

$$d_{\text{proto}} = \| \mathbf{z}_{\text{patch}} - \mathbf{p}_c \|_2.$$

Because the embedding vectors were L2-normalized, Euclidean distance in the embedding space is equivalent to cosine distance; thus, our prototype-distance measure captures pattern geometry rather than amplitude effects.

To quantify the relationship between prototype distance and recognition accuracy, we used Pearson correlation. This analysis tests a clear continuous-variable relationship—namely, whether prototype distance linearly predicts recognition accuracy. Pearson correlation provides a direct estimate of the direction and strength of this linear association.

Because local patches were cropped from the same set of full images used to define category prototypes, we conducted a control to ensure that the reported effects were not driven by this stimulus-set coupling. Specifically, we implemented a cross-validated (disjoint-set) prototype estimation procedure: prototypes were computed from a subset of whole images, and prototype distances and all downstream analyses were evaluated on held-out whole images and their corresponding patches, ensuring that prototype anchors were estimated from data independent of the test stimuli. The cross-validated results were highly consistent with the original analyses, supporting the robustness of our conclusions (Supplementary Figure S5).

Behavioral accuracy and prototype-distance heatmaps: To visualize the spatial distribution of behavioral performance and prototype distance over image space, we constructed a behavioral-accuracy heatmap and a prototype-distance heatmap for each whole image. Let $R_{i,k} \subset \Omega$ denote the pixel set covered by patch k cropped from image i, where Ω is the set of pixel coordinates in the whole image. For each patch, we obtained prototype distance $d_{i,k}^{\mathrm{proto}}$ from the CEBRA embedding and behavioral accuracy $p_{i,k}^{\mathrm{beh}}$ from the active task (averaged across animals and repeated trials). For any pixel location $\mathbf{x} \in \Omega$, we averaged values over all patches covering that pixel:

$$H_i^{\text{proto}}(\mathbf{x}) = \frac{1}{|\mathcal{K}(\mathbf{x})|}\sum_{k\in\mathcal{K}(\mathbf{x})} d_{i,k}^{\text{proto}}, H_i^{\text{beh}}(\mathbf{x}) = \frac{1}{|\mathcal{K}(\mathbf{x})|}\sum_{k\in\mathcal{K}(\mathbf{x})} p_{i,k}^{\text{beh}},$$

where $\mathcal{K}(\mathbf{x}) = \{k|\mathbf{x} \in R_{i,k}\}$indexes patches covering pixel $\mathbf{x}$. Spatial coupling between prototype distance and behavioral performance was quantified using Pearson correlation, Spearman correlation, and mutual information computed over paired pixel-wise samples from $H_i^{\text{proto}}$and H_i^{beh}. This analysis treats the two maps as paired two-dimensional fields and asks whether they exhibit global spatial concordance across corresponding locations. Because heatmap values can be non-normally distributed, include outliers, and show relationships that may be non-linear yet monotonic, we report both Pearson and Spearman correlations. Pearson correlation captures linear agreement in pixel-wise intensities, whereas Spearman correlation captures rank-order agreement and is more robust to scale differences and outliers. In addition, mutual information provides a complementary, model-free measure of statistical dependence that can capture non-linear associations not reflected by correlation alone. Together, these metrics provide a robust and interpretable characterization of spatial coupling between H_i^{proto} and H_i^{beh}.

Deep network modeling

Baseline model (AlexNet): We used ImageNet-pretrained AlexNet (Krizhevsky et al., 2012), ResNet (He et al., 2016), ConvNeXt (Liu et al., 2022), and Vision Transformer (ViT; Dosovitskiy et al., 2021), together with the biologically motivated CORnet model (Kubilius et al., 2019), as baseline architectures for comparison. AlexNet served as a “mammal-style” reference model, as its higher-level features have been widely used as analogs of primate ventral-stream and inferotemporal (IT) representations (Yamins et al., 2014; Khaligh-Razavi

& Kriegeskorte, 2014; Güçlü & van Gerven, 2015; Yamins & DiCarlo, 2016; Rose et al., 2021). To cover a broader range of inductive biases and computational mechanisms, we additionally included modern architectures: ResNet and ConvNeXt as representative convolutional networks (residual learning and contemporary ConvNet design) to assess robustness of conclusions across the CNN family; ViT as a patch-token, global self-attention framework to contrast architecture-dependent local-to-global integration; and CORnet as a recurrent, neuroscience-oriented model with an explicit ventral-stream stage mapping, providing a closer proxy for biological visual circuitry.

In this study, all networks were fine-tuned on the whole-image stimulus set. We replaced the original classification head with a 4-way classifier and optimized it using a cross-entropy loss. During fine-tuning, the backbone was trained with a learning rate of 1×10^{-5}, whereas the classification head used a learning rate of 3×10^{-4}. After fine-tuning, the same whole images and local patches used in the behavioral and MVL experiments were preprocessed following standard deep-network conventions and fed into each model. We then extracted activation vectors from representative layers at different processing stages (Table 1) as the primary features for representational analyses. These features were subsequently input to CEBRA-time to obtain a "network embedding space," within which we defined "network prototypes" and computed patch-to-prototype distances using exactly the same pipeline as for MVL.

Bird-like network with neural alignment and prototype constraints: After identifying systematic differences between deep-network and MVL representational geometry, we further asked whether, while keeping the front-end network feature extractor unchanged, the network

could be transformed into a more MVL-like "bird-like" model in representational geometry by introducing local prototype learning (Chen et al., 2019). This approach is related both to prior efforts to "neural-tune" deep networks using representational similarity, RDM alignment, or behavioral objectives (Yamins & DiCarlo, 2016; Schrimpf et al., 2018; Nonaka et al., 2021; Jacob et al., 2021), and to prototype- and metric-based representation learning in machine vision, where class-aware anchors or matching spaces are used to facilitate few-shot or fine-grained recognition (Tang et al., 2020, 2022, 2023). In contrast to directly enforcing representational alignment as an explicit optimization target, our goal here is to test whether a prototype-based inductive bias, formulated as a mechanism-level learning rule, can reproduce key aspects of the neural geometry observed in MVL, without claiming that the biological system necessarily implements the same algorithmic form.

We again used the pretrained deep-networks frontend as a feature extractor $f_{\text{front}}(\cdot)$. Given an input image $\mathbf{x}$, we extracted the feature vector $\mathbf{h} = f_{\text{front}}(\mathbf{x}) \in \mathbb{R}^D$ and divided into $7 * 7$local feature vectors. We then appended trainable local prototype bank $P \in \mathbb{R}^{C \times K}$, where Cdenotes the number of categories and $K$the number of local prototypes per category. During training, each local feature was compared against the class-specific prototypes to compute similarity scores; the class with the highest matching score was taken as the final classification output, and the prototype bank was updated via backpropagation together with the rest of the trainable parameters. During inference, the prototype bank was fixed and predictions were obtained directly from the matching between local patches and the prototype bank. Training used only a cross-entropy loss. The backbone was optimized with a learning rate of 1×10^{-5},

whereas the prototype bank and classification head used a learning rate of 3×10^{-4}. The model was trained using only whole images and evaluated on local patches.

All models were implemented in PyTorch, with the runtime environment Python 3.9, torch 2.7.1, and torchvision 0.22.1. CORnet-S was loaded using its public GitHub implementation, whereas all other pretrained models were instantiated from the official torchvision.models implementations.

Histology

After experiments, pigeons were euthanized by overdose injection of 20% urethane, followed by transcardial perfusion. The circulatory system was flushed with phosphate-buffered saline (PBS) and fixed with 4% paraformaldehyde (PFA). Brains were removed and post-fixed in 4% PFA for ~12 h, then cryoprotected in 30% sucrose at 4°C for 48 h. Tissue was frozen and coronally sectioned at 20 μm thickness. Sections were stained for cytochrome oxidase and examined under a microscope to verify electrode tracks and lesion sites (Supplementary Figure S1D).

