UNCOVERING THE NEURAL REPRESENTATION OF MULTIPLE DIMENSIONS OF OBJECT CATEGORIZATION IN HUMAN VISUAL CORTEX

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Marius Cătălin Iordan
May 2016

This dissertation is online at: http://purl.stanford.edu/fq245zt7119

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Fei-Fei Li, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Stefano Ermon

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Kalanit Grill-Spector

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Diane Beck

Approved for the Stanford University Committee on Graduate Studies.

Patricia J. Gumport, Vice Provost for Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format. An original signed hard copy of the signature page is on file in University Archives.

Abstract

We rely on vision more than on any other sensory modality to interact with and make sense of the world. Our behavior and culture, as well as the data we generate all rely strongly on visual information to index and capture salient relationships in the world. Within this realm, categorization is a fundamental building block of our visual experience. It is due to this marvelous generalization process that we take the problem of perceiving and understanding trillions of entities in our world (objects and scenes) and reduce it to a more manageable magnitude by binning virtually everything we see into a few tens of thousands of categories. Thus, it becomes a fundamental problem in understanding human vision to elucidate the mechanisms by which our visual cortices extract such complex information from a noisy sea of colored dots encoded by our retinas when we look out into the world.

But what represents a 'good' category and why do these distinctions emerge the way they do? Cognitively, useful distinctions between groups of items simultaneously maximize within-category similarity and between-category dissimilarity. The underlying hypothesis behind the work we put forward in this dissertation is that this key idea of similarity maximization also extends to the instantiation of neural patterns of representation in visual cortex. To this end, we use computational approaches in the context of several functional neuroimaging (fMRI) experiments to explore how behaviorally pervasive dimensions of object categorization, such as hierarchical organization and typicality, are represented in the brain and how they help us build a coherent picture of the world. Finally, we propose and test a model of neural object category processing based on the hypothesis that the cognitive utility of category structure partly drives information processing in visual cortex.

Für Elise

Acknowledgements

Well, this has been an incredible journey, mostly due to all the wonderful people I've had the privilege to get to know and spend time with here. First and foremost, I want to thank the the Stanford Vision Lab and the Illinois Attention and Perception Lab – thank you for being the wonderful people you all are.

Among them, I want to single out my collaborators and especially thank Michelle Renee Greene and Christopher Anthony Baldassano for their invaluable help and support. Michelle has been a wonderful mentor and close collaborator on virtually everything I did in grad school. She is awesome beyond words. Chris is awesome, too – he's an amazing person, a great researcher, and one of the strongest and most collected people I have ever had the privilege of calling my friend. I also want to thank Armand Joulin and Clara Fannjiang, not just for their invaluable contributions to our research, but also for providing perspective and support throughout our time together in the lab – they're great researchers and even greater people.

To all my friends and proxy mentors from UC Berkeley, a great thank you is due for your support, advice, and for welcoming me into your community – the list is too large to write down here, but I especially want to thank Michael Silver, David Whitney, Allison Yamanashi Leib, and the entire Vision Science department.

Thank you to my reading and oral examination committee members, Kalanit Grill-Spector, Stefano Ermon, and Russell Poldrack for being great role models both in your constructive scientific comments and your friendly, welcoming personalities.

Many thanks are due to the entire Stanford Vision community, as well. I feel very fortunate to have been part of it throughout my time here – thank you for the great teaching, advice, and practice talk feedback throughout the years! Among them,

special thanks should go Kalanit Grill-Spector and Brian Wandell. Every time I had a conversation with either of them, it felt like I was leveling up in my understanding of how human vision works and how to think like a true scientist – for that I will be eternally grateful.

Many, many thanks are also due to Diane Beck and Fei-Fei Li, my indefatigable guides through the graduate school labyrinth. I am truly honored and privileged to have had the both of you as my advisors. I am very grateful to you for everything you have done for me throughout the past six years, for your mentorship, for your support, for putting up with me and indulging me even though I sometimes didn't deserve it. Thank you for helping me become a scientist.

I would like to profusely thank my family for supporting everything I do, especially my mom, Monica Iordan, who throughout my childhood encouraged me to ask questions all the time and to treat finding an answer to them as a challenge worthy of my pursuits. I would not be here today had it not been for their sacrifices, support, and unconditional love.

And, of course, there's one more person I have to thank. Someone who has been my partner in life for a decade, although to be honest, I don't really remember what it's like to not share my life with her. My truly better half, Elise Ann Piazza. I can't really say that going through grad school didn't rob me of most of my sanity, but the only reason I have any left is because of her love and support. Thank you, Elise. I love you!

Contents

List of Tables

List of Figures

Chapter 1

Introduction

We rely on vision more than any other sensory modality to interact with and make sense of the world. Fortunately, our visual world is highly structured: most objects have clear boundaries; they appear in places we expect them to, and the relationships between them obey the rules of intuitive physics we learned to predict. In a conceptual sense, we make heavy use of that fact in our understanding of what we see, yet still about a third of our brains (or half for non-human primates) is recruited in service of vision, either directly or indirectly [52]. Even so, it's unclear how our brain extracts information about the visual world or even what the building blocks of this representation are across visual brain areas.

But there is hope: our ability to probe the inner workings of neural circuits and of large-scale neural representations has steadily and rapidly grown in the last few decades, especially because of the key insight of merging the discovery power of state-of-the-art statistical learning algorithms with the sensitivity of modern functional neuroimaging techniques (e.g. fMRI). It is this coupling that gave us the ability to reliably predict what a person is seeing [58] (or dreaming about! [64]) based solely on their neural activity patterns. This trend has accelerated with advances in both fields and will continue to do so. It's easy to imagine, due in part to this trajectory, that in our lifetimes we will see unprecedented access to the basic functions and computations that make us able to see, that enable us to interpret the world and link it to our past knowledge of it, that make us human.

We are, nevertheless, still early on the path towards this goal and there are many unanswered questions about how we extract and organize information about the world. Fortunately, our visual system does a great job in leveraging the correlational structure of the world to build a coherent picture of what our environment. It is due to this marvelous generalization process that we take the problem of perceiving and understanding trillions of distinct objects in our world and reduce it to a more manageable magnitude by binning virtually everything we see into a few tens of thousands of categories [11]. This phenomenon renders categorization as a fundamental building block of our visual experience, one which takes visually distinct entities and groups them together according to their many shared characteristics and affordances.

Although great strides have been made in understanding and describing the complexity of our object category structure, the neural underpinnings of many of its aspects remain elusive. For example, when asked about an object in their environment, people often use a mid-level of generality to describe it (e.g. dog), although other equally valid labels exist, both more general and more specific (e.g. object, animal, mammal, collie, Mr. Woof) [116]. More interestingly, even within their category, not all dogs are created equal: most people would agree that a Golden Retriever is more representative of the concept 'dog' than a Chihuahua [115]. However, it is unclear how these pervasive aspects of category are represented in visual cortex and how they arise as a consequence of computations performed in the brain.

Addressing these problems will be the main focus of the work we describe in this dissertation. Going forward, we show the first neural evidence that preferentially extracting information at the mid-level of generality (e.g. dog) may be an emergent property of the human visual system and, moreover, that such categorization may be part of visual processing from its very early stages (Chapter 2). Similarly, our work shows that everyday typicality judgments are correlated with neural distance between categories in object-selective regions of our brains (Chapter 3). These findings provide the first glimpse into the neural underpinnings of processes we've known about and built cognitive models for over the course of forty years, but have until now remained elusive neurally. Yet, this is not an endpoint, but a stepping stone into a rich space of questions that can help us understand how fluid our representation of the world is and

how our brains adapt their processing to ever-changing task demands. Consequently, we build upon insights provided by these findings to put forward a model of object category processing in human visual cortex based on the hypothesis that cognitive utility aspects of our category structure drive successive computations across the ventral visual stream (Chapter 4).

Moreover, our brains solve visual recognition through the interplay of computational, representational, and physical levels of interpretation of input from our eyes [92]. Concurrently with investigating the mechanisms of object perception, we also developed tools that help us better understand this key relationship between the function of neural circuits and their position on the cortical surface, a relationship that is currently not well understood (Chapter 5).

Finally, we gather together the key findings of our work in Chapter 6 and discuss their implications for uncovering how information is organized and how it flows in a structured manner hroughout human visual cortex.

Chapter 2

Basic Level Category Structure Emerges Gradually Across Human Occipito-Temporal Cortex

We begin our investigation by searching for the neural underpinnings of one of the most pervasive phenomenons of perceptual categorization: Although objects can be simultaneously categorized at multiple levels of specificity ranging from very broad ("natural object") to very distinct ("Mr. Woof"), it is a mid-level of generality (basic level: "dog") that often provides the most cognitively useful distinction between categories. Indeed, most objects are identified and named faster at the basic level [71, 90, 98, 99, 116, 123, 125], basic-level category names are the first learned by children[6, 14, 65, 94, 116], are used nearly exclusively when people freely name an object[116], tend to be shorter and more frequently used in language [14, 98, 116], and at least some basic-level categories seem to be culturally universal [10, 114]. Thus, a preponderance of evidence suggests that this basic-level advantage captures something fundamental about human perceptual categorization. Yet surprisingly, it is unknown how it (or, more broadly, the hierarchical representation of object categories) is achieved in the brain.

To address this question, we used multi-voxel pattern analyses to examine how

well each taxonomic level (superordinate, basic, and subordinate) of real-world object categories is represented across human occipito-temporal cortex. We found that although in early visual cortex objects are best represented at the subordinate level (an effect mostly driven by low-level feature overlap between objects in the same category), this advantage diminishes compared to the basic level as we move up the visual hierarchy, disappearing in object-selective regions of occipito-temporal cortex (LOC). This pattern stems from a combined increase in within-category similarity (category cohesion) and between-category dissimilarity (category distinctiveness) of neural activity patterns at the basic level, relative to both subordinate and superordinate levels, suggesting that successive visual areas may be optimizing basic level representations. This chapter is joint work with Michelle R. Greene, Diane M. Beck, and Fei-Fei Li, and was previously published as [67].

2.1 Introduction

Humans can distinguish between thousands of object categories in the real world with impressive speed and accuracy. Understanding how the brain represents categories across visual cortex is a key step in elucidating the complex cognitive mechanism by which categorization is achieved.

The mapping of category information across human visual cortex has been a major effort of modern neuroimaging studies, uncovering specific cortical regions specialized for broad stimulus categories such as faces, scenes, objects, and bodies [35, 40, 72, 91], as well as organizational principles corresponding to broad attribute dimensions, including animacy [19, 23, 76, 81] and real-world object size [76, 77]. Furthermore, many studies have demonstrated that category information is recoverable from distributed representations [25, 38, 58, 61, 66]. However, most previous studies have glossed over a fundamental property of real-world categories: specifically, any particular object may belong to multiple categories simultaneously, ranging from very broad ("natural object", "animal") to very distinct ("pug", "Mr. Woof"). Indeed, it is yet unknown how this hierarchical representation is achieved in the brain.

We thus focus our investigation on assessing how category representations at different taxonomic levels (subordinate, basic, superordinate) change over the span of the human ventral visual cortex. While under certain conditions, category levels are flexible and may change with context, typicality, and degree of expertise [71, 89, 125], most often human observers categorize objects faster and more accurately at a mid-level of specificity (i.e. basic level [6, 14, 65, 90, 94, 98, 116]. Thus, in our work, we restricted our analysis to sets of categories where these three levels are clearly differentiated behaviorally.

Concurrently, in characterizing the neural representation of this category hierarchy, we were inspired by Rosch et al.'s [116] seminal work on categorization, which argued that a good category simultaneously maximizes within-category similarity (cohesiveness) and between-category dissimilarity (distinctiveness). In our work, we applied this principle to multi-voxel fMRI patterns as a concrete measure of the strength of category representations across visual cortex. In particular, we ran two functional imaging (fMRI) experiments in which participants were shown objects from hierarchies comprising three behaviorally normed taxonomic levels (superordinate, basic, and subordinate) and we employed several multi-voxel pattern analyses (MVPA) to characterize the similarity and dissimilarity of activity patterns across these separate levels. Here, high cohesion (positive correlation between activity patterns) would indicate that information content is similar within that particular category. Similarly, high distinctiveness (zero or negative correlation between activity patterns) would indicate that categories are distinguishable from one another; therefore categories are more separable in that space.

In visual cortex, because two subordinate level exemplars (e.g. two pugs) should be most visually similar to each other, one might predict that categories adhere most strongly to a subordinate level representation. On the other hand, one might also expect superior superordinate level adherence (e.g. natural vs. man-made objects) since these categories might best reflect organization at the coarse scale of fMRI voxels. Finally, a wealth of behavioral evidence points to a mid-level of generality (basic level: "dog") as being privileged in providing the most cognitively useful distinction between categories: objects are learned, and recognized faster at this intermediate

level than at all other levels [6, 14, 65, 71, 90, 94, 98, 99, 116, 125]. This suggests that we may see evidence for superior basic level representations in visual cortex.

Thus, a sub-goal of our work is to ask whether a particular behaviorally relevant taxonomic level is better represented in visual cortex. Here, we show that for a set of object categories that exhibit a clear basic level advantage, category representations change as a function of taxonomic level as we move up the visual cortical hierarchy, progressively favoring the basic level relative to other levels of specificity. Thus, although objects are best represented at the subordinate level in early visual cortex, the basic level matches the quality of this representation in high-level object-selective regions, as well as dominates superordinate representations throughout visual cortex. This provides evidence that basic level structure may be an emergent property of the human visual system.

2.2 Materials and Methods

2.2.1 Experiment 1: Two Superordinate Categories - Natural and Man-Made

To investigate how categories are represented across multiple levels of specificity, we ran a functional imaging experiment in which participants were shown objects from a three-tiered taxonomy (superordinate, basic, and subordinate levels) and we employed several multi-voxel pattern analyses (MVPA) to characterize the similarity of activity patterns across these separate levels. To verify that our putative taxonomic levels are representative of real-world category organization, we first ran two behavioral experiments that assess the perceptual and semantic differences in recognizing and categorizing objects across these taxonomic levels.

2.2.1.1 Stimuli

We constructed a three-tiered taxonomic hierarchy comprising 2 superordinate level (natural, man-made), 4 basic level (dog, flower, plane, shoe), and 32 subordinate level categories. These included 8 breeds of dogs: Komondor, Chihuahua, Pug, Malamute,

Mastiff, Schnauzer, Welsh Corgi, Schipperke; 8 types of planes: airliner, biplane, fighter, delta plane, stealth, glider, gyroplane, seaplane; 8 types of flowers: blue daisy, ice poppy, sunflower, orchid, chrysanthemum, cosmos, violet, toadflax; 8 types of shoes: slippers, cowboy boots, running shoes, pumps, loafers, flip-flops, clogs, cleats. We had 32 instances of each of our 32 subordinate level categories for a total of 1,024 color photographs collected from the ImageNet online database [32]. Photos were tightly cropped in a square region around the object of interest, resized to 400 x 400 pixels, and included their natural background (Fig. 2.1 A).

2.2.1.2 Behavioral Experiment: Match-to-Category Verification

Participants

Twelve participants (7 female, ages 18–30, including one of the authors) took part in the experiment. All subjects had normal or corrected-to-normal vision, were financially compensated, and provided informed written consent in compliance with procedures approved by the Stanford University Institutional Review Board.

Materials

Stimuli were presented on a 21-inch CRT monitor, approximately 30 cm away from the observer. Images were shown centrally, subtending 16 x 16 degrees of visual angle. The experiment was implemented in MATLAB (http://www.mathworks.com), using the Psychophysics toolbox extension [12, 106].

Experimental Procedure

Each observer viewed 1,024 images for 200 ms each, followed by a category query term. Query terms matched the image's category on half of the trials, and were drawn from a random other category on the other half of trials. Query terms were drawn equally from superordinate level (e.g. "natural" or "man-made"), basic level ("plane", "dog", "flower", "shoe"), or subordinate level category (e.g. "Chihuahua" or "Chrysanthemum"). Participants were instructed to respond as quickly and accurately as possible as to whether the query term matched the image they had just

seen. Performance feedback (accuracy and RT) was displayed at the end of each trial. Immediately before the experimental trials, participants were shown example images of each of the 32 subordinate categories, along with each of the three valid category labels affixed to that category.

Data Analysis

Reaction times less than 200 ms and greater than 2 s were discarded from analysis (1% of data, no more than 5% from any one participant). One participant was discarded due to high numbers of rejected trials (46%) and errors (37%). Reaction times were transformed into z-scores. To test for a basic level advantage, we examined the reaction times for verifying an image as a member of a superordinate-, basic- or subordinate level category, both overall and for each basic level category in particular. We also computed a measure of basic level advantage for each of the 32 subordinate level categories, defined as the reaction time difference (in z-scores) of basic level categorization compared to subordinate and superordinate level categorization.

2.2.1.3 Behavioral Experiment: Same-Different Categorization

Participants

Twelve participants (5 female, ages 18–30) took part in the experiment. All subjects had normal or corrected-to-normal vision, were financially compensated, and provided informed written consent in compliance with procedures approved by the Stanford University Institutional Review Board. One participant also took part in the fMRI experiment.

Materials

Stimuli were presented on a 21-inch CRT monitor, approximately 30 cm away from the observer. Images were shown side by side, each subtending 16 x 16 degrees of visual angle, with 3 degrees between them. The experiment was implemented in MATLAB (http://www.mathworks.com), using the Psychophysics toolbox extension [12, 106].

Experimental Procedure

Each observer viewed 1,024 trials, with 512 trials showing pairs of images drawn from the same subordinate level category, and 512 trials showing image pairs from two different subordinate level categories (16 pairs per subordinate per taxonomic level, randomly drawn for each participant). Participants were instructed to respond as quickly and accurately as possible whether both images were from the same subordinate category. Image pairs remained on the screen until response, and reaction time and accuracy feedback were given after each response.

Data Analysis

Reaction times less than 200 ms and greater than 2 s were discarded from analysis (2% of data, no more than 11% from any one participant). Reaction times were transformed into z-scores relative to each participant's mean RT. We computed the average time required to reject a pair of images as being from the same subordinate level category and used this as a category distance measure in the context of a classical multi-dimensional scaling analysis (criterion: metric stress).

2.2.1.4 fMRI Experiment

Participants

10 volunteers (2 females, ages 23–28, including author M.C.I.) with no past history of psychiatric or neurological disorders and normal or corrected-to-normal vision participated in this experiment. Participants gave informed written consent in compliance with procedures approved by the Stanford University Institutional Review Board. Except for the participating author, all subjects received financial compensation.

Scanning Parameters and Preprocessing

Imaging data were acquired with a 3 Tesla G.E. Healthcare scanner. A gradient echo, echo-planar sequence was used to obtain functional images (volume repetition time (TR), 2 s; echo time (TE), 30 ms; flip angle, 80 degrees; matrix 128 x 128 voxels; FOV 20 cm; 29 oblique 3 mm slices with 1 mm gap; in-plane resolution, 1.56 x 1.56 mm).

We also collected a high-resolution (1 x 1 x 1 mm voxels) structural scan (SPGR; TR 5.9 ms; TE 2.0 ms; flip angle 11 degrees) in each scanning session. The functional data were spatially aligned to compensate for motion during acquisition and each voxel's intensity was converted to percent signal change relative to the temporal mean of that voxel using the AFNI software package [26]. To perform our analyses, we computed the average voxel activity for each block. We did not use a GLM analysis and did not perform any smoothing.

Experimental Procedure

Images were presented centrally subtending 21 x 21 degrees of visual angle and were superimposed on an equiluminant gray background. We used a back-projection system (Optoma Corporation) operating at a resolution of 1024 x 768 pixels at 75 Hz. Participants performed 8 runs, with 16 blocks per run and 8 images per block. Each block consisted of a 500 ms fixation cross presented centrally, followed by 8 consecutive stimulus presentations from the same subordinate level category, with a 12 s gap between the blocks. Each image was presented for 160 ms, followed by a 590 ms blank gray screen. Subjects were asked to maintain fixation at the center of the screen, and respond via button-press whenever an image was repeated (one-back task, 0–2 repetitions per block). Over the course of the experiment, each participant viewed 4 blocks from each of the 32 subordinate level categories, for a total of 128 blocks. The order of blocks, the number of repetitions in each block, and the images in each block were counter-balanced across runs and between subjects.

Regions of Interest (ROIs)

The positions and extents of each participant's functional ROIs (LOC, TOS, PPA, RSC, and FFA) brain were obtained using standard localizer runs conducted in a separate fMRI session. For functional ROIs, subjects observed two runs, each with 12 blocks drawn equally from six categories: child faces, adult faces, indoor scenes, outdoor scenes, objects (abstract sculptures with no semantic meaning), and phase-scrambled objects. Blocks were separated by 12 s fixation cross periods and comprised 12 image presentations, each of which consisted of images presented for 900 ms,

followed by a 100 ms fixation cross. Each image was presented exactly once, with the exception of two images during each block that were repeated twice in a row. Subjects were asked to maintain fixation at the center of the screen and respond via button press whenever an image was repeated. To avoid any issues related to intrinsic variability in signal reliability across our participant pool, we selected fixed-volume ROIs across all our participants. The volume of each region in mm^3 was chosen conservatively, based on sizes previously reported in the literature, accounting for resolution differences between studies [51, 133]: LOC: 500 voxels; TOS: 200 voxels; PPA: 300 voxels; RSC: 200 voxels; FFA: 100 voxels. LOC was defined as the top 500 voxels bilaterally near the inferior occipital gyrus that responded to an Objects > Scrambled Objects GLM contrast. PPA was defined as the top 300 voxels bilaterally near the parahippocampal gyrus that responded to a Scenes > Objects GLM contrast. TOS was defined as the top 200 voxels bilaterally near the trans-occipital sulcus that responded to a Scenes > Objects GLM contrast. RSC was defined as the top 200 voxels bilaterally near retrosplenial cortex that responded to a Scenes > Objects GLM contrast. FFA was defined as the top 100 voxels bilaterally near the fusiform gyrus that responded to a Faces > Objects GLM contrast. All ROIs were identified bilaterally, except for some participants' FFA (RH only: 3/10 for Experiment 1; 5/17 for Experiment 2).

To determine the locations of early visual areas V1, V2, V3v, and hV4, we used a standard retinotopic mapping protocol in a separate experiment, in which a checkerboard pattern undergoing contrast reversals at 5 Hz moved through the visual field in discrete increments [119]. First, a wedge subtending an angle of 45 degrees from fixation was presented at 16 different polar angles for 2.4 s each. Next, an annulus subtending 3 degrees of visual angle was presented at 15 different radii for 2.4 s each. Each subject passively observed two runs of 6 cycles in each condition, yielding 512 timepoints per subject. The locations and extents of early visual areas were delineated on a flattened cortical surface for each subject, using a horizontal vs. vertical meridian general linear test, which gave the boundaries between retinotopic maps.

We aligned the positions of the ROIs to the experimental sessions using the AFNI software package [26], by first aligning the structural scans between sessions with

sub-millimeter precision, and then applying the alignment transformation to the ROI positions. Percent signal change was then extracted for each voxel in each ROI and these vectors were submitted to the similarity and classification analyses described next.

2.2.1.5 fMRI Data Analysis

Within-Category Similarity (Cohesion) and Between-Category Similarity (Distinctiveness)

These analyses are defined identically to quantities used in [81]: Cohesion is within-category similarity; Distinctiveness is between-category dissimilarity. For each category at each taxonomic level (subordinate, basic, superordinate), we computed category cohesion as the average correlation between neural patterns elicited by within-category pairs of blocks (4 per subordinate category, 32 per basic category, 64 per superordinate category) at that taxonomic level. For example, at the basic level, cohesion for "dogs" is defined as the average correlation between voxel activations for any two blocks where any type of dog was shown. Similarly, we computed category distinctiveness as the average correlation between neural patterns elicited by between-category pairs of blocks at each taxonomic level. For example, at the basic level, distinctiveness for "dogs" is defined as average correlation between voxel activations for a block where dogs where shown and another block where, for example, flowers where shown. We performed each of these analyses for each subject and ROI separately. To show that the effects we obtain are not solely due to low-level image features, we also computed cohesion and distinctiveness in an analogous fashion for image descriptor features extracted from our stimulus images: color histograms, GIST [103], HOG [27], SIFT [87].

Category Boundary Effect

To quantify the interplay between cohesion and distinctiveness and how they give rise to category distinctions, we also defined the category boundary effect identically to [81] as the difference between cohesiveness and distinctiveness across a taxonomic

level, averaged across all categories from that level. This quantity provides a measure of how well categories are separated at each taxonomic level. For each ROI, we also compute category boundary effect differences between the basic level versus the subordinate and superordinate level representations. These analyses were also repeated for the image descriptor feature representations of our stimuli.

Correlation Classifier

To assess the amount of information present in the neural patterns at each taxonomic level, we implemented a standard MVPA correlation classifier to predict stimulus categories from neural patterns of activation at all three levels in our taxonomy (subordinate, basic, and superordinate). For each participant, we performed cross-validation by using 2 out of 8 runs for testing (1 block from each subordinate category) and the remaining 6 runs for training (3 blocks from each subordinate category). We averaged the results across cross-validation folds to obtain classification accuracies for each participant and ROI. To compare classification results between different taxonomic levels, we normalized the decoding accuracy using the formula (x ? c)/(1 ? c), where x is the accuracy obtained at a given level and c is the chance value (c is 12.5% for the subordinate level, 25% for the basic level, and 50 for the superordinate level). To control for the number of training examples at the basic- and superordinate levels, we matched the number of training and testing points to those at the subordinate level (3 blocks for training, 1 block for testing) by randomly sampling blocks 1,000 times with replacement. We performed one-tailed t-tests to identify results that were significantly different from chance levels (defined above) and two-tailed t-tests within each area to identify when decoding accuracy at the basic level is significantly greater than accuracy at the other levels in the taxonomy. We also obtained ROI confusion matrices by extracting subordinate level confusion matrices for each subject and averaging them together. A row of a confusion matrix records the probability of classifying the corresponding subordinate category as each of the 32 subordinate categories in the columns.

2.2.2 Experiment 2: Three Superordinate Categories - Vehicles, Furniture, and Musical Instruments

Experiment 1 used object categories that straddle the boundaries of two main dimensions of selectivity known to affect the responses to objects in occipito-temporal cortex: animacy [23] and real-world size [76]. Furthermore, by including naturalistic backgrounds together with our objects of interest in Experiment 1, it is possible that this factor may influence the observed category grouping. To ensure this is not the case, as well as to demonstrate the generalizability of our results from Experiment 1 to additional categories, we constructed a new three-tiered taxonomic hierarchy comprising exclusively big, inanimate objects. We first generated a putative taxonomy comprising 36 subordinate level categories and used a match-to-category behavioral experiment to eliminate 9 members with ambiguous category status (defined as weak basic level advantage over the subordinate). Similarly to Experiment 1, we then used a same-different subordinate categorization behavioral experiment to further verify that our new putative taxonomic levels are representative of real-world category organization. Finally, we conducted a second fMRI experiment using the new stimuli and replicated our analyses from Experiment 1.

2.2.2.1 Stimuli

We constructed a new three-tiered taxonomic hierarchy comprising exclusively big inanimate objects: 3 superordinate level categories (vehicles, furniture, musical instruments), 9 basic level categories (cars, airplanes, ships, chairs, beds, tables, drums, guitars, pianos), and 36 subordinate level categories (4 types of each of the 9 basic level categories listed above: cars ? sports car, sedan, antique car, station wagon; airplanes–airliner, biplane, fighter, stealth plane; ships–ice breaker, cargo ship, battleship, cruise ship; chairs–folding chair, armchair, straight chair, Eames chair; beds–canopy bed, sleigh bed, platform bed, bunk bed; tables–dining table, coffee table, pedestal table, folding table; drums–bass drum, snare drum, timpani, bongos; guitars–flamenco, Stratocaster, dreadnaught, Les Paul; pianos–grand piano, Hammond organ, upright piano, synthesizer). We had 40 instances of each of our 36 subordinate level

categories for a total of 1,440 color photographs collected from the ImageNet online database [32]. Images were cropped tightly around each object of interest and we replaced the original background with pixel-wise full-color 1/f noise. The resulting images were 400 x 400 pixels, ensuring that all images stimulated the same retinal area.

2.2.2.2 Behavioral Experiment: Match-to-Category Verification

The first aim of this behavioral experiment was to finalize our category taxonomy by assessing category status in general, and basic level advantage in particular. This ensured that categories we included in our taxonomy are representative of the relationships present in real-world taxonomies. We tested the category taxonomy listed above and then eliminated members with ambiguous category status. The taxonomy was pruned of 9 subordinate categories (one for each basic), by eliminating those subordinates with the lowest behavioral basic level advantage (see Data Analysis): antique car, stealth plane, battleship, Eames chair, bunk bed, folding table, bongos, Les Paul guitar, synthesizer). We used the resulting taxonomy (27 subordinate categories, Fig. 2.4 A) for all subsequent analyses.

Participants

Ten participants (6 female, ages 18–35, including authors M.C.I. and M.R.G.) participated in the first behavioral experiment. All volunteers had normal or corrected-to-normal vision, and provided informed consent in compliance with procedures approved the Stanford University Institutional Review Board. Non-author participants were compensated for their time.

Materials and Experimental Procedure

Analogous to Experiment 1.

Data Analysis

Reaction times less than 200 ms and greater than 2 s were discarded from analysis (2% of data, no more than 10% of trials from any participant). Reaction times

for correct trials (84% of trials) were transformed into z-scores. To test for a basic level advantage, we examined the differences in reaction times to correctly verify an image as a member of a superordinate-, basic- or subordinate level category. We also defined basic level advantage to be the reaction time difference (in z-scores) of basic level categorization compared with subordinate level categorization, and used this metric to reject the subordinate level categories in each branch of the hierarchy with the weakest basic level effects. Only one of the remaining 27 basic level categories (biplane) had a negative basic level advantage, possibly because this less-typical plane is better categorized at the subordinate level [71].

2.2.2.3 Behavioral Experiment: Same-Different Categorization

Participants

Twenty individuals (9 females, ages 18–35) with normal or corrected to normal vision participated in this experiment. None of the participants took part in the fMRI experiment or in the first behavioral experiment. All provided informed consent in compliance with procedures approved by the Stanford University Institutional Review Board and were compensated for their time.

Materials and Experimental Procedure

Analogous to Experiment 1.

Data Analysis

Reaction times less than 200 ms and greater than 2 s were discarded from analysis (¡1% of data, no more than 8% from any one participant). Reaction times were transformed into z-scores. We computed the average time required to reject a pair of images as being from the same subordinate level category and used this as a category distance measure in the context of a classical multi-dimensional scaling analysis (criterion: metric stress).

2.2.2.4 fMRI Experiment

Participants

17 volunteers (4 females, ages 23–31, including authors M.C.I. and M.R.G.) with no past history of psychiatric or neurological disorders and normal or corrected-to-normal vision participated in this experiment. Participants gave informed written consent in compliance with procedures approved by the Stanford University Institutional Review Board. Except for the participating authors, all subjects received financial compensation.

Scanning Parameters, Preprocessing, Experimental Procedure, and Regions of Interest (ROIs)

The second fMRI experiment was conducted similarly to Experiment 1. Participants performed 5 runs, with 27 blocks per run and 8 images per block. Over the course of the experiment, each participant viewed 5 blocks from each of the 27 subordinate level categories, for a total of 135 blocks. The order of blocks, the number of repetitions in each block, and the images in each block were counter-balanced across runs and between subjects.

2.2.2.5 fMRI Data Analysis

Within-Category Similarity (Cohesion) and Between-Category Dissimilarity (Distinctiveness), Category Boundary Effect, and Correlation Classifier

Performed analogously to Experiment 1.

2.2.3 Statistical Analyses

For all our experiments, we used paired two-tailed t-tests when comparing observed effects against chance and when establishing whether a significant difference exists between two observed effects. We used Kolmogorov-Smirnov tests to establish that no significant deviation from normality exists for the distributions of all effects to which t-tests were applied. Because statistical tests are made on a single number

derived from the pattern of voxels within an ROI per condition of interest, and these conditions are relatively few, we did not correct for multiple comparisons within our ROI analyses.

We also used Friedman non-parametric tests to investigate whether trends exist in data where the dependent variable is ordinal, but not continuously organized. All statistical tests were implemented in MATLAB.

2.3 Results

2.3.1 Experiment 1: Two Superordinate Categories–Natural and Man-Made

2.3.1.1 Behavioral Experiments

In our first experiment, we used a three-tiered taxonomic hierarchy comprising 2 superordinate level (natural, man-made), 4 basic level (dog, flower, plane, shoe), and 32 subordinate level categories (e.g. Chihuahua, stealth plane) (Fig. 2.1 A).

To verify that our putative basic level categories reflect entry-level concepts, we first conducted a delayed match-to-category behavioral experiment. As predicted, participants were significantly faster to verify category membership at the basic level (662 ms; s.e.m. 36 ms) than at the superordinate (747 ms; s.e.m. 43 ms) or subordinate levels (782 ms; s.e.m. 44 ms) (Fig. 1C; Basic > Superord. $t_1 8 = 5.1$, $p < 0.001$; Basic > Subord. $t_1 8 = 8.6$, $p < 0.001$). We also computed a measure of basic level advantage for each of the 32 subordinate level categories, defined as the reaction time difference (in z-scores) of basic level categorization compared to subordinate and superordinate level categorization. All 32 categories showed a basic level advantage over the superordinate level (Fig. 1D), and all except 3 categories (cowboy boots, clogs, and sunflowers) showed a basic level advantage over the subordinate level of the taxonomy (Fig. 1E). These few exceptions most likely represent less prototypical exemplars of their basic level category [71].

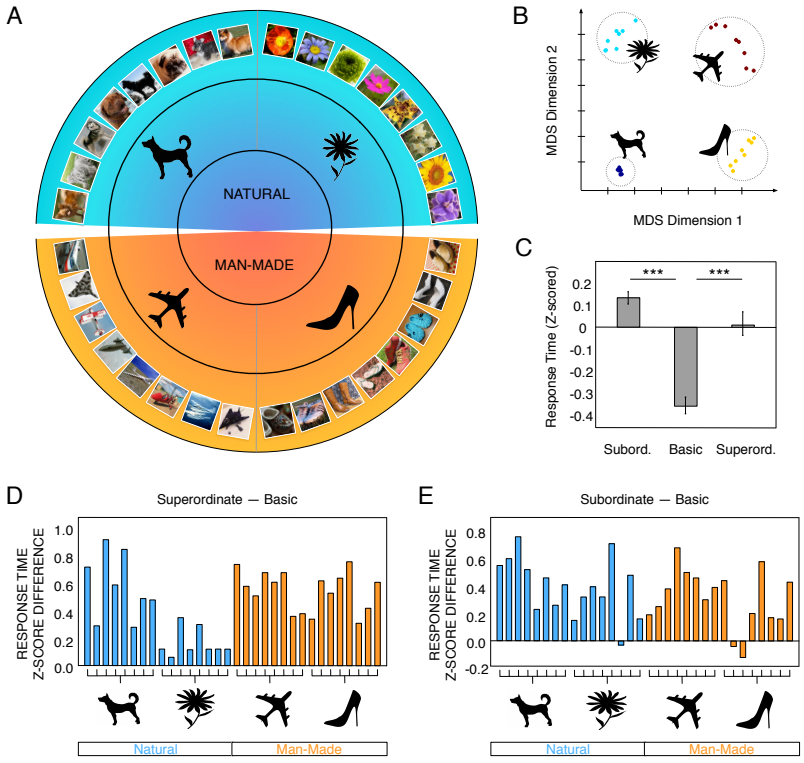


Figure 2.1: **Stimulus set and behavioral results for Experiment 1.** (A) The stimulus set was organized according to a three-level taxonomic hierarchy comprising 32 subordinate level (most specific, outside layer), four basic level (middle layer), and two superordinate level (most general, center) categories. Each subordinate category comprised 32 color photographs, with a representative image shown. (B) Same-different subordinate level categorization behavioral experiment. We applied classical MDS to the perceptual distance between subordinate categories measured as z-scored RTs. In a two-dimensional solution, the four basic level categories formed separate clusters. (C–E) Match-to-category behavioral experiment. (C) Participants verified category membership significantly faster at the basic level than at the superordinate or subordinate levels. (D–E) RT difference between basic and subordinate / superordinate categorization conditions. Positive values indicate basic level advantage. Participants identified all stimulus categories faster at the basic level than at the subordinate / superordinate level. There are only three exceptions: "sunflowers", "clogs", and "cowboy boots", perhaps reflecting the atypicality of these stimuli [71]. $***$ $p < .001$. Error bars: 95% confidence interval.

To map categories in terms of their behavioral similarity, we next used a same-different subordinate level categorization experiment to measure the perceptual distance between all pairs of subordinate categories. Reasoning that images in similar categories will take longer to reject than images from dissimilar categories, we used response times to pairs of images in the "different" condition to generate a distance metric between our subordinate categories. Consistent with prior work [116], participants found objects within the same basic category to be more similar to each other than to stimuli in other basic categories ($t_2 2 = 4.7$, $p < 0.001$). Furthermore, classical multidimensional scaling (MDS) applied to this distance metric revealed that in a two-dimensional solution the four basic level categories form separate clusters (Fig. 1B), with the first MDS dimension separating the natural and man-made categories (superordinate level).

These results replicated Rosch et al.'s [116] original findings for our object categories by demonstrating that our taxonomy exhibits a clear basic level advantage and as such is representative of hierarchically organized real-world categories.

2.3.1.2 Neural Category Boundaries Favor Basic Level Representations

Having verified the taxonomy behaviorally, we scanned participants viewing these same 32 categories to find out how neural category representations change across taxonomic levels and across human ventral visual cortex. Because task may influence entry-level categorization [57, 89, 90], we asked participants to perform a one-back repetition task in the scanner (i.e. no explicit categorization task) used solely to ensure they maintained attention and alertness during the experiment. Our analyses focused on object- (lateral occipital complex (LOC)), scene- (parahippocampal place area (PPA), retrosplenial cortex (RSC), trans-occipital sulcus (TOS)), and face-selective regions (fusiform face area (FFA)), as well as early visual cortex areas (V1, V2, V3v, hV4).

Our first task was to assess the strength of category representations at each taxonomic level in terms of their cohesion and distinctiveness. According to [116], categories form such that they concurrently maximize within-category similarity (cohesion) and between-category dissimilarity (distinctiveness). To quantify the interplay

between cohesion and distinctiveness and how they give rise to category distinctions, we defined the category boundary effect [81] as the difference between cohesiveness and distinctiveness across a taxonomic level, averaged across all categories from that level. We computed the category boundary effect for each taxonomic level (subordinate, basic, superordinate) in each brain region of interest.

We found that the category boundary effect is generally higher at the subordinate and basic levels compared to the superordinate level across visual cortex, especially in higher visual areas (Fig. 2A; Subordinate > Superordinate: V1: $t_9 = 2.5$, $p = 0.032$; V2 $t_9 = 2.5$, $p = 0.035$; V3v: $t_9 = 1.9$, $p = 0.089$; hV4: $t_9 = 1.7$, $p = 0.133$; LOC: $t_9 = 5.6$, $p < 0.001$; FFA: $t_9 = 4.0$, $p = 0.003$; PPA: $t_9 = 4.7$, $p = 0.001$; TOS: $t_9 = 2.1$, $p = 0.067$; RSC: $t_9 = 3.5$, $p = 0.007$; Basic > Superordinate: V1: $t_9 = 2.2$, $p = 0.058$; V2: $t_9 = 2.8$, $p = 0.022$; V3v: $t_9 = 2.8$, $p = 0.020$; hV4: $t_9 = 2.8$, $p = 0.021$; LOC: $t_9 = 7.6$, $p < 0.001$; FFA: $t_9 = 3.6$, $p = 0.006$; PPA: $t_9 = 6.5$, $p < 0.001$; TOS: $t_9 = 3.3$, $p = 0.009$; RSC: $t_9 = 4.8$, $p = 0.001$). Moreover, the category boundary effect increased in LOC compared to early visual areas at all levels of the taxonomy (Subordinate: LOC > V1: $t_9 = 5.7$, $p < 0.001$; LOC > V2: $t_9 = 7.3$, $p < 0.001$; LOC > V3v: $t_9 = 9.7$, $p < 0.001$; LOC > hV4: $t_9 = 5.7$, $p < 0.001$; Basic: LOC > V1: $t_9 = 6.5$, $p < 0.001$; LOC > V2: $t_9 = 7.3$, $p < 0.001$; LOC > V3v: $t_9 = 8.3$, $p < 0.001$; LOC > hV4: $t_9 = 5.5$, $p < 0.001$; Superordinate: LOC > V1: $t_9 = 4.7$, $p = 0.001$; LOC > V2: $t_9 = 5.4$, $p = 0.001$; LOC > V3v: $t_9 = 5.7$, $p = 0.001$; LOC > hV4: $t_9 = 2.5$, $p = 0.032$). Taken together, these results suggest that categories become more sharply distinguishable as we move up the visual hierarchy, and that throughout ventral visual cortex activity patterns adhere better to subordinate and basic level categories than to the more general (superordinate) levels of representation.

To characterize the difference between our taxonomic levels more clearly, we looked at the difference between the category boundary effect at the basic level compared to the other two levels (Figs. 2C–D). We found that category boundary is always higher at the basic level than the superordinate across early visual areas and LOC. Moreover, subordinate category boundary started out with advantage over the basic level (generally negative values for V1, Fig. 2C), but this advantage disappeared as we move up the visual cortical hierarchy (generally positive values for LOC, Fig. 2C).

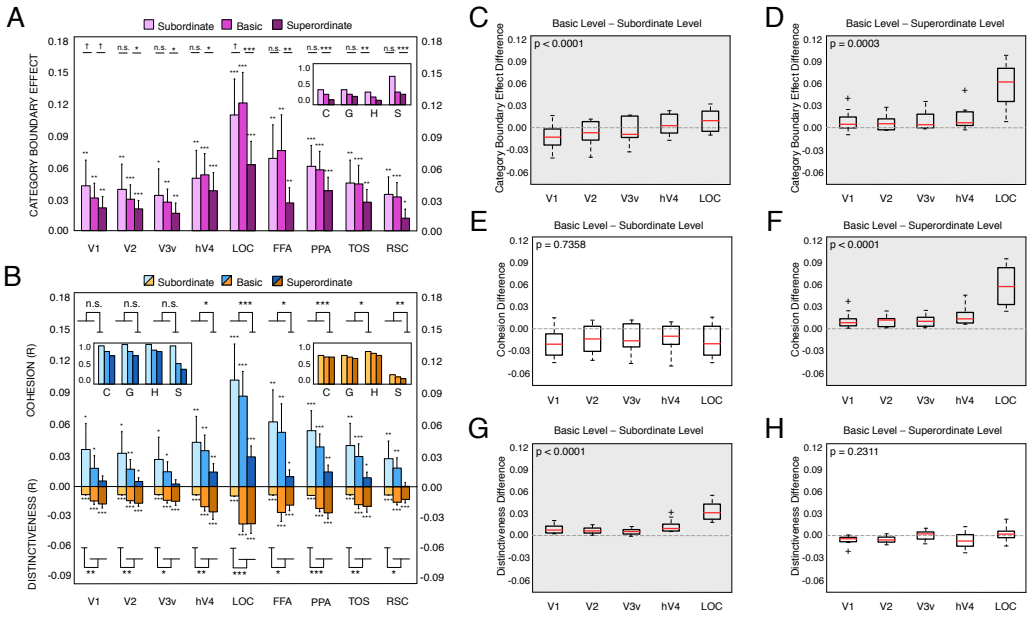


Figure 2.2: **Neural category boundaries favor basic level representations in Experiment 1.** (A) Category boundary effect for neural activity patterns at each taxonomic level and in each ROI. Inset shows same analysis for image feature descriptors: C = color histograms; G = GIST features; H = HOG features; S = SIFT features. The subordinate and basic levels were together strongly represented, with the former being especially emphasized in early visual cortex, whereas the latter becoming more prominent in LOC. (B) Cohesion and distinctiveness for neural activity patterns at each taxonomic level and in each ROI. Inset shows same analyses for image feature descriptors. (C–D) Category boundary effect difference between basic level and subordinate and superordinate levels. We uncovered a gradual trade-off between the subordinate and basic levels, which appeared to develop as we moved up the visual hierarchy, with a trending basic level advantage arising in object-selective cortex. (E-F) Cohesion difference between basic level and subordinate and superordinate levels. (G–H) Distinctiveness difference between basic level and subordinate and superordinate levels. The category boundary difference appears to be driven by separate components of the category boundary effect, depending on taxonomic level. $*$ $p < .05$, $**$ $p < .01$, $***$ $p < .001$, † $p < .10$, n.s. = not significant. Error bars: 95% confidence interval. Shaded graphs indicate a significant increase from V1 to LOC.

Interestingly, the basic level gained an advantage over both the subordinate and the superordinate levels as we move up the visual hierarchy from V1 to LOC (increasing trends in category boundary effect difference from V1 to LOC for: Basic - Subord. $p < 0.001$; Basic - Superord. $p < 0.001$; Friedman non-parametric tests).

Overall, our results suggest that the neural representation of object categories in occipito-temporal cortex is highly dynamic across the taxonomic spectrum. First, we find evidence supporting both our initial predictions: the subordinate and basic levels are both strongly represented, with the former being especially emphasized in early visual cortex, while the latter becoming more prominent in object-selective cortex LOC. Second, we uncover a gradual trade-off between the subordinate and basic levels, which appears to develop as we move up the visual hierarchy, with a basic level advantage arising in object-selective cortex.

2.3.1.3 Category Cohesion and Distinctiveness Across Occipito-Temporal Cortex

The category boundary effect provides an intuitive measure of how categories group at each taxonomic level. However, this effect comprises contributions from both cohesion and distinctiveness, which describe the similarity of object representations within and between categories, respectively. Historically, it has been hypothesized that basic categories provide the best behavioral differentiation between concepts because they combine the strengths, but not the weaknesses of both subordinate and superordinate categories [113]: members of subordinate categories, although very similar to each other (high cohesion), share too many features that overlap with members of other categories (low distinctiveness), and exemplars of superordinate categories, although very different from one another (high distinctiveness), share too few features in common with each other to successfully generalize across the entire category (low cohesion). To determine whether activity patterns in ventral visual cortex conform to this principle, we computed the average cohesion and distinctiveness of the activity patterns evoked by our stimuli for each taxonomic level.

We found that cohesion generally decreased with level of specificity across all ROIs (Fig. 2.2 B, top), and was significantly weaker at the superordinate level compared to

the other two levels in the taxonomy in all high-level areas and hV4 (Superordinate < Basic & Subordinate LOC: $t_9 = 7.3$, $p < 0.001$; FFA: $t_9 = 4.4$, $p = 0.005$; PPA: $t_9 = 6.9$, $p < 0.001$; TOS: $t_9 = 4.1$, $p = 0.009$; RSC: $t_9 = 4.8$, $p = 0.003$; hV4: $t_9 = 3.9$, $p = 0.011$). This result is consistent with the expectation that objects share more low-level features in common at the subordinate level [79, 81]. Concurrently, between-category dissimilarity (distinctiveness) generally increased with taxonomic level (Fig. 2.2 B, bottom), and was significantly weaker at the subordinate level compared to the basic and superordinate levels in all ROIs (Subordinate < Basic & Superordinate LOC: $t_9 = 7.7$, $p < 0.001$; FFA: $t_9 = 4.7$, $p = 0.004$; PPA: $t_9 = 6.6$, $p < 0.001$; TOS: $t_9 = 5.2$, $p = 0.002$; RSC: $t_9 = 3.9$, $p = 0.011$; V1: $t_9 = 4.7$, $p = 0.003$; V2: $t_9 = 5.9$, $p < 0.001$; V3v: $t_9 = 4.1$, $p = 0.008$; hV4: $t_9 = 5.7$, $p = 0.001$). In other words, these results are in general agreement with the assertion that the basic level may be privileged because it strikes the best balance between category cohesion and distinctiveness [113].

Although the general pattern of higher cohesion for subordinate and basic level categories and higher distinctiveness for basic and superordinate level categories held across our ROIs, the degree of cohesion and distinctiveness changed across visual areas. Interestingly, category cohesion increased in LOC compared to V1 at all levels of the taxonomy (LOC > V1: Subord. $t_9 = 5.8$ $p < 0.001$; Basic $t_9 = 6.5$, $p < 0.001$; Superord. $t_9 = 4.9$, $p < 0.001$), suggesting that object representations become overall more homogenous within their category in later visual areas. Furthermore, distinctiveness increased in LOC compared to V1 at all levels of the taxonomy (LOC > V1: Subord. $t_9 = 3.5$, $p = 0.006$; Basic $t_9 = 6.5$, $p < 0.001$; Superord. $t_9 = 4.5$, $p = 0.002$), which suggests that object representations become better differentiated across categories in later visual areas. Thus, in keeping with Rosch et al.'s [116] assertion that good object categories are represented such that they maximize within-category similarity and between-category dissimilarity, our results suggest that LOC appears to be producing stronger category representations than earlier visual areas.

Do these changes in cohesion and distinctiveness favor the basic level? To assess this we compared cohesion and distinctiveness across both taxonomic levels and visual areas (Fig. 2.2 E–H). The advantage of the basic level over the subordinate in

later visual areas was mainly due to the sharp increase in distinctiveness between early visual areas and LOC (Fig. 2.2 G, increasing trend in distinctiveness difference from V1 to LOC for Basic - Subord. $p < 0.001$; Friedman non-parametric test). Cohesion, on the other hand, was fairly stable across the same visual areas (Fig. 2.2 E, no increasing trend in cohesion difference from V1 to LOC for Basic - Subord. $p = 0.736$; Friedman non-parametric test). Conversely, the advantage of the basic level over the superordinate was mainly due to the sharp increase in cohesion between early visual areas and LOC (Fig. 2.2 F, increasing trend in cohesion difference from V1 to LOC for Basic - Superord. $p < 0.001$; Friedman non-parametric test), whereas distinctiveness remained relatively unchanged (Fig. 2.2 H, no increasing trend in distinctiveness difference from V1 to LOC for Basic - Superord. $p = 0.231$; Friedman non-parametric test). This pattern of results aligns well with both theoretical considerations of category, as well as intuitions about subordinate and superordinate categories. As predicted, we show that a trade-off exists between category cohesion and category distinctiveness at the two extremes of our taxonomy (subordinate and superordinate levels), with the basic level potentially striking the best balance between these two quantities by encompassing both strong within-category similarity and strong between-category dissimilarity. In short, our data suggests that the basic level simultaneously gains an advantage over both the subordinate and the superordinate levels as we move up the visual hierarchy from V1 to LOC.

2.3.1.4 The Contribution of Low-Level Visual Features

The changes in cohesion and distinctiveness across visual cortex suggest that LOC may be optimizing both of these two components of what constitutes a good category. To determine the extent to which the patterns of results obtained in LOC are captured by low-level image features, we computed category boundary effect, cohesion, and distinctiveness in an analogous fashion for image descriptor features extracted from our stimulus images: color histograms, GIST [103], HOG [27], SIFT [87].

We found that all image descriptor category boundaries clearly favored the subordinate level (Fig. 2.2 A, inset). As such, these boundaries were similar to early visual cortex representations, but they did not capture category representations in LOC.

By contrast, neural patterns in LOC exhibited a trend for reversing the preference of subordinate and basic levels, favoring the latter (Basic > Subordinate LOC: $t_9 = 2.0$, $p = 0.072$).

Furthermore, we found that for all our feature descriptors, cohesion has high positive values for all levels of the taxonomy. However, concomitantly, between-category similarity was also very high (Fig. 2.2 B, insets), indicating poor distinctiveness at the image descriptor level. In other words, a high degree of similarity exists between all our stimulus images in terms of their low-level features, irrespective of category, and across all levels of the taxonomy (i.e. distinctions between all categories are very slight). Thus, while image features may partly explain category cohesion, they do a poor job at characterizing the distinctiveness between object categories we observe in the neural data. This lack of distinctiveness makes low-level image features a poor candidate for explaining the results we obtained in LOC which show the basic level gaining an advantage compared to the other two levels in our taxonomy.

Our results are consistent with the predictions put forth by Rosch et al. [116] based on behavioral observations: object categories are represented such that they maximize within-category similarity and between-category dissimilarity. This property is not solely due to low-level image features, it holds across multiple levels of category generality (subordinate, basic, superordinate), and is, in fact, enhanced as we move up the ventral visual stream: cohesion and distinctiveness increase in object-selective areas compared to early visual cortex.

2.3.1.5 Correlation Classification Shows Basic Level Advantage in LOC

Our analyses so far suggest that in order to understand category organization in visual cortex, we must consider cohesion and distinctiveness together. Furthermore, the category boundary analysis used here and by others [81] assumes cohesion and distinctiveness combine linearly to give rise to category distinctions. This linearity assumption may not be strictly true, raising the possibility that we are underestimating (or overestimating) the degree to which the activity patterns adhere to a particular taxonomic level. Thus, to complement our category boundary analysis, we also used a data-driven method that weighs cohesion and distinctiveness automatically, without

any prior knowledge provided by the experimenters.

In particular, we implemented an MVPA correlation classifier to decode category identity from each ROI at each taxonomic level (subordinate, basic, superordinate). We found that the information present in voxel-level neural patterns was sufficient to distinguish between categories at all taxonomic levels and in all brain regions considered above-chance: object-, scene-, and face-selective areas (LOC, FFA, PPA, RSC, TOS), as well as early visual areas (V1, V2, V3v, hV4) (Fig. 2.3 A).

Critically, however, we also found that information about object category did not increase monotonically with category generality (taxonomic level) in all brain areas. In LOC (and to a lesser extent in FFA and RSC) accuracy was highest at the basic level and we saw a significant drop in decoding for both the subordinate and the superordinate levels, compared to the basic level (LOC: Basic > Subord. $t_9 = 11.1$, $p < 0.001$, Basic > Superord. $t_9 = 4.5$, $p = 0.002$; FFA: Basic > Subord. $t_9 = 4.1$, $p = 0.003$, Basic > Superord. $t_9 = 3.0$, $p = 0.014$; RSC: Basic > Subord. $t_9 = 3.9$, $p = 0.004$, Basic > Superord. $t_9 = 2.6$, $p = 0.028$). Moreover, we found that in all regions, when classification errors did occur, the confusions were more likely to be within the same basic level than between basic levels (i.e. breeds of dogs were commonly confused with other breeds of dogs, but not with types of flowers, shoes, or planes; Fig. 2.3 B), with the effect most salient in LOC (Within Basic Confusions > Between Basic Confusions: LOC: $t_9 = 15.9$, $p < 0.001$; TOS: $t_9 = 5.6$, $p < 0.001$; PPA: $t_9 = 5.9$, $p < 0.001$; RSC: $t_9 = 4.2$, $p = 0.002$; FFA: $t_9 = 4.1$, $p = 0.003$; V1: $t_9 = 3.1$, $p = 0.013$; V2: $t_9 = 4.5$, $p = 0.001$; V3v: $t_9 = 4.9$, $p < 0.001$; hV4: $t_9 = 6.3$, $p < 0.001$).

The trends observed in the correlation classifier decoding results suggest that basic level categories are more clearly delineated at the voxel population level in object-selective areas, compared to the other two levels in our taxonomy. This result provides a quantitative validation to the intuition provided by the category boundary analysis that the basic level represents an optimal level of specificity in object taxonomy in object-selective cortex.

Finally, the basic level is most distinguishable in LOC using the MVPA analysis, but not using the category boundary effect analysis. This finding suggests that MVPA

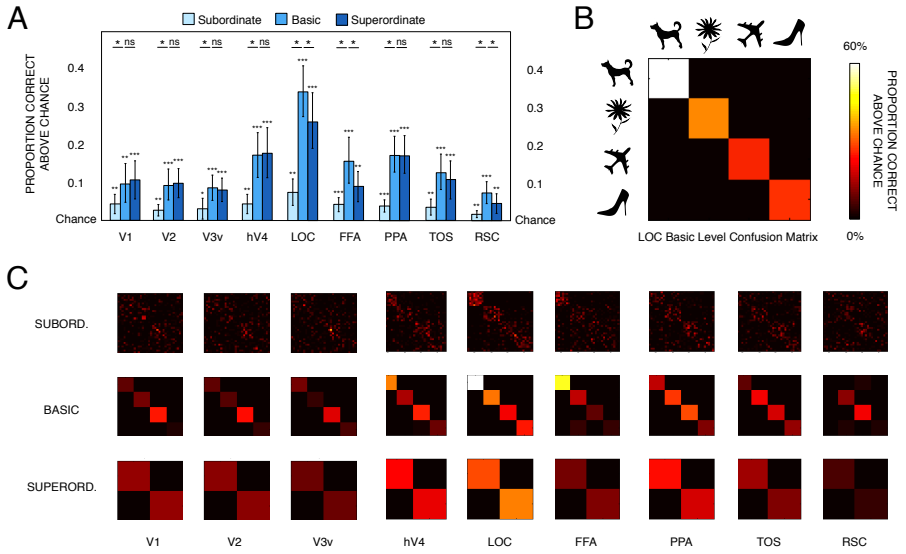


Figure 2.3: **MVPA classification reveals that object categories are most distinct at the basic level in LOC in Experiment 1.** (A) Proportion above chance of correct decoding responses for all levels of the taxonomy (chance is zero): subordinate, basic, and superordinate. Top insets denote whether differences between adjacent bars are significant. Category information was discernible significantly above chance at all taxonomic levels and in all ROIs, with higher visual areas generally showing larger values. Decoding at the basic level was easier than at the subordinate and superordinate levels in LOC, RSC, and FFA (shaded), but not in any of the other brain areas considered. (B) Confusion matrix example: LOC basic level classification. Basic categories were ordered on the axes according to the pictograms: dogs, flowers, planes, and shoes. At the subordinate level, within each basic category, the eight corresponding subordinates were listed alphabetically. At the superordinate level, the "natural object" category was listed first, and the "man-made object" category was listed second. (C) Confusion matrices for decoding analysis in A: top = subordinate level; middle = basic level; bottom = superordinate level. In all regions, when classification errors did occur, the confusions were more likely to be within the same basic level than between basic levels with the effect most salient in LOC. The basic level matrices show that confusions become more common within the basic level as we move up the visual hierarchy. $*$ $p < .05$, $**$ $p < .01$, $***$ $p < .001$, n.s. = not significant. Error bars: 95% confidence interval. SUBORD. = subordinate; SUPERORD. = superordinate.

did not weight cohesion and distinctiveness equally when assigning category labels to neural activations, and thus cohesion and distinctiveness might not contribute equally to generating category boundaries in LOC.

2.3.2 Three Superordinate Categories - Vehicles, Furniture, and Musical Instruments: Removing the Contribution of Real-World Size, Animacy, and Natural Backgrounds

2.3.2.1 Behavioral Experiments

The stimulus set used in Experiment 1 comprised categories which straddle the boundaries of two main dimensions of selectivity known to affect the responses to objects in occipito-temporal cortex: animacy and real-world size. To ensure that these dimensions have no effect on our results, as well as to demonstrate the generalizability of our results from Experiment 1 to additional categories, we constructed a new three-tiered taxonomic hierarchy comprising exclusively big and inanimate objects: 3 superordinate level categories (vehicles, furniture, musical instruments), 9 basic level categories (cars, airplanes, ships, chairs, beds, tables, drums, guitars, pianos), and 36 subordinate level categories (4 types of each of the 9 basic level categories). In addition, to ensure that our effects were driven by objects and not by their naturalistic backgrounds, we superimposed our new stimuli on meaningless 1/f noise backgrounds.

To finalize our category taxonomy, as well as to verify that our putative basic level categories reflect entry-level concepts, we conducted a delayed match-to-category behavioral experiment, similar to the one used in Experiment 1. Our strategy was to test our initial category taxonomy, and then eliminate members with ambiguous category status. To prune our taxonomy, we defined the basic level advantage to be the reaction time difference (in z-scores) of basic level categorization compared to subordinate level categorization (Fig. 2.4 E). We then used this metric to reject the subordinate with the weakest basic level advantage out of the four putative subordinate level categories in each basic (eliminating 9 subordinate categories total out of

the initial 36), resulting in 27 total subordinate level categories (Fig. 2.4 A). Only one of the remaining subordinates (biplane) had a negative basic level advantage, possibly because this less-typical plane is better categorized at the subordinate level [71].

To test for the strength of the basic level advantage in our pruned taxonomy, we examined the differences in reaction times to correctly verify an image as a member of a subordinate, basic or superordinate level category. We observed strong basic level effects overall (Fig. 2.4 C): participants were significantly faster to verify category membership at the basic level (566 ms; s.e.m 36 ms) than at the superordinate level (623 ms; s.e.m 39 ms, Basic > Superord. $t_1 8 = 6.8$, $p < 0.001$). Similarly, basic level categorization was faster than subordinate level categorization (618 ms, s.e.m 36 ms, Basic > Subord. $t_1 8 = 6.2$, $p < 0.001$).

To map categories in terms of their behavioral similarity and dissimilarity, we next used a same-different subordinate level categorization experiment to measure the perceptual distance between all pairs of subordinate categories. Reasoning that images in similar categories will take longer to reject than images from dissimilar categories, we used response times to pairs of images in the "different" condition to generate a distance metric between our subordinate categories. Consistent with prior work [116], participants found objects within the same basic category to be more similar to each other than to stimuli in other basic categories ($t_7 4 = 39.0$, $p < 0.001$). Furthermore, classical multidimensional scaling (MDS) applied to this distance metric revealed that in a two-dimensional projection each of the nine basic level categories are clearly separated from one another (Fig. 2.4 B).

Similarly to Experiment 1, these results replicated Rosch et al.'s [116] original findings for our new set of object categories by demonstrating that our second taxonomy also exhibits a clear basic level advantage after removing the contribution of animacy, image backgrounds, and real-world size as described by others [76, 77].

2.3.2.2 Neural Category Boundaries Equally Favor Subordinate and Basic Level Representations

We scanned participants viewing these same three superordinate level categories to assess how neural category representations change across taxonomic levels and across

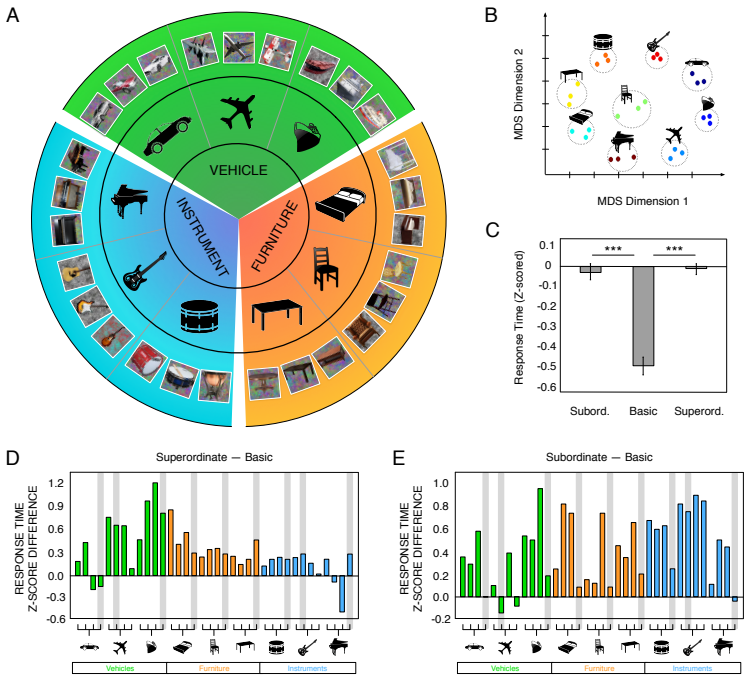


Figure 2.4: **Stimulus set and behavioral results for Experiment 2.** (A) The stimulus set was organized according to a three-level taxonomic hierarchy comprising 27 subordinate level (most specific, outside layer), nine basic level (middle layer), and three superordinate level (most general, center) categories. Each subordinate category consisted of 40 color photographs, with a representative image shown. (B) Same-different subordinate level categorization behavioral experiment. We applied classical MDS to the perceptual distance between subordinate categories measured as z scored RTs. In a two-dimensional solution, all nine basic level categories form separate clusters. (C–E) We used a match-to-category behavioral experiment to finalize our category taxonomy by assessing category status in general and basic level advantage in particular. We tested a larger category taxonomy (36 subordinate categories) and then eliminated members with ambiguous category status. (C) Participants verified category membership significantly faster at the basic level than at the superordinate or subordinate levels. (D–E) RT difference between basic and subordinate / superordinate categorization conditions. Positive values indicate basic level advantage. Participants identified almost all stimulus categories faster at the basic level than at the subordinate / superordinate level. We used this metric to reject the subordinate with the weakest such effect of the putative four subordinate level categories in each basic level (shaded categories were eliminated). $***$ $p < .001$. Error bars: 95% confidence interval. Subord. = subordinate; Superord. = superordinate.

human ventral visual cortex. As in Experiment 1, participants performed a one-back repetition task in the scanner (i.e. no explicit categorization task). Again, our analyses focused on object- (lateral occipital complex (LOC)), scene- (parahippocampal place area (PPA), retrosplenial cortex (RSC), trans-occipital sulcus (TOS)), and face-selective regions (fusiform face area FFA), as well as early visual cortex areas (V1, V2, V3v, hV4).