RESULTS

Behavioral performance and MVL population decoding validate the experimental paradigm

We first established the reliability of our paradigm at both the behavioral and neural levels. In the active task, pigeons' choices gradually improved from near-chance performance during training and eventually stabilized at accuracies well above chance across multiple consecutive sessions (Supplementary Figure S3). Here, chance level was defined as 25% for the four-way categorization (human/pigeon/ship/tree) and 50% for the animal vs non-animal binary categorization. Upon entering the test phase, overall accuracy decreased relative to training but remained above chance, indicating that the animals had formed robust category representations for this stimulus set. Final performance in the training and test phases is summarized in Figure 1D, providing the empirical basis for quantitatively linking behavior to neural representational structure in subsequent analyses.

To verify at the neural level that MVL carries sufficient category information, we recorded population responses in left MVL from 10 chronically implanted pigeons during passive viewing while presenting the same stimulus set as in the behavioral task. After spike sorting and unit quality control, we obtained 257 well-isolated units for analysis. Neural features were constructed as each neuron's mean firing rate within the evoked response window (50–300 ms after stimulus onset). We then performed category decoding using a linear support vector machine (linear SVM). Importantly, decoding analyses were conducted separately for whole images and for local patches (i.e., training and testing within each stimulus type), allowing us

368 to assess category information under full-view versus fragmentary viewing conditions. For
369 each pigeon, we randomly subsampled increasing numbers of neurons, repeated decoding to
370 quantify variability, and evaluated performance with cross-validation as described in Methods.
371 Decoding accuracy increased monotonically with the number of neurons and gradually
372 approached saturation at larger population sizes (Figure 1E). This characteristic rise-to-
373 saturation profile indicates that MVL population activity contains substantial category
374 information that is encoded in a distributed manner: single neurons contribute limited
375 discriminative power, whereas recruiting larger neuronal ensembles markedly improves
376 classification.

377 Single-unit analyses further supported the quality and stability of the recorded signals. Spike-
378 sorting quality metrics, waveform features, category selectivity indices, and representative
379 raster plots all indicated high signal-to-noise ratios and reliable visually evoked responses
380 (Supplementary Figure S4A–D). Together, robust behavioral performance in the active task
381 and decodable category-specific activity in MVL during passive viewing demonstrate that our
382 paradigm elicits reliable decisions and measurable category representations at the population
383 level, thereby establishing a solid foundation for the subsequent analyses of representational
384 geometry and prototype organization in MVL.

385 **MVL forms animal-category prototypes that couple to behavior**

386 Having established the validity of the task and recordings, we next examined the geometric
387 organization of MVL population representations. Our goal was not merely to determine
388 whether category information could be decoded, but to characterize how MVL organizes

category structure in representational space, and whether animal versus non-animal categories show systematic geometric differences. To this end, we selected three pigeons with the highest recording quality and embedded their population time-series responses to all whole images and local patches from the passive-viewing task into a low-dimensional space using CEBRA-time, yielding an embedding that preserves temporal structure while capturing condition-dependent differences.

In the resulting CEBRA space, the category prototypes were well separated (Figure 2A, right; stars denote category prototypes). For animal-category stimuli, MVL representations exhibited compact, prototype-centered clustering, forming distinct clouds around their respective prototypes. In contrast, non-animal stimuli were distributed more diffusely in the same space and showed a weaker clustering tendency (Figure 2A; Supplementary Figure S6A). Building on this observation, we quantified euclidean prototype distance for each local patch relative to its corresponding category prototype, where prototypes were defined from MVL responses to whole images. Animal-category patches were, on average, closer to the category center, whereas non-animal patches showed larger distances and a broader distribution (Figure 2B; Supplementary Figure S6B). Thus, both the embedding geometry and the distance statistics indicate that MVL representations for animal categories are more compact and prototype-centered, whereas non-animal categories are encoded in a more heterogeneous and dispersed manner.

We next linked prototype distance to behavioral performance in the active categorization task. Within each pigeon, animal-category prototype distance was negatively correlated with

behavioral accuracy: patches closer to the prototype were more likely to be correctly classified, whereas accuracy declined as distance increased (Figure 2C; Supplementary Figure S6C; correlation coefficients and *p* values are reported in the figure legends). In contrast, the same within-subject analysis for non-animal categories did not reveal a stable relationship between prototype distance and behavior, with correlations near zero or showing only weak trends. This dissociation indicates that MVL prototypes are not only geometrically more salient for animal categories but are also functionally meaningful, such that distance to the prototype indexes decision difficulty. Collectively, these results support a prototype-centered decision geometry in MVL, at least for animal categories.

Prototype distance primarily reflects diagnostic semantic parts rather than low-level visual features

The observation that prototype distance predicts behavioral accuracy for animal categories raises a central question: what does MVL prototype geometry encode? Specifically, does prototype distance primarily reflect low-level visual statistics (e.g., contrast or texture), or higher-level information about diagnostic semantic parts? To address this, we analyzed the relationship between prototype distance and stimulus content at three complementary levels: spatial mapping, qualitative visualization, and quantitative annotation.

First, we constructed prototype-distance heatmaps by assigning each patch's prototype distance to all pixels covered by that patch and averaging values in overlapping regions. This yielded a pixel-wise map of prototype distance over each whole image (Figure 3A, bottom). In parallel, we constructed behavioral-accuracy heatmaps from the active task using the same

mapping procedure (Figure 3A, top). For animal categories, these two maps showed a pronounced spatial coupling: regions associated with higher behavioral accuracy tended to exhibit smaller prototype distances to the category prototype, whereas low-accuracy regions tended to coincide with larger prototype distances. This coupling was markedly weaker for non-animal categories. We quantified coupling for each image by computing mutual information, Pearson correlation, and Spearman correlation between the paired heatmaps, and then statistically evaluated these image-level values across the stimulus set. This analysis revealed consistently stronger coupling for animal categories (Figure 3B; Supplementary Figure S7A), further supporting the functional relevance of prototype distance in animal recognition.

To more directly interpret what image content corresponds to different distances, we next visualized animal-category patches that were closest and farthest from the neural prototype (left 10 vs right 10; Figure 3C; non-animal examples in Supplementary Figure S7B). Close-to-prototype patches typically contained highly diagnostic semantic parts—most prominently the head/face region and its immediate context—whereas far-from-prototype patches more often consisted of body boundaries or highly localized texture fragments that are unlikely to support reliable categorization in isolation.

We then formalized this observation using explicit semantic annotation. Each local patch was labeled with a binary indicator denoting whether it contained key diagnostic parts (head/face). As prototype distance increased, the fraction of patches containing these diagnostic semantic parts decreased (Figure 3D). Thus, patches nearer the prototype were more likely to

carry semantic evidence critical for animal categorization, whereas patches farther from the prototype were more likely to lack such evidence. To validate this interpretation, we excluded head/face-containing patches and re-ran the analysis pipeline; the prototype–behavior coupling effect was indeed attenuated (Supplementary Figure S8). No analogous a priori semantic-part label was imposed for the non-animal categories, because no comparably validated local semantic criterion was available for ship or tree patches in the current stimulus set.

Finally, to rule out confounds from low- and mid-level image properties, we further quantified several patch-level attributes, including Sobel edge density, image entropy, LBP entropy, as well as curvature distributions and contour organization, and examined their relationships with prototype distance (Figure 3E; Supplementary Figure S7C; Supplementary Figure S9). Overall, these low-level metrics showed little to no association with prototype distance. Taken together, these results support the interpretation that MVL prototype distance primarily reflects whether an animal patch contains diagnostic semantic parts (for animals, head/face), and that the key effect is not driven by coarse low-level visual statistics.

Baseline models show a non-MVL prototype geometry

The semantics-sensitive prototype geometry observed in MVL, together with its tight coupling to behavior, raises a key question: is this organization a generic property of any high-performing visual classifier, or a distinctive feature of the avian high-level visual system? To address this, we selected a set of representative modern convolutional and transformer architectures—AlexNet (Krizhevsky et al., 2012), ResNet (He et al., 2016), ConvNeXt (Liu et al., 2022), and the Vision Transformer (ViT; Dosovitskiy et al., 2021)—as well as the

biologically motivated CORnet model (Kubilius et al., 2019) as baseline comparators.

Using the same stimulus set, we extracted feature vectors from multiple layers of the fine-tuned models corresponding to early, intermediate, and late processing stages (Table 1), for both whole images and local patches, and constructed a low-dimensional representational space using the same CEBRA embedding pipeline as in the MVL analyses, to ensure methodological comparability. Because the MVL analyses were based on time-series responses ($T = 25$ bins) whereas model features are static ($T = 1$), we additionally performed a matched control analysis by averaging the MVL representations over time ($T = 1$) and re-running the full pipeline. Although the point-cloud distribution in the embedding space changed, the key results—including prototype distances and their correlations with behavior—remained consistent with the original analyses (Supplementary Figure S10-11).

In the CEBRA embedding space, for all baseline models, only the late-stage representation vectors yielded clearly separable category information, indicating that the deep layers provide sufficient discriminative capacity for this task, whereas early and intermediate layers are not sufficient to support complex categorization (Fig. 4A; Supplementary Figure S12-15). In terms of the point-cloud geometry in the embedding space, the baseline models also spontaneously exhibit clustered versus dispersed configurations. However, unlike MVL, they show a pronounced tendency to form a compact cluster for the tree category while dispersing other categories, potentially because canopy-related visual features occupy a large proportion of both whole images and local patches, making tree stimuli highly similar in feature space (Fig. 4A; Supplementary Figure S12-15). We next defined an “model prototype” for each category as

the mean representation of that category's whole images, and computed each patch's euclidean distance to its corresponding prototype. Strikingly, the resulting prototype-distance distributions showed the opposite pattern to MVL: animal-category patches tended to lie farther from their prototypes, whereas non-animal patches tended to lie closer (Figure 4A,B; Supplementary Figure S16A,B; Supplementary Figure S12-15). Thus, even when prototype definition and distance metrics were matched to the neural analyses, all baseline models exhibited a fundamentally different prototype geometry from pigeon MVL.

Critically, model prototype distance also failed to reproduce the functional coupling observed in MVL. We related patch-to-prototype distance to the network's own decision variable, defined as the probability for the correct class produced by the trained classification head. In this analysis, animal stimuli exhibited a trend opposite to MVL, with larger prototype distance associated with higher correct-class probability, whereas non-animal stimuli showed the reverse (Figure 4C). Moreover, when examined within individual categories, the correlation between patch prototype distance and correct-class probability was near zero (Supplementary Figure S16C; Supplementary Figure S12-15). Together, these results argue against a single, domain-general prototype-distance axis in these baseline models that consistently captures recognition difficulty across semantic domains. Instead, the network's "prototypes" appear more akin to discriminative centers shaped by the classification objective, rather than a stable, semantically unified geometry that predicts decision difficulty across categories, as observed in MVL.

To further contrast baseline models with MVL in image space, we constructed prototype-

distance heatmaps from AlexNet representations (Figure 4E). Unlike MVL-based heatmaps, which showed pronounced part-specific structure (Figure 3A, bottom), AlexNet heatmaps were comparatively uniform within images and were dominated by between-category offsets in overall distance magnitude (with animal categories exhibiting larger prototype distances, i.e., lighter colors). Quantifying the relationship between AlexNet prototype-distance heatmaps and pigeons' behavioral-accuracy heatmaps using mutual information, Pearson correlation, and Spearman correlation revealed no reliable animal–non-animal dissociation; in some cases, correlations even became negative (Figure 4D; Supplementary Figure S7D). This convergent evidence further indicates that deep-network lacks the stable and unified semantic prototype structure expressed by MVL population codes.

This contrast is informative in two ways. First, it demonstrates that the semantic prototype geometry observed in MVL is not an inevitable consequence of high classification performance, but instead reflects a specific representational organization that does not automatically emerge in a canonical DNN. Second, under the present task setting, mammal-style DNNs and pigeon MVL differ in how local evidence is organized relative to category prototypes: the former is more strongly shaped by the training objective and dataset statistics, whereas the latter exhibits a more prototype-centered organization associated with diagnostic semantic parts and behavioral performance. These observations motivate our subsequent approach of introducing local prototype learning to transform a standard DNN toward MVL-like representational geometry.

A local prototype learning mechanism reproduces MVL-like prototype geometry

Having identified systematic differences between MVL and DNN in representational geometry and prototype organization, we next asked whether a standard DNN could be endowed with a prototype-based inductive bias such that its internal geometry would more closely resemble several features observed in avian MVL. To this end, we designed a local prototype bank to learn category prototypes, enabling end-to-end training of the entire network. Training was performed using only whole images, whereas evaluation was conducted on local patches, and the objective was standard cross-entropy classification loss (Figure 5A). We performed this procedure across different backbone baselines; during training, the loss decreased steadily and classification accuracy showed an overall increasing trend with iterations (Figure S17A; Supplementary Figure S18A), indicating effective convergence.

After training, we first evaluated performance on local patches. The modified network achieved higher classification accuracy on patch stimuli than the original baseline (Figure 5B; Supplementary Figure S18B), suggesting that introducing the local prototype learning mechanism improves the model's ability to utilize local information under fragmentary viewing.

We then examined whether the modified network also captured MVL-like representational geometry. We applied the same CEBRA analysis to the outputs of the model and repeated the prototype-based analyses used for MVL. Notably, animal-category stimuli in the modified network formed compact, prototype-centered clusters in embedding space, reproducing several global features observed in the MVL CEBRA embedding (Figure 5C; Supplementary Figure

S17B, S19A). Consistent with this shift, prototype-distance statistics also moved toward the MVL pattern: animal-category patches showed shorter prototype distances than non-animal samples (Figure 5D; Supplementary Figure S17C; S19B). Moreover, patch-to-prototype distance exhibited a clear negative relationship with the model's prediction confidence, quantified as the probability assigned to the correct class, closely matching the prototype distance–behavior coupling observed in MVL (Figure 5E; Supplementary Figure S17D; S19C). Notably, the category-wise analysis in Supplementary Figure S17D showed negative trends for both animal and non-animal categories, suggesting that the network may also form within-category prototype structure for non-animal classes.

Finally, we projected both prototype distance and accuracy back onto the image plane to generate heatmaps (Supplementary Figure S20). We quantified the coupling between the network's prototype-distance heatmaps and pigeons' behavioral-accuracy heatmaps using mutual information, Pearson correlation, and Spearman correlation. The resulting pattern was broadly similar to that observed for MVL (Figure 5F; Supplementary Figure S17E; cf. Figure 3B), indicating that the modified network captures MVL-consistent structure not only in embedding space but also in the spatial organization of prototype distance across object images.

Together, these results show that imposing appropriate local prototype learning on a standard deep-network can simultaneously improve fragment-based recognition and reshape internal representational geometry toward the semantics-sensitive, prototype-centered organization expressed by MVL population codes. More broadly, it suggests that MVL prototype structure and its behavioral coupling can be abstracted into implementable computational principles and

injected into artificial networks via differentiable objective functions.

DISCUSSION

In this study, we combined behavior, MVL population coding, and deep-network modeling to characterize the computations supporting pigeons' prototype-based recognition under fragmentary viewing. At the neural level, MVL population activity revealed a clear geometric signature: animal categories formed compact clusters in CEBRA embedding space, with well-defined neural prototypes given by the mean response to whole images. Crucially, the distance from a local patch to its category prototype not only differentiated animal versus non-animal domains at the level of representational geometry, but also strongly predicted behavioral accuracy within animal categories. Semantic annotation and control analyses further showed that this prototype distance primarily reflected whether a patch contained diagnostic parts (especially head/face), rather than low-level image statistics such as contrast, edge density, or texture complexity. Together, these results support the view that MVL exhibits a prototype-centered geometry for animal recognition, and that distance to prototype constitutes a useful axis along which decision difficulty is expressed. Importantly, this semantic interpretation is currently strongest for animal categories, for which diagnostic local parts could be defined a priori. By contrast, the weaker non-animal effects should not be taken as evidence that non-animal objects lack informative local cues; rather, they indicate that an analogous semantic anchor was not established under the present stimulus set and annotation scheme. This finding aligns with the broader perspective that high-level visual systems organize information in abstract representational spaces where distances predict behavior and computational readouts (Kriegeskorte et al., 2008; Nili et al., 2014; Devereux et al., 2013; Cichy et al., 2016;

Kriegeskorte, 2015). It also extends prior work on avian MVL that largely emphasized single-unit tuning and decoding-based selectivity (Azizi et al., 2019; Anderson et al., 2020; Clark et al., 2022; Pusch et al., 2023) by revealing a population-level geometry in which animal prototypes are organized around diagnostic semantic parts.