Our first task for the new taxonomy was to reassess the strength of category representations at each taxonomic level. As such, we computed the category boundary effect for each taxonomic level (subordinate, basic, superordinate) and each brain region of interest. We found that, as in Experiment 1, the category boundary effect was largest at the subordinate level in early visual areas, but this trend disappeared in higher visual areas compared to the basic level (Fig. 2.5 A; Subordinate > Basic: V1: $t_{16} = 5.7$, $p < 0.001$; V2: $t_{16} = 4.2$, $p < 0.001$; V3v: $t_{16} = 3.8$, $p = 0.002$; hV4: $t_{16} = 1.4$, $p = 0.186$; LOC: $t_{16} = 0.3$, $p = 0.754$; FFA: $t_{16} = 0.7$, $p = 0.525$; PPA: $t_{16} = 2.6$, $p = 0.018$; TOS: $t_{16} = 0.1$, $p = 0.920$; RSC: $t_{16} = 0.6$, $p = 0.583$; Subordinate > Superordinate: V1: $t_{16} = 3.5$, $p = 0.003$; V2: $t_{16} = 3.0$, $p = 0.008$; V3v: $t_{16} = 2.7$, $p = 0.016$; hV4: $t_{16} = 1.5$, $p = 0.161$; LOC: $t_{16} = 2.1$, $p = 0.053$; FFA: $t_{16} = 0.6$, $p = 0.533$; PPA: $t_{16} = 0.7$, $p = 0.501$; TOS: $t_{16} = 2.5$, $p = 0.022$; RSC: $t_{16} = 1.3$, $p = 0.203$). Moreover, the category boundary effect increased in LOC compared to V1 at all levels of the taxonomy (LOC > V1: Subord. $t_{16} = 2.9$, $p = 0.011$; Basic $t_{16} = 4.1$, $p < 0.001$; Superord. $t_{16} = 2.6$, $p = 0.021$). These results again suggest that categories become more sharply distinguishable as we move up the visual hierarchy, and furthermore, early visual areas appear to favor subordinate distinctions, while in later areas, this difference disappears between subordinate and basic levels.

This trend was, in fact, an enhanced version of our findings in Experiment 1: when comparing the difference between category boundaries at the basic level versus the other two levels, it became clear that the relative difference between the basic and subordinate levels decreased much more sharply and ultimately disappeared as we move up the visual hierarchy from V1 to LOC. Simultaneously, the difference between basic and superordinate levels strongly increased along the visual hierarchy (Fig. 2.5 C–D; increasing trends in category boundary effect difference from V1 to

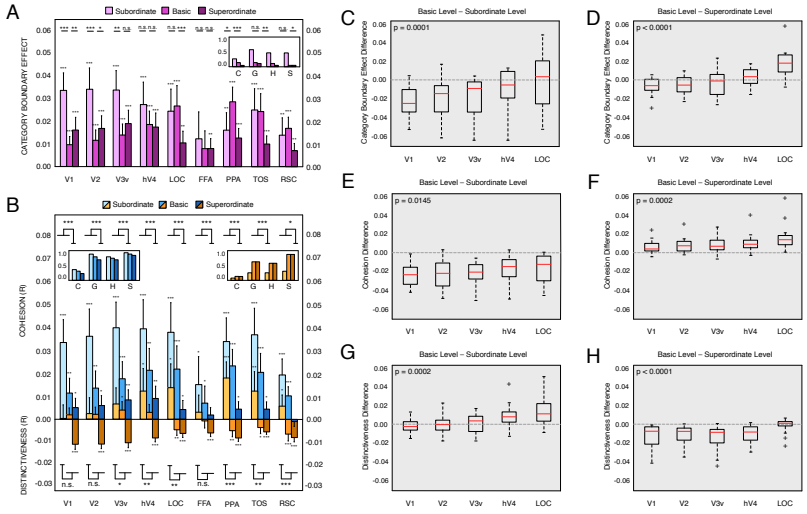


Figure 2.5: **After controlling for animacy, real-world size, and naturalistic backgrounds, neural category boundaries still show that basic level representations gain an increasing advantage as we move up the ventral visual stream.** (A) Category boundary effect for neural activity patterns at each taxonomic level and in each ROI. Inset shows same analysis for image feature descriptors: C = color histograms; G = GIST features; H = HOG features; S = SIFT features. Early visual areas favored subordinate distinctions, whereas, in later areas, this difference disappeared between subordinate and basic levels. (B) Cohesion and distinctiveness for neural activity patterns at each taxonomic level and in each ROI. Inset shows same analyses for image feature descriptors. Cohesion generally decreased with taxonomic level and was significantly weaker at the superordinate level compared to the other two levels in all ROIs. Distinctiveness generally increased with taxonomic level and was significantly weaker at the subordinate level compared to the basic and superordinate levels in all ROIs, except for FFA, V1, and V2. (C–D) Category boundary effect difference between basic level and subordinate and superordinate levels. We observed an enhanced version of our findings in Experiment 1: the basic level gains an advantage over both the subordinate and superordinate levels as we move up the visual hierarchy from V1 to LOC. (E–F) Cohesion difference between basic level and subordinate and superordinate levels. (G–H) Distinctiveness difference between basic level and subordinate and superordinate levels. In contrast to Experiment 1, the category boundary difference appears to be driven by both components of the category boundary effect. $*$ $p < .05$, $**$ $p < .01$, $***$ $p < .001$, n.s. = not significant. Error bars: 95% confidence interval. Shaded graphs indicate a significant increase from V1 to LOC.

LOC for Basic - Subord. $p < 0.001$; Basic - Superord. $p < 0.001$; Friedman non-parametric tests). Additionally, category boundary was generally higher at the basic level than the superordinate throughout visual cortex, and, interestingly, subordinate category boundary started out with advantage over the basic level (generally negative values for V1, Fig. 2.5 C), but this advantage disappeared as we moved up the visual cortical hierarchy (zero and slightly positive values for LOC, Fig. 2.5 C).

Overall, the second experiment confirms our initial results: the subordinate and basic levels are both strongly represented, with the former being especially emphasized in early visual cortex, while the latter gaining an equally strong representation in LOC. Moreover, we notice a gradual trade-off between the subordinate and basic levels, which becomes apparent as we move up the visual hierarchy.

2.3.2.3 Category Cohesion and Distinctiveness Across Occipito-Temporal Cortex

Next, we replicated the analyses that investigate each component of the category boundary effect separately (cohesion and distinctiveness). Again, we found that cohesion decreased with taxonomic level across all ROIs (Fig. 2.5 B, top), such that is was significantly weaker at the superordinate level compared to the other two levels in the taxonomy in all ROIs (Superordinate < Basic & Subordinate V1: $t_{16} = 6.9$, $p < 0.001$; V2: $t_{16} = 6.2$, $p < 0.001$; V3v: $t_{16} = 6.9$, $p < 0.001$; hV4: $t_{16} = 6.5$, $p < 0.001$; LOC: $t_{16} = 6.0$, $p < 0.001$; FFA: $t_{16} = 2.5$, $p = 0.026$; PPA: $t_{16} = 7.1$, $p < 0.001$; TOS: $t_{16} = 6.4$, $p < 0.001$; RSC: $t_{16} = 6.3$, $p < 0.001$). Concurrently, between-category dissimilarity (distinctiveness) generally increased with taxonomic level (Fig. 2.5 B, bottom), such that it was significantly weaker at the subordinate level compared to the basic and superordinate levels in all ROIs, except for FFA, V1, and V2 (Subordinate < Basic & Superordinate V1: $t_{16} = 1.5$, $p = 0.142$; V2: $t_{16} = 1.9$, $p = 0.080$; V3v: $t_{16} = 2.6$, $p = 0.018$; hV4: $t_{16} = 3.1$, $p = 0.007$; LOC: $t_{16} = 3.3$, $p = 0.005$; FFA: $t_{16} = 1.6$, $p = 0.130$; PPA: $t_{16} = 6.4$, $p < 0.001$; TOS: $t_{16} = 3.5$, $p = 0.003$; RSC: $t_{16} = 4.2$, $p < 0.001$). In other words, Experiment 2 confirms our initial findings: the ventral visual cortex is optimizing category representations and object categories are represented such that they maximize within-category similarity and

between-category dissimilarity, with the basic level striking the best balance between category cohesion and distinctiveness.

Furthermore, in contrast to Experiment 1 where cohesion increased in LOC compared to V1 at all levels of the taxonomy, in Experiment 2 we observed this effect only at the basic level, but not at the subordinate or superordinate levels (LOC > V1: Subord. $t_{16} = 1.2$, $p = 0.249$; Basic $t_{16} = 2.9$, $p = 0.012$; Superord. $t_{16} = 0.6$, $p = 0.551$), suggesting that object representations become overall more homogenous within their basic category in later visual areas. Note that this phenomenon cannot be explained by animacy, real-world size, or image backgrounds, as our stimulus set in this experiment did not vary across these factors. As in Experiment 1, distinctiveness increased in LOC compared to V1 at all levels of the taxonomy (LOC > V1: Subord. $t_{16} = 3.3$, $p = 0.005$; Basic $t_{16} = 5.7$, $p < 0.001$; Superord. $t_{16} = 4.1$, $p < 0.001$), which suggests that object representations become better differentiated across categories in later visual areas, regardless of taxonomic level. In other words, basic level category representations benefit from both increased cohesion and distinctiveness as we move up the ventral visual stream, whereas subordinate and superordinate categories only exhibit increased distinctiveness, suggesting a potential advantage for the basic level in higher visual areas.

To further investigate whether changes in cohesion and distinctiveness favor the basic level, we compared these quantities across both taxonomic level and visual areas (Fig. 2.5 E–H). The advantage of the basic level over the subordinate level in later visual areas was again mainly due to the sharp increase in distinctiveness between early visual areas and LOC (Fig. 2.5 G, Basic - Subord. distinctiveness increase from V1 to LOC: $p < 0.001$; Friedman non-parametric test). Interestingly, however, cohesion also exhibited a slight increase as we moved up the visual stream, albeit much less so than distinctiveness (Fig. 2.5 E, Basic - Subord. cohesion increase from V1 to LOC: $p = 0.015$; Friedman non-parametric test). This suggests that although the contribution of cohesion to the difference between subordinate and basic level representations is small, this component nonetheless exerts a quantifiable influence in the category representations we observed.

Whereas in Experiment 1 the advantage of the basic level over the superordinate

was mostly due to an increase in cohesion, here the same advantage was due to an increase in both cohesion and distinctiveness between early visual areas and LOC (Fig. 2.5 F,H; increasing trends in cohesion and distinctiveness difference from V1 to LOC for: Cohesion Basic - Superord. $p < 0.001$; Distinctiveness Basic - Superord. $p < 001$; Friedman non-parametric tests). It is possible that in Experiment 1 we were unable to detect this emerging distinctiveness advantage for the basic level over the superordinate because our "natural object" category included both animate and inanimate stimuli, which were overall more distinctive, and thus obscured a more subtle change between levels.

Overall, our results replicate our findings in Experiment 1 which show that a trade-off exists between category cohesion and category distinctiveness at the two extremes of our taxonomy (subordinate and superordinate levels), with the basic level potentially striking the best balance between these two quantities by encompassing both strong within-category similarity and strong between-category dissimilarity. In short, our data suggests that the basic level simultaneously gains an advantage over both the subordinate and the superordinate levels as we move up the visual hierarchy from V1 to LOC.

2.3.2.4 The Contribution of Low-Level Visual Features

Similarly to Experiment 1, we sought to show that the patterns of results obtained in LOC were not attributed to low-level image features. As such, we computed category boundary effect, cohesion, and distinctiveness in an analogous fashion for image descriptor features extracted from our stimulus images: color histograms, GIST [103], HOG [27], SIFT [87]. Here, we found an enhanced version of our findings from Experiment 1: descriptor category boundaries strongly favored the subordinate level thus more closely capturing early visual cortex representations (Fig. 2.5 A, inset). By contrast, neural patterns in LOC, TOS, PPA, and RSC exhibited a trend for reversing the preference of subordinate and basic levels, favoring the latter.

Furthermore, we once again found that for all our feature descriptors, both cohesion and distinctiveness had high positive values for all levels of the taxonomy. This implies a high degree of similarity exists between all our stimulus images in terms

of their low-level features, even among categories that were highly distinctive in our neural data. Thus, while image features may partly explain cohesion, they do a poor job at characterizing the distinctiveness between object categories we observe in the neural data.

2.3.2.5 Correlation Classification Shows Basic Level Advantage in LOC

Finally, as we did in Experiment 1, we used a more data driven approach to assess category boundaries by implementing an MVPA correlation classifier to decode category identity from each ROI at each taxonomic level (subordinate, basic, superordinate). We found that the information present in voxel-level neural patterns was sufficient to distinguish above-chance between categories at all levels in the hierarchy and in all brain regions considered: object-, scene-, and face-selective areas (LOC, FFA, PPA, RSC, TOS), as well as early visual areas (V1, V2, V3v, hV4) (Fig. 2.6 A).

Critically, however, we again found that information about object category did not increase monotonically with category generality (taxonomic level) in all brain areas. In LOC, accuracy was highest at the basic level and we saw a significant drop in decoding for both the subordinate and the superordinate levels, compared to the basic level (LOC: Basic > Subord. $t_{16} = 2.4$, $p = 0.031$, Basic > Superord. $t_{16} = 4.4$, $p < 0.001$). Moreover, we found that in all regions, when classification errors did occur, the confusions were more likely to be within the same basic level than between basic levels (i.e. types of cars were commonly confused with other types of cars, but not with types of ships, for example; Fig. 2.6 B), with the effect most salient in LOC (Within Basic Confusions > Between Basic Confusions: V1: $t_{16} = 5.3$, $p < 0.001$; V2: $t_{16} = 5.7$, $p < 0.001$; V3v: $t_{16} = 6.3$, $p < 0.001$; hV4: $t_{16} = 5.1$, $p < 0.001$; LOC: $t_{16} = 5.5$, $p < 0.001$; FFA: $t_{16} = 4.4$, $p < 0.001$; TOS: $t_{16} = 6.3$, $p < 0.001$; PPA: $t_{16} = 7.7$, $p < 0.001$; RSC: $t_{16} = 6.9$, $p < 0.001$).

Our correlation classifier decoding results mirror the findings from Experiment 1, which suggest that the basic level represents an optimal level of specificity in object taxonomy in object-selective cortex. Furthermore, we again see evidence that MVPA did not weigh cohesion and distinctiveness equally when assigning category labels to neural activations, since decoding produces a stronger advantage for the basic level

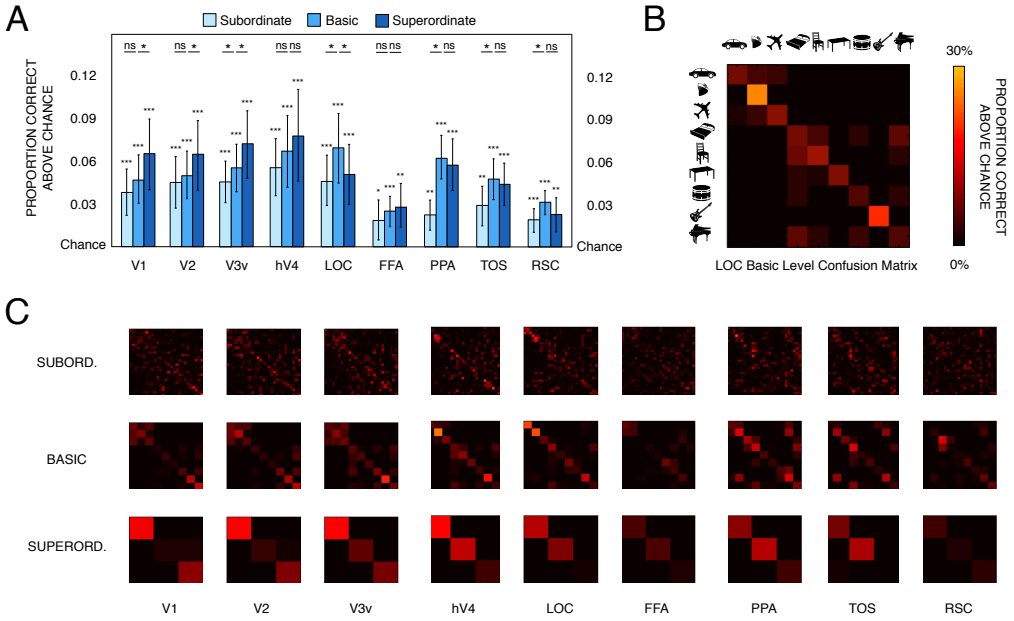


Figure 2.6: **After controlling for animacy, real-world size, and naturalistic backgrounds, MVPA classification reveals that object categories are most distinct at the basic level in LOC.** (A) Proportion above chance of correct decoding responses for all levels of the taxonomy (chance is zero): subordinate, basic, and superordinate. Top insets denote whether differences between adjacent bars are significant. Category information was discernible significantly above chance at all taxonomic levels and in all ROIs. Decoding accuracy at the basic level was higher than both at the subordinate and superordinate levels in LOC, but not in any of the other brain areas considered. (B) Confusion matrix example: LOC basic level classification. Basic categories were ordered on the axes according to the pictograms: cars, ships, planes, beds, chairs, tables, drums, guitars, and pianos. At the subordinate level, within each basic category, the three corresponding subordinates were listed alphabetically. At the superordinate level, the "vehicle" category was listed first, the "furniture" category was listed second, and the "musical instrument" category was listed last. (C) Confusion matrices for decoding analysis in A: top = subordinate level; middle = basic level; bottom = superordinate level. In all regions, when classification errors did occur, the confusions were more likely to be within the same basic level than between basic levels with the effect most salient in LOC. The basic level matrices show that confusions become more common within the basic level as we move up the visual hierarchy. $*$ $p < .05$, $**$ $p < .01$, $***$ $p < .001$, n.s. = not significant. Error bars: 95% confidence interval. SUBORD. = subordinate; SUPERORD. = superordinate.

that the category boundary analysis. This suggests that cohesion and distinctiveness might not contribute equally to generating category boundaries in LOC.

Together, our two experiments show that category representations change as a function of taxonomic level across the span of the human ventral visual processing stream: initially, subordinate categories are more distinguishable in early visual areas, but this advantage diminishes in later areas, and this is due to changes in both category cohesion and distinctiveness between visual areas. Most importantly, by testing two separate taxonomies, each representative of real-world hierarchical organization of objects, we show that this effect is robust, generalizable, not fully explained by low-level visual features, and persists after eliminating image backgrounds and removing the contribution of animacy and real-world size.

2.4 Discussion

Our work establishes a link between the neural representation of object categories in occipito-temporal cortex and human object taxonomy. We achieve this by showing that for two category taxonomies which exhibit a clear behavioral basic level advantage, category representations change as a function of taxonomic level as we move up the ventral visual cortical hierarchy. This provides evidence that basic level structure may be an emergent property of the human visual system.

Consistent with the tenets of categorization theory [116], patterns in high-level visual cortex adhere to the principle of simultaneously maximizing within-group similarity and between-group dissimilarity. Moreover, our results provide the first neural support for the hypothesis that the basic level strikes the best balance between these two measures, whereas the subordinate and superordinate levels appear to each optimize similarity along one dimension over the other [113]. Moreover, our data underscore the importance of considering the joint contribution of both aspects that give rise to the concept of a category in visual cortex: within-category similarity (cohesion) may be an intuitive candidate for what makes a good category, but our work shows that, in fact, distinctiveness is just as important in establishing neural category boundaries and actually varies more sharply between early visual areas and

object-selective cortex than cohesion. Most importantly, this organizational principle emerges gradually as we move up the visual cortical hierarchy and is not present in either low-level image features or early visual cortex activations (Fig. 2.2 and Fig. 2.5). This suggests that objects in the world do not group naturally by basic level category in terms of their appearance, but instead successive levels in the visual system may be optimizing basic level categorizations.

Previous studies have reported that information decoding in early visual areas using a linear classifier is at chance levels when retinal location, viewing angle, or size is altered [20, 38, 39]. Perhaps surprisingly, our results show a predilection for early visual areas to group objects strongly at the subordinate level, with this grouping diminishing gradually in favor of the basic level only in later visual processing regions. This effect can be well understood if we consider that objects within the same subordinate level category share more low-level features in common with each other than with members of other (subordinate, basic, or superordinate) categories, as evidenced by the low-level feature analyses in Fig. 2.2 A,B and Fig. 2.5 A,B (insets). Thus, given that subordinates share more overall low-level features in common, we expect to observe greater subordinate level category cohesion, especially in early visual areas. High cohesion here may also be partially explained by the fact that our stimuli were all presented centrally, allowing the low-level features to overlap despite the localized information processing and the small receptive fields in these areas [20, 38, 39]. Consequently, our results do not imply that fine-grained category distinctions are most strongly represented in early visual cortex; instead, early visual cortex is simply the region where low-level features best drive similarity of activity patterns. In fact, decoding performance in early visual cortex indicates that subordinate categories are less distinguishable than the other two, presumably reflecting their low distinctiveness. Altogether, this suggests that the principle of maximizing within-category similarity and between-category dissimilarity is necessary, but not sufficient for a good category representation: for example, in early visual areas we see strong category boundaries, but they are likely due to low-level features.

Our results support the hypothesis that fine-grained categories become more separable in higher visual areas at the scale of neural response afforded to us by fMRI.

This trend is illustrated best in Fig. 2.5 B: initially, activity patterns elicited by subordinates are not distinguishable (distinctiveness near zero in V1), but they become increasingly anti-correlated (significantly positive distinctiveness in hV4 and LOC). This indicates that fine-grained distinctions increase with complexity of visual processing and that the high category boundary effect values observed for subordinate categories in early visual areas are mainly driven by high cohesion due to low-level feature overlap. Finally, while low-level features of our stimuli may, in part, contribute to the overall trend we observe for subordinate categories, prior evidence suggests that, indeed, visual features may be inextricably linked to categorical representations [42, 74].

Anatomically, prior evidence suggests that large-scale smooth selectivity gradients for semantic category groupings [66] and object attributes, such as animacy [19, 23, 76, 81] and real-world size [76, 77] underlie object category responses in the human visual system. By leveraging similarity in cortical activity patterns, our work complements this view by revealing what may be an important principle of categorization in the brain: fine-grained representations trade off with more general basic level representations after early visual areas. By analyzing the similarity between the patterns of category responses, we uncovered a tendency for object-selective cortex (LOC) to amplify basic level category boundaries compared to those at other taxonomic levels. Although this tendency is strongest in LOC, other high level areas exhibit similar trends compared to early visual cortex (albeit much less so than LOC). These include both scene-selective (PPA, TOS, RSC) and face-selective areas (FFA). The fact that object-related activity behaves similarly in these high-level visual areas, including those are that are not typically associated with object processing (PPA, TOS, RSC), suggests that these areas may share common computations; computations whose byproduct is to clarify and separate categories.

The behavioral basic level advantage is mainly supported by evidence that most objects are categorized faster at the basic level [71, 90, 98, 99, 116, 123, 125] (but not for domain-level naming [127] and that basic level labels are used nearly exclusively when people freely name an object [116]. Our results offer a plausible neural explanation for these aspects of the basic level perceptual advantage. Under our

proposed model, the basic level advantage arises due to cortical computations that increase the efficacy of basic level category boundaries between early visual cortex and object-selective cortex. If LOC primarily enhances category representations between objects at the basic level, then areas which use its afferents as input (temporal [93] or frontal [49, 95]) would require less computation and thus less time to extract or construct categorical information at the basic level of specificity. Information about basic category is easily linearly separable in LOC, whereas further computations would be required to access subordinate and superordinate representations. Consequently, basic level information is mostly available from polling object-selective areas at little additional computational cost, and thus voluntarily expressed faster, which is consistent with prior behavioral findings [116]. Consequently, our results are also consistent with the hypothesis that an enhanced basic level advantage for neural patterns of activity might arise at a post-perceptual level of representation, such as in high level semantic areas (e.g. pMTG, ITG) that likely represent and build amodal representations of object categories [15, 22, 43]. Indeed, we believe the search for such an area and representation constitutes an interesting avenue for future study.

Although the basic level advantage is a well-accepted phenomenon, there is some controversy surrounding its robustness: some behavioral studies report that either the subordinate, or the superordinate level is accessed first, rather than the basic level of specificity [71, 89, 90, 125, 127]. In our behavioral experiments, we found a strong basic level advantage for virtually all categories we investigate in terms of speeded categorizations, thus confirming that the entry-level for our taxonomy lies at the basic level. Nonetheless, the real world contains several orders of magnitude more categories embedded in a much deeper hierarchical tree than the three-tiered taxonomy we used. Thus, the results reported here do not preclude the possibility that a carefully picked stimulus set (e.g. containing less typical members of basic categories [71]), a different task (e.g. ultra-fast categorization [89]), or a set of participants who possess expertise in the categories being tested [125] may change the level of the taxonomy at which neural patterns may group object stimuli. Furthermore, our results are highly generalizable across two separate hierarchies where the superordinate level is defined at different specificity distance from the basic level (arguably natural and

man-made are farther from the basic level than vehicles, musical instruments, and furniture). We are agnostic, however, whether other possible superordinates may fare differently against our basic level categories. Nevertheless, we would then predict that such effect would also be reflected in behavior. As such, all the above manipulations provide interesting avenues of further inquiry.

More broadly, our data suggest an alternative hypothesis to the view that categorical distinctions emerge mainly from processing in anterior temporal or frontal areas of the brain [49, 93, 95], a view also mirrored by models that strongly encapsulate vision from cognition [48, 108, 110]. Instead, our work shows that clear category separations emerge gradually as early as occipito-temporal regions and in the absence of an explicit categorization task, suggesting that categorization may be part of visual processing. This view is consistent with recent behavioral results that show that categories alter perception [50], even when categorization is task-irrelevant [88].

The basic level advantage is a pervasive phenomenon that captures something fundamental about human cognition. As such, it has influenced many fields of knowledge, ranging from psychology and neuroscience, to molecular biology, engineering, and the humanities. In fact, Rosch's original finding was cited over 4,000 times across these disciplines. Our work provides a long overdue understanding of why the basic level might be privileged: the human brain appears to build basic level categories over successive visual areas. Such an understanding is key to answering the broader question about how the human brain extracts and organizes information from our visual world.

2.5 Acknowledgments

This work was funded by a William R. Hewlett Stanford Graduate Fellowship (to M.C.I.), an NRSA Grant NEI F32 EY019815 (to M.R.G.), and a National Institutes of Health Grant 1 R01 EY019429 (to D.M.B and L.F.-F.).

Chapter 3

Typicality Sharpens Category Representations in Object-Selective Cortex

As we argued in the previous chapter, the purpose of categorization is to identify generalizable classes of objects whose members can be treated equivalently. Within a category, however, some exemplars are more representative of that concept than others and considerable evidence suggests that the typicality of a particular item is reflected in how fast and how accurately we perceive it in our daily lives [109, 112, 115]. However, despite such long-standing behavioral effects, little is known about how typicality influences the neural representation of real-world objects from the same category.

To address this question, we used a functional neuroimaging experiment where we showed participants 64 subordinate object categories (exemplars) grouped into 8 basic categories. Typicality for each exemplar was assessed behaviorally and we used several multi-voxel pattern analyses to characterize how typicality affects the pattern of responses elicited in early visual and object-selective areas: V1, V2, V3v, hV4, LOC. We found that in LOC, but not in early areas, typical exemplars elicited activity more similar to the central category tendency and created sharper category

boundaries than less typical exemplars, suggesting that typicality enhances within-category similarity and between-category dissimilarity. Additionally, we uncovered a brain region (cIPL) where category boundaries favor less typical categories. Our results suggest that typicality may constitute a previously unexplored principle of organization for intra-category neural structure and, furthermore, that this representation is not directly reflected in image features describing natural input, but rather built by the visual system at an intermediate processing stage. This chapter is joint work with Michelle R. Greene, Diane M. Beck, and Fei-Fei Li, and was previously published as [68].

3.1 Introduction

The purpose of categorization is to identify generalizable classes of objects whose members can be treated equivalently. Within a category, however, some exemplars are more representative of that concept than other members of the same category. This typicality effect usually manifests behaviorally as increased speed of recognition, as well as lower error rates for verifying category membership of the more typical item [109, 112, 115]. Despite well-studied behavioral effects, little is known about how typicality influences the neural representation of objects from the same category: for example, why are some dog exemplars more representative of the category "dog" than others and where can we find evidence for this distinction in the brain?

Previous investigations of the neural basis for typicality have employed category learning paradigms over artificially constructed categories [2, 29–31, 138]. By contrast, our environment contains tens of thousands of distinct object categories [11, 32]. Furthermore, considerable evidence suggests that perceived typicality is reflected in how fast and how accurately we perceive many such real-world objects and categories [109, 112, 115]. Thus, the overarching goal of our present work is to investigate how the typicality of real-world object categories affects their representation in human visual cortex.

Many theories and cognitive models have been proposed for the instantiation of

typicality as a dimension of object representation in human categorization (for reviews, see e.g. [1, 8, 9, 96], however, a clear neural correlate of these models has yet to be identified. Nevertheless, in virtually all such models, distinct objects are defined as points in a multidimensional psychological space and similarity (in terms of features or properties) between such items belonging to the same or different categories represents the defining characteristic by which typicality (and categorization itself) are instantiated. In the spirit of this observation, we set out to test one of the earliest and most fundamental hypotheses regarding the instantiation of typicality relationships between exemplars in a given category: the family resemblance hypothesis first put forward by Rosch and Mervis [115]. Their proposed model states that highly typical members of a category are those that share most features in common with other members of that category (i.e. a typical subordinate level exemplar, such as a Golden Retriever, is highly representative of the basic level category "dog"), while simultaneously sharing the fewest features in common with other categories in a similar semantic space (i.e. with other basic level categories within the same superordinate category; e.g. Golden Retrievers would share very few features in common with cats).

Investigating hypotheses such as this one is challenging in the real-world domain mainly because the sheer number of categories in our environment is estimated to be in the tens of thousands [11] and because controlling for the features of natural visual stimuli is notoriously difficult. In our present experiment, we put forward the first attempt to push beyond small-scale, artificial, hand-designed datasets for investigating how typicality modulates neural representations by leveraging a large-scale taxonomically structured image database (ImageNet [32]), along with employing a method for obtaining high-throughput behavioral rankings (the Amazon Mechanical Turk platform). As such, we are now able to test directly whether brain regions exist where the family resemblance hypothesis represents a guiding principle for the neural intra-class organization of a large set of real-world object categories and, furthermore, compare this organization against the corresponding low-level visual feature representation of the over one thousand images we used as stimuli in our study.

To this end, we performed a passive viewing fMRI experiment in which participants viewed color photographs from 64 subordinate level object categories grouped

into 8 basic level categories. The typicality of each subordinate category (hitherto referred to as an "exemplar") within its corresponding basic category (hitherto referred to as a "category") was ranked behaviorally. The family resemblance model was originally defined using a semantic feature space: e.g. the category "dog" is exemplified by features such as "has-tail", "wags-tail", and "is-furry"; and an exemplar which possesses more of these features would be rated as more typical. Although Rosch's family resemblance hypothesis has been well received, it has been difficult to find definitive evidence for it primarily because the feature space used by the brain is unknown. Here, we set out to investigate this question in the domain of neural activation patterns, where we can remain agnostic as to the nature of the feature spaces, semantic or otherwise, in which object categories are represented. Multi-voxel pattern analyses allow us to characterize the similarity between neural patterns elicited by these categories throughout human visual cortex, without making any explicit assumptions regarding the building blocks of the feature spaces themselves. As such, we found that in object-selective regions of occipito-temporal cortex, but not in early visual areas, typical exemplars were more similar to the central tendency of the category and created significantly sharper category boundaries than less typical exemplars, suggesting that typicality enhances category cohesion (within-category similarity) and category distinctiveness (between-category dissimilarity). Thus, we present the first evidence that typicality modulates neural representations of real-world object categories in object-selective cortex in a manner consistent with the family resemblance hypothesis. Interestingly, using a whole-brain analysis, we also uncovered the first evidence of a brain region where category boundaries favor less typical categories (cIPL). Taken together, these findings suggests that the two extremes of the behavioral typicality continuum may simultaneously exert separate influence on the neural representation of real-world object categories across human visual cortex, and moreover, that typicality may constitute a previously unexplored principle of organization for intra-category neural structure, one that is likely built by the visual system at an intermediate processing stage, rather than inherited from low-level features of our input.

3.2 Materials and Methods

3.2.1 Constructing a Behaviorally-Normed Category Set

The goal of our experiment was to test the family resemblance hypothesis [115] which posits that highly typical members of a category share the most features in common with other members of that category, while simultaneously sharing the fewest features in common with members of semantically related categories. To test this model appropriately, we required a set of basic level categories (e.g. dog, car), each comprising multiple subordinate level categories (exemplars, e.g. Chihuahua, sedan) for which perceived typicality could be assessed behaviorally.

In our experiment, we started with a four-tiered taxonomic hierarchy comprising the following putative levels: two domain level categories (natural, man-made), four superordinate level categories (animals, plants, musical instruments, vehicles), sixteen basic level categories (e.g. bird, cat, dog, fish for "animals"), and one hundred and twenty-eight subordinate level categories (e.g. Chihuahua, stealth plane, parsley). Subsequently, we assessed the entry levels in each of our four superordinate tiers. We performed a match-to-category behavioral experiment in which we asked participants to verify whether each image belonged to its subordinate, basic, superordinate, or domain level category. We found that, of our four putative superordinate categories, "animals" and "vehicles" were the only ones who adhered strongly to the putative hierarchy, whereas plants and musical instruments varied across disparate taxonomic tiers and, for some of their categories, the basic level was situated either at a more general or more specific tier than their putative designation (e.g. putative basic levels "wind instruments", "string instruments", "garden plants" closer to superordinate level; putative superordinate level "plants" closer to basic level; putative superordinate "musical instruments" closer to domain level; see Appendix B, Fig. B.1, Fig. B.2). Therefore, to maintain a consistent, verified hierarchy, we selected a subset of our original dataset comprising eight basic level categories (dogs, cats, birds, fish; cars, boats, planes, trains) and sixty-four subordinates (eight for each basic category, e.g. Chihuahua, stealth plane, etc.). This hierarchy has the added advantage that it contains equal numbers of natural / animate and man-made/inanimate categories,

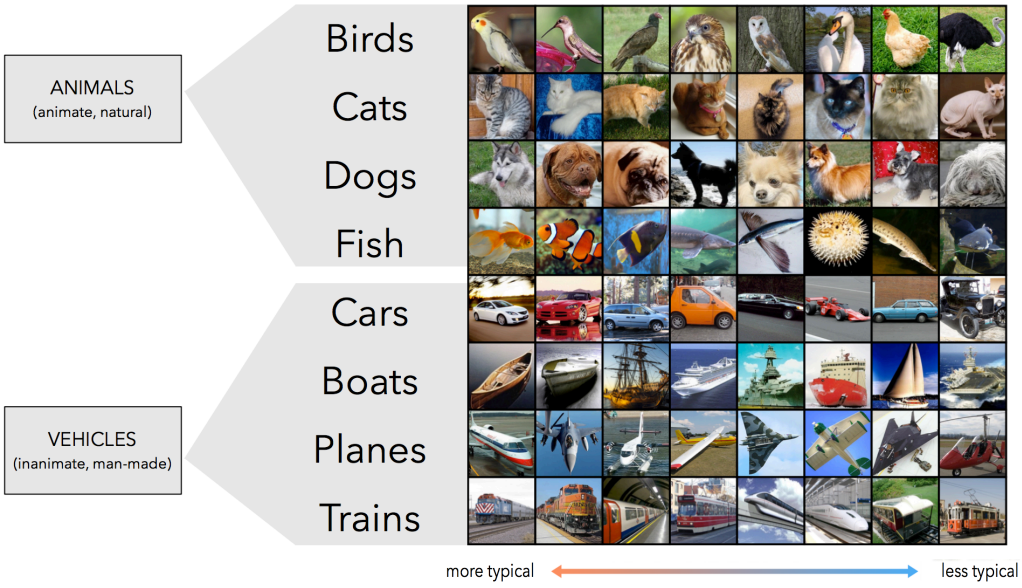


Figure 3.1: **Typicality ranked stimulus set.** Our stimulus set comprised 8 subordinate level exemplars from each of 8 basic level categories. Participants were shown 16 images from each exemplar, varying in pose and color (only one representative image is shown above). Within each basic category, exemplars are organized according to behavioral typicality from the most typical (left) to the least typical (right): e.g. airliners (rank 1) and fighter planes (rank 2) were judged to be much more typical examples of planes than stealth planes (rank 7) and gyrocopters (rank 8).

a distinction known to affect representations of object categories in human visual cortex [23, 76].

Subsequently, we used ImageNet [32] to collect 16 distinct images containing objects of interest from each of our sixty-four subordinate level categories; i.e. if the subordinate category is pugs, then we showed 16 distinct photographs of pugs. Pictures were cropped to feature the objects prominently and centrally within a square region (400 x 400 pixels in size) and included their natural background. Within each subordinate category, the images varied greatly in color and pose. Representative images from each of our 64 categories are shown in Fig. 3.1.

3.2.2 Behavioral Experiment: Typicality Rankings

3.2.2.1 Participants and Materials

40 participants were recruited on Amazon's Mechanical Turk platform (AMT) from a pool of trusted US-based participants with at least 2,000 previously accepted AMT results at a minimum of 98% approval. Participants completed the study from their own personal computing device.

3.2.2.2 Experimental Procedures

Each of the AMT hits contained 28 trials comprising each possible pairwise comparison between the eight subordinate categories within a particular basic category. In each trial, participants viewed a randomly drawn image from two subordinate categories and were asked to indicate by clicking which image was the most typical of its corresponding basic category. Ten individual participants ranked each basic category, with each participant ranking a median of six basic level categories overall. Participants were compensated \$0.50 per hit and each hit took an average of 88 seconds to complete.

3.2.2.3 Data Analysis

Pairwise typicality rankings for the eight subordinates in each basic category were obtained. We computed the percentage of times each subordinate was chosen as the more typical item in a pair and used this quantity to order subordinates according to their typicality in each basic category independently. We also recorded a high value for the inter-subject reliability of the collected typicality rankings (75% $\pm$ 2%, mean $\pm$ s.e.m.; see Appendix B, Fig. B.3).

3.2.3 fMRI Experiment

3.2.3.1 Participants

12 volunteers (2 females, ages 24–32, including authors M.C.I. and M.R.G.) with no past history of psychiatric or neurological disorders and normal or corrected-to-normal vision participated in this experiment. Participants gave informed written consent in compliance with procedures approved by the Stanford University Institutional Review Board. Except for the participating authors, all subjects received financial compensation. One participant was subsequently rejected from our analyses due to our inability to satisfactorily identify their regions of interest using the localizer scanning procedures detailed in the corresponding section below.

3.2.3.2 Scanning Parameters and Preprocessing

Imaging data were acquired with a 3 Tesla G.E. Healthcare scanner. A gradient echo, echo-planar sequence was used to obtain functional images (volume repetition time (TR) 2 s; echo time (TE) 30 ms; flip angle 80 degrees; matrix 128 x 128 voxels; FOV 20 cm; 29 oblique 3 mm slices with 1 mm gap; in-plane resolution 1.56 x 1.56 mm). We also collected a high-resolution (1 x 1 x 1 mm voxels) structural scan (SPGR; TR 5.9 ms; TE 2.0 ms; flip angle 11 degrees) in each scanning session. The functional data were spatially aligned to compensate for motion during acquisition and each voxel's intensity was converted to percent signal change relative to the temporal mean of that voxel using the AFNI software package [26]. To perform our analyses, we computed the average voxel activity for each block. We did not perform any smoothing.

3.2.3.3 Experimental Procedures

Images were presented centrally subtending 21 x 21 degrees of visual angle and were superimposed on an equiluminant gray background. We used a back-projection system (Optoma Corporation) operating at a resolution of 1024 x 768 pixels at 75 Hz. Participants performed 2 sessions, 8 runs each, with 16 blocks per run and 8 images per block. Each block consisted of a 500 ms fixation cross presented centrally,

followed by 8 consecutive stimulus presentations from the same subordinate level category, with a 12 s gap between the blocks. Each image was presented for 160 ms, followed by a 590 ms blank gray screen. Subjects were asked to maintain fixation at the center of the screen, and respond via button-press whenever an image was repeated (one-back task, 0–2 repetitions per block). Over the course of the experiment, each participant viewed 2 blocks from each of the subordinate level categories. The order of blocks, the number of repetitions in each block, and the images in each block were counter-balanced across runs and between subjects.

3.2.3.4 Regions of Interest (ROIs)

The positions and extents of each participant's lateral occipital complex (LOC) were obtained using standard localizer runs conducted in a separate fMRI session. Participants completed two runs, each with 12 blocks drawn equally from six categories: child faces, adult faces, indoor scenes, outdoor scenes, objects (abstract sculptures with no semantic meaning), and phase-scrambled objects. Blocks were separated by 12 s fixation cross periods and comprised 12 image presentations, each of which consisted of images presented for 900 ms, followed by a 100 ms fixation cross. Each image was presented exactly once, with the exception of two images during each block that were repeated twice in a row. Subjects were asked to maintain fixation at the center of the screen and respond via button press whenever an image was repeated. To avoid any issues related to intrinsic variability in signal reliability across our participant pool, we selected fixed-volume ROIs across all our participants. The volume of LOC in mm^3 was chosen conservatively, based on sizes previously reported in the literature, accounting for resolution differences between studies [51, 67, 133]. Accordingly, LOC was defined as the top 500 voxels bilaterally near the inferior occipital gyrus that responded to an Objects > Scrambled Objects GLM contrast.

To determine the locations of early visual areas V1, V2, V3v, and hV4, we used a standard retinotopic mapping protocol in a separate experiment, in which a checkerboard pattern undergoing contrast reversals at 5 Hz moved through the visual field in discrete increments [119]. First, a wedge subtending an angle of 45 degrees from fixation was presented at 16 different polar angles for 2.4 s each. Next, an annulus

subtending 3 degrees of visual angle was presented at 15 different radii for 2.4 s each. Each subject passively observed two runs of 6 cycles in each condition, yielding 512 timepoints per subject. The locations and extents of early visual areas were delineated on a flattened cortical surface for each subject, using a horizontal vs. vertical meridian general linear test, which gave the boundaries between retinotopic maps.

We aligned the positions of the ROIs to the experimental sessions using the AFNI software package [26], by first aligning the structural scans between sessions with sub-millimeter precision, and then applying the alignment transformation to the ROI positions. Percent signal change was then extracted for each voxel in each ROI and these vectors were submitted to the similarity analyses described next.

3.2.4 fMRI Data Analysis

3.2.4.1 Correlation Advantage

First, we assessed whether the most or the least typical exemplars in each category were more similar to the central category tendency. To this end, for each basic category, we used the average neural patterns of all exemplars as a proxy for the central category tendency representation. This definition is similar to that of a putative prototype for that category [121]. We then computed the correlation (Pearson's r) between this category central tendency, on the one hand, and the most and least typical subordinates in each basic category, on the other hand. We hypothesized that if the family resemblance hypothesis is upheld, then the most typical subordinate will be more similar (correlated in its elicited pattern of activation) to the central category tendency than the least typical subordinate. Additionally, we computed a version of this analysis where we omitted from the computation of the central tendency the most typical and least typical exemplars (leaving only the six middle-typicality exemplars in each category). Results were similar, regardless of the method used to compute the central category tendency. Throughout our analyses, we chose to focus on Pearson correlation as a straightforward, scale-invariant measure of similarity of neural patterns, which has the ability to normalize across differences in mean activation level between stimuli and is therefore less susceptible to such variation across a large set

of object categories.

3.2.4.2 Category Boundary Effect

Next, we assessed whether typical exemplars share fewer features in common with other categories than less typical exemplars. Here, we refer to neural features (as measured by voxel activity levels) and we make no assumption that the features are semantic or otherwise [22], only that multi-voxel patterns reflect some underlying feature space. By measuring similarity of brain activity patterns we aim to bridge the gap between the two types of features, positing that similarity in one descriptive space (voxels) is a good proxy for similarity in the other (internal feature representation). We hypothesized that if this is the case, then categories defined solely by relatively higher typicality exemplars would be more distinguishable from one another than categories comprising only less typical exemplars. To this end, for each ROI and each subject, we split our dataset into two halves comprising the four most typical and four least typical exemplars, respectively, from each category. We then computed a category boundary effect measure separately for each of the two halves of our dataset. We defined the category boundary effect identically to previous work [67, 81] as the difference between within-category similarity and between-category similarity, averaged across all categories considered. For each basic level category, we computed within-category similarity as the average correlation (Pearson's r) between neural patterns elicited by within-category pairs of blocks (e.g. for "dogs", this quantity is defined as the average correlation between voxel activations for any two blocks where any type of dog was shown). Similarly, we computed between-category similarity as the average correlation between neural patterns elicited by between-category pairs of blocks across basic level categories (e.g. for "dogs", this quantity is defined as average correlation between voxel activations for a block where dogs were shown and another block where, for example, planes were shown). We performed each of these analyses for each subject and ROI separately. We used this measure to quantify how well categories are separated in the neural space of representation, given their behavioral typicality.

3.2.4.3 Low-Level Feature Analysis

To show that the effects in the correlation advantage and category boundary effect analyses above, are not solely due to low-level image features, we also performed analogous computations for image descriptor features extracted from our stimulus images: LAB color histograms, GIST [103], and multi-scale Gabor wavelet features [73]. Color histograms were represented using LAB color space. For each image, we created a two-dimensional histogram of the a$*$ and b$*$ channels using 64 bins per channel. We then averaged these histograms over each of the 16 distinct stimuli in each subordinate category, such that each subordinate was represented as a 4,096-length vector representing the averaged colors its corresponding images. For GIST, we used the descriptor features first proposed by Oliva and Torralba [103]. This model provides a summary statistic representation of the dominant orientations and spatial frequencies at multiple scales coarsely localized on the image plane. We used spatial bins at 4 cycles per image and 8 orientations at each of 4 spatial scales for a total of 3.072 filter outputs per image. We averaged the GIST descriptors for each of the 16 distinct stimuli in each subordinate category to arrive at a 3,072-dimensional representation of each of our 64 subordinates. For wavelet features, we represented each image in our stimulus set as the output of a bank of multi-scale Gabor filters. This type of representation has been used to successfully model the representation in early visual areas [73]. Each image was converted to grayscale, downsampled to 128 x 128 pixels, and represented with a bank of Gabor filters at three spatial scales (3, 6, and 11 cycles per image with a luminance-only wavelet that covers the entire image), four orientations (0, 45, 90, and 135 degrees), and two quadrature phases (0 and 90 degrees). An isotropic Gaussian mask was used for each wavelet, with its size relative to spatial frequency, such that each wavelet has a spatial frequency bandwidth of one octave and an orientation bandwidth of 41 degrees. Wavelets were truncated to lie within the borders of the image. Thus, each image is represented by $3 * 3 * 2 * 4 + 6 * 6 * 2 * 4 + 11 * 11 * 2 * 4 = 1,328$ total Gabor wavelets. We created the wavelet representation of each of our 64 subordinate categories by averaging over the representation of the 16 distinct images associated with each of them.

3.2.4.4 Whole-Brain Searchlight Analysis

For each participant's brain, we extracted all grey matter voxels and placed a sphere of radius 4 voxels at every other voxel location (step size: 2 voxels). We excluded all locations where half or more of the voxels in the proposed cube did not overlap with grey matter. For each cube, we computed a local category boundary effect (CBE) for responses to the most typical and the least typical half of our dataset, similar to the analysis procedure described above. We then used these values to identify brain regions where category boundaries were stronger between more typical categories (More Typical Half CBE > Less Typical Half CBE) and vice versa (More Typical Half CBE < Less Typical Half CBE). Individual subject results were transformed into group space by aligning to the Talairach atlas and averaging the aligned maps together. To establish statistical significance for our results, we thresholded the group maps for each analysis by using a false discovery rate (FDR) of 0.05, which was determined by computing 1,000 simulated group maps, obtained by permuting the category labels without replacement in each voxel cube searchlight.

3.2.5 Statistical Analyses

For all our experiments, we used paired two-tailed t-tests when comparing observed effects against chance and when establishing whether a significant difference exists between two observed effects. We used Kolmogorov-Smirnov tests to establish that no significant deviation from normality exists for the distributions of all effects to which t-tests were applied. All statistical tests were implemented in MATLAB.

3.3 Results

3.3.1 Typical Exemplars Are More Neurally Similar to Category Central Tendency

Using two separate behavioral experiments (see Materials and Methods), we established a dataset of eight verified basic level categories (4 natural / animate and 4

man-made / inanimate), each of which comprised eight subordinate level categories normed according to their typicality. Henceforth, we will use the term "category" to refer to one of our eight basic level categories and the term "exemplar" to refer to one of our sixty-four subordinate level categories. To investigate whether the family resemblance hypothesis is upheld in visual cortex neural patterns of activation, we scanned participants viewing our sixty-four exemplars (16 visually different images per exemplar, see Materials and Methods). Since psychological representations of categories are influenced by factors such as task, learning, and attention [57, 86, 101], we asked participants to perform a one-back repetition task in the scanner (i.e. no explicit categorization or typicality judgment task) used solely to ensure they maintained alertness during the experiment. Our analyses focused on object-selective cortex (lateral occipital complex (LOC)) and early visual areas (V1, V2, V3v, hV4).

First, we assessed the intra-class component of the family resemblance hypothesis, namely that more typical exemplars in a category share more features in common with the central category tendency that do atypical exemplars. To test this, within each of our eight categories, we compared how similar (using Pearson's r) the most typical and least typical exemplars were to the central category tendency, defined here by averaging together the neural patterns corresponding to all exemplars in each category. This definition is similar to that of a putative prototype for that category [121].

Here, we hypothesized that if family resemblance provides a good model for the organization of neural patterns of activation elicited by real-world objects with respect to their typicality, then more typical items should sit closer to the center of this space and hence be more similar to the central category tendency, than the atypical exemplars. Indeed, we found that highly typical exemplars were by far more similar to the category average than less typical exemplars in object-selective cortex (i.e. LOC), but not in early visual areas (Fig. 3.2 A; LOC: high $>$ low $t_{10} = 3.8$, $p = 0.003$; V1: high $>$ low $t_{10} < 1$, $p = 0.491$; V2: high $>$ low $t_{10} = 1.3$, $p = 0.228$; V3v: high $>$ low $t_{10} < 1$, $p = 0.468$; hV4: high $>$ low $t_{10} = 1.2$, $p = 0.261$). Additionally, these results replicate using a version of the analysis where we omitted from the computation of the central tendency the most typical and least typical exemplars

(leaving only the six middle-typicality exemplars in each category, see Appendix B, Fig. B.4). Interpreted differently, an equivalent prediction of the family resemblance hypothesis is that the degree of similarity of each subordinate within a basic category to the most typical subordinate in that category should consistently decrease with the typicality rating given to that particular subordinate. Indeed, we found that this alternative prediction mirrors our results above: similarity is highest between the two most typical subordinates within a basic category and drops successively as typicality for a given subordinate decreases (see Appendix B, Fig. B.5). Together, these findings show that intra-class structure of real-world categories is consistent with the family resemblance hypothesis in LOC and provides evidence that the representation of object categories shares key properties in common with prototype- and norm-based representations (see e.g. [1, 83, 120].

To show that the effects we observed cannot be explained solely on the basis of the low-level properties of the stimuli themselves, we replicated our similarity analysis using several sets of descriptor features extracted from our images: LAB color histograms, GIST [103], and multi-scale Gabor wavelet features [73] (see Materials and Methods for details on how each of the features was computed). We found that all features show similar numerical correlations between the most typical and least typical exemplars with the central category tendency. Additionally, for GIST and wavelet features, we saw an opposite pattern to our LOC results, namely that correlation with the central category tendency was numerically higher for exemplars ranked as less typical (GIST high r = 0.86, low r = 0.84; Wavelet: high r = 0.60, low r = 0.64; Color: high r = 0.88, low r = 0.84). For color histograms, a small trend is observed for typical exemplars to be more correlated with the central category tendency, however this trend disappears (and in fact reverses) when excluding the most and the least typical exemplars from the computation of the central category tendency (middle-six exemplars analysis: GIST: high r = 0.87, low r = 0.89; Wavelet: high r = 0.54, low r = 0.60; Color: high r = 0.91, low r = 0.93; see Appendix B, Fig. B.4). Overall, this implies that low-level features alone cannot fully account for the pattern of results we observe in object-selective cortex, and further suggests the human visual system likely constructs (or, at the very least, strongly amplifies) feature descriptions of our

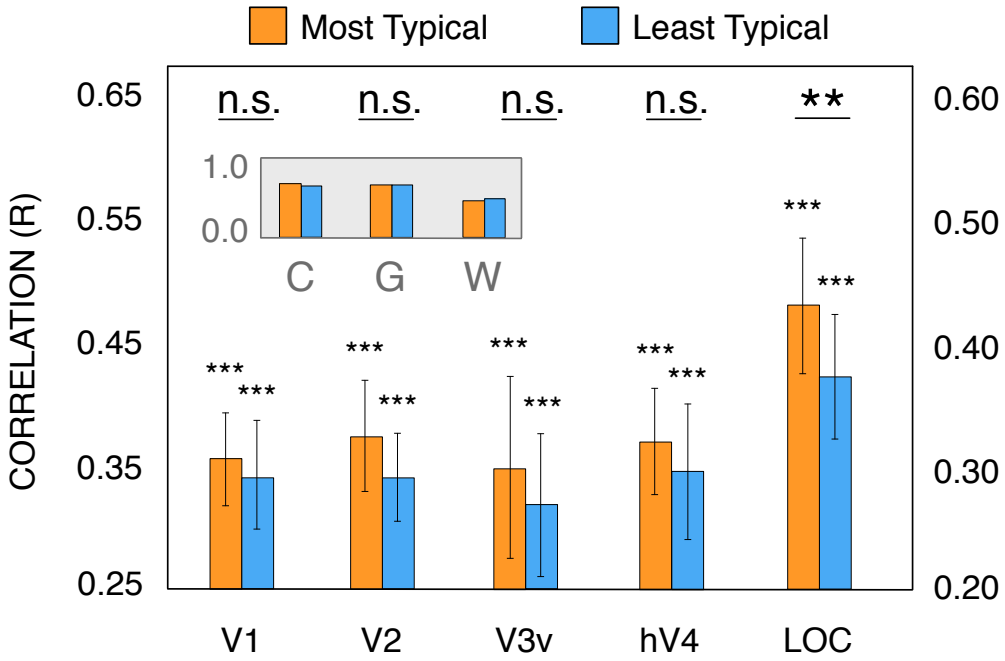


Figure 3.2: **Typical exemplars are more correlated with category central tendency than less typical exemplars in object-selective cortex.** Correlation between category central tendency and most typical exemplar in each category (orange) or least typical exemplar in each category (blue), averaged across all 8 basic level categories. In object-selective cortex (LOC), typical categories are more similar to the average category representation than less typical categories and this effect is not present in early visual areas. (Inset) We performed a similar analysis using the image-level features from our stimulus set: LAB color histograms (C), GIST features (G), and multi-scale Gabor wavelet features (W). All features show similar values for both highly typical and less typical exemplar correlations, with the GIST and wavelet features exhibiting an opposite trend to our LOC results (higher correlation for less typical exemplars). Therefore, low-level stimulus features cannot solely explain our results in object-selective cortex. $***$ $p < .001$, $**$ $p < .01$, n.s. - not significant. Error bars: 95% confidence interval.

visual input that correlate with behavioral typicality judgments later on.

3.3.2 Typical Exemplars Exhibit Stronger Inter-Category Boundaries

We saw that typicality is correlated with how similar an exemplar is to its central category tendency. Next, we investigated whether typicality affects the second dimension of the family resemblance hypothesis: are typical exemplars more dissimilar to other categories than atypical ones? We hypothesized that if this is the case, then categories defined solely by relatively higher typicality exemplars would be more distinguishable from one another than categories comprising only less typical exemplars. As such, we split our dataset into two halves, corresponding to the most typical and least typical exemplars from each category. We subsequently computed the category boundary effect [67, 81] for each of the two halves of the dataset as the difference between within-category similarity and between-category similarity, averaged across our eight basic level categories. We predicted that if the family resemblance hypothesis holds, then the category boundary effect would be stronger when computed on the half of the dataset comprising the four most typical exemplars from each category than when computed on the half of the dataset consisting of the least typical four exemplars from each category. Using this measure of how separable categories are in the space of neural patterns of activation, we found that typical exemplars are more easily distinguishable than less typical exemplars in object-selective cortex (Fig. 3.3 A; LOC: most typical > least typical, $t_{10} = 3.0$, $p = 0.013$). By contrast, typicality does not modulate how separable categories are in the space of neural activations in early visual areas (V1: most typical > least typical, $t_{10} < 1$, $p = 0.597$; V2: most typical > least typical, $t_{10} = 1.5$, $p = 0.167$; V3v: most typical > least typical, $t_{10} = 1.1$, $p = 0.298$; hV4: most typical > least typical, $t_{10} = 1.9$, $p = 0.092$).

Analogously to our previous analysis, we asked whether low-level features of our stimulus set are sufficient to explain the pattern of results we observed in object-selective cortex. Accordingly, we computed the category boundary effect on feature descriptors (LAB color histograms, GIST, and multi-scale Gabor wavelet features)

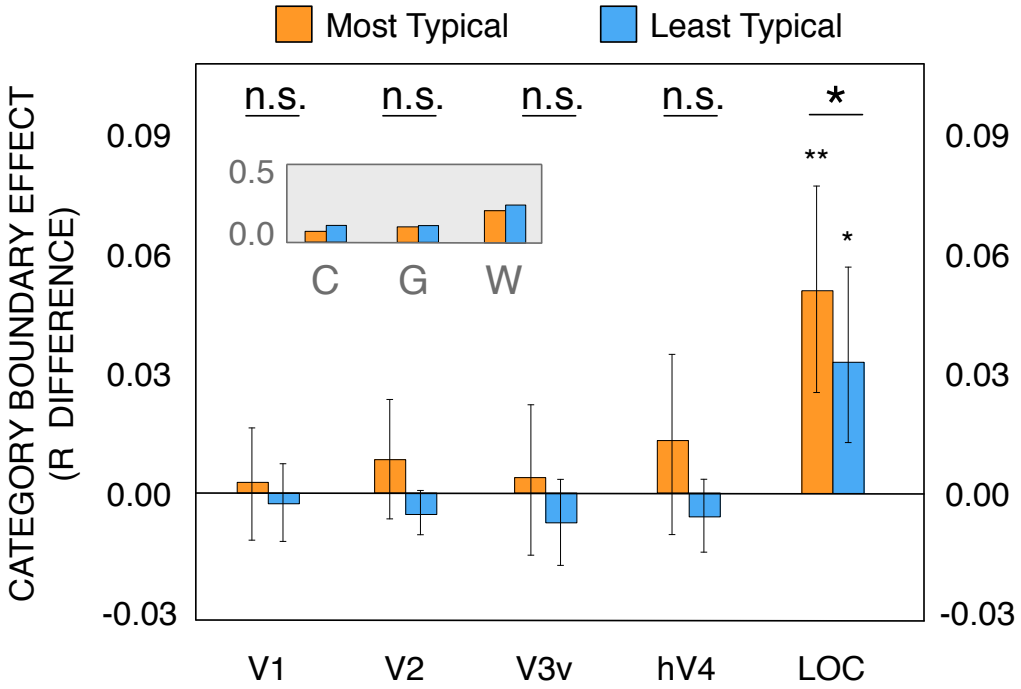


Figure 3.3: **Category boundaries are stronger for highly typical exemplars in object-selective cortex.** Category boundary effect for the two halves of our dataset comprising the most typical 4 exemplars from each category (orange) and the least typical 4 exemplars from each category (blue). In object-selective cortex (LOC), typical exemplars from one category are more distinguishable from exemplars of other categories, an effect not reflected in early visual areas' patterns of activation. (Inset) We performed a similar analysis using the image-level features from our stimulus set: LAB color histograms (C), GIST features (G), and multi-scale Gabor wavelet features (W). All of the feature representations show an opposite trend to that observed in LOC (stronger category boundaries for less typical items) and therefore cannot fully explain our results in object-selective cortex. $**$ $p < .01$, $*$ $p < .05$, n.s. - not significant. Error bars: 95% confidence interval.

extracted from the most typical half and least typical half of our dataset. For all of our feature representations, we found an opposite effect to the one present in LOC: numerically more pronounced category boundaries for the less typical half of our dataset, compared to the most typical half (high vs. low category boundary: Color 0.09 vs. 0.14; GIST 0.13 vs. 0.14; Wavelet 0.27 vs. 0.32). These results, together with the finding that category boundaries are identical in early visual areas for the two halves of our dataset, provide evidence that it is unlikely that low-level features are directly responsible for the emergence of the typicality effect we observe in object-selective regions. In short, this suggests that typical exemplars become more separated in their neural representation in LOC, and that this effect is not purely driven by the visual appearance of our exemplars and categories, but instead is a direct result of sequential processing along the ventral visual stream.

Finally, the category boundary effect is a compound measure that relies on both within-category similarity (category cohesion) and between-category dissimilarity (category distinctiveness) [67, 81]. To investigate the contributions of each of these components of category representation on the strength of the typicality effect we observed, we computed these measures separately for our two halves of the dataset comprising the most and least typical categories, respectively. In all visual areas, we observed no significant differences in cohesion or distinctiveness between the two halves of our dataset (cohesion: LOC: most typical > least typical, $t_{10} = 1.7$, $p = 0.120$; V1: most typical > least typical, $t_{10} < 1$, $p = 0.564$; V2: most typical > least typical, $t_{10} = 1.5$, $p = 0.153$; V3v: most typical > least typical, $t_{10} < 1$, $p = 0.631$; hV4: most typical > least typical, $t_{10} < 1$, $p = 0.763$; distinctiveness: LOC: most typical > least typical, $_{10} < 1$, $p = 0.736$; V1: most typical > least typical, $t_{10} < 1$, $p = 0.735$; V2: most typical > least typical, $t_{10} < 1$, $p = 0.537$; V3v: most typical > least typical, $t_{10} < 1$, $p = 0.760$; hV4: most typical > least typical, $t_{10} = -1.2$, $p = 0.247$). Considering our main finding that a significant difference exists between category boundaries elicited by more and less typical exemplars in LOC, the lack of a significant effect for cohesion and distinctiveness suggests that neither within-category similarity, nor between-category similarity differences drive our effects on their own, but rather it is their combined effect (difference) that separates typical

and atypical exemplars in this brain region.

An analogous prediction of this second aspect of the family resemblance hypothesis indicates that if typical subordinates are indeed more separable from other categories, then they should sit farther from a putative fixed category boundary between two basic categories compared to less typical categories. Indeed, a separate analysis that defined fixed support-vector-machine (SVM) boundaries between every pair of basic categories indicated that, on average, the most typical four subordinates in each category exhibited larger distances to their corresponding boundary than the four least typical subordinates in LOC, but not in early visual regions (Appendix B, Fig. B.6).

Overall, our findings provide strong evidence in favor of the neural plausibility of the family resemblance hypothesis in LOC. In this brain region, typical exemplars are more similar to the average category representation and are more separable (as conferred by their larger category boundary effect) across categories than atypical exemplars, which suggests that typicality exerts a measurable and consistent modulatory effect on the nature of the distributed patterns of neural representation of real-world object categories in object-selective cortex.

3.3.3 Whole-Brain Analysis

So far, we have limited our analyses to functionally defined cortical areas. However, it may be the case that activity in other brain areas beyond our pre-selected ROIs may favor the representation or dissociation of typical and atypical exemplars from the same category. To investigate this hypothesis, we performed a whole-brain searchlight analysis [80] where we computed the category boundary effect for the most typical half of the dataset and the least typical half of the dataset for equally spaced spheres of voxels tiling the entire gray matter surface of our participants' brains. This analysis identifies brain regions where typicality organizes the neural representation space according to the family resemblance hypothesis (typical exemplars more similar to central category tendency, while maximizing distance to other categories). More interestingly, by performing the reverse contrast, we may also uncover brain regions where the opposite is true: since we know that even atypical exemplars are still

identified as members of their respective categories, it is likely that computations exist which are meant to ensure differentiation between these exemplars and thus enable correct assignment into their purported categories.

Consistent with our previous ROI results, we found that typicality modulates the strength of category distinctions in right LOC and to a lesser extent in a region adjacent to right hV4 (Fig. 3.4, right). This finding indicates that, indeed, typicality modulates representation of object categories in object-selective cortex and that this effect is strongest in this region, not simply a late vs. early visual cortex difference in representation.

Interestingly, we also uncovered an advantage for neural patterns of activation distinguishing best between atypical exemplars, compared to highly typical exemplars, in the caudal inferior parietal lobule (cIPL; Fig. 3.4, left). This region has been previously implicated in contextual processing [75] and category learning [138], which raises the possibility that enhanced category boundaries for atypical categories here may be due to additional or specialized processing required to disambiguate between less typical exemplars and subsequently assign them a correct category label.

Taken together, our results suggest that typicality is linked to the neural representation of object categories across several brain regions, with its effects extending to both intra-class and inter-class organization. Our results provide neural confirmation for both predictions of the family resemblance hypothesis in object-selective cortex [115] and, furthermore, we provide the first evidence that typicality provides a concrete dimension of neural organization for real-world object categories in both object-selective cortex (LOC) and cIPL, but outside of early visual cortex, which further suggests that this representation is not directly reflected in image features describing natural input, but rather built by the visual system at an intermediate processing stage.