A second implication concerns how MVL geometry relates to pigeons' well-documented preference for local information in whole–part tasks. Comparative studies have repeatedly shown that pigeons can categorize fragmented and occluded stimuli and often rely on local cues, consistent with the tectofugal pathway operating at relatively small spatial scales (Qadri & Cook, 2015; Soto & Wasserman, 2012; Clark & Colombo, 2022; Niu et al., 2022). Our results refine this interpretation by suggesting that the relevant "local" information in MVL is not encoded merely as undifferentiated texture patches. Instead, at least for animal categories, local fragments appear to be organized relative to the category prototype in a manner linked to diagnostic semantic parts, such that patches carrying diagnostic information are embedded nearer to the category prototype and yield higher accuracy. Thus, an apparent "local bias" need not imply a low-level, feature-bound representation; rather, MVL can support semantic prototype coding at the local scale. This local–semantic compromise is consistent with ecological demands in visually crowded environments, where rapid detection and recognition of socially and behaviorally relevant parts (e.g., faces, eyes) can be advantageous (Pusch et al., 2023; Wasserman et al., 2024).

In sharp contrast, deep-networks exhibited fundamentally different representational geometry on the same stimulus set. Although all baseline models' features supported clear

category separability in CEBRA space, neither the clustering pattern nor the "network prototype" defined from model captured MVL-like organization. Animal patches tended to lie farther from network prototypes, whereas non-animal patches were closer; moreover, the relationship between prototype distance and the network's decision variable was inconsistent across domains and near-zero within individual categories. Heatmap analyses reinforced this dissociation: model prototype-distance maps lacked robust part-specific structure and showed weak or even negative coupling with pigeons' behavioral maps. Overall, these results suggest that, under our task setting, standard DNNs can achieve adequate categorization with their high-level features (Baker et al., 2018; Baker & Elder, 2022), yet their local–global representational geometry differs from that observed in avian MVL, providing a more fine-grained entry point for cross-species comparisons of model representations.

Importantly, the divergence between MVL and modern deep-network was not merely descriptive. By introducing local prototype learning, we constructed a model that reproduced several key MVL-consistent signatures. The modified network improved patch-level classification relative to the baselines and, critically, captured the clustering geometry in embedding space together with the negative coupling between patch-to-prototype distance and correct-class probability. These results suggest that MVL's prototype geometry can be captured by an implementable computational bias: enforcing local prototype learning can move a DNN toward several MVL-like geometric signatures while maintaining task performance. From a cognitive computational neuroscience perspective, this provides a concrete example of how goal-driven modeling and biologically inspired constraints can be integrated. At the same time,

because the prototype-based model is category-general by construction, the present modeling results are best interpreted as supporting a useful prototype-based computational account, rather than as a complete explanation of why the biological data show stronger effects for animal than non-animal categories.

Several limitations point to clear directions for future work. First, our recordings focused on MVL, a high-level node in the tectofugal pathway, but we did not directly measure upstream or parallel regions (e.g., ENTO, NCL). As a result, we cannot yet determine where along the avian visual hierarchy semantic prototype geometry emerges, or how it is transformed across stages (Azizi et al., 2019; Pusch et al., 2023). Systematic recordings across multiple areas could test whether prototypes arise gradually through feedforward processing or appear prominently only at specific higher-level regions. Second, our paradigm used static natural images and cropped patches. Although CEBRA-time captures temporal structure in neural responses, the task does not probe the full richness of natural vision, including temporal prediction, motion invariance, or multi-step decision-making (Cichy et al., 2016; Schneider et al., 2023). Extending the paradigm to dynamic videos and naturalistic viewing could reveal how prototype geometry evolves over time and interacts with motion signals. Third, although introducing the local prototype learning mechanism made the networks exhibit MVL-consistent geometric signatures, the effects varied across backbones. This backbone dependence limits the strength of any model-specific mechanistic interpretation and suggests that the present results are better understood at the level of representational principle than definitive circuit implementation. Fourth, while we identified clear diagnostic parts for animal categories (head/face), we did not

establish an analogous a priori semantic-part definition for the non-animal categories. Therefore, the weaker non-animal effects should be interpreted cautiously and should not be taken as evidence that non-animal objects lack informative local cues. Rather, this remains an unresolved issue under the current stimulus set and annotation framework. Finally, the space of "mid-level" features is inherently broad. In the present study we adopted a set of representative, interpretable, and reproducible proxies; future work could incorporate stronger models of mid-level structure (e.g., metrics based on learnable shape primitives or hierarchical part parsing) to further broaden the scope of controls.

In summary, our results identify a population-level representational principle in pigeon MVL: a semantic part-related prototype geometry that predicts behavior under occlusion-like, fragmentary viewing. By translating this principle into differentiable constraints, we provide a practical bridge between neural description and executable computation, offering a biologically grounded computational account of avian high-level vision and concrete guidance for neurally constrained, brain-inspired visual modeling.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

SUPPLEMENTARY DATA

Supplementary Figures

COMPETING INTERESTS

The authors declare that they have no conflict of interest.

AUTHOR CONTRIBUTIONS

X.Z., S.L., S.Z and L.S. conceived and designed the study. X.Z. and S.L. performed the surgery and experiments. X.Z., S.L. and Y.L. analysed data. P.H. developed the visual stimulus procedure. X.Z., S.L., S.Z and L.S. wrote the manuscript. All authors read and approved the final version of the manuscript.

ACKNOWLEDGEMENTS

We thank Yanyan Peng and Juncai Zhu, ph.D. students at Zhengzhou University, for helpful discussions.

TABLES AND FIGURES

Table 1. Layers used for representational analyses across baseline models.

Model	Early layer	Mid layer	Late layer
ConvNeXt-T	stage1[-1]	stage2[-1]	stage4[-1]
CORnet-S	V1	V4	IT
ResNet18	layer1[-1]	layer2[-1]	layer4[-1]
ViT-B/16	block4	block8	block12

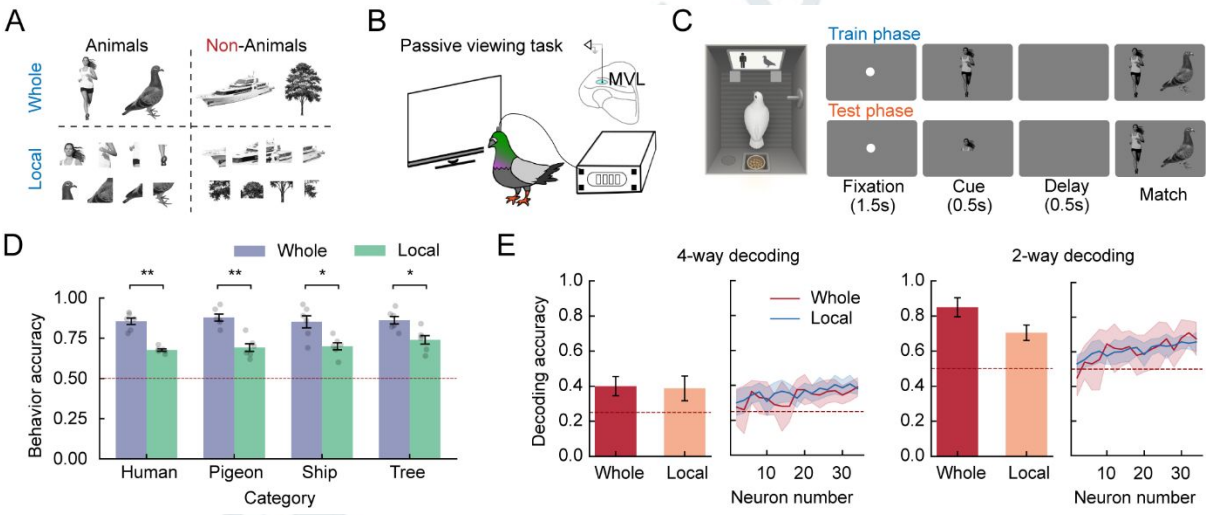


Figure 1 Overview of the task design, behavioral performance, and MVL population decoding. A: Schematic of the stimulus set: four categories (Human, Pigeon, Ship, Tree) presented as whole images (Whole) and local patches (Local); Human and Pigeon are grouped as Animals, whereas Ship and Tree are grouped as Non-Animals. **B:** Illustration of the passive viewing experiment and recording site: MVL population activity was recorded while pigeons viewed the stimuli. **C:** Passive viewing task timeline (train/test phases): Fixation (1.5 s), Cue (0.5 s), Delay (0.5 s), Match. **D:** Behavioral categorization accuracy comparing Whole versus

Local across the four categories; Wilcoxon rank-sum test, *: $P<0.05$, **: $P<0.01$. **E:** MVL population decoding: 4-way category decoding and 2-way superordinate decoding (Animals vs Non-Animals), comparing decoding accuracy between Whole and Local conditions, and summarizing performance as a function of the number of sampled neurons.