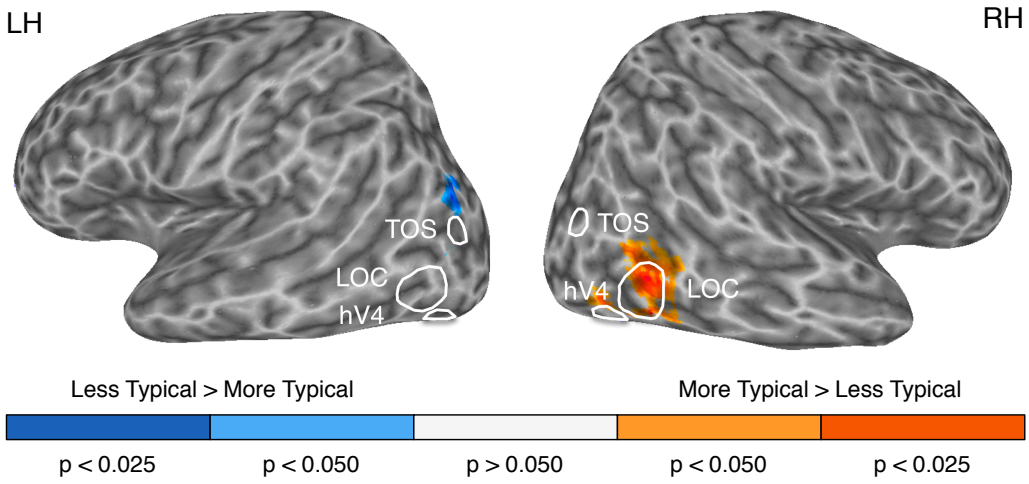


Figure 3.4: **Whole-brain searchlight analysis uncovers brain regions where category boundaries are stronger between most typical and least typical exemplars.** We performed a whole-brain searchlight analysis where we computed the difference between the category boundary effects obtained for the most typical half of our dataset and the least typical half of our dataset. Figure shows group map results, corrected for multiple comparisons using an FDR measure (see Materials and Methods for details). Regions shown in orange (right LOC, right hV4) showed a significant effect of typicality: highly typical exemplars were more distinguishable from exemplars of other categories. Conversely, regions shown in blue (left cIPL) showed the opposite trend: less typical exemplars were more easily distinguishable form members of other categories. This cortical region has been previously implicated in category learning [138] and contextual processing [75], which suggests the possibility that it may aid in the categorization of atypical items, perhaps through mediating contextual facilitation of recognition.

3.4 Discussion

Typicality is a ubiquitous, yet often overlooked property of virtually all objects we interact with in our visual environment. Despite well-studied and long-standing behavioral effects associated with typicality, such as increased speed of recognition and lower error rates for identifying the category membership of more typical items [109, 112, 115], little is known about how typicality relates to the neural representation of objects from the same category. Our work is the first to address this fundamental question using a large array of real-world stimuli. As such, we provide the first neural test of the predictions of the family resemblance hypothesis for real-world object categories: namely, that highly typical exemplars share most features in common with other members of their category (e.g. "Golden retriever" is a highly representative dog), while simultaneously sharing the fewest features in common with other exemplars from semantically-related categories (e.g. Golden retrievers share fewer features with cats than less typical exemplars such as Chihuahuas). Using several similarity-based multivariate pattern analyses, which make no explicit assumptions regarding the nature of the neural feature space in which objects are represented, we found that this conception of category structure describes the organization of neural patterns better in object-selective regions than in early visual areas of the brain. Coupled with the fact that this representation is not directly reflected in image features describing natural input, these data suggest that such a representation is not given in the input, but rather built by the visual system at an intermediate processing stage. In the current set of experiments, we exclusively investigated how typicality affects the neural representation of a set of carefully normed, hierarchically organized object categories. While there is no reason to believe that a separate collection of categories (i.e. one not possessing a taxonomic relationship) would behave differently within the context of neural typicality measures as exemplified in our results, such an experiment remains an interesting question for future work.

The neural basis of typicality has been previously investigated almost exclusively using learning paradigms over artificially constructed category spaces (see e.g. [2, 31, 120, 121]. One of the main advantages of using artificial categories is the tremendous

degree of control one possesses over the instantiation of the feature space, as well as the stimuli themselves. Additionally, synthetic category spaces remove all potential confounds related to object properties that may be directly linked to typicality itself, such as familiarity, discriminability, and expertise. Nevertheless, these idealized and impoverished spaces not only noticeably lack the complexity of visual stimuli we encounter in our everyday environment, but participants' experience with them are necessarily more limited, leaving open the question as to what degree such findings generalize to the real world and to categories that are overlearned. By testing the predictions of the family resemblance hypothesis on real-world categories directly, our current experiment provides long overdue concrete evidence for a typicality-based organization of the neural representation space for such categories in human visual cortex. In our experiment, we not only found that highly typical objects generate stronger category boundaries in object-selective cortex, but we also uncovered the first evidence for a brain region where the opposite is true: in the caudate inferior parietal lobule (cIPL), we see atypical exemplars becoming more differentiated by neural patterns of activity than their highly typical counterparts. This region is superior to the trans-occipital sulcus and the functionally defined scene-selective region TOS (or OPA) [34, 53], likely overlapping with functionally defined area IPS0 [122]. A representation of objects is known to exist in posterior parietal cortex (PPC), independent of action planning, and this cortical region has been shown to exhibit adaptation to object properties, including shape and size [75]. Furthermore, the PPC has also been implicated in the learning of new categories [138], in the recall of words and objects, provided the stimuli are associated with strong memory of source context [70, 107, 130, 131], as well as in the representation of perceptual decision variables [62, 128]. Taken together, these findings raise the possibility that this cortical region may aid in the categorization of atypical items, perhaps through mediating contextual facilitation of recognition. Intuitively, processing category boundaries both in terms of typical and atypical exemplars are both potentially necessary for arriving at a unified percept of a category: to recognize a "dog" in our visual interaction with the world, our brain must understand both what a dog usually looks like (typicality), as well as what degree of deviation from this representation should place our percept outside of

that particular category.

Nevertheless, caution is necessary in interpreting these results, especially in dorsal stream regions: given that typicality is a subjective measure that subsumes multiple dimensions and features of object categories (including e.g. frequency of occurrence in the world and familiarity with such objects), the possibility exists that our findings may have been influenced by differences in the allocation of attentional resources across such dimensions (e.g. if participants paid more attention to blocks containing less familiar subordinate categories). However, our searchlight analysis identified regions where the category boundary effect (computed via the similarity of multi-voxel patterns) differs consistently between typical and atypical members of our categories, which indicates the presence of discriminable category information in these brain regions. Thus, if attention plays a role in our findings, then it would necessarily have to be operating on the category representations themselves, bringing within category members closer in neural space and pulling between category members apart. Additionally, previous work has shown that two parallel and hierarchically organized neural systems for object representation exist along the ventral and dorsal pathways [13, 75, 129, 134] and our results in cIPL are consistent with such an account.

Recent work has shown that distance from an inferred category boundary constructed from patterns of neural activation in human inferotemporal cortex can be used to successfully predict behavioral categorization [17, 111]. This distance-based model of category representation is consistent with our results in LOC, where we show that category boundaries are stronger between highly typical exemplars than between less typical exemplars, with the latter sitting farther from the category central tendency. Relatedly, many distance metrics have been previously employed for characterizing the similarity of neural patterns of activity in human visual cortex in general, and typicality in particular, ranging from overall cortical activity level [83, 104] to Pearson correlation (e.g. [28, 31, 58, 67]) and Euclidean or city block distance [9, 120]. Of these, we chose to focus on Pearson correlation as a straightforward, scale-invariant measure of similarity of neural patterns [28]. This is especially relevant, given that we perform a large-scale experiment using 64 real-world categories and prior evidence has shown that objects from different categories have the potential to

elicit consistently different univariate activity profiles both within and between brain regions (e.g. animate vs. inanimate categories [23, 76], small vs. big objects [76, 77]). Moreover, our decision is consistent with analyses used in many recent experiments investigating the underpinnings of object categorization and typicality in humans and non-human primates(e.g. [23, 31, 58, 67, 81]).

Several cognitive theories have been proposed that suggest we may expect real-world object categories to have a strong prototype-dominated cortical representation [9, 102], with typical exemplars closer in neural distance to the basic level prototype (category central tendency) and less typical exemplars generating a more distinct neural pattern of activation (i.e. larger neural distance from prototype). Indeed, previous work involving artificially constructed face stimuli suggests that both feature-based and neural distance from a category central tendency are usually correlated with perceived typicality [31, 83, 84, 120]. Prototype theory is typically contrasted with exemplar theory, which proposes that we represent categories with respect to several emblematic exemplars (or perhaps all exemplars) in each category, which serve to map that particular category's representational space [9, 100, 102]. This theory has also received some support; recent work has shown that exemplar models explain a comparable amount of variance in human performance on category generalization and prediction tasks [1] and even surpass prototype models in performance using data from humans and monkeys categorizing cartoon depictions of faces and fish [120]. In our work, we find brain areas that separately emphasize characteristics from both of these putative representational models, raising the possibility that the human brain may use both strategies for forming categories. First, we show that, in object-selective regions, typical categories are closer to the central category tendency and category boundaries are sharpened between typical and atypical exemplars, a finding that is consistent with the family resemblance hypothesis, as well as with a prototype-based encoding of category structure (but see [9] for an alternate explanation of how exemplar theory may also account for such a prediction). Conversely, we also find that atypical exemplars exhibit stronger category boundaries in cIPL. One potential explanation for this finding is that real-world categories, especially due to their inherent intra-class complexity, may not be fully or accurately captured by a

single prototype per category. Thus, while a prototype representation would imply that the intra-class distribution of subordinate categories within a basic is less important compared to the location of the category central tendency (i.e. prototype), by contrast an exemplar representation would predict a much heavier reliance on less typical subordinates for differentiating between basic categories, which may be the case in cIPL. Taken together, these two contrasting patterns of results suggest that the human brain may, in fact, use both exemplar and prototype models to structure category representations, albeit in different brain regions. Such a position could reconcile the seemingly contradictory behavioral and modeling results that have yet to eliminate either model as the sole framework for intra-category organization [83, 120]. Critically, our results provide clear evidence that LOC and cIPL are strong candidates for future investigations attempting to elucidate the contributions of these individual models in explaining the eventual emergence of perceptual typicality.

Over the past two decades, evidence has been uncovered for specific cortical regions selective for broad stimulus classes such as faces, scenes, objects, and bodies [35, 40, 72, 91], as well as organizational principles corresponding to broad attribute dimensions, including animacy [19, 23, 76, 81] and real-world object size [76, 77]. Furthermore, many studies have demonstrated that category information is recoverable from distributed representations [25, 38, 39, 58, 60, 61, 66], yet what constitutes a category representation in the high-dimensional space of neural patterns of activity is still poorly understood. Here, we show that perceived typicality, a high-level cognitive property of objects, directly modulates the representation of exemplars and categories fairly early in visual processing. Our results raise the possibility that the same theoretical principles that guide the cognitive formation of categories (cognitive usefulness and feature correlation constraints present in the environment [116]) may, in fact, fundamentally and sequentially guide the processing of visual input from its very early cortical stages. Indeed, previous work from our lab has already shown that this early link to cognition also holds for hierarchical organization of category structure, whose influence on the organization of neural patterns becomes apparent as early as lateral occipito-temporal cortex [67]. In the process of building category

representations, the inclusion of such principles would improve the utility and flexibility of eventually generated categories by emphasizing better boundaries between them and by allowing distinctions between individual exemplars and multiple levels of generality to emerge gradually from the neural representation. Furthermore, such principles constitute important signposts for recent work whose goal is to map the layers of deep learning models for visual categorization onto successive stages of the ventral visual hierarchy [16, 135, 136]. Most such computational models include few, if any, high-level cognitive constrains on their internal representation aside from categorization itself as an end-goal. Moving forward, we argue that attempts to build models of visual processing that more accurately mirror the human visual processing hierarchy would benefit from incorporating (either explicitly or at a verification stage) other high-level properties such as typicality, which we have presently identified as having a measurable impact on the feature spaces of visual regions strongly involved in object and category recognition (e.g. LOC).

Together, these findings solidify our understanding of how we define and describe boundaries between category representations in the brain, and moreover, put forward a new hypothesis for the organization and goals of intermediate visual processing: it is not simply focused on isolating and identifying primitives such as shapes, objects, or scenes, and their interplay, but also on employing cognitively relevant principles of category organization (of which typicality and hierarchical organization are two examples) to directly guide the development of the neural representation, for the ensuing purpose of improved and more flexible categorization, action, and cognition.

3.5 Acknowledgments

This work was funded by the William R. Hewlett Stanford Graduate Fellowship (to M.C.I.), the William and Adeline Hendess Phi Beta Kappa Graduate Fellowship (to M.C.I.), an NRSA Grant from the National Eye Institute NEIF32EY019815 (to M.R.G.), and a National Institutes of Health Grant 1 R01 EY019429 (to D.M.B and L.F.-F.).

Chapter 4

Category Boundaries and Typicality Warp the Neural Representation Space of Real-World Objects in Human Ventral Visual Cortex

Categories create cognitively useful generalizations by leveraging the correlational structure of the world. Although previous work has shown that object categories possess both hierarchical structure (basic- and entry-level effects [71, 116] and typicality structure [112, 115], still little is known about the neural underpinnings of how these processes help give rise to our object category structure.

To address this goal, we extend upon on our findings described in Chapters 2 and 3 to propose a new model of object processing in human ventral visual cortex based on the hypothesis that sequential computations across brain regions optimize specifically for cognitively useful aspects of category structure, such as the emergence of category boundaries and typicality gradients within a category. Using two fMRI experiments employing ninety-four object categories, we found strong evidence that between-category distinctions and within-category typicality structure both warped

neural representations sequentially across the ventral stream: category distinctions slowly pushed representations apart between early and mid-level visual areas and, simultaneously, perceived typicality of category members modulated the internal neural category space so that in later processing stages highly typical items became more similar to one another and less typical items were pushed away from the category central tendency. This suggests that eventual cognitive goals of visual categorization directly guide the feature transformations underlying sequential neural processing of visual input along the ventral visual stream hierarchy of brain regions from early visual cortex to inferotemporal cortex. This chapter is joint work with Michelle R. Greene, Diane M. Beck, and Fei-Fei Li.

4.1 Introduction

Categorization is a fundamental building block of cognitive experience whose goal is to generalize across similar objects and assign them a cognitively useful label. Within visually selective cortex, numerous regions show preferential activation for broad stimulus classes such as faces, scenes, objects, and bodies have been found across human visual cortex [35, 40, 53, 72, 91]. Furthermore, category information for scenes, faces, and various objects is widespread and linearly decodable across most visually selective cortex [23, 66, 67, 133], suggesting that neural activity across many of these regions may contribute to the ultimate goal of separating seemingly distinct visual stimuli into interpretable, actionable categories further down the processing stream.

When investigating the emergence of visual category information in the brain, an underlying hypothesis is that in the retina and early visual areas, categories start out as tangled surfaces in a high dimensional space of low-level features. As such, we can think of individual stimuli (such as my Chihuahua, Mr. Woof) being represented as points in a multidimensional space of neural activity patterns. A category (e.g. dog) then becomes a surface joining together the points corresponding to its members, with the implicit assumption that the resulting manifold is continuous. Concordantly, as we go up the ventral visual stream, we expect that sequential processing would slowly

disentangle these manifolds of representation such that regions of inferotemporal cortex and beyond are able to access and represent invariant category information[33, 44, 110, 126]. It is unknown, however, whether this process happens in a stepwise fashion across visual cortex, and if so, what are the transformations that occur at each step.

The prevalent view of object categorization suggests that information present in posterior occipito-temporal cortex does not reflect cognitive constraints (such as generating explicit and unequivocal category distinctions), which are instead enforced and instantiated later on in the processing stream (e.g. anterior temporal and frontal regions [49, 93, 95]). This perspective is also mirrored by models that strongly encapsulate vision from cognition [45, 48, 108, 110]. Here, we propose a competing model of object categorization where sequential computations in visually selective cortex optimize specifically for cognitively useful aspects of category structure. More specifically, we posit that the eventual cognitive goals of visual categorization directly guide the feature transformations underlying sequential neural processing along the ventral visual stream hierarchy of brain regions involved in object categorization.

To address this hypothesis, we focus our investigation on two distinct aspects of category structure and subsequently test how they modulate the representation of visual stimuli. First, categories arise cognitively such that they simultaneously maximize similarity between members of the same category and dissimilarity with members of other categories [109, 116]. If categorization as an end-goal drives aspects of this sequential computation, then we could quantify and measure this process stepwise across the ventral visual hierarchy: accordingly, we predict that objects belonging to the same category should become increasingly similar, while objects belonging to different categories should become increasingly distinctive, regardless of their low-level feature properties. Second, within a category, not all members are created equal and this has cognitive implications for categorization. Indeed, most concrete object categories are described internally by a graded typicality structure present among their members that directly influences speed and accuracy of recognition [112, 115]. Motivated by the cognitive organization of exemplars in a category according to their

perceived typicality, we predict that this graded representation should become increasingly apparent in the activity patterns elicited by category members as we measure stepwise changes in neural representation going up the ventral visual hierarchy.

To test these predictions, we conducted two fMRI experiments in which participants were shown color photographs of 15 subordinate level categories from each of two basic level categories (Experiment 1: dogs and cars), and 8 subordinate-level categories from each of eight basic level categories (Experiment 2: birds, cats, dogs, fish, airplanes, boats, cars, trains). Typicality for each subordinate within its basic category was also assessed behaviorally. We then used several multi-voxel pattern analyses to measure whether the multidimensional neural representation of each of these categories warps in a principled way in relationship to their cognitive structure across the span of the human ventral visual processing hierarchy.

In both of our experiments, we found strong evidence that both aspects of category structure we investigated warped the neural representation directly and sequentially across the ventral stream: category distinctions slowly pushed representations apart as we moved between early and mid-level visual areas, and simultaneously, perceived typicality of category members rearranged the internal neural category space so that in later processing stages highly typical items became more similar to one another and less typical items were pushed away from the category central tendency.

4.2 Materials and Methods

4.2.1 Experiment 1: Two Basic Level Categories - Thirty Subordinate Categories

4.2.1.1 Constructing a Normed Category Set

The goal of our study was to investigate whether the cognitive structure of our category space exerts a measurable influence on the sequential processing of stimuli in visual cortex. To test this hypothesis, we focused on category distinctions and

perceived typicality as critical elements of human category structure elements. Accordingly, we first chose two distinct basic level categories which are well differentiable based on the neural patterns of activity they elicit in visual cortex: dogs and cars [68]. From each of these two basic level categories, we first obtained typicality ratings for twenty-four subordinate level categories (e.g. pug, jeep) from each basic-level category. We then used these ratings to select fifteen subordinates from each basic category grouped into three tiers: five highly typical subordinates (dogs: golden retriever, beagle, Saint Bernard, mastiff, collie; cars: Ford Mustang, Chevrolet Crossfire, BMW Z4, Rolls Royce, Lamborghini), five middle typicality subordinates (dogs: Doberman, pug, schnauzer, sheepdog, schipperke; cars: Cadillac, Mini Cooper, Mitsubishi Miev, Land Rover, Nissan Cube), and five low typicality subordinates (dogs: Airedale, poodle, Chihuahua, Afghan hound, Komondor; cars: antique car, Jeep Wrangler, Ford Ranger, limousine, Hummer). Typicality ratings for all twenty-four subordinate categories within their basic level category are shown in Appendix C. Subsequently, we used ImageNet [32] and Google image search to collect 28 distinct images containing objects of interest from each of our resulting thirty subordinate level categories for the purpose of showing these pictures to participants during an fMRI experiment; i.e. for the subordinate category "pug", we obtained 28 distinct photographs of pugs. Pictures were cropped to feature the objects prominently and centrally within a square region (400 x 400 pixels in size) and included their natural background. Within each subordinate category, the images varied greatly in color and pose. Representative images from each of our 30 categories, together with their respective typicality ratings are shown in Fig. 4.1 A. Representative images from our initial set of 48 categories, together with their respective typicality ratings are shown in Appendix C, Fig. C.1.

4.2.1.2 Behavioral Experiment: Typicality Rankings

Participants and Materials

40 participants were recruited on Amazon's Mechanical Turk platform (AMT) from a pool of trusted US-based participants with at least 2,000 previously accepted AMT

results at a minimum of 98% approval. Participants completed the study from their own personal computing device.

Experimental Procedure

Each of the AMT hits contained 300 trials comprising each possible pairwise comparison between the twenty-four subordinate categories within a particular basic category. In each trial, participants viewed a randomly drawn image from two subordinate categories and were asked to indicate by clicking which image was the most typical of its corresponding basic category. Participants were compensated $0.50 per hit and each hit took an average of 925 seconds to complete.

Data Analysis

Pairwise typicality rankings for the twenty-four subordinates in each basic category were obtained. We computed the percentage of times each subordinate was chosen as the more typical item in a pair and used this quantity to order subordinates according to their typicality in each basic category independently.

4.2.1.3 fMRI Experiment

Participants

14 volunteers (5 females, ages 21–34, including authors M.C.I. and M.R.G.) with no past history of psychiatric or neurological disorders and normal or corrected-to-normal vision participated in this experiment. Participants gave informed written consent in compliance with procedures approved by the Stanford University Institutional Review Board. Except for the participating authors, all subjects received financial compensation.

Scanning Parameters and Preprocessing

Imaging data were acquired with a 3 Tesla G.E. Healthcare scanner. A gradient echo, echo-planar sequence was used to obtain functional images (volume repetition time (TR) 2 s; echo time (TE) 30 ms; flip angle, 77 degrees; matrix, 80 x 80 voxels; FOV

23.2 cm; 42 oblique 2.9 mm slices; in-plane resolution 2.9 x 2.9 mm). We also collected a high-resolution (0.9 x 0.9 x 0.9 mm voxels) structural scan (BRAVO; TR 7.24 ms; TE 2.78 ms; flip angle, 12 degrees) in each scanning session. The functional data were spatially aligned to compensate for motion during acquisition and each voxel's intensity was converted to percent signal change relative to the temporal mean of that voxel using the AFNI software package [26]. To perform our analyses, we computed the average voxel activity for each block (see below for block design details). We did not use a GLM analysis and did not perform any smoothing.

Experimental Procedure

Images were presented centrally subtending 12 x 12 degrees of visual angle and were superimposed on an equiluminant gray background using the PsychToolbox [12, 106] extension of MATLAB (Mathworks, Natick, MA). We used an LCD display (Resonance Technology) operating at a resolution of 640 x 480 at 240 Hz, visible from a mirror within the head-coil. Participants performed one session comprising 8 runs, with 15 blocks per run and 8 images per block. Each block consisted of a 500 ms fixation cross presented centrally, followed by 8 consecutive stimulus presentations from the same subordinate level category, with a 12 s gap between the blocks. Each image was presented for 160 ms, followed by a 590 ms blank gray screen. Subjects were asked to maintain fixation at the center of the screen, and respond via button-press whenever an image was repeated (one-back task, 0–2 repetitions per block, totaling 28 unique image presentations and 4 repetitions per subordinate category per subject during the experiment). Over the course of the experiment, each participant viewed 4 blocks from each of the subordinate level categories. The order of blocks, the number of repetitions in each block, and the images in each block were counter-balanced across runs and between subjects.

Regions of Interest (ROIs)

The positions and extents of each participant's functional ROIs (LOC, TOS, PPA, and FFA) were obtained using standard localizer runs conducted in a separate fMRI session. For functional ROIs, subjects observed two runs, each with 12 blocks drawn

equally from six categories: child faces, adult faces, indoor scenes, outdoor scenes, objects (abstract sculptures with no semantic meaning), and phase-scrambled objects. Blocks were separated by 12 s fixation cross periods and comprised 12 image presentations, each of which consisted of images presented for 900 ms, followed by a 100 ms fixation cross. Each image was presented exactly once, with the exception of two images during each block that were repeated twice in a row. Subjects were asked to maintain fixation at the center of the screen and respond via button press whenever an image was repeated. To avoid any issues related to intrinsic variability in signal reliability across our participant pool, we selected fixed-volume ROIs across all our participants. The volume of each region in mm^3 was chosen conservatively, based on sizes previously reported in the literature, accounting for resolution differences between studies [51, 67, 68, 133]: LOC: 210 voxels (5,100 mm^3); TOS: 85 voxels (2,100 mm^3); PPA: 125 voxels (3,000 mm^3); FFA: 50 voxels (1,200 mm^3). LOC was defined as the top 210 contiguous voxels bilaterally near the inferior occipital gyrus that responded to an Objects > Scrambled Objects GLM contrast. PPA was defined as the top 125 contiguous voxels bilaterally near the parahippocampal gyrus that responded to a Scenes > Objects GLM contrast. TOS was defined as the top 85 contiguous voxels bilaterally near the trans-occipital sulcus that responded to a Scenes > Objects GLM contrast. FFA was defined as the top 50 contiguous voxels bilaterally near the fusiform gyrus that responded to a Faces > Objects GLM contrast.

To determine the locations of early visual areas V1, V2, V3v, and hV4, we used a standard retinotopic mapping protocol in a separate experiment, in which a checkerboard pattern undergoing contrast reversals at 5 Hz moved through the visual field in discrete increments [119]. First, a wedge subtending an angle of 45 degrees from fixation was presented at 16 different polar angles for 2.4 s each. Next, an annulus subtending 3 degrees of visual angle was presented at 15 different radii for 2.4 s each. Each subject passively observed two runs of 6 cycles in each condition, yielding 512 timepoints per subject. The locations and extents of early visual areas were delineated on a flattened cortical surface for each subject, using a horizontal vs. vertical meridian general linear test, which gave the boundaries between retinotopic maps.

We aligned the positions of the ROIs to the experimental sessions using the AFNI

software package [26], first aligning the structural scans between sessions with sub-millimeter precision, and then applying the alignment transformation to the ROI positions. Percent signal change was then extracted for each voxel in each ROI and these vectors were submitted to the multi-voxel pattern analyses described next.

4.2.1.4 fMRI Data Analysis

Category Warping

For each ROI, we computed the correlation distance (1 - Pearson's r) between the average voxel-wise patterns elicited by each pair of subordinate categories. Average patterns for a subordinate category were obtained by z-scoring the percent signal change vectors described above and averaging them across all blocks corresponding to that particular subordinate category (e.g. the four blocks in which pugs were shown to participants in the scanner). To examine whether the range of correlation (as a proxy for the span of the representational space) changes as we move up the ventral visual stream, we first computed and plotted the ranges of the raw correlation values for within-category distances (e.g. between two breeds of dogs or two types of cars) and between-category distances (e.g. between a breed of dog and a type of car). Subsequently, to investigate the relative change in neural distances across brain regions, we generated histograms of z-scored within- and between-category distances for each ROI. Finally, to investigate how distributions of distances change across brain regions, for each pair of ROIs, we plotted the z-scored distances corresponding to each pair of subordinate categories against one another on a 2D plot. Here, when a point representing a subordinate category pair is above the diagonal, then the distance between those two subordinates increases between the representational spaces of the X-axis brain region and the Y-axis brain region. Similarly, when a point is below the diagonal, the distance it represents decreases between the representational spaces of the X-axis brain region and the Y-axis brain region. Accordingly, for each plot, we computed a "category warping index" that measures how many subordinate category pairs sit above versus below the diagonal in the graph plots corresponding to each pair of ROIs as the difference between these quantities.

Typicality Warping

Analogous to the previous analysis, we computed the z-scored correlation distances (1 - Pearson's r) within and between the high typicality and low typicality tiers of subordinate categories in each of our two basic categories. We generated histograms for these distances in each ROI and plotted these distances in each pair of ROIs against each other. Finally, following a similar reasoning to above, we also generated a "typicality warping index" that measures how many subordinate category pairs sit above versus below the diagonal in the graph plots corresponding to each pair of ROIs as the difference between these quantities.

4.2.2 Experiment 2: Eight Basic Level Categories - Sixty-Four Subordinate Categories

Experiment 1 used two basic level categories, each comprising fifteen subordinate categories, to investigate whether the cognitive structure of our category space exerts a measurable influence on the sequential processing of stimuli in visual cortex. Our environment, however, contains thousands of distinct object categories [11, 32]. To show that our findings represent a generalizable principle of category representation in visual cortex, one that would be applicable beyond our choice of stimuli in Experiment 1, we conducted a second experiment where we constructed a larger and much more varied stimulus set comprising eight basic level categories (birds, cats, dogs, fish, boats, cars, planes, trains), which are well differentiable based on the neural patterns of activity they elicit in visual cortex (Iordan et al. in press) and which span natural, man-made, animate, and inanimate superordinate category boundaries. From each of these eight basic categories, we chose eight subordinate level categories (e.g. pug, jeep) and used an identical behavioral experiment as in Experiment 1 to obtain typicality ratings for each subordinate category within its basic. We then used these ratings to group the subordinates from each basic category into two tiers: four highly typical subordinates (e.g. cats: Egyptian, Angora, Manx, Abyssinian) and four low typicality subordinates (e.g. cats: Tortoiseshell, Siamese, Persian, Sphinx). Typicality rankings and full names for all sixty-four subordinate categories are shown in Appendix C.

Subsequently, we used ImageNet [32] and Google image search to collect 16 distinct images containing objects of interest from each of our resulting sixty-four subordinate level categories for the purpose of showing these pictures to participants during an fMRI experiment; i.e. for the subordinate category "pug", we showed 16 distinct photographs of pugs. Pictures were cropped to feature the objects prominently and centrally within a square region (400 x 400 pixels in size) and included their natural background. Within each subordinate category, the images varied greatly in color and pose. Representative images from each of our sixty-four categories, together with their respective typicality ratings are shown in Fig. 4.1 C.

4.2.2.1 Behavioral Experiment: Typicality Rankings

Participants and Materials

40 participants were recruited on Amazon's Mechanical Turk platform (AMT) from a pool of trusted US-based participants with at least 2,000 previously accepted AMT results at a minimum of 98% approval. Participants completed the study from their own personal computing device.

Experimental Procedure and Data Analysis

Analogous to Experiment 1.

4.2.2.2 fMRI Experiment

Participants

10 volunteers (4 females, ages 23–31, including authors M.C.I. and M.R.G.) with no past history of psychiatric or neurological disorders and normal or corrected-to-normal vision participated in this experiment. Participants gave informed written consent in compliance with procedures approved by the Stanford University Institutional Review Board. Except for the participating authors, all subjects received financial compensation.

Scanning Parameters and Preprocessing

Imaging data were acquired with a 3 Tesla G.E. Healthcare scanner. A gradient echo, echo-planar sequence was used to obtain functional images (volume repetition time (TR) 2 s; echo time (TE) 30 ms; flip angle 80 degrees; matrix 128 x 128 voxels; FOV 20 cm; 29 oblique 3 mm slices with 1 mm gap; in-plane resolution 1.56 x 1.56mm). We also collected a high-resolution (1 x 1 x 1 mm voxels) structural scan (SPGR; TR 5.9 ms; TE 2.0 ms; flip angle 11 degrees) in each scanning session. The functional data were spatially aligned to compensate for motion during acquisition and each voxel's intensity was converted to percent signal change relative to the temporal mean of that voxel using the AFNI software package [26]. To perform our analyses, we computed the average voxel activity for each block. We did not use a GLM analysis and did not perform any smoothing.

Experimental Procedure

Images were presented centrally subtending 21 x 21 degrees visual angle and were superimposed on an equiluminant gray background. We used an LCD display (Resonance Technology) operating at a resolution of 640 x 480 at 240 Hz, visible from a mirror within the head-coil. Participants performed 8 runs, with 16 blocks per run and 8 images per block Each block consisted of a 500 ms fixation cross presented centrally, followed by 8 consecutive stimulus presentations from the same subordinate level category, with a 12 s gap between the blocks. Each image was presented for 160 ms, followed by a 590 ms blank gray screen. Subjects were asked to maintain fixation at the center of the screen, and respond via button-press whenever an image was repeated (one-back task, 0-2 repetitions per block). Over the course of the experiment, each participant viewed 2 blocks from each of the subordinate level categories. The order of blocks, the number of repetitions in each block, and the images in each block were counter-balanced across runs and between subjects.

Regions of Interest (ROIs)

The positions and extents of each participant's functional ROIs (LOC, TOS, PPA, and FFA) were obtained using standard localizer runs conducted in a separate fMRI

session. The details are analogous to Experiment 1. The volume of each region in mm^3 was chosen conservatively, based on sizes previously reported in the literature, accounting for resolution differences between studies [51, 67, 68, 133]: LOC: 500 voxels; TOS: 200 voxels; PPA: 300 voxels; FFA: 100 voxels. To determine the locations of early visual areas V1, V2, V3v, and hV4, we used a standard retinotopic mapping protocol in a separate experiment, in which a checkerboard pattern undergoing contrast reversals at 5 Hz moved through the visual field in discrete increments [119], analogously to Experiment 1.

We aligned the positions of the ROIs to the experimental sessions using the AFNI software package [26], first aligning the structural scans between sessions with submillimeter precision, and then applying the alignment transformation to the ROI positions. Percent signal change was then extracted for each voxel in each ROI and these vectors were submitted to the multi-voxel pattern analyses described next

4.2.2.3 fMRI Data Analysis

Category Warping and Typicality Warping

All analyses were performed analogously to Experiment 1. To determine category warping effects, we measured distances within each basic category (e.g. pug and malamute) and compared them to distances between subordinates belonging to different basic categories (e.g. pug and jeep). To determine typicality warping effects, we measured distances between the four most typical subordinates in each basic category (see Behavioral Experiment details above) and compared them to distances between the four least typical subordinates in each basic category.

4.2.3 Statistical Analyses

For all our analyses, we used paired two-tailed t-tests when comparing observed effects against chance and when establishing whether a significant difference exists between two observed effects or between the means of two distributions. We used Kolmogorov-Smirnov tests to establish that no significant deviation from normality exists for the distributions of all effects to which t-tests were applied. Because statistical tests were

applied to a single number derived from the pattern of voxels within an ROI per condition of interest, and these conditions are relatively few, we did not correct for multiple comparisons within our ROI analyses. All statistical tests were implemented in MATLAB.

4.3 Results

4.3.1 Experiment 1: Two Basic Level Categories - Thirty Subordinate Categories

4.3.1.1 Category Representations Become More Separable Across the Ventral Visual Stream

The prevalent model for representing and processing visual information in visual cortex posits that object categories start out as tangled surfaces in a high dimensional space of low-level features and sequential processing across the ventral stream slowly disentangles these manifolds of representation such that regions of inferotemporal cortex and beyond are able to access and represent invariant category information [33, 44, 110, 126]. It is unknown, however, whether this process happens in a stepwise fashion across visual cortex, and if so, what are the transformations that occur at each step.

To address this question, we conducted a passive-viewing fMRI experiment where we showed participants pictures from two basic categories (dogs and cars), each comprising 15 subordinate categories with 28 distinct images per subordinate that varied substantially in pose and color (Fig. 4.1 A). Using average subordinate-category level responses elicited by our stimuli across the visual cortex of our participants, we first set out to test whether the a central tenet of most categorization models is upheld: do neural representations of categories become more separable as we go up the ventral visual stream, moving from feature spaces that favor low-level properties (e.g. V1) to feature spaces that favor higher, more abstract properties of the input (e.g. hV4, object-selective cortex LOC).

Given that we have little insight into the composition and principles guiding the

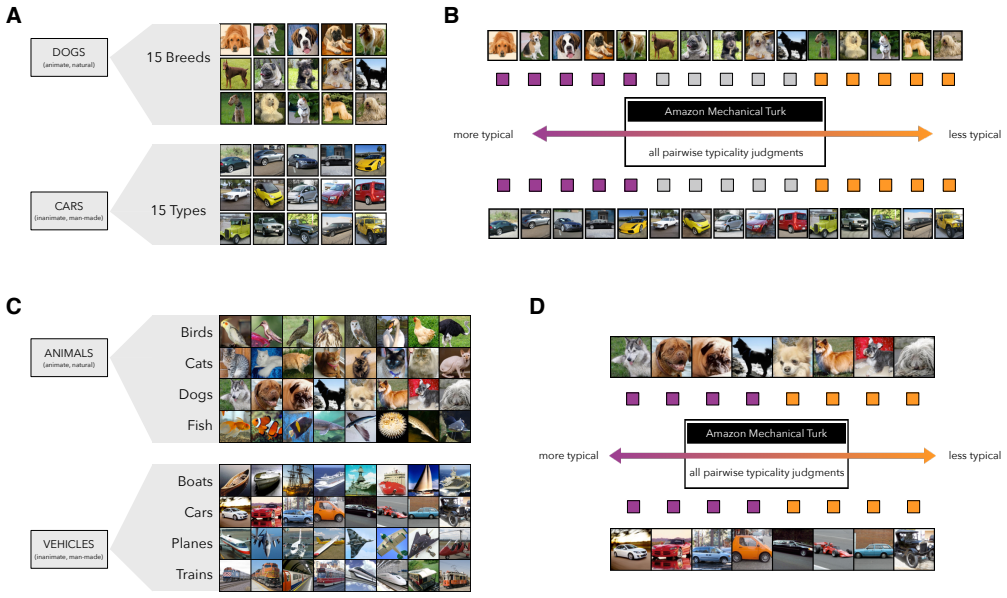


Figure 4.1: **Stimulus Sets and Corresponding Typicality Rankings.** (A) The Experiment 1 stimulus set comprised 15 subordinate categories from each of 2 basic level categories (dogs and cars). Participants were shown 28 images per subordinate, varying in pose and color (only one representative image shown for each subordinate). (B) Typicality rankings for Experiment 1 were obtained using a behavioral experiment conducted on the Amazon Mechanical Turk crowd-sourcing platform. Within each basic category, subordinates are ordered according to typicality from the most typical (golden retriever and BMW Z4 on left) to the least typical (Komondor and Hummer on right). (C-D) The Experiment 2 stimulus set comprised 8 subordinate categories from each of 8 basic level categories (birds, cats, dogs, fish, boats, cars, planes, and trains). Participants were shown 16 images per subordinate, varying in pose and color (only one representative image shown for each subordinate). Typicality rankings for Experiment 2 were obtained using a behavioral experiment conducted on the Amazon Mechanical Turk crowd-sourcing platform. Within each basic category in part C, subordinates are ordered according to typicality from the most typical (e.g. malamute on left) to the least typical (Komondor on right). Categories marked with purple squares in panels (B) and (D) were used as high typicality subordinates in the subsequent "typicality warping" analyses. Similarly, subordinates marked with orange squares in panels (B) and (D) were used as low typicality subordinates in the same analyses.

organization of these feature spaces across virtually all regions in the ventral stream, it would be difficult to attempt to model such spaces directly (although attempts have been made to establish relationships between distinct layers of neural network models and visual processing stages, see e.g. [63]. Instead, we reasoned that if activity patterns elicited by distinct categories in a brain region can be thought as points in a multi-dimensional space (e.g. the space defined by all the voxels in that region), then we can focus on measuring the similarity between representations of such categories as a proxy for gaining insights about how the space is organized within that particular cortical region. Consequently, we computed the Pearson correlation between patterns of activity corresponding to each pair from our thirty subordinate categories across multiple ventral visual brain regions known to be involved in representing information about object category representations [66, 67]: early visual cortex (V1, V2, V3v, hV4), object-selective (lateral occipital complex LOC), scene-selective (parahippocampal place area PPA, trans-occipital sulcus TOS / OPA), and face-selective areas (fusiform face area FFA). Consistent with recent prior work [63], we found that although we expect category representations to become more invariant as we move up the ventral stream, the absolute range of the similarity space remains at least as large or slightly increases in intermediate-level object selective regions, compared to early visual regions (V1: r range = 1.22 $\pm$ 0.12 r; LOC: r range = 1.37 $\pm$ 0.15; LOC > V1: $t_{13} = 3.6$, $p = 0.003$) (Fig. 4.2 A). This suggests that visual processing doesn't exclusively emphasize generalization of representation across the span of each category, but instead maintains and perhaps enhances specificity of information extracted from the visual input at the level of each individual category.

Nevertheless, a central tenet of both our model, as well as many previous ones [33, 44, 110, 126] is that as feature spaces become more complex going up the ventral visual hierarchy, stimuli belonging to the same category should become increasingly similar in how they are represented in patterns of activity, while simultaneously becoming more dissimilar to stimuli from other categories. To test this hypothesis, we computed the similarity distances (1 - Pearson's r) between all pairs of subordinates, conditioned on whether the two subordinates in a pair belonged to the same basic category (Wth, e.g. pug and Chihuahua) or to distinct basic categories (Btw, e.g. pug and jeep). To

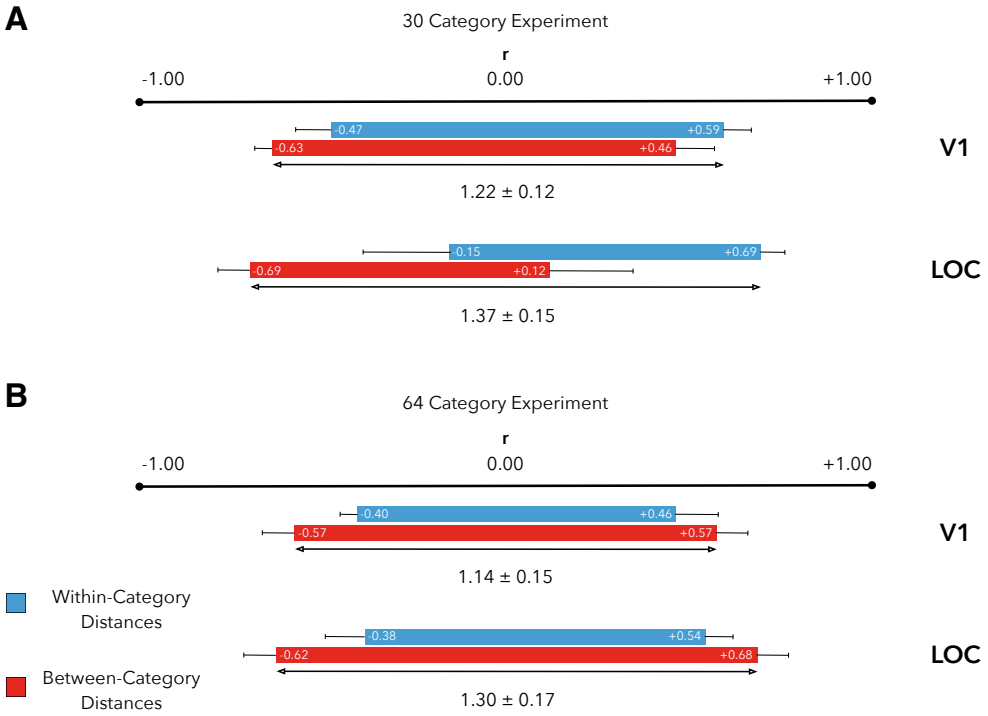


Figure 4.2: **Correlation Ranges for Within- and Between-Category Distances in V1 and LOC.** Pearson correlation ranges for within-category distances (blue) and between-category distances (red) for V1 and LOC. Consistent with prior work [63], as we move up the ventral stream, the absolute range of the similarity space remains at least as large or slightly increases in intermediate-level object selective regions (LOC), compared to early visual regions (V1). (A) Experiment 1: 30 subordinate categories. (B) Experiment 2: 64 subordinate categories.

normalize between different ranges of similarity across brain regions, we z-scored the similarity distances within each brain region and compared the histograms of these within-basic-category and between-basic-category distances both within brain regions (which is indicative of the nature of the local representation), as well as between brain regions (which gives as a measure of the effect each subsequent computation has on the organization of the feature space). The resulting histograms of distances are shown in Fig. 4.3. Beginning in V1, we already see a reasonable degree of separation and a significant difference between the means of the two distributions (V1: Wth < Btw mean diff. = 0.9, $t_{13} = 8.2$, $p < 0.001$). This suggests that our two basic categories (dogs and cars) are already quite distinct even in the space of low-level features. At the next step of visual computation, V2, the distributions of distances look very similar to V1 and, again, we see a significant difference between the mean within-category and mean between-category distances (V2: Wth < Btw mean diff. = 1.1, $t_{13} = 8.8$, $p < 0.001$). Interestingly, however, when we examine the two endpoints of the visual object ventral stream pathway, namely how the representation changes between V1 and object-selective cortex, LOC, we immediately see a striking difference between these two regions: the within- and between-category distances are much more strongly separated in LOC than V1, which is consistent with patterns of activity organizing by category much better in LOC than V1 (LOC: Wth < Btw mean diff. = 1.7, $t_{13} = 19$, $p < 0.001$). By contrast, measurements in intermediate and scene-selective visual regions show a similar representation to early visual regions (hV4: Wth < Btw mean diff. = 1.0, $t_{13} = 7.9$, $p < 0.001$; PPA: Wth < Btw mean diff. = 0.8, $t_{13} = 9.6$, $p < 0.001$; TOS: Wth < Btw mean diff. = 0.8, $t_{13} = 6.2$, $p < 0.001$). This suggests that a sharp qualitative change in the structure of the feature space may arise between hV4 and LOC, which is not mirrored in other stimulus selective regions of occipito-temporal cortex. A potential exception to this trend may be FFA, which shows an intermediate degree of increased separation between the within- and between-category distances, compared to V1 and LOC (FFA: Wth < Btw mean diff. = 1.4, $t_{13} = 15$, $p < 0.001$), although this effect may be driven by the presence of animal faces in the dog basic category. Taken together, our results suggest that the human visual system likely constructs (or, at the very least, strongly amplifies) feature descriptions of our

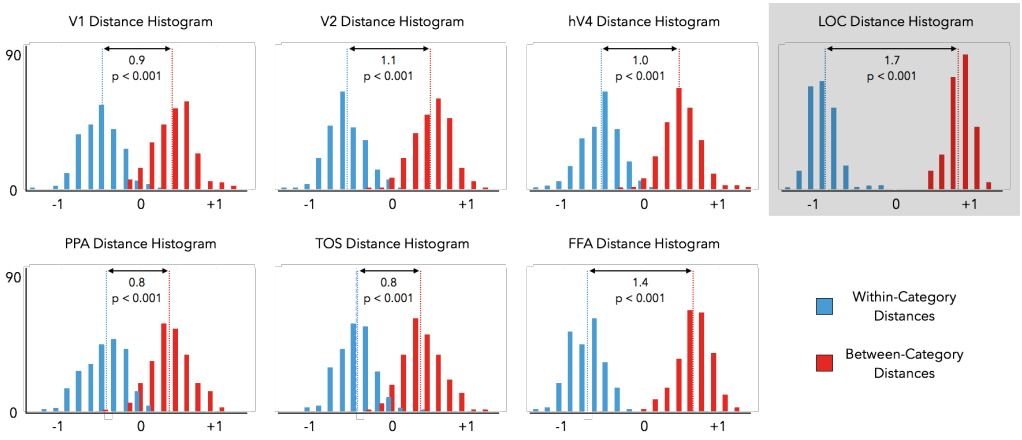


Figure 4.3: **Experiment 1 Category Distance Histograms.** Graphs show z-scored Pearson correlation distance histograms for within-category distances (blue) and between-category distances (red) for early visual (V1, V2, hV4), object-selective (LOC), scene-selective (PPA, TOS), and face-selective (FFA) regions. The basic categories "dog" and "car" are reasonably separable in virtually all brain regions considered with the highest distinction arising in LOC (top right, grey). This suggests that a sharp qualitative change in the structure of the feature space may arise between hV4 and LOC, which is not mirrored in other stimulus selective regions of occipito-temporal cortex.

visual input that correlate with category boundaries.

4.3.1.2 Proposed Model of Category Disentanglement

Our initial findings above provide evidence for our hypothesis that as we move up the ventral visual stream the complexity of the feature spaces increases and in these subsequent spaces categories become more internally cohesive (lower within-category distances) and more mutually distinctive (higher between-category distances). This is consistent with findings showing that categories become more easily separable in object-selective regions of inferotemporal cortex (e.g. LOC), compared to early visual regions (e.g. V1) [63, 66, 67]. However, it remains unclear how the within- and between-category information change step by step as visually selective brain regions increasingly farther from V1 process visual stimuli. Moreover, a common assumption of object categorization frameworks suggests that information present in early visual

regions and posterior occipito-temporal cortex does not reflect cognitive constraints, which are instead enforced and instantiated later on in the processing stream (e.g. anterior temporal and frontal regions [49, 93, 95]). This perspective is also mirrored by models that strongly encapsulate vision from cognition [47, 108, 110]. Here, we propose a competing model of object categorization where sequential computations in visually selective cortex optimize specifically for cognitively useful aspects of category structure. More specifically, we posit that the eventual cognitive goals of visual categorization directly guide the feature transformations underlying sequential neural processing along the ventral visual stream hierarchy of brain regions involved in object categorization.

One of the most straightforward predictions of our model is that that as feature spaces become more complex going up the ventral visual hierarchy, we should observe stimuli belonging to the same category become increasingly similar in how they are represented in patterns of activity, while simultaneously becoming more dissimilar to stimuli from other categories, which is directly supported by our previously described within- and between-category distance histograms results (Fig. 4.3). Furthermore, if categorization as an eventual goal of the processing underlying the hierarchy of visual brain regions indeed drives aspects of this processing, it may play a role not only in the grouping of categorical stimuli within each feature space, but also in the manner in which the feature spaces themselves are organized. To address this possibility, we examined how the relative organization of the representational spaces changes across the ventral visual stream by plotting the distances elicited by the same pairs of subordinates (and basic categories) in two different brain regions against each other (Fig. 4.4).

Using this framework and assuming a straightforward sequential increase in distinguishability of category representations (i.e. categories slowly become more separable, but the different feature spaces across visual cortex represent them in a consistent manner), we can put forward an initial prediction of how we expect category representations to evolve as we move up the ventral visual stream (Fig. 4.5). Here, we propose that categories would start out partially overlapping, and this representation would be slow to change in the first few stages of visual processing (e.g. going

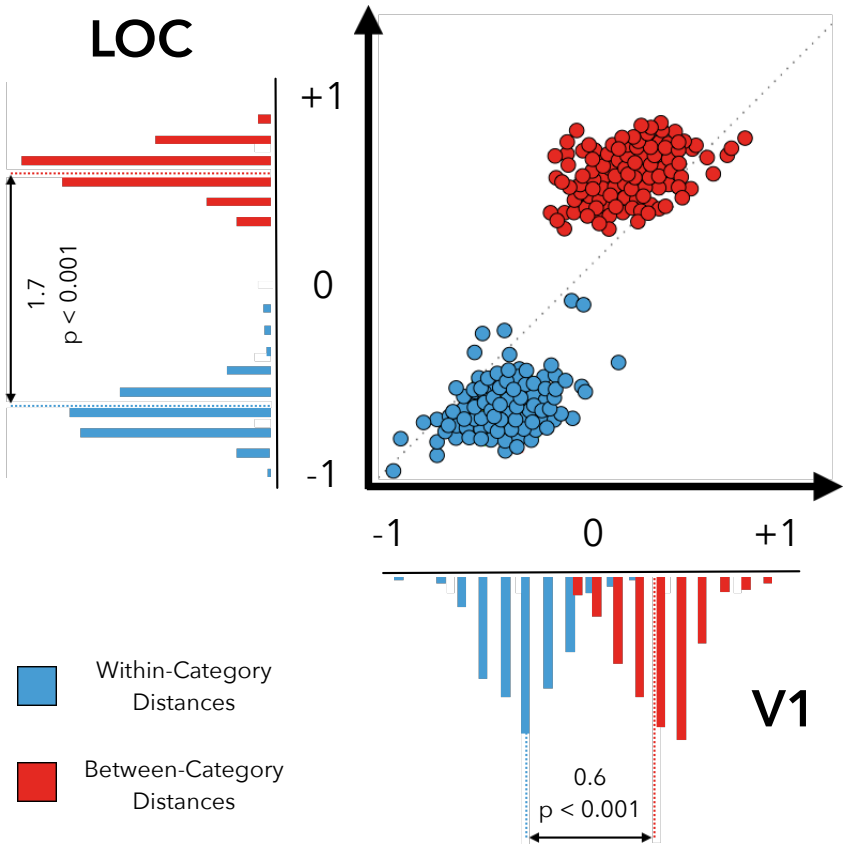


Figure 4.4: **Evolution of Relative Category Distances across Brain Regions.** Each pair of subordinate categories is plotted as a point in a two dimensional representation, where the X and Y axes are defined as the Pearson correlation distance between the two subordinates in each of two separate brain regions (in the example above: V1 and LOC). Projecting the resulting distribution onto either of the axes recovers the corresponding category distance histogram for that particular brain region represented on the axis (cf. Fig. 4.3). By examining the position of the subordinate category pairs (i.e. points in the graph) relative to the diagonal, we can identify similarities and differences between the representational spaces of the two brain regions. For example, if all points are close to the diagonal, then representations change very little between the two brain areas; however, if there is a significant deviation from the diagonal, then this indicates that the representational space changes in a principled way from one brain are to another (as seen above between V1 and LOC; see text for details).

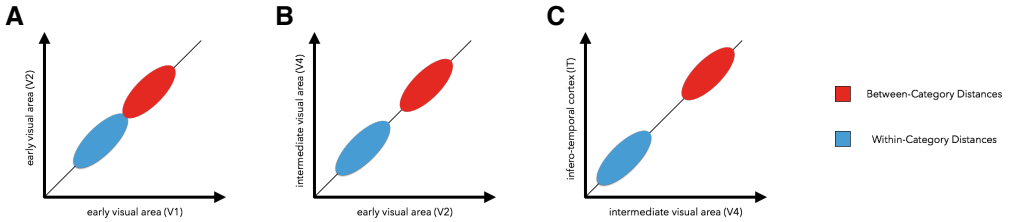


Figure 4.5: **Initial Model for Evolution of Category Representations across Ventral Visual Stream.** We propose that categories would start out partially overlapping, mainly due to overlap in low-level features (A). As we move up the ventral visual stream, computations in successive intermediate visual brain regions would contribute to incrementally shrinking the distances within categories and expanding the distances between categories (B). Finally, at the apex of ventral stream computation (inferotemporal cortex), this process reaches its peak in generating fully dissociable category representations with the least amount of distribution overlap (C).

from V1 to V2, Fig. 4.5 A). Then, as we proceed up the ventral stream, sequential computations would incrementally shrink distances within and expand distances between eventual categories (e.g. hV4, Fig. 4.5 B). Finally, in inferotemporal cortex (e.g. LOC, Fig. 4.5 C), distances between objects in different categories become significantly larger than distances between objects within the same category and the two distributions of distances are strongly separable.

Going forward, we tested the predictions of this model using our set of thirty subordinates and two basic level categories. In the resulting graphs, if a point is close to the diagonal, then the distance between the two subordinates it represents is highly similar across the two visual areas being compared. However, any potential deviation from the diagonal indicates a relative shift in the distance between that pair of subordinates as we go from one brain area to another. To foreshadow our upcoming results, the presence of consistent shifts across the ventral visual hierarchy correlated with category boundaries between our stimuli would provide evidence for our hypothesis that the goal of eventually building categories (i.e. the organization of responses elicited by visually distinct stimuli into coherent, self-similar groups of points close together in the corresponding multidimensional space) influences the manner in which the intermediate feature spaces are organized.

4.3.1.3 Category Boundaries Warp Neural Distances in Occipito-Temporal Cortex

Our main hypothesis states that as feature spaces become more complex going up the ventral visual hierarchy, we should observe stimuli belonging to the same category become increasingly similar in how they are represented in patterns of activity, while simultaneously becoming more dissimilar to stimuli from other categories. Using the between-brain-region analysis (Fig. 4.6), we see that this change is slow to occur as processing starts out in early visual regions. In the V1 - V2 plot, our two conclusions from earlier become immediately apparent: distances within categories (in blue) are usually smaller than distances between categories (in red). Moreover, virtually all the points sit close to the diagonal, which suggests that the representations of these categories are quite similar in the feature spaces of V1 and V2, which is recognizable from comparing the histograms of within- and between-category distances in Fig. 4.3.

Next, by looking across a larger extent of the object processing pathway, namely how the representation changes between V1 and object-selective cortex, LOC (Fig. 4.6, gray box), we see a striking difference from the V1 - V2 step: not only are distance distributions farther apart, but they each sit on different sides of the diagonal. In the large step between V1 - LOC, within-category distance pairs sit below the diagonal, thus within-category distances in LOC are smaller relative the span of their representation space, compared to the within-category distances in V1. Essentially, the feature space in LOC, whatever its properties and primitives, shrinks relative distances between a pair of dogs or a pair of cars compared to their original representation in the feature space of V1. Conversely, between-category distances sit above the diagonal, so they are relatively larger in the representational space of LOC than in V1. This indicates that the feature space in LOC expands the relative distance between subordinates belonging to separate categories (e.g. a dog and a car) compared to their original V1 representation. This process is akin to a principled relative warping of the representational space between V1 and LOC towards grouping information in this space such that category boundaries are emphasized. To measure this change in the structure of the representational space, we first computed the proportion of points

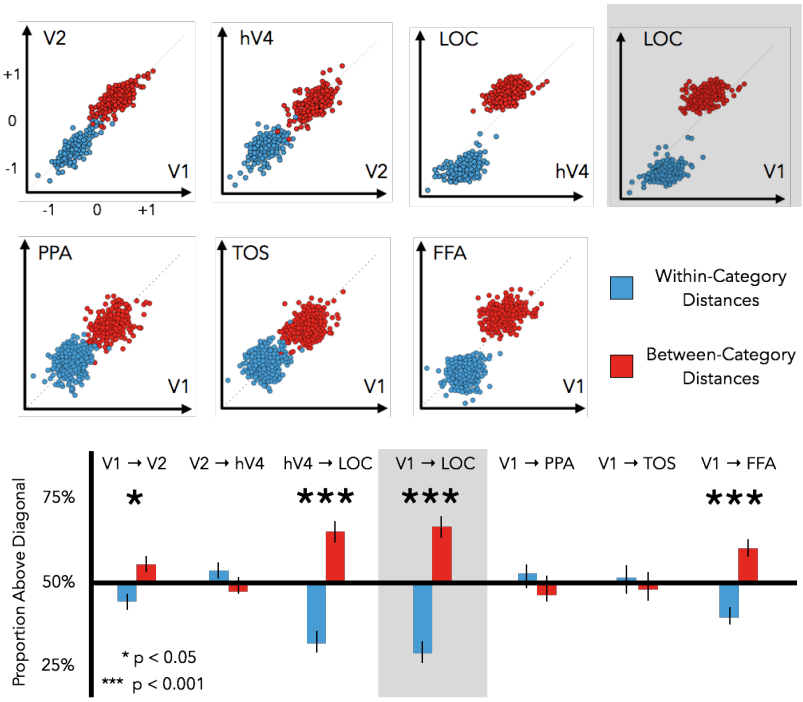


Figure 4.6: **Category Boundaries Warp Neural Representations in Occipito-Temporal Cortex.** (Top, Middle) Graphs show how representations of distances corresponding to subordinate category pairs change as we move up the ventral visual stream. Axes represent z-scored distances between pairs of categories in the corresponding brain region. Representations are relatively stable between early visual regions (V1, V2, hV4), as well as between early visual cortex and scene-selective regions (PPA, TOS). However, we see a striking shift in the quality of the representation as we move between hV4 and LOC. Here, within-category distance pairs lie below the diagonal, while between-category distance pairs sit above the diagonal, which indicates that the feature space of LOC shrinks relative distances within categories and expands relative distances between categories, compared to the feature space of V1. This effect is also present to a lesser extent between early visual regions and face-selective cortex (FFA), likely due to the presence of faces in the "dog" basic level category. (Bottom) We measured this "category warping" effect quantitatively by computing the proportion of within- and between-category distance pairs that sit above the diagonal. Concordantly, we see that across the ventral stream, a significant category warping effect exists not just between hV4 and LOC, but also between V1 and V2, indicating that visual processing proceeds in a manner that sequentially facilitates the emergence of categorical distinctions.

corresponding to pairwise distances between subordinates from a particular condition (e.g. within-category distances) relative to the diagonal in each graph (Fig. 4.6, bottom). Here, a proportion of 50% for within-category distances would signify that these distances have an equal probability of being stretched or compressed when going from the feature space of the brain region on the X-axis (e.g. V1) to the feature space of the brain region on the Y-axis (e.g. V2). Concordantly, a proportion approaching 100% indicates that virtually all distances of that particular type increase between the two feature spaces, while a proportion near 0% implies that most distances of that kind shrink. Finally, we can define a "category warping coefficient" as the difference between proportions in the two conditions (i.e. Wth and Btw). By our reasoning above, this quantity measures whether a principled warping of the representational space occurs between two given brain regions, such that within-category and between category distances are affected in an asymmetrical manner from one another. In our experiment, we observed principled warping effects to occur across multiple region pairs throughout the span of the classical ventral stream processing path (V1 - V2 - hV4 - LOC), with strongest effects at the hV4 - LOC boundary, but also present fairly early on, at the boundary between V1 and V2 (Fig. 4.6, bottom row; V1 - V2: Wth - Btw warp = 11.4%, $t_{13} = 2.9$, $p = 0.013$; V2 - hV4: Wth - Btw warp = -4.1%, $t_{13} = 0.8$, $p = 0.441$; hV4 - LOC: Wth - Btw warp = 33.9%, $t_{13} = 8.2$, $p < 0.001$; V1 - LOC: Wth - Btw diff. = 34.1, $t_{13} = 8.3$, $p < 0.001$). Additionally, a category warping effect was present between early visual cortex and face-selective regions (V1 - FFA: Wth - Btw warp = 17.%, $t_{13} = 4.7$, $p < 0.001$), but not between early visual regions and scene-selective areas (V1 - PPA: Wth - Btw warp = -4.1%, $t_{13} = 0.8$, $p = 0.447$; V1 - TOS: Wth - Btw warp = -1.0%, $t_{13} = 0.2$, $p = 0.831$). This is likely due to the presence of faces in the dog basic category and suggests that representational space warping may not be exclusive to objects, but instead may underlie a more general process related to the processing of visual stimuli across human occipito-temporal cortex.

Interestingly, when significant category warping did occur, it occurred exclusively in one direction, compressing within category distances and expanding between-category distances between early visual cortex and intermediate, stimulus selective

regions of occipito-temporal cortex (e.g. LOC, FFA). Taken together, our results suggest that as we move up the ventral stream, the neural representation space of object categories, at the population level, warps by making exemplars more similar within a category and more dissimilar between categories, and this happens in discrete steps, as we see here between V1 and V2 and between hV4 and LOC. In turn, this provides evidence for our original hypothesis that visual processing proceeds in a manner that sequentially facilitates the emergence of categorical distinctions.

4.3.1.4 Typicality Warps Neural Distances in Occipito-Temporal Cortex

The category warping results above provide strong evidence that categorization as a primary goal of computations across the ventral visual stream is indeed intimately tied into the organization of feature spaces of visually selective cortical areas. However, from prior work, we also have reasons to believe that another cognitively useful dimension of objects, namely their typicality, plays an important role in how categories are internally represented across the feature spaces of successive visual cortical regions [31, 68]. Motivated by neural evidence that typicality sharpens category boundaries in occipito-temporal cortex [68], we predict in the context of our model that this graded representation should become increasingly apparent in the activity patterns elicited by category members as we measure stepwise changes in neural representation going up the ventral visual hierarchy.

To investigate how typicality modulates the internal structure of categories, we behaviorally assessed the typicality of each of our 15 subordinate categories within their corresponding basic category, thus allowing us to identify the five most typical (Fig. 4.1 B, purple indicators) and five least typical (Fig. 4.1 B, orange indicators) breeds of dogs and types of cars from our stimulus set. We then used these two typicality tiers to contrast the representation of distances between pairs of highly typical subordinates and pairs of less typical subordinates within and across visually selective brain regions across the ventral stream. If cognitively useful dimensions of object representation are indeed mediated by a common mechanism in their influence over the organization of feature spaces in visual cortex, then we predict that our two separate typicality tiers would behave in a similar fashion to category distinctions

themselves. More specifically, we hypothesize that as we go up the ventral visual stream, typicality would modulate intra-category representation by bringing highly typical items closer to the category center and emphasizing the relative dissimilarity of atypical items to the rest of the category. Previous work suggests that this is a plausible hypothesis for intra-category organization [68], however it remains unclear whether this process is analogous to the category warping effect we observed earlier and, more importantly, how this process proceeds stepwise in the hierarchy of visual processing.

First, to test whether distances between the highly typical and less typical pairs of subordinates become increasingly separated as we go up the ventral stream, we constructed corresponding histograms of these quantities for our brain regions of interest (Fig. 4.7), analogously to the category boundary histograms in our previous analysis. Consistent with prior work [68], we saw that histograms of distances for typical and less typical pairs of subordinates showed a moderate degree of overlap in early visual cortex (V1: High Typ. $<$ Low Typ. mean diff. $= 0.3$, $t_{13} = 2.2$, $p = 0.051$; V2: High Typ. $<$ Low Typ. mean diff. $= 0.3$, $t_{13} = 1.8$, $p = 0.096$; hV4: High Typ. $<$ Low Typ. mean diff. $= 0.2$, $t_{13} = 1.8$, $p = 0.084$), which upholds the theory that typicality is not strongly linked to low-level feature representations in these regions. Additionally, as predicted, we observed a strong qualitative difference in the representation of these typicality tiers in LOC: the means of the two distributions are much more robustly separated (LOC: Wth $<$ Btw mean diff. $= 0.6$, $t_{13} = 5.3$, $p < 0.001$), which indicates that neural distances in object-selective cortex are usually smaller between highly typical items than between less typical items. Finally, we found that typicality does not significantly modulate the representation of object categories in scene-selective regions (PPA: High Typ. $<$ Low Typ. mean diff. $= 0.0$, $t_{13} = 0.0$, $p = 0.965$; TOS: High Typ. $<$ Low Typ. mean diff. $= 0.1$, $t_{13} = 0.6$, $p = 0.538$), and only moderately affects face-selective cortex (FFA: High Typ. $<$ Low Typ. mean diff. $= 0.3$, $t_{13} = 2.8$, $p = 0.016$), again potentially due to the presence of animal faces in our "dog" basic category.