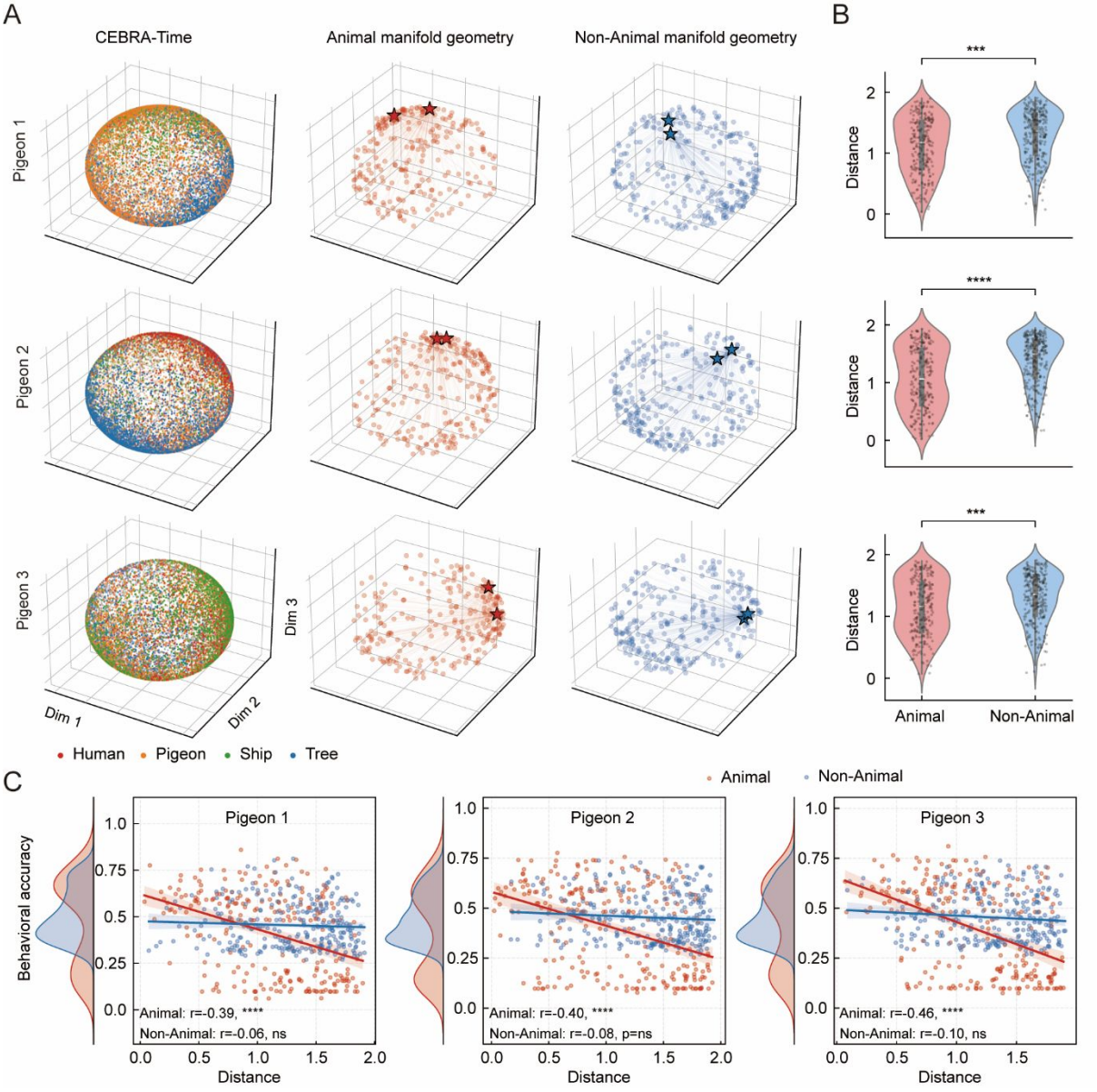


Figure 2 Prototype geometry in the MVL embedding space and its prediction of local-fragment recognition. A: Low-dimensional embedding of MVL population time series using CEBRA-time (Dim 1–3), shown for three pigeons with points colored by four basic categories

(Human, Pigeon, Ship, Tree). The embedded responses are further summarized into Animal versus Non-Animal manifold geometry to compare category-level structure; star markers denote category neural prototypes defined from whole-image responses. **B:** Distributions of prototype distance (distance from each local patch to its corresponding category prototype defined from whole-image responses) for Animal vs Non-Animal patches; Wilcoxon rank-sum test, ***: $P<0.001$, ****: $P<0.0001$. **C:** Relationship between prototype distance and behavioral accuracy: correlations were quantified using Pearson's correlation. Animal patches show robust negative correlations (Pigeon 1: $r=-0.39$, Pigeon 2: $r=-0.40$, Pigeon 3: $r=-0.46$, ****: $P<0.0001$), whereas Non-Animal patches show weak, non-significant correlations (Pigeon 1: $r=-0.06$, Pigeon 2: $r=-0.08$, Pigeon 3: $r=-0.10$, ns: Not significant).

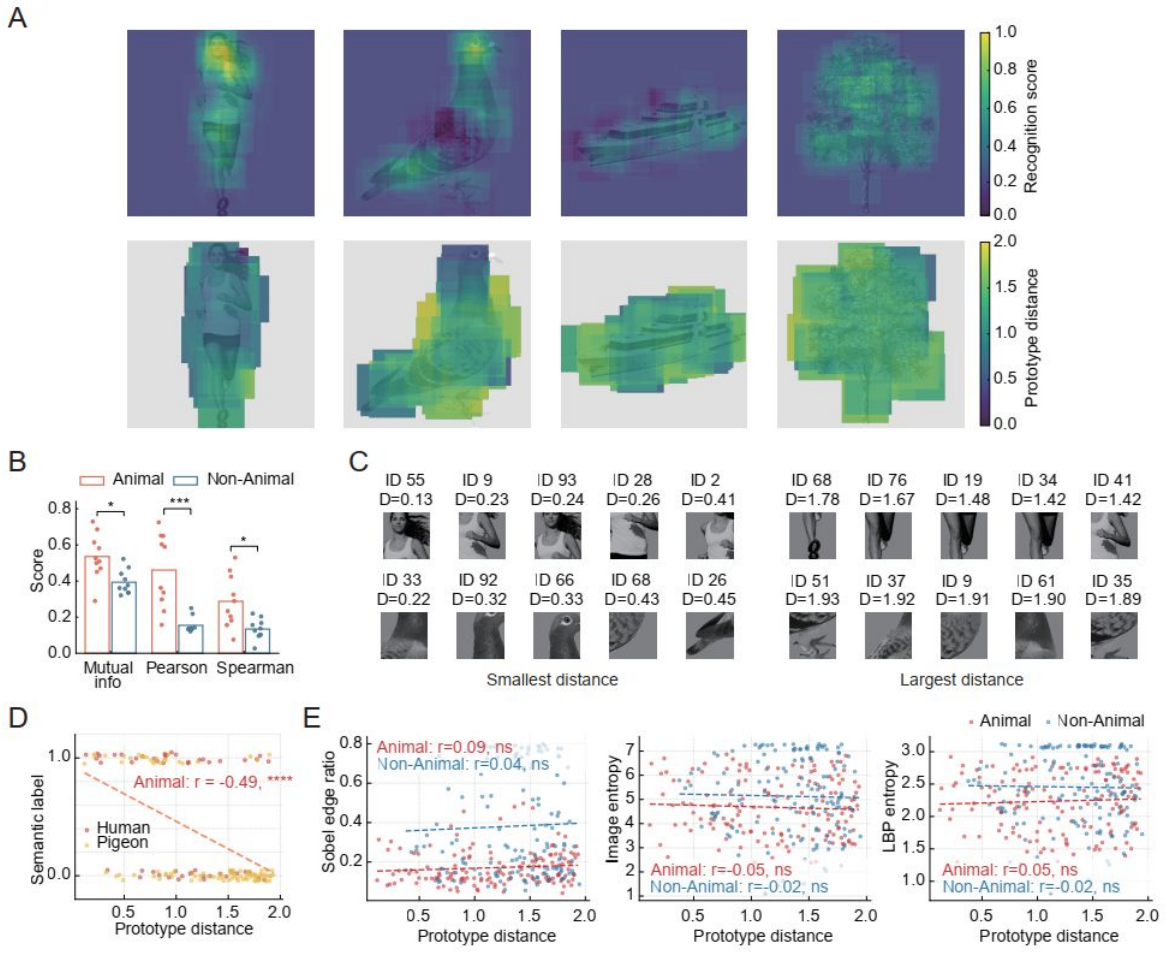
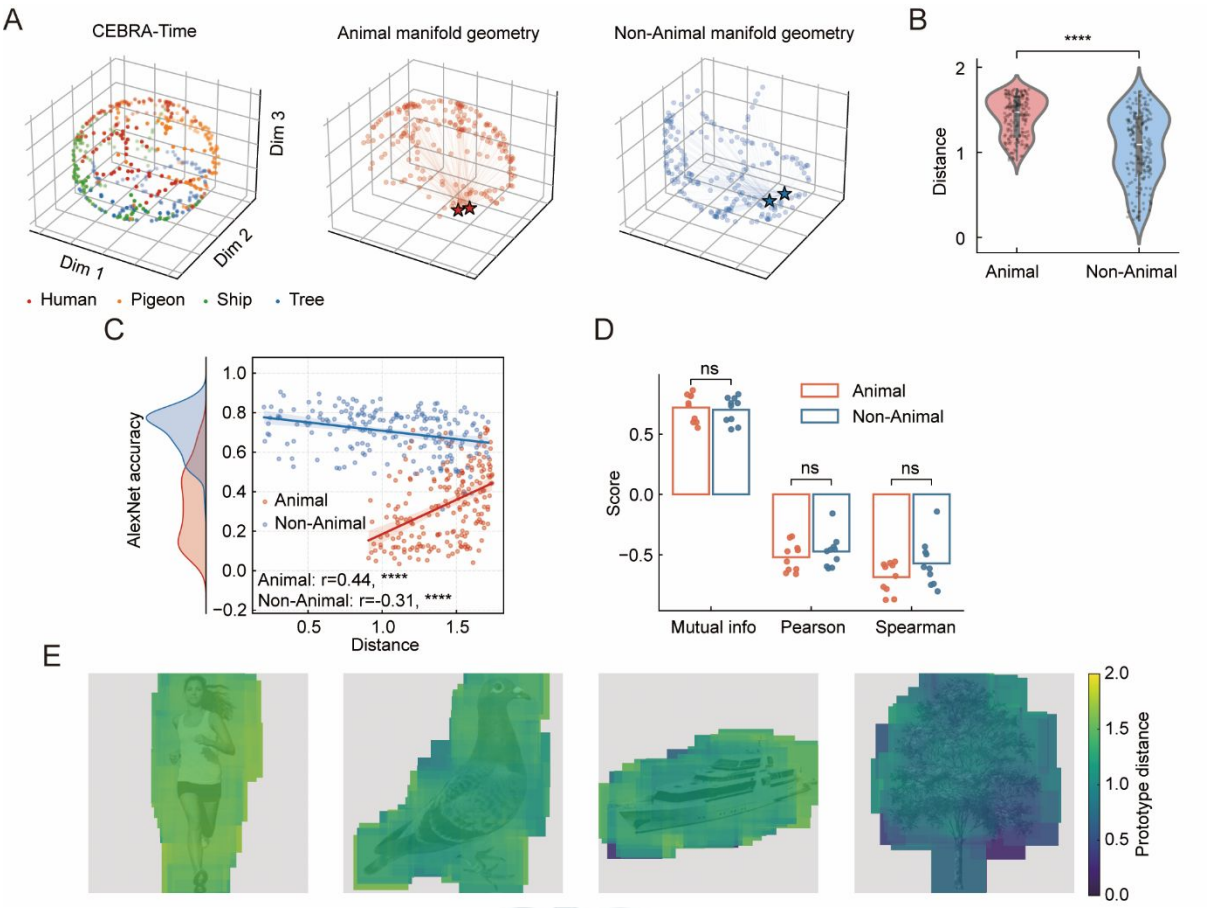


Figure 3 Linking prototype distance to local recognition scores and controlling for semantic and low-level visual factors. A: Example visualizations of recognition score and prototype distance mapped back onto the corresponding whole images (columns: Human, Pigeon, Ship, Tree). Top: recognition-score heatmaps (color bar 0–1). Bottom: spatial maps of prototype distance (color bar 0–2). **B:** The association between prototype distance and recognition score was quantified using mutual information, Pearson correlation, and Spearman correlation, and compared between Animal and Non-Animal patches. Bars summarize group statistics and dots indicate individual results; Wilcoxon rank-sum test, *: $P<0.05$, ***: $P<0.001$. **C:** Representative local patches with the smallest versus largest prototype distances (patch IDs and distances D are labeled), illustrating the content differences between near-prototype and far-from-prototype fragments. **D:** Relationship between semantic annotation (binary semantic label) and prototype distance: for Animal patches, the semantic label indicates whether a patch contains diagnostic parts such as the head/face, and it shows a significant negative correlation with prototype distance (Pearson $r=-0.49$, ****: $P<0.0001$). **E:** Low-level feature controls: prototype distance was tested against Sobel edge ratio, image entropy, and LBP entropy. No significant correlations were observed for either Animal or Non-Animal patches (see reported coefficients and ns labels, ns: Not significant).



741

742 **Figure 4 Prototype geometry in AlexNet feature space and its links to model behavioral**

743 **performance. A:** Three-dimensional embedding of AlexNet feature representations using

744 CEBRA-time, with points colored by the four basic categories (Human, Pigeon, Ship, Tree).

745 Manifold geometry is summarized for Animal and Non-Animal superordinate groups; star

746 markers denote category prototypes defined from whole-image features. **B:** Distributions of

747 prototype distance, showing a significant difference between Animal and Non-Animal

748 conditions; Wilcoxon rank-sum test, ****: $P<0.0001$. **C:** Relationship between prototype

749 distance and AlexNet patch-level classification accuracy: Animal patches show a significant

750 positive correlation ($r=0.44$, ****: $P<0.0001$), whereas Non-Animal patches show a significant

751 negative correlation ($r=-0.31$, ****: $P<0.0001$); the left panel shows the marginal distribution

752 of accuracy. **D:** The association between prototype distance and recognition score (Score) was

quantified using mutual information, Pearson's correlation, and Spearman's correlation, and compared between Animal and Non-Animal patches; correlations are overall negative, while the Animal vs Non-Animal difference is not significant (Wilcoxon rank-sum test, ns: Not significant). Dots indicate individual results. **E:** Spatial visualization of AlexNet prototype distance by mapping patch-level distances back onto the corresponding whole images for each category; colors indicate prototype distance (color bar at right).