Subsequently, we used the between-brain-region analysis (see Fig. 4.4 for an example in the category boundary case) to test whether the emergent influence of typicality

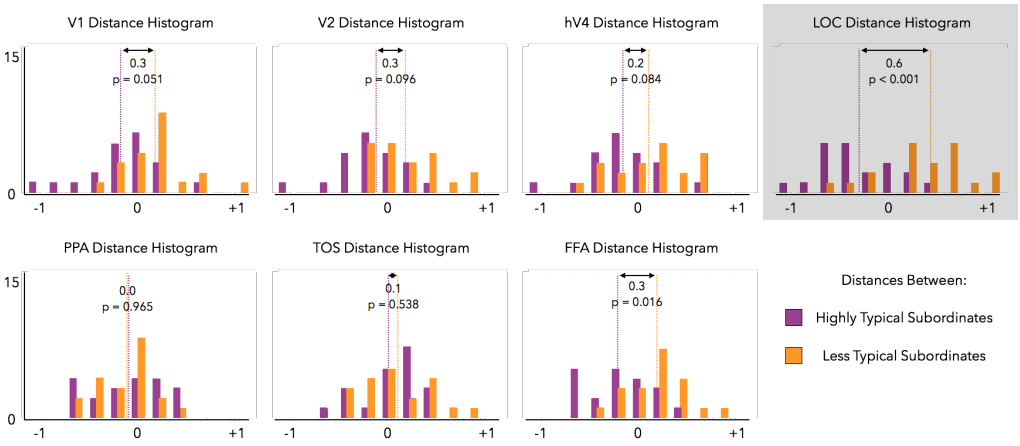


Figure 4.7: **Experiment 1 Typicality Distance Histograms.** Graphs show Z-scored Pearson correlation distance histograms for within-highly-typical-subordinates distances (purple) and within-less-typical-subordinates distances (orange) for early visual (V1, V2, hV4), object-selective (LOC), scene-selective (PPA, TOS), and face-selective (FFA) regions. In early visual regions and scene-selective regions, typicality does not significantly modulate the representation of real-world objects. By contrast, typical and less typical subordinates are strongly separable in LOC (top right, grey), which suggests a sharp qualitative change in the structure of the feature space may arise between hV4 and LOC, which is not mirrored in other stimulus selective regions of occipito-temporal cortex.

on the organization of object categories in LOC is mirrored by a similar warping effect we observed for category boundaries. In effect, per our hypothesis, we set out to test whether category boundaries and typicality share a common effect on the organization of the feature space across the ventral visual stream. Here, inter-category typicality warping would manifest as a trend for distances between less typical subordinates (in orange) to lie mainly above the diagonal and / or distances between highly typical subordinates to fall below the diagonal. Consistent with the histogram results, typicality warping is not present in early visual regions, but arises mostly in a stepwise fashion between hV4 and LOC (Fig. 4.8, top row; V1 - V2 warp = -3.0%, t_{13} = 0.6, p = 0.569; V2 - hV4 warp = -2.8%, t_{13} = 0.3, p = 0.792; hV4 - LOC warp = 26.7%, t_{13} = 3.1, p = 0.008; V1 - LOC warp = 20.9%, t_{13} = 3.3, p = 0.006). This suggests that between hV4 and LOC (and implicitly between V1 and LOC), the feature space in which these objects are represented changes in a way such that less typical items stand out more from the rest of the category representation. In effect, they are pushed away from the category central tendency. We also see a trend for distances between highly typical subordinates to shrink (purple points lie mostly below the diagonal between these brain regions). Interestingly, this effect does not extend to scene- and face-selective areas or (V1 - TOS warp = -15.7%, t_{13} = 1.2, p = 0.241; V1 - FFA warp = 2.0%, t_{13} = 0.2, p = 0.852), and in fact an opposite effect is observed in the most anterior scene-selective visual region PPA (Fig. 4.8, bottom row; V1 - PPA warp = -21.5%, t_{13} = 2.4, p = 0.034), such that less typical exemplars become more similar to each other compared to highly typical exemplars. A potential explanation for this finding would be that PPA is not only known to possess discriminative object information citeiordan2015a, but also heavily involved in representing scene context [85] and, as such, may tend to homogenize across all members of an object category, regardless of typicality (i.e. equalizing distances between all exemplars in a category by shrinking distances between less typical items and enhancing distances between highly typical items).

Taken together, these results suggest that as we go up the ventral visual stream, intra-category neural representation space warps to increase similarity between highly typical subordinates and distinguish more strongly between atypical subordinates and

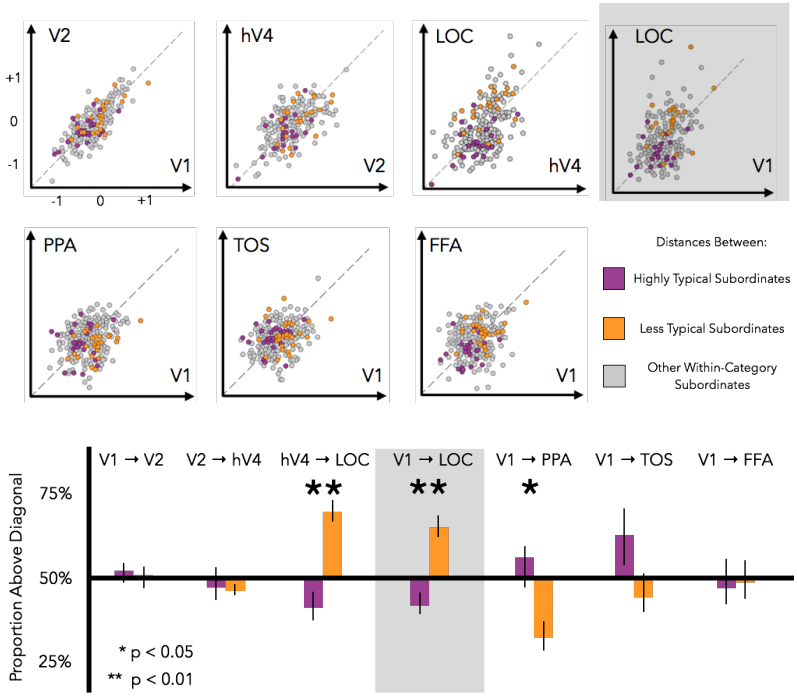


Figure 4.8: **Typicality Warps Neural Distances Across Occipito-Temporal Cortex.** (Top, Middle) Graphs show how representations of z-scored distances corresponding to subordinate category pairs of high (purple), low (orange), and intermediate (gray) typicality change as we move up the ventral visual stream. Representations are relatively stable between early visual regions (V1, V2, hV4), as well as between early visual cortex and face-selective regions (FFA). However, we see a striking shift in the quality of the representation as we move between hV4 and LOC. Here, high typicality subordinate category pairs exhibit a tendency to lie below the diagonal, while low typicality subordinate category pairs sit above the diagonal, which indicates that the feature space of LOC shrinks relative distances between typical exemplars within a category and expands relative distances between low typicality exemplars, compared to the feature space of V1. The opposite effect is present to a lesser extent between early visual regions and scene-selective cortex (PPA). (Bottom) We measured the "typicality warping" effect quantitatively by computing the proportion of high and low typicality subordinate category pairs that sit above the diagonal. Concordantly, we see that across the ventral stream, the main significant category warping effect occurs not between hV4 and LOC, suggesting a sharp shift in the modulation of object representations by typicality at this stage in visual processing.

the rest of the category, with a sharp jump between hV4 and LOC. Moreover, our results suggest that in the successive transformations operating over the relative representations of objects across multiple brain regions, typicality has a similar warping effect on the intra-category space as we observed earlier with category distinctions themselves affecting on inter-category differentiation.

4.3.1.5 Updated Model of Visual Processing in Ventral Visual Cortex

Our results above suggest the presence of widespread effects of category and typicality warping throughout occipito-temporal cortex, and as such provide evidence for our hypothesis that eventual cognitive goals of visual categorization directly guide the feature transformations underlying sequential neural processing along the ventral visual stream hierarchy.

Going back to our original posited model of category processing (Fig. 4.5), we not only confirmed that it is consistent with the tenets of how visual information is represented across the ventral visual stream, but we are additionally in a position to update and enrich it according to our new found evidence. Accordingly, we propose an updated model that posits how both category distinctions (as an inter-category principle) and typicality (as an intra-category principle) guide category processing across successive areas of visual cortex (Fig. 4.9). Here, we see that in the early stages of visual processing, representations of objects belonging to different categories start out partially overlapping, likely due to similarities and differences in their low-level features (e.g. V1, Fig. 4.9 A). As we move up the ventral stream, however, categories become more separated, yet typicality still plays little role in the organization of intra-category structure (e.g. hV4, Fig. 4.9 B). Finally, this representation undergoes an abrupt change as we go from intermediate visual areas (hV4) to object-selective regions (LOC): not only do categories become increasingly separated in this space, but also become internally organized according to how typical their members would eventually be perceived (Fig. 4.9 C). Critically, these two processes don't simply rearrange the representations of objects in a static feature space, but appear to warp and reshape the feature spaces themselves contrasted to earlier visual processing regions: the representational space of object-selective cortex becomes doubly warped

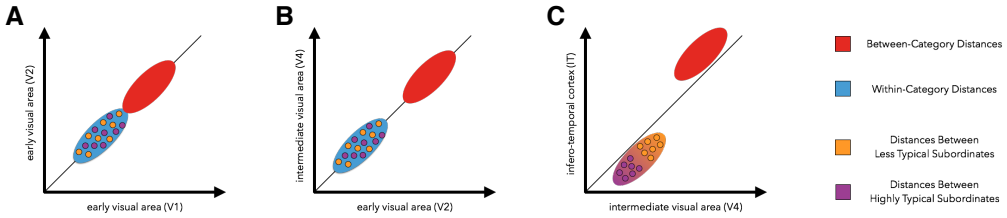


Figure 4.9: **Updated Model for Evolution of Category Representations across Ventral Visual Stream.** We propose that categories would start out partially overlapping, mainly due to overlap in low-level features (A). As we move up the ventral visual stream, computations in successive intermediate visual brain regions would contribute to incrementally shrinking the distances within categories and expanding the distances between categories (B). At both these initial stages, typicality plays little role in the intra-category organization of visual objects. However, at the apex of ventral stream computation (inferotemporal cortex), this process would reach its peak in generating fully dissociable category representations with the least amount of distribution overlap and furthermore organize exemplars within each category such that highly typical members gravitate closer to one another and less typical members are pushed away (C). Critically, these two processes also fundamentally warp the feature spaces themselves contrasted to earlier visual processing regions: the representational space of object-selective cortex becomes doubly warped to, on a global scale, relatively decrease within-category distances and inflate between-category distances (i.e. category warping) and, on a local scale, bring highly typical items closer to one another within the same category and push less typical items away from the category center (i.e. typicality warping).

to, on a global scale, relatively decrease within-category distances and inflate between-category distances (i.e. category warping) and, on a local scale, bring highly typical items closer to one another within the same category and push less typical items away from the category center (i.e. typicality warping). Given that these processes occur simultaneously and are both most evident as a sharp change in the representation between hV4 and LOC, this suggests that category and typicality warping may, in fact, represent two-tiered facets of the same mechanism or phenomenon. As such, they provide the first evidence for a hierarchically organized cognitive-structure-driven hypothesis of visual object information across the human ventral visual stream.

4.4 Experiment 2: Eight Basic Level Categories - Sixty-Four Subordinate Categories

To show that the warping with respect to within and between category distances represents a generalizable principle of category representation in visual cortex, one that would be applicable beyond our choice of stimuli in Experiment 1, we conducted a second fMRI experiment where we constructed a larger and much more varied stimulus set comprising eight basic level categories (birds, cats, dogs, fish, boats, cars, planes, trains), which are well differentiable based on the neural patterns of activity they elicit in visual cortex [68] and which span natural, man-made, animate, and inanimate superordinate categories (Fig. 4.1 C). From each of these eight basic categories, we chose eight subordinate level categories (e.g. pug, jeep) and, analogously to Experiment 1, obtained typicality ratings for each subordinate category within its basic (Fig. 4.1 D). Using data from this subsequent fMRI experiment, we set out to investigate whether the predictions of our model (Fig. 4.9) hold at a large scale and thus provide evidence that the cognitive structure of our category space exerts a measurable influence on the sequential processing of stimuli in visual cortex.

4.4.1 Category Representations Become More Separable and Warp the Neural Representation Space Across the Ventral Visual Stream

As a first step, we tested whether basic categories become more separable as we move up the ventral stream and whether, in doing so, they maintain a representational range large enough to potentially accommodate enhanced specificity of information, rather than suppress it. Indeed, by computing the Pearson correlation between patterns of activity corresponding to each pair from our sixty-four subordinate categories across visually selective cortex (V1, V2, hV4, LOC, PPA, TOS / OPA, FFA), we found that the absolute range of the similarity space increases in intermediate-level object selective regions, compared to early visual regions (V1: r range = 1.14 $\pm$ 0.15 r; LOC: r range = 1.30 $\pm$ 0.17; LOC > V1: $t_9 = 3.5$, $p = 0.007$) (Fig. 4.2 B), consistent

with prior work [63] and our initial findings using two basic level categories above. This suggests that increased separability between categories and developing putative invariant representations with increased proximity to inferotemporal cortex does not entail a shrinking of the space in absolute terms, but rather the opposite; by showing that the correlation ranges increase from early visual cortex to object-selective regions we leave open the possibility that low-level and detailed information about the stimuli may be preserved and, perhaps, even enhanced as we go up the ventral stream, rather than abstracted away.

Nevertheless, the representational space of higher-level visual areas should necessarily be different and likely more complex than that of early visual regions. In our model, we posit that this increase in complexity brought forth through feature transformations underlying sequential neural processing along the ventral visual stream hierarchy is closely tied to eventual cognitive goals of visual categorization, such as generating strong boundaries between category representations and a salient intra-category organization of its constituent members (e.g. typicality relationships). For category boundaries, this hypothesis predicts that in later stages of the ventral stream exemplars belonging to the same category should become increasingly similar in how they are represented in patterns of activity, while simultaneously becoming more dissimilar to exemplars from other categories. We tested this prediction on the large-scale, diverse set of real-world categories employed in Experiment 2 by comparing the normalized similarity distance (1 - Pearson's r) between pairs of subordinates from the same (Wth, e.g. pug and Chihuahua) and distinct basic categories (Btw, e.g. pug and jeep) both within and across visually selective brain regions. The resulting histograms of distances are shown in Fig. 4.10. Similarly to Experiment 1, patterns of activity from different basic level categories start out highly overlapped, yet not indistinguishable from one another in early and intermediate visual cortex (V1: Wth $<$ Btw mean diff. $= 0.1$, $t_9 = 4.7$, $p = 0.001$; V2: Wth $<$ Btw mean diff. $= 0.1$, $t_9 = 5.6$, $p < 0.001$; hV4: Wth $<$ Btw mean diff. $= 0.2$, $t_9 = 5.8$, $p < 0.001$). This suggests that our eight basic categories (bird, cat, dog, fish, boat, car, plane, train) exhibit enough variability in their low-level features to elicit differentiable patterns of activity from the very beginning of visual processing. As we moved up the ventral stream to

object-selective regions, we observed a marked difference in the nature of the within- and between-category distance histograms: separability between categories increased dramatically in LOC, compared to early visual regions (LOC: Wth < Btw mean diff. = 0.5, $t_9 = 8.1$, $p < 0.001$). In a modest departure from the results of Experiment 1, here we also noticed an increased degree of separability between our basic level categories in the most anterior scene-selective region compared to V1 (PPA: Wth < Btw mean diff. = 0.4, $t_9 = 6.2$, $p < 0.001$), but not in dorsal scene areas (TOS: Wth < Btw mean diff. = 0.2, $t_9 = 5.0$, $p < 0.001$) or face-selective regions (FFA: Wth < Btw mean diff. = 0.2, $t_9 = 5.0$, $p < 0.001$). There results likely reflect the diverse composition of our second stimulus set, where contextual effects may play a role in the differentiation observed in PPA and the relatively smaller proportion and strength of face-like stimuli in Experiment 2 compared to Experiment 1 is a plausible explanation for the weakening of the effect observed in FFA. All in all, analogously to Experiment 1, our results suggest that the sequential computation across the human visual system likely amplified, rather than conserved feature descriptions of our visual input that correlate with category boundaries.

Our model also predicts that not only do category representations become more easily separable in later visual regions, but also that the representational space employed in these regions changes in an expected manner given the eventual categorical distinctions that are drawn perceptually between collections of visual stimuli. To test whether this hypothesis holds in the context of a large collection of categories, we used a between-brain-region analysis (see Fig. 4.4 for a detailed example in the context of Experiment 1) to plot the evolution of distances between pairs of subordinate categories across the span of the ventral visual stream (Fig. 4.11). Here, we immediately noticed a higher degree of variability in the span of within- and between-category distances across our stimulus set compared to Experiment 1. Nevertheless, consistent with our initial results, most distances remained close to the diagonal in the steps between V1 - V2 and V2 - hV4, suggesting that the representational spaces change slowly in early and intermediate visual regions. Critically, the largest change arose at the hV4 - LOC boundary, where the distribution of within-category distances begins to visibly shift below the diagonal. This cumulative effect across the span of the first

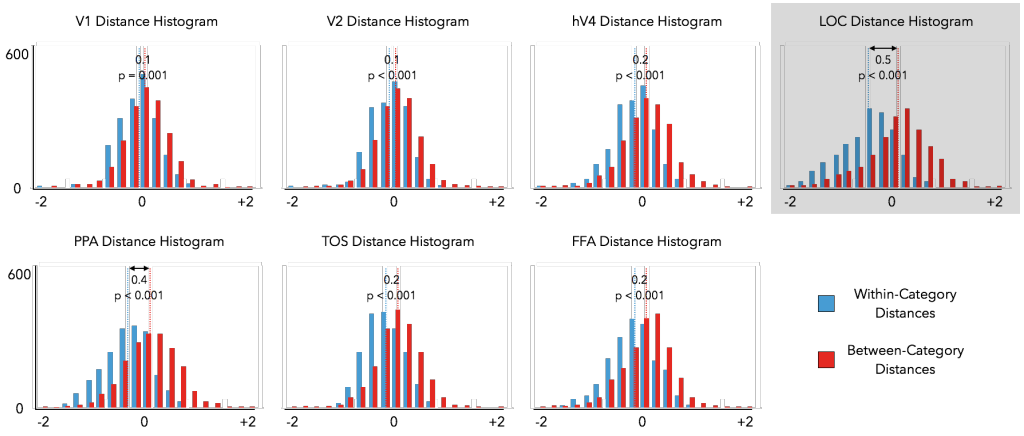


Figure 4.10: **Experiment 2 Category Distance Histograms.** Graphs show z-scored Pearson correlation distance histograms for within-category distances (blue) and between-category distances (red) for early visual (V1, V2, hV4), object-selective (LOC), scene-selective (PPA, TOS), and face-selective (FFA) regions. The eight basic categories: bird, cat, dog, fish, boat, car, plane, and train are reasonably separable in virtually all brain regions considered with the highest distinction arising in LOC (top right, grey). This suggests that a sharp qualitative change in the structure of the feature space may arise between hV4 and LOC, which is not mirrored in other stimulus selective regions of occipito-temporal cortex.

few stages of the ventral visual stream is summarized in the large step between V1 - LOC (Fig. 4.11, top right, grey): between early visual regions and object-selective cortex, the feature space warps to minimize the distance between members of the same category in an asymmetrical way compared to the span of the entire representational space, which, surprisingly here, has less of an impact on the inter-category space compared to Experiment 1. As foreshadowed in the distance histogram analysis above, we also observed a strong warping effect for object categories in PPA, but less so for TOS and FFA, the former of which may be due to contextual effects, as our stimuli comprised a centrally presented object surrounded by naturalistic background.

To quantify the change in the nature of the representational space across the visual hierarchy, we used a similar "category warping" coefficient to Experiment 1 for each condition, which we defined as the difference in the proportion of points that a particular condition (i.e. Wth: within-category subordinate pairs; Btw: between-category subordinate pairs) possesses above the diagonal. Concordantly, a high proportion (50-100%) denoted a trend for the feature space of the latter region to expand distances in that particular condition, compared to the earlier region. Conversely, a low proportion coefficient (0-50%) is indicative of a relative shrinking of the distances in the given condition as we go from the area on the X-axis to the area on the Y-axis. Therefore, a significantly positive or negative category warping coefficient signifies that the representational space affects within- and between-category distances asymmetrically as we go from one brain area to another. In our data, we observed principled category warping effects across multiple region pairs throughout the span of the classical ventral stream processing path (V1 - V2 - hV4 - LOC), with strongest effects at the hV4 - LOC boundary, but also present one step earlier, at the boundary between V2 and hV4 (Fig. 4.11, bottom row; V1 - V2: Wth - Btw diff. = 1.0%, $t_9 = 1.1$, $p = 0.281$; V2 - hV4: Wth - Btw diff. = 5.0%, $t_9 = 2.9$, $p = 0.017$; hV4 - LOC: Wth - Btw diff. = 15.4%, $t_9 = 7.3$, $p < 0.001$; V1 - LOC: Wth - Btw diff. = 16.8%, $t_9 = 7.5$, $p < 0.001$). Category warping was also present between early visual cortex and the most anterior scene-selective region PPA (V1 - PPA: Wth - Btw diff. = 10.5%, $t_9 = 5.5$, $p < 0.001$) and to a lesser extent between early visual cortex and dorsal scene-selective cortex (V1 - TOS: Wth - Btw diff. = 3.8%, $t_9 = 2.4$, $p = 0.041$),

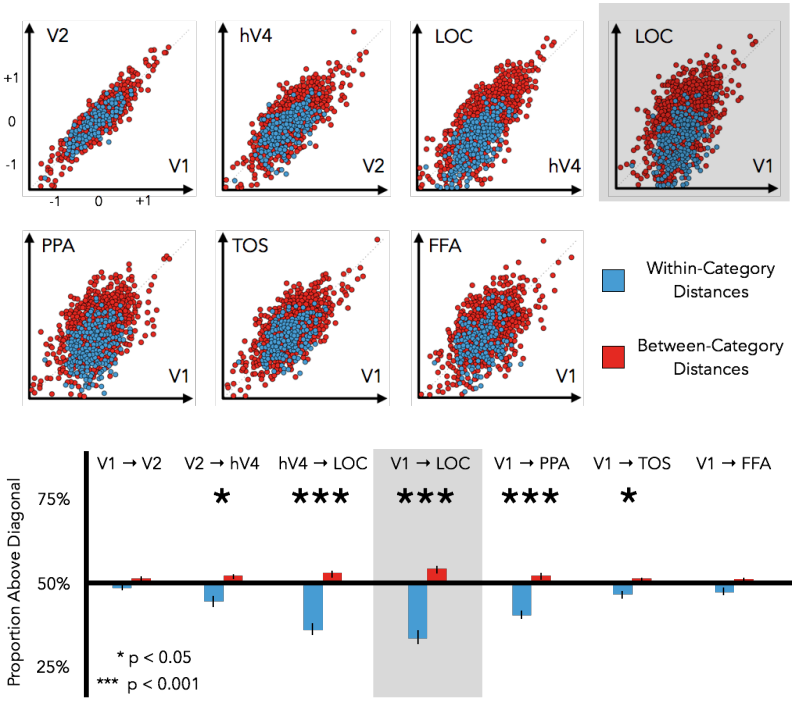


Figure 4.11: **Category Boundaries Warp Neural Representations in Occipito-Temporal Cortex for a Large Array of Real-World Basic Categories.** (Top, Middle) Graphs show how representations of distances corresponding to subordinate category pairs change as we move up the ventral visual stream. Axes represent z-scored distances between pairs of categories in the corresponding brain region. Representations were relatively stable between early visual regions (V1, V2, hV4), as well as between early visual cortex and face-selective regions (FFA). However, we saw a striking shift in the quality of the representation as we moved between hV4 and LOC. Here, within-category distance pairs lied below the diagonal, while between-category distance pairs sat above the diagonal, which indicated that the feature space of LOC shrinks relative distances within categories and expands relative distances between categories, compared to the feature space of V1. This effect is also present to a lesser extent between early visual regions and scene-selective areas (PPA, TOS), likely due to contextual effects. (Bottom) We measured this "category warping" effect quantitatively by computing the proportion of within- and between-category distance pairs that sit above the diagonal. Concordantly, we see that across the ventral stream, a significant category warping effect exists not just between hV4 and LOC, but also between V2 and hV4, indicating that visual processing proceeds in a manner that sequentially facilitates the emergence of categorical distinctions.

but not between V1 and face-selective regions (V1 - FFA: Wth - Btw diff. = 3.6%, $t_9 = 1.9$, $p = 0.089$). Critically, analogously to Experiment 1, all observed category warping occurred exclusively in one direction, compressing within-category distances and expanding between-category distances between early visual cortex and intermediate, stimulus selective regions of occipito-temporal cortex (e.g. LOC, PPA). Taken together, our findings suggest that as we move up the ventral stream, the neural representation space of object categories warps by making items more similar within a category and more dissimilar between categories, and this occurs in discrete steps, as we see here between V2 and hV4 and between hV4 and LOC. Moreover, our results in scene-selective PPA raise the possibility that the warping of the representational space may not be exclusive to objects, but instead may underlie the existence of a more general process across human occipito-temporal cortex, one in which visual processing proceeds in a manner that sequentially facilitates the emergence of categorical distinctions across a variety of classes of input.

4.4.2 Typicality Warps the Intra-Category Neural Representation Space in Object-Selective Cortex

Our findings using a large category set support the predictions of our model that categorization as a primary goal of visual processing is indeed intimately tied into the organization of feature spaces of visually selective cortical areas. Next, we set out to test whether the secondary set of claims put forward by our model generalize beyond the two-category case where they were first observed, namely that typicality as a high-level cognitively useful property of intra-category organization also plays a role in shaping the representational spaces in visual cortex.

To investigate how typicality modulates the internal structure of categories, we used a large-scale Amazon Mechanical Turk behavioral experiment to assess the typicality of each of our sixty-four subordinate categories within their corresponding eight basic categories, thus allowing us to identify the four most typical (Fig. 4.1 D, purple indicators) and four least typical (Fig. 4.1 D, orange indicators) subordinate from each basic. Subsequently, we measured and contrasted the distances between pairs of

highly typical subordinates and pairs of less typical subordinates within and across visually selective brain regions across the ventral stream, reasoning that if typicality as a cognitively useful dimension of object representation exerts a measurable influence over the organization of the feature spaces across the ventral visual stream, then we should observe a similar trend to our first experiment where highly typical exemplars are brought closer together and less typical exemplars are pushed away from the category center as we move up the visual hierarchy.

Our first test of this hypothesis relied on constructing histograms for such highly typical or less typical subordinate pairs for each of our visual regions of interest (Fig. 4.12). Consistent with our results in Experiment 1, we observed that typicality is not strongly linked to category representations in early visual regions (V1: High Typ. $<$ Low Typ. mean diff. $= 0.0$, $t_9 = 0.9$, $p = 0.407$; V2: High Typ. $<$ Low Typ. mean diff. $= 0.1$, $t_9 = 2.1$, $p = 0.069$; hV4: High Typ. $<$ Low Typ. mean diff. $= 0.1$, $t_9 = 0.9$, $p = 0.386$), but is instead strongly represented as a differentiating factor in the activity patterns elicited by our stimuli in object-selective cortex (LOC: High Typ. $<$ Low Typ. mean diff. $= 0.2$, $t_9 = 3.2$, $p = 0.010$). In effect, in LOC distances between highly typical subordinate pairs were significantly smaller than distances between pairs of less typical subordinates, which suggests a reorganization of the representational space towards emphasizing typicality distinctions. Furthermore, we found that typicality had little modulating effect on the representations in scene-selective regions (PPA: High Typ. $<$ Low Typ. mean diff. $= 0.0$, $t_9 = 0.4$, $p = 0.668$; TOS: High Typ. $<$ Low Typ. mean diff. $= 0.1$, $t_9 = 1.3$, $p = 0.241$) or face-selective regions (FFA: High Typ. $<$ Low Typ. mean diff. $= 0.1$, $t_9 = 0.9$, $p = 0.405$). All in all, our results suggest that typicality is indeed a high-level property of objects, as representations of the stimuli in early visual cortex cannot fully explain the pattern of results we observe in object-selective cortex.

Subsequently, we used our between-brain-region analysis to investigate whether typicality influences the representation of object categories stepwise across the ventral visual hierarchy, and if so, to what extent at each step. Analogously to Experiment 1, we defined a typicality warping coefficient which measured the propensity for distances between highly typical subordinates to become compressed across brain regions and

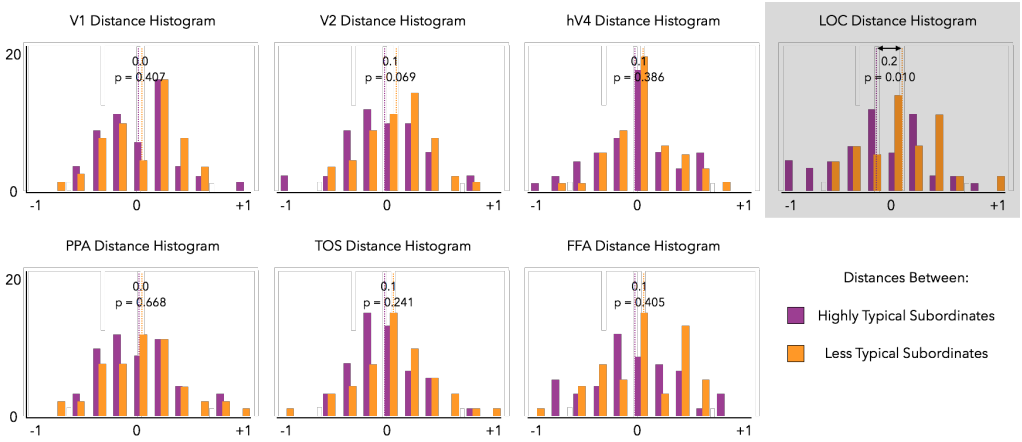


Figure 4.12: **Experiment 2 Typicality Distance Histograms.** Graphs show z-scored Pearson correlation distance histograms for within-highly-typical-subordinates distances (purple) and within-less-typical-subordinates distances (orange) for early visual (V1, V2, hV4), object-selective (LOC), scene-selective (PPA, TOS), and face-selective (FFA) regions. In early visual regions, scene- and face-selective regions, typicality does not significantly modulate the representation of real-world objects. By contrast, typical and less typical subordinates are strongly separable in LOC (top right, grey), which suggests a sharp qualitative change in the structure of the feature space may arise between hV4 and LOC, which is not mirrored in other stimulus selective regions of occipito-temporal cortex.

the tendency for distances between low typicality subordinates to expand. Consistent with our initial results, typicality warping was not present in early visual regions, but arose mostly in a stepwise fashion between hV4 and LOC (Fig. 4.13; V1 - V2 diff. = 4.4%, $t_9 = 2.1$, $p = 0.069$; V2 - hV4 diff. = -2.0%, $t_9 = 0.4$, $p = 0.704$; hV4 - LOC diff. = 9.7%, $t_9 = 2.3$, $p = 0.044$; V1 - LOC diff. = 12.1%, $t_9 = 3.0$, $p = 0.015$). This suggests that between V1 and LOC (and implicitly between hV4 and LOC), the feature space in which these objects are represented changes in a way such that less typical subordinates are more dissimilar to the rest of their basic category. In effect, they are pushed away from the category central tendency. We also see a trend for distances between highly typical subordinates to shrink (purple points lie mostly below the diagonal between these brain regions), although this trend was not significant (Fig. 4.13, bottom row). Additionally, in contrast to Experiment 1, we did not observe a significant modulation of typicality on the relative representation of distances between early visual cortex and scene- and face-selective regions (V1 - PPA diff. = -1.1%, $t_9 = 0.3$, $p = 0.775$; V1 - TOS diff. = 2.8%, $t_9 = 0.9$, $p = 0.368$; V1 - FFA diff. = 2.2%, $t_9 = 0.4$, $p = 0.668$). This finding is consistent with prior work showing that typicality effects on real-world object category representations in occipito-temporal cortex are strongest in object-selective cortex [68].

Finally, to ensure that our results are not solely due to a superordinate distinction within our category set, we also successfully replicated all our analyses above (category and typicality warping, distance histograms, and between-brain-region analyses) independently on each of the two halves of our large stimulus set from Experiment 2 corresponding to two broad superordinate distinctions: natural / animals and man-made / vehicles (Appendix C, Figs. C.2–C.9). Moreover, our results from Experiment 1 also replicated thoroughly when solely using the "dog" and "car" exemplars shown to our participants in the second fMRI study, which comprise a quarter subset of the latter's eight basic level categories (Appendix C, Figs. C.10–C.13).