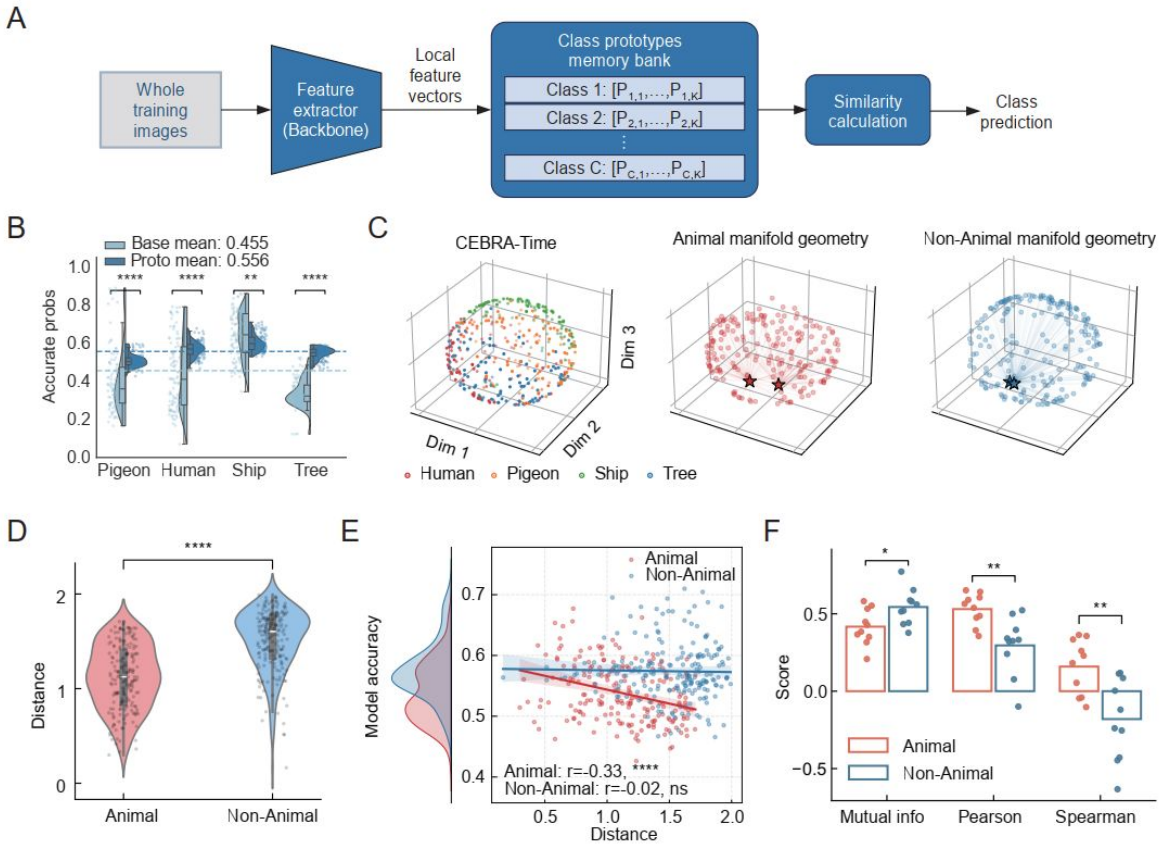


Figure 5 Architecture, performance gains, and restored prototype geometry in the neurally constrained network. A: Schematic of the local prototype learning model. Whole training images are processed by a feature extractor (backbone) to obtain local feature vectors, which are matched to a class-prototype memory bank; class prediction is produced from the

765 similarity scores. **B:** Classification performance: accurate prediction probability of CORnet vs
766 CORnet with local prototype learning across four basic categories (Human, Pigeon, Ship, Tree)
767 with significance annotations; Wilcoxon rank-sum test, **: $P<0.01$, ****: $P<0.0001$. Dashed
768 lines indicate mean accuracy (Base mean = 0.455; Proto mean = 0.556). **C:** Restored prototype
769 geometry: network representations are embedded into 3D using CEBRA-time, showing
770 category-wise distributions and the manifold geometry summarized for Animal and Non-
771 Animal groups; stars denote category prototypes defined from whole-image responses. **D:**
772 Prototype-distance distributions: distances from local patches to their corresponding category
773 prototypes differ significantly between Animal and Non-Animal conditions; Wilcoxon rank-
774 sum test, ****: $P<0.0001$. **E:** Prototype distance predicts model performance: prototype distance
775 is significantly negatively correlated with patch-level model accuracy (Pearson correlation;
776 Animal: $r=-0.33$, ****: $P<0.0001$; Non-Animal: $r=-0.02$, ns: Not significant), with marginal
777 distributions of accuracy shown. **F:** Quantifying association strength: mutual information,
778 Pearson correlation, and Spearman correlation are used to quantify the association between
779 model prototype distance and pigeon behavioral accuracy, comparing Animal vs Non-Animal
780 patches; Wilcoxon rank-sum test, ns: Not significant; *: $P<0.05$, **: $P<0.01$.
781

REFERENCES

Posner MI, Keele SW. 1968. On the genesis of abstract ideas. *Journal of Experimental Psychology* **77**(3):353–363.

Rosch E. 1975. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General* **104**(3):192–233.

Nosofsky RM. 1992. Similarity scaling and cognitive process models. *Annual Review of Psychology* **43**(1):25–53.

Leopold DA, O'Toole AJ, Vetter T, et al. 2001. Prototype-referenced shape encoding revealed by high-level aftereffects. *Nature Neuroscience* **4**(1):89–94.

Kahn DA, Harris AM, Wolk DA, et al. 2010. Temporally distinct neural coding of perceptual similarity and prototype bias. *Journal of Vision* **10**(10):12.

Iordan MC, Greene MR, Beck DM, et al. 2016. Typicality sharpens category representations in object-selective cortex. *NeuroImage* **134**:170–179.

Orlov T, Zohary E. 2018. Object representations in human visual cortex formed through temporal integration of partial shape information. Journal of Neuroscience, 38(3): 659−678.

Tang H, Schrimpf M, Lotter W, Moerman C, Paredes A, Ortega Caro J, et al. 2018. Recurrent computations for visual pattern completion. Proceedings of the National Academy of Sciences of the United States of America, 115(35): 8835−8840.

Rajaei K, Mohsenzadeh Y, Ebrahimpour R, Khaligh-Razavi SM. 2019. Beyond core object recognition: Recurrent processes account for object recognition under occlusion. PLoS

802 Computational Biology, 15(5): e1007001: 1−30.

803 Kar K, Kubilius J, Schmidt K, Issa EB, DiCarlo JJ. 2019. Evidence that recurrent circuits are
804 critical to the ventral stream's execution of core object recognition behavior. Nature
805 Neuroscience, 22(6): 974−983.

806 Geirhos R, Rubisch P, Michaelis C, Bethge M, Wichmann FA, Brendel W. 2019. ImageNet-
807 trained CNNs are biased towards texture; increasing shape bias improves accuracy and
808 robustness. In: International Conference on Learning Representations (ICLR 2019). 1−22.

809 Kortylewski A, He J, Liu Q, Yuille AL. 2020. Compositional convolutional neural networks:
810 A deep architecture with innate robustness to partial occlusion. In: Proceedings of the
811 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020).
812 18939−18948.

813 Yamins DLK, DiCarlo JJ. 2016. Using goal-driven deep learning models to understand sensory
814 cortex. Nature Neuroscience, 19: 356−365.

815 Schrimpf M, Kubilius J, Lee MJ, Murty NAR, Ajemian R, DiCarlo JJ. 2020. Integrative
816 benchmarking to advance neurally mechanistic models of human intelligence. Neuron, 108(3):
817 413−423.

818 Maniquet T, Op de Beeck H, Costantino AI. 2025. Recurrent issues with deep neural network
819 models of visual recognition. Scientific Reports, 15: 36344: 1−17.

820 Tang H, Yuan C, Li Z. 2022. Learning attention-guided pyramidal features for few-shot fine-
821 grained recognition. *Pattern Recognition*, 130: 108792. DOI: 10.1016/j.patcog.2022.108792.

Tang H, Li Z, Peng Z, Tang J. 2020. BlockMix: Meta regularization and self-calibrated inference for metric-based meta-learning. In: *Proceedings of the 28th ACM International Conference on Multimedia*. 610–618. DOI: 10.1145/3394171.3413884.

Tang H, Liu J, Yan S, Yan R, Li Z, Tang J. 2023. M3Net: Multi-view encoding, matching, and fusion for few-shot fine-grained action recognition. In: *Proceedings of the 31st ACM International Conference on Multimedia*. 1719–1728. DOI: 10.1145/3581783.3612221.

Tang H, He S, Qin J. 2025. Connecting giants: Synergistic knowledge transfer of large multimodal models for few-shot learning. In: *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)*. 6227–6235. DOI: 10.24963/ijcai.2025/693.

Zhou T, Wang W. 2024. Prototype-based semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10): 6858–6872. DOI: 10.1109/TPAMI.2024.3387116.

Wang C, Liu Y, Chen Y, Liu F, Tian Y, McCarthy DJ, Frazer H, Carneiro G. 2023. Learning support and trivial prototypes for interpretable image classification. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2062–2072. DOI: 10.1109/ICCV51070.2023.00197.

Wasserman EA, Turner BM, Güntürkün O. 2024. The pigeon as a model of complex visual processing and category learning. *Neuroscience Insights*, 19: 1−5.

Clark W, Colombo M. 2022. Seeing the forest for the trees, and the ground below my beak:

global and local processing in the pigeon's visual system. *Frontiers in Psychology*, 13: 888528: 1−7.

Qadri MAJ, Cook RG. 2015. Experimental divergences in the visual cognition of birds and mammals. *Comparative Cognition and Behavior Reviews* **10**:73–105.

Azizi AH, Pusch R, Koenen C, et al. 2019. Emerging category representation in the visual forebrain hierarchy of pigeons (*Columba livia*). *Behavioural Brain Research* **356**:423–434.

Anderson C, Parra RS, Chapman H, et al. 2020. Pigeon nidopallium caudolaterale, entopallium, and mesopallium ventrolaterale neural responses during categorisation of Monet and Picasso paintings. *Scientific Reports* **10**:15971.

Clark W, Chilcott M, Azizi A, et al. 2022. Neurons in the pigeon visual network discriminate between faces, scrambled faces, and sine grating images. *Scientific Reports* **12**:589.

Karten HJ, Hodos W. 1967. A stereotaxic atlas of the brain of the pigeon (*Columba livia*). Baltimore: The Johns Hopkins Press.

Güntürkün O, Hahmann U. 1994. Visual acuity and hemispheric asymmetries in pigeons. *Behavioural Brain Research* **60**(2):171–175.

Güntürkün O, Diekamp B, Manns M, et al. 2000. Asymmetry pays: visual lateralization improves discrimination success in pigeons. *Current Biology* **10**(17):1079–1081.

Wang ZH, Zhu JC, Zhu MJ, et al. 2025. Multidimensional feature encoding and functional organization in the pigeon entopallium. *Zoological Research*. DOI:10.24272/j.issn.2095-8137.2025.217.

862 Pachitariu M, Steinmetz NA, Kadir SN, et al. 2016. Fast and accurate spike sorting of high-
863 channel count probes with KiloSort. In: *Advances in Neural Information Processing Systems*.
864 **29**:4448–4456.

865 Rossant C, Kadir SN, Goodman DFM, et al. 2016. Spike sorting for large, dense electrode
866 arrays. *Nature Neuroscience* **19**(4):634–641.

867 Ditz HM, Fechner J, Nieder A. 2022. Cell-type specific pallial circuits shape categorical tuning
868 responses in the crow telencephalon. *Communications Biology* **5**(1):269.

869 Trainito C, von Nicolai C, Miller EK, et al. 2019. Extracellular spike waveform dissociates
870 four functionally distinct cell classes in primate cortex. *Current Biology* **29**(18):2973–2982.e5.

871 Majaj NJ, Hong H, Solomon EA, et al. 2015. Simple learned weighted sums of inferior
872 temporal neuronal firing rates accurately predict human core object recognition performance.
873 *Journal of Neuroscience* **35**(39):13402–13418.

874 Schneider S, Lee JH, Mathis MW, et al. 2023. Learnable latent embeddings for joint
875 behavioural and neural analysis. *Nature* **617**(7960):360–368.

876 Krizhevsky A, Sutskever I, Hinton GE. 2012. ImageNet classification with deep convolutional
877 neural networks. In: *Advances in Neural Information Processing Systems*. **25**:1097–1105. 2012.

878 He K, Zhang X, Ren S, Sun J. 2016. Deep residual learning for image recognition. In:
879 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR
880 2016). 770−778.

881 Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S. 2022. A ConvNet for the 2020s. In:

882 Proceedings - 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition
883 (CVPR 2022). 11966−11976.

884 Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. 2021. An
885 image is worth 16×16 words: Transformers for image recognition at scale.
886 arXiv:2010.11929v2. 1−22.

887 Kubilius J, Schrimpf M, Kar K, Hong H, Majaj NJ, Issa EB, et al. 2019. Brain-like object
888 recognition with high-performing shallow recurrent ANNs. arXiv:1909.06161v2. 1−17.

889 Yamins DLK, Hong H, Cadieu CF, et al. 2014. Performance-optimized hierarchical models
890 predict neural responses in higher visual cortex. *Proceedings of the National Academy of*
891 *Sciences of the United States of America* **111**(23):8619–8624. 2014.

892 Khaligh-Razavi SM, Kriegeskorte N. 2014. Deep supervised, but not unsupervised, models
893 may explain IT cortical representation. *PLoS Computational Biology* **10**(11):e1003915.

894 Güçlü U, van Gerven MAJ. 2015. Deep neural networks reveal a gradient in the complexity of
895 neural representations across the ventral stream. *The Journal of Neuroscience: The Official*
896 *Journal of the Society for Neuroscience* **35**(27):10005–10014.

897 Rose O, Johnson J, Wang B, et al. 2021. Visual prototypes in the ventral stream are attuned to
898 complexity and gaze behavior. *Nature Communications* **12**:6723.

899 Chen C, Li O, Tao C, Barnett AJ, Su J, Rudin C. 2019. This looks like that: Deep learning for
900 interpretable image recognition. arXiv:1806.10574v5: 1−12.

901 Schrimpf M, Kubilius J, Hong H, Majaj NJ, Rajalingham R, Issa EB, et al. 2018. Brain-Score:

902 Which artificial neural network for object recognition is most brain-like? *bioRxiv*, 1−9.

903 Nonaka S, Majima K, Aoki SC, et al. 2021. Brain hierarchy score: which deep neural networks
904 are hierarchically brain-like? *iScience* **24**(9):103013.

905 Jacob G, Pramod RT, Katti H, et al. 2021. Qualitative similarities and differences in visual
906 object representations between brains and deep networks. *Nature Communications* **12**(1):1872.

907 Kriegeskorte N, Mur M, Bandettini P. 2008. Representational similarity analysis: connecting
908 the branches of systems neuroscience. *Frontiers in Systems Neuroscience* **2**:4.

909 Nili H, Wingfield C, Walther A, et al. 2014. A toolbox for representational similarity analysis.
910 *PLoS Computational Biology* **10**(4):e1003553.

911 Devereux BJ, Clarke A, Marouchos A, et al. 2013. Representational similarity analysis reveals
912 commonalities and differences in the semantic processing of words and objects. *The Journal*
913 *of Neuroscience: The Official Journal of the Society for Neuroscience* **33**(48):18906–18916.

914 Cichy RM, Khosla A, Pantazis D, et al. 2016. Comparison of deep neural networks to spatio-
915 temporal cortical dynamics of human visual object recognition reveals hierarchical
916 correspondence. *Scientific Reports* **6**:27755.

917 Kriegeskorte N. 2015. Deep neural networks: a new framework for modeling biological vision
918 and brain information processing. *Annual Review of Vision Science* **1**:417–446.

919 Pusch R, Clark W, Rose J, et al. 2023. Visual categories and concepts in the avian brain. *Animal*
920 *Cognition* **26**:153–173.

921 Soto FA, Wasserman EA. 2012. Visual object categorization in birds and primates: integrating

922 behavioral, neurobiological, and computational evidence within a “general process”
923 framework. *Cognitive, Affective, & Behavioral Neuroscience*, **12**(1): 220–240.

924 Niu X, Jiang Z, Peng Y, et al. 2022. Visual cognition of birds and its underlying neural
925 mechanism: a review. *Avian Research*, **13**(2): 100023.

926 Baker N, Lu H, Erlikhman G, et al. 2018. Deep convolutional networks do not classify based
927 on global object shape. *PLoS Computational Biology* **14**(12):e1006613.

928 Baker N, Elder JH. 2022. Deep learning models fail to capture the configural nature of human
929 shape perception. *iScience* **25**(9):104913.