Taken together, our results suggest the presence of widespread effects of category and typicality warping throughout occipito-temporal cortex, and as such provide evidence for our hypothesis that eventual cognitive goals of visual categorization directly guide the feature transformations underlying sequential neural processing along the

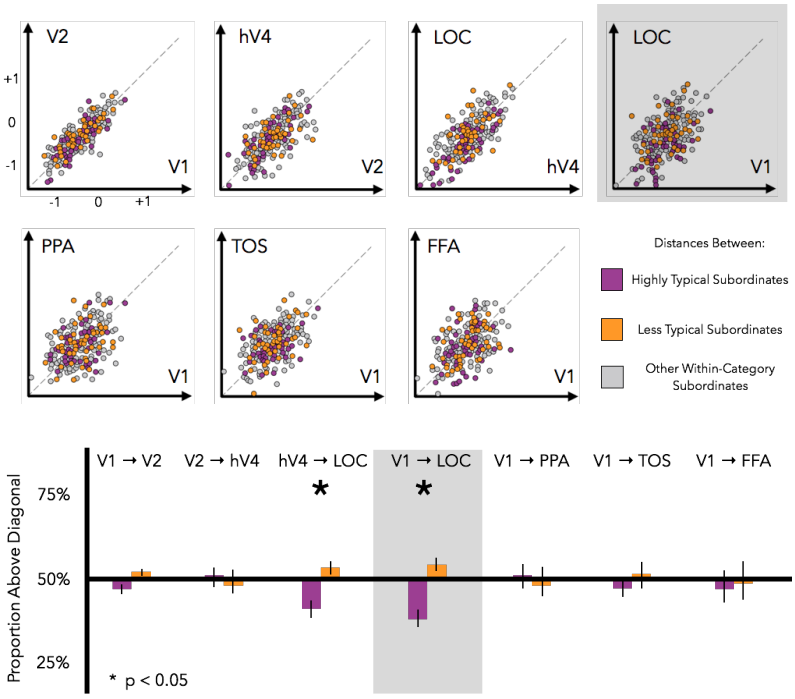


Figure 4.13: **Typicality Warps Neural Distances in Object-Selective Cortex.** (Top, Middle) Graphs show how representations of z-scored distances corresponding to subordinate category pairs of high (purple), low (orange), and intermediate (gray) typicality change as we move up the ventral visual stream. Representations are relatively stable between early visual regions (V1, V2, hV4), as well as between early visual cortex and scene- (PPA, TOS) and face-selective regions (FFA). However, we see a striking shift in the quality of the representation as we move between hV4 and LOC. Here, high typicality subordinate category pairs exhibit a tendency to lie below the diagonal, while low typicality subordinate category pairs sit above the diagonal, which indicates that the feature space of LOC shrinks relative distances between typical exemplars within a category and expands relative distances between low typicality exemplars, compared to the feature space of V1. (Bottom) We measured this "typicality warping" effect quantitatively by computing the proportion of high and low typicality subordinate category pairs that sit above the diagonal. Concordantly, we see that across the ventral stream, the main significant category warping effect occurs not between hV4 and LOC, suggesting a sharp shift in the modulation of object representations by typicality at this stage in visual processing.

ventral visual pathway. This allowed us to put forward a two-tiered model of object category processing (Fig. 4.9) that posits that object representations start out highly overlapping in early visual cortex and gradually become more separated along category boundaries as we move up the ventral stream. Eventual category distinctions, in effect, become an organizing principle for the feature spaces of successive visual regions in the object processing pathway: representational spaces warp in later areas compared to early regions to bring together items that belong to the same category and further enhance separation between categories. This process is mirrored in how typicality reorganized the intra-category organization of objects in these feature spaces by shrinking distances between highly typical members and pushing less typical items away from the category center. Given that these processes occur simultaneously and are both most evident as a sharp change in the representation between hV4 and LOC, this suggests that category and typicality warping may, in fact, represent distinct facets of the same two-tiered mechanism or phenomenon. As such, they provide the first evidence that the cognitive structure of our perceptual object space may directly modulate the hierarchical processing of visual information across the human ventral visual pathway.

4.5 Discussion

Prior work has emphasized the important role that broad categorical distinctions play in the functional organization of visual cortex, as numerous regions showing preferential activation for broad stimulus classes such as faces, scenes, objects, and bodies have been uncovered across human visual cortex [35, 40, 53, 72, 91], suggesting that neural activity across many of these regions may contribute to the ultimate goal of separating visual stimuli into interpretable, actionable categories further down the processing stream. In agreement with this view, many prevailing models of the ventral visual processing pathway propose that it contains mechanisms designed to take intertwined high dimensional representations of visual stimuli (low-level feature collections in early visual areas) and systematically disentangle and rearrange them

into separable, invariant category representations later on (e.g. in inferotemporal cortex [33, 44, 110, 126]. The details behind this process have so far remained unclear, as well as whether this untangling occurs in a stepwise fashion across visual cortex, and if so, what are the transformations that occur at each step. Our work addressed these questions by putting forward a principled framework for investigating how information processing related to the eventual perception of object categories proceeds sequentially across human ventral visual cortex. The central tenet of our model proposes that this sequential computation is driven in part by and specifically optimizes for the fulfillment of eventual perceptual goals of visual processing, such as generating clear category distinctions and organizing members of a category along a typicality gradient.

The prevalent view of object categorization, however, posits that information present in posterior occipito-temporal cortex does not reflect cognitive constraints, which are instead enforced and instantiated later on in the processing stream (e.g. anterior temporal and frontal regions [49, 93, 95]. This perspective is also mirrored by models that strongly encapsulate vision from cognition [45, 47, 108, 110]. In direct contrast to this view, in the current study we propose a competing model of object processing in human ventral visual stream based on the hypothesis that sequential computations in visually selective cortex optimize specifically for cognitively useful aspects of category structure. More specifically, in both of our fMRI experiments, we found strong evidence that both aspects of category structure we investigated (generating category distinctions and engendering typicality distinctions between members of a category) warped the neural representation directly and sequentially across the ventral stream: category distinctions slowly pushed their representations apart as we moved between early and mid-level visual areas, and simultaneously, perceived typicality of category members rearranged the internal neural category space so that in later processing stages highly typical items became more similar to one another and less typical items were pushed away from the category central tendency. Thus, our results suggest that these eventual cognitive goals of visual categorization directly guide the feature transformations underlying the sequential neural processing of visual input along the ventral visual stream hierarchy of brain regions from early visual

cortex to inferotemporal cortex.

Moreover, recent neuroimaging work has shown that the representation space in occipito-temporal cortex is highly fluid and can be influenced in an online fashion by learning [21], attention [18], or shifting task demands [57]. By contrast, our study investigated the representation of objects throughout the ventral stream while explicitly attempting to eliminate the influence of such factors: throughout our experiments stimuli are never repeated and we employ an image-level 1-back task not related to categorization or typicality, solely for ensuring participant alertness. This suggests that our findings are likely not due to explicit transient top-down constraints imposed on the visual system by higher level processing regions (e.g. anterior temporal and frontal regions [49, 93, 95]), but instead reflect properties of a fundamental mechanism of object processing across the ventral visual pathway.

Concordantly, our proposed model has direct implications for recent avenues of research whose goal is to understand the sequential stepwise computation in visual cortex via parallels to emergent properties observed in layers of deep artificial neural networks trained for solving specific visual tasks [16, 135, 136]. Such deep networks are trained to optimize performance on a single specific end-goal task, usually basic-level categorization. By showing that multiple cognitively useful goals of object perception influence the representation of visual input throughout the human ventral stream, our work puts forth the possibility that achieving a strong connection between deep artificial models and biological vision may require incorporating (either explicitly or at a verification stage) other high-level properties such as typicality, which we have presently identified as having a measurable impact on the feature spaces of visual regions strongly involved in object and category recognition (e.g. LOC). Furthermore, it remains unknown whether inter-category boundaries and intra-category graded typicality relationships represent isolated or independently arising properties of category structure. Our findings suggest that the same warping mechanism by which items which are sought to be made more similar are brought together and items which are sought to be made dissimilar are pushed away relative to each other across successive representational spaces may, in fact, be shared between typicality gradients and category boundaries, between inter-category and intra-category scales

of object representation. This raises the question of whether multiple cognitive utility constraints operating on the visual processing hierarchy engage in mutual interaction, be it interference or facilitation. While it is virtually impossible to test this question in the context of our current study, using complex modeling approaches afforded to us by advances in deep neural networks designed to solve visual tasks may be able to approach this issue through imposing distinct learning objectives on the same architecture known to eventually develop a representation correlated with that of primate visual cortex [16].

Looking beyond object categorization, our results also raise the possibility that a generalized mechanism based on cognitive utility of category structure may drive processing across a larger swath of functionally selective occipito-temporal cortex. Prior work from our lab has shown that computations in this brain region optimize for basic-level categorization and they don't do so exclusively in object-selective regions, but to a lesser degree across scene- and face-selective regions, as well [67]. Similarly, in the current study we show that category warping is not limited to LOC, but again extends to other category-selective regions such as PPA, TOS, and FFA (Fig. 4.6 and Fig. 4.11). Moreover, prior work has shown that while they exhibit activation preference for a particular stimulus type above all others (e.g. faces for FFA, scenes for PPA and TOS / OPA), nevertheless a graded activation profile pervades such functional regions such that they respond to [34, 40, 72, 97] and likely process information about other non-preferred stimuli, as well [67, 97]. Taken together, this suggests that the warping mechanism may be indicative of a larger type of processing constraint pervading intermediate visual areas: cognitive utility as a default force driving refinement of visual information representation as we go up the ventral visual stream. If such a mechanism indeed exists, it would predict similar warping effects for other high-level stimulus classes, such as scenes and faces, but with PPA / TOS and respectively FFA as the highest level visually selective areas of their corresponding computational pathways. Testing these predictions represents an exciting avenue of research for future studies on the effect of cognitive utility constraints on processing in visual cortex.

Although fundamental to our understanding of visual processing and eventual

perception, the primitives underlying feature spaces of high-level visual regions are unknown. For object-selective cortex, we have evidence that the representation space is modulated by high-level properties of objects, such as animacy [19, 23, 76] and real-world size [76, 77], by structural features, such as shape [54, 78] and parts [41, 56], and may, in fact, include cross-modal information beyond visual object details [4, 5, 105]. Our work identifies a separate level of description for this representational space, one that addresses its function and its organization relative to the spaces from which it derives most of its feed-forward input. By proposing and testing a model for how representations of objects change across regions involved in object processing, we established a novel way to divine information about axes of positive variance in this space without having direct access to the underlying representation itself. Moreover, our results imply that this space organizes itself in a manner driven by even higher-level constraints, which are likely many synapses removed from the computations performed by this particular region of the cortex: eventual contribution of its efferent output to the later instantiation of cognitively useful percepts and decisions about the world. Indeed, by showing that category boundaries and typicality gradients modulate processing as early as posterior occipito-temporal cortex, we broaden the space of potential properties which may constitute relevant dimensions of organization for the feature spaces of stimulus selective regions in visual cortex and beyond.

We started our endeavor with a simple hypothesis that represents a logical extension to previous neural models of object category processing: cognitively useful aspects of category structure are not just the end-goal of the computation, but, in fact, guide it sequentially as we go up the ventral visual stream. Consequently, we proposed a new model of category processing and uncovered evidence for a two-tiered organizational principle mitigated by cognitive utility: category boundaries and typicality simultaneously warp the neural representation space across the span of visual cortex by reorganizing the internal structure of the successive feature spaces to emphasize these two eventual cognitively useful aspects of real-world objects. By examining the makeup of this space in an indirect fashion, we showed that the guiding principles for its internal organization might be more accessible than previously thought, if only at a higher, more complex level than hitherto expected. And with these alternative

descriptions in hand, we may be able to eventually recover the lower level primitives that we have so arduously sought to define since we first uncovered selectivity to particular stimulus classes in visual cortex.

4.6 Acknowledgments

This work was funded by the William R. Hewlett Stanford Graduate Fellowship (to M.C.I.), the William and Adeline Hendess Phi Beta Kappa Graduate Fellowship (to M.C.I.), and an ONR MURI Award No. N00014-14-1-0671 (to D.M.B and L.F.-F.).

Chapter 5

Locally-Optimized Inter-subject Prediction of Functional Cortical Regions

Our brains solve visual recognition through the interplay of computational, representational, and physical levels of interpretation of input from our eyes [92]. Concurrently with investigating the mechanisms of object perception, we also seek to develop tools that help us better understand this key relationship between the function of neural circuits and their position on the cortical surface, a relationship that is currently not well understood.

To address this goal, we approach the problem of inter-subject registration of cortical areas, as a necessary step in functional imaging (fMRI) studies for making inferences about equivalent brain function across a population. The main challenge behind successfully predicting the location of cortical regions in never-before-seen cortical sheets it that many high-level visual brain areas are defined as peaks of functional contrasts whose cortical position is highly variable. As such, most alignment methods fail to accurately map functional regions of interest (ROIs) across participants. To address this problem, we propose a locally optimized registration method that directly predicts the location of a seed ROI on a separate target cortical sheet by maximizing the functional correlation between their time courses, while simultaneously

allowing for non-smooth local deformations in region topology. Our method outperforms the two most commonly used alternatives (anatomical landmark-based AFNI alignment and cortical convexity-based FreeSurfer alignment) in overlap between predicted region and functionally-defined LOC. Furthermore, the maps obtained using our method are more consistent across subjects than both baseline measures. Critically, our method represents an important step forward towards predicting brain regions without explicit localizer scans and deciphering the poorly understood relationship between the location of functional regions, their anatomical extent, and the consistency of computations those regions perform across people. This chapter is joint work with Armand Joulin, Diane M. Beck, and Fei-Fei Li, previously published as [69].

5.1 Introduction

A common and reasonable assumption of modern neuroscience is that virtually all human brain areas, whether functionally or anatomically defined, are shared across the vast majority of the population and a correspondence of processing role exists between such equivalent areas. However, no two brains have the same anatomical shape or folding pattern, and thus finding a precise correspondence between locations in two separate cortical surfaces is a highly non-trivial problem.

Currently, most state-of-the-art cortical prediction and alignment methods define transformations between entire cortical volumes that attempt to preserve anatomical landmarks, cortical curvature, or functional connectivity, and subsequently check whether specific regions of interest (ROIs) are accurately matched between subjects [24, 118, 137]. However, many high-level visual brain areas are defined as peaks of functional contrasts (e.g. higher activation for objects versus scrambled objects for lateral occipital complex LOC [55]) and it is usually difficult to identify clear anatomical landmarks and boundaries for these areas, due to large variability in their cortical position [3, 132] and functional response [7] (Fig. 5.1, left). As a consequence, although they provide a reasonable global matching, previous methods usually fail to accurately map such functional ROIs across participants.

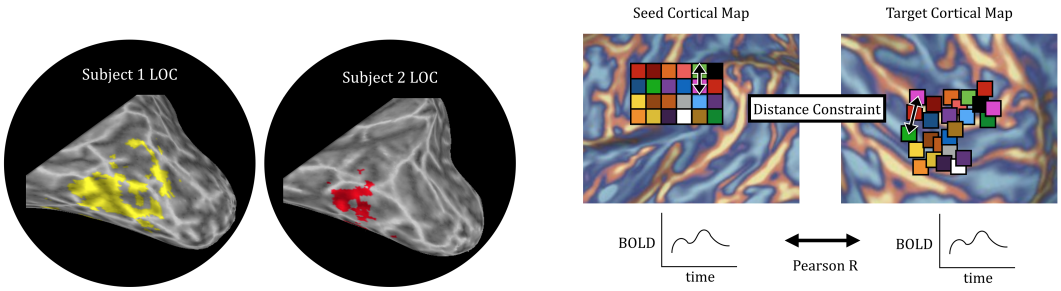


Figure 5.1: **(Left) LOC variability.** Location and extent of lateral occipital complex (LOC) is highly variable across subjects, even when using the same localizer experiment, same scanner, and same analysis pipeline. **(Right) Schematic representation of our proposed method.** Our algorithm tiles the seed region with smaller sub-regions and finds the best functional match for each of them in the target map. The sub-regions are allowed to move independently from one another, provided only that the distance between any two initially adjacent sub-regions does not increase by more than a set threshold.

Thus, our goal is to increase the reliability of inter-subject mapping for these cortical functional peaks, as well as for the visual areas they define, using fMRI data. To address this problem, we describe a general-purpose method for predicting the location of functional areas across people that we apply to the problem of localizing object-selective cortex, LOC.

5.2 Related Work

Our problem can be thought of as a special instance of cortical alignment, where the main goal becomes accurate prediction of a particular region's location, rather than finding a complete correspondence between entire brain volumes. By comparison, virtually all extant alignment methods ([24, 26, 46, 59, 117, 118, 124, 137]) define transformations across full cortical volumes and subsequently check whether specific regions of interest (ROIs) are accurately matched between subjects.

Anatomical alignment relies on large scale correspondences between all human brains, including the reliable presence and the relatively consistent position of primary features such as major sulci and gyri on the cortical surface (e.g. Talairach [124],

AFNI [26]). Additionally, given that the main obstacle in aligning the cortical surface between subjects is its folding variability, methods have been proposed that warp gray matter meshes using local curvature properties of the cortex (e.g. FreeSurfer [46]). These methods, as well as recent extensions [137] also suffer from significant shortcomings in matching functional areas.

A recent method incorporates functional connectivity constraints in the mapping [24] and shows improved ability to align intertwined networks in the brain (i.e. default mode network). However, many functional areas are not usually a strong part of these networks and thus receive little benefit from this approach.

Finally, another class of alignment methods uses functional correlation constraints. For example, hyperalignment [59, 117] and other methods that rely on low-dimensional embeddings of functional responses (e.g. [82]) usually offer improvements over commonly used anatomical alignment methods (e.g. Talairach [124], AFNI [26]). Nevertheless, such methods represent a point in the target map as a linear combination of (possibly) all voxels in the other map, and thus are not directly amenable to transferring the location of one contained area across maps without explicit additional knowledge, such as post-hoc labeling. Another promising recent method [118] starts with FreeSurfer alignment and maximizes local functional correlation across the cortical surface to nudge the vertices of the surface map. This method performs well for early visual areas, but shows limited ability to match functional regions as distance from the occipital pole increases. In contrast to [118], we enforce maximal alignment and prediction specificity to a single region of interest and, furthermore, we allow for locally non-smooth deformations in our mapping, which bypasses the (otherwise ubiquitous in previous work) expectation of using continuous maps between cortical sheets or volumes.

5.3 Locally-Optimized Cortical Region Prediction

Our goal is to predict the location of functionally-defined high-level visual areas between participants. To compute a correspondence between equivalent functional regions, we reasoned that although two cortical surfaces (corresponding to two separate

subjects) must express the same necessary computational units that give rise to observed function, these units might not be perfectly equivalent or identically distributed spatially across the two ROIs [66]. Thus, a key design principle behind our method is to allow a small degree of non-smoothness in the local deformations afforded by the mapping between the two cortical surfaces.

Our method was inspired by a computer vision object co-localization technique first discussed in [36, 37] and takes as input pairs of flattened cortical surfaces from participants who previously took part in an arbitrary fMRI experiment that exposed them to complex, varying stimuli (e.g. visual categorization [67]). We standardized the cortical surfaces by resampling the multidimensional functional data of each subject to a regular square grid at a resolution of 2 x 2 mm. Each point in the resulting grids has a functional time course associated with it which corresponds to the estimated response of that point to the stimuli shown across the entire duration of the fMRI experiment (e.g. using a 512 TR fMRI experiment as input implies a 512-dimensional representation for each point in the resulting standardized cortical maps). Then, for each possible pair of participants, one of them is selected as the seed and the other as the target (for our final results, each participant in each pair is, in turn, selected as the seed and target, and performance is averaged across both these configurations). The location of the functionally defined region of interest in the seed subject is then tiled with a grid of n x n patches, where each patch is associated with a small area on the brain surface (e.g. 5 x 5 voxels). Finally, the algorithm seeks to find maximal functional correspondences between each seed patch and an equivalent region in the target map by maximizing the sum of time course correlations across all the patches, while enforcing that the distances between adjacent patches change by less than a specified amount in each direction (i.e. $\rho = 4$ voxels) between the seed and target maps. An example seed ROI parcellation and target matching are shown in Fig. 5.1 (right). The optimization problem can be written as:

$$\begin{aligned} \underset{M}{\text{minimize}} \quad & \textstyle\sum_i d_F(F_i, F_{m_i}) \\ \text{subject to} \quad & d_s(p_{m_i}, p_{m_j}) \leq \rho, \end{aligned}$$

where $M = \{(i, m_i)\}$ is the collection of correspondences between seed (i) and target patches (m_i); d_F is the feature distance between the patches in each correspondence, computed as 1 - Pearson's r; d_s is the cortical distance difference between the original and mapped configuration of each pair of patches (patch i mapped to patch j) in the two maps; and ρ is the maximum allowable distance change between neighboring patches across maps. We solve the optimization problem above using a deterministic grid search through the space of all possible patch jitter permutations.

5.3.1 Advantages Over Previous Methods

Our method presents several key advantages over other alignment methods, which render it more general and more precise. First, virtually all previous methods compute a complete correspondence code between entire cortical surfaces. Afterwards, the location of functional areas is obtained second-hand, e.g. by aligning a contrast map and re-thresholding. Here, we instead focus on maximizing the quality of the mapping for a single, specific seed ROI. Furthermore, other alignment methods usually generate a smooth manifold transformation between cortices. However, this entails a very strong assumption that activation profiles vary smoothly and with the same spatial distribution across subjects. We forgo this assumption by allowing locally-non-smooth deformations in the topology of the predicted ROI. Finally, cortical registration methods are usually described by highly complex optimization problems that can only be solved up to a local minimum, and are thus highly sensitive to parameter initialization. By contrast, our method has a global optimum solution to which we converge deterministically and is therefore much more robust.

5.4 Experiments

5.4.1 fMRI Dataset and Baselines

We tested our method by predicting the location of a difficult to match, functionally defined, object-selective ROI (lateral occipital complex LOC) between subjects using data from a block design passive-viewing fMRI experiment where participants ($n = 7$)

Figure 5.2: **Stimulus set for fMRI experiment used to perform and evaluate the cortical prediction algorithm.** During the experiment, participants were shown images from 32 object categories: 8 breeds of dogs, 8 types of flowers, 8 types of planes, 8 types of shoes (32 images per category; 1,024 images total).

were shown 1,024 images of objects from 32 categories (Fig. 5.2, see [67] for details about the procedure and preprocessing). We computed the position of each participant's LOC using standard localizer runs conducted in a separate fMRI session [51, 119]. We then used the AFNI-SUMA software package [26] to project and interpolate the data from the 3D volume onto a 2D flattened regular grid cortical map.

We compared our algorithm against the two most commonly used cortical registration methods: anatomical landmark-based AFNI 3dvolreg [26] and cortical convexity-based FreeSurfer [46]. AFNI uses information about overall brain shape and automatically defined anatomical points of interest to warp cortical volumes across subjects. FreeSurfer also uses brain shape, as well as information about cortical curvature (sulci and gyri locations, distribution of normals to the gray matter surface) to iteratively distort one cortical surface into another.

5.4.2 Results

To test how well our method predicts the location of LOC across subjects, we used two metrics: accuracy and consistency. Accuracy represents the percentage of overlap between functionally-defined LOC and predicted LOC after mapping from a different

subject's brain. Consistency is defined as the amount of overlap between predicted regions from multiple subjects aligned to the same target map. For both metrics, overlap is computed as intersection over union.

We show results for the two baselines, as well as our method in Fig. 5.3. Our registration method vastly outperformed the two canonical baselines in overlap between predicted region and ground truth LOC: baselines 10-11%, ours 24-25%. Furthermore, the maps obtained using our method are more consistent across subjects than both baseline measures (overlap of region commonly mapped from 3+ subjects: baselines 9-11%, ours 26%).

Qualitatively, the cortical maps further showcase the strength of our results compared to the AFNI and FreeSurfer baselines. In the first two panels of Fig. 5.3 (bottom left) we see that functional regions in other subjects are mapped with a high degree of variance onto the target subject cortical sheet. Often, there is little overlap with our localizer-defined ROI and, most importantly, the mapping may place the region several centimeters away from its desired location, often on a different gyrus. By contrast, our method (Fig. 5.3, bottom right) shows much less variance in the predicted area, with the peak of the prediction fully contained within our localizer-defined region.

These results suggest that our registration technique significantly increases the reliability of transferring the location of functional ROIs between subjects.

5.5 Conclusion

In this paper, we proposed a locally optimized registration method that predicts the location of a seed region of interest (ROI) on a separate target cortical sheet by maximizing the functional correlation between regions and simultaneously constraining the global structure of the mapping, while allowing for non-local deformations in its topology.

Our method vastly outperforms two canonical alignment baselines (anatomical landmark based AFNI [26] and cortical curvature based FreeSurfer [46]) in both precision and consistency. By improving the quality and reliability of matching and

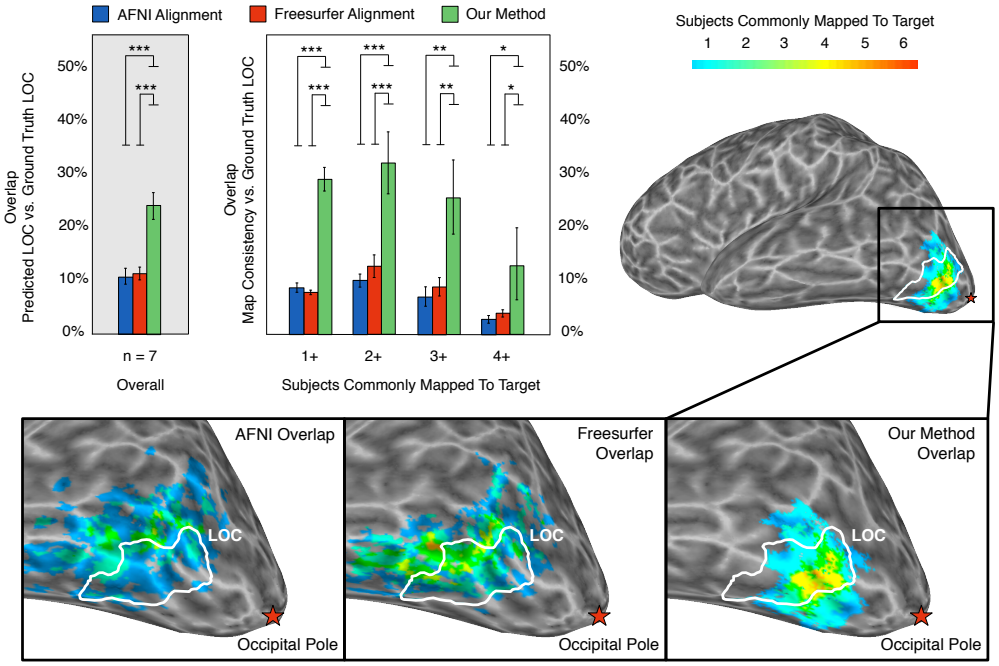


Figure 5.3: **Alignment Results: Accuracy and Consistency** ($n = 7$ subjects). For every target subject, we align LOC from all other 6 subjects to the target cortical surface using functional data from the above experiment. (Top Left) Overlap between predicted LOC and LOC defined using separate standard localizer procedure, measured as intersection over union of surfaces. (Top Right) We select the voxels predicted consistently in the target map for $n+$ subjects and compute the overlap between this restricted region and ground truth LOC for $n \in \{1, 2, 3, 4\}$. (Bottom) Consistency of predicted LOC obtained from aligning using AFNI 3dvolreg, FreeSurfer, and Our Method for a representative subject. Heatmap indicates how many subjects' LOC were mapped to that voxel on the target surface. White outline indicates LOC boundaries defined using separate standard localizer procedure.

transferring the location of functional ROIs across subjects, our technique represents an important step towards obviating the need for running separate time- and resource-consuming localizer scans for every functional brain region. Instead, we envision an eventual solution where a single 'localizer' experiment is performed using a high variance stimulus (i.e. natural movie [66]), which is then used to define all functional ROIs, including potential regions which have yet to be identified. Such a mapping is also useful in settings where one needs to compare analyses and hypotheses between datasets where functional localizers are missing and gathering extra sessions of data is either expensive (large number of participants) or impossible (unavailability of former subjects).

Finally, the relationship between peaks of functional contrasts and the computation performed by the cortex surrounding them is not well understood. Since our method improves the quality of functional ROI mapping between subjects, it becomes especially useful for investigating the key complex relationship between anatomy, functional contrast peaks, and cortical computation.

Chapter 6

Conclusion

How does our visual system take a noisy sea of colored dots encoded by our retinas and generate salient labels for everything we see? What's more, implicit in this process is a marvelous generalization step: we almost invariably refer to, act upon, and think about concrete entities in our world through their *category*, rather than their individuality; we gloss over their visual discrepancies to collect them into self-similar bins whose members share features, affordances, and meaning. Within the realm of our vision, categorization is a fundamental building block of our perceptual experience.

But there are many components to our shared human category structure, most of which remain hidden in how they are built and extracted by our visual cortex from raw photons impinging upon our sensorium. The main goal of this dissertation has been to use computational approaches to explore how pervasive, yet poorly understood, dimensions of object categorization are represented in our brain and how they contribute to our building a coherent picture of the world.

For example, our category space is hierarchically organized: the same picture can be simultaneously interpreted as an animal, a dog, a collie, or "Mr. Woof"; although most people would prefer the basic-level label "dog". While a preponderance of evidence suggests that this basic-level advantage captures something fundamental about human perceptual categorization [6, 10, 14, 65, 71, 90, 94, 98, 99, 114, 116, 123, 125], it is surprisingly unknown how it (or, more broadly, the hierarchical representation

of object categories) is achieved in the brain. We address this question at length in Chapter 2, where we provide the first neural evidence that preferentially extracting information at a mid-level of generality (e.g. "dog", the most privileged in our daily interaction with the world) may be an emergent property of the human visual system and that such categorization may be part of visual processing from its very early stages.

On the other hand, even within their "category", not all dogs are created equal: most people would agree that a Golden Retriever is more representative of the concept "dog" than a Chihuahua. Indeed, in our interaction with the world we enjoy the benefits of generalization (i.e. categories), but the meaning and characteristics of individual objects is far from lost. Interestingly, this preference for particular members of a category is not trivial or purely descriptive, as considerable evidence suggests that the typicality of a particular item is reflected in how fast and how accurately we perceive it in our daily lives [109, 112, 115]. And yet, little is known about how typicality influences the neural representation of real-world objects from the same category. We address this question in Chapter 3, where we show that everyday typicality judgments are correlated with neural distance between categories in object-selective regions of our brain. As such, our results suggest that typicality may constitute a previously unexplored principle of organization for intra-category neural structure in high-level visual cortex.

Ultimately, it falls on us to attempt to use our new found insights into the neural underpinnings of category structure to describe in a principled way how the process of extracting category information from our visual input is accomplished by the brain. Consequently, in Chapter 4, we built upon our previous findings to put forward a model of object category processing in human visual cortex based on the hypothesis that cognitive utility aspects of our category structure drive successive computations across the ventral visual stream. By testing the predictions of our model, we showed that category distinctions slowly pushed representations apart between early and mid-level visual areas and, simultaneously, perceived typicality of category members modulated the internal neural category space so that in later processing stages highly typical items became more similar to one another and less typical items were pushed

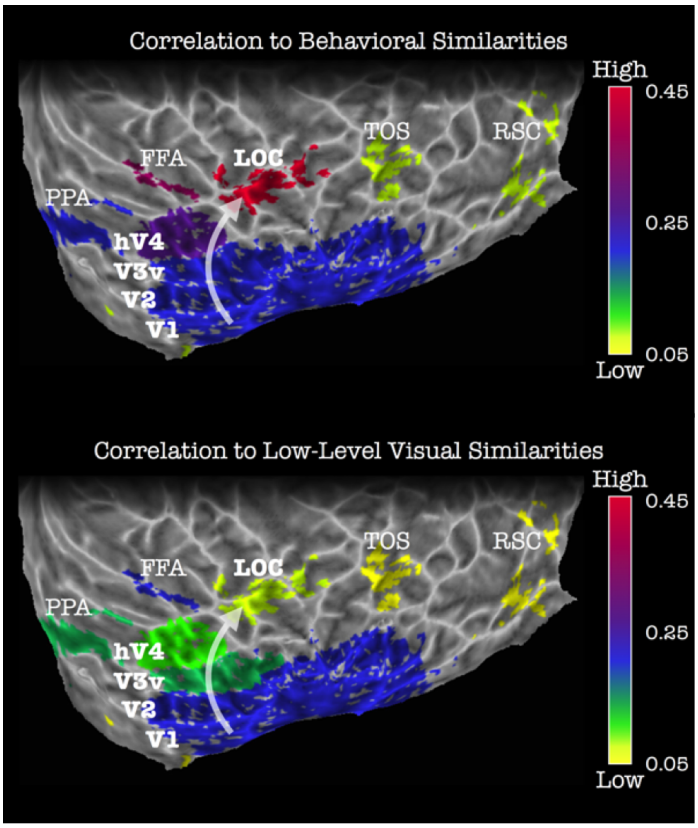
away from their category central tendency. This provides the first glimpse into the neural underpinnings of processes we've known about and built cognitive models for over the course of forty years, but have until now remained elusive in the brain. Yet, this is not an endpoint, but a stepping stone into a rich space of questions that can help us understand how fluid our representation of the world is and how our brain adapts its processing to task demands.

Finally, our brain solves visual recognition through the interplay of computational, representational, and physical levels of interpretation of input from our eyes. Concurrently with investigating the mechanisms of object perception, we also strive to develop tools that help us better understand this key relationship between the function of neural circuits and their position on the cortical surface, a relationship that is currently not well understood. In Chapter 5 we put forward a novel algorithm aimed at predicting the location of functional visual regions across people, thus helping to decipher the poorly understood relationship between the location of functional regions, their anatomical extent, and the consistency of computations those regions perform across individuals.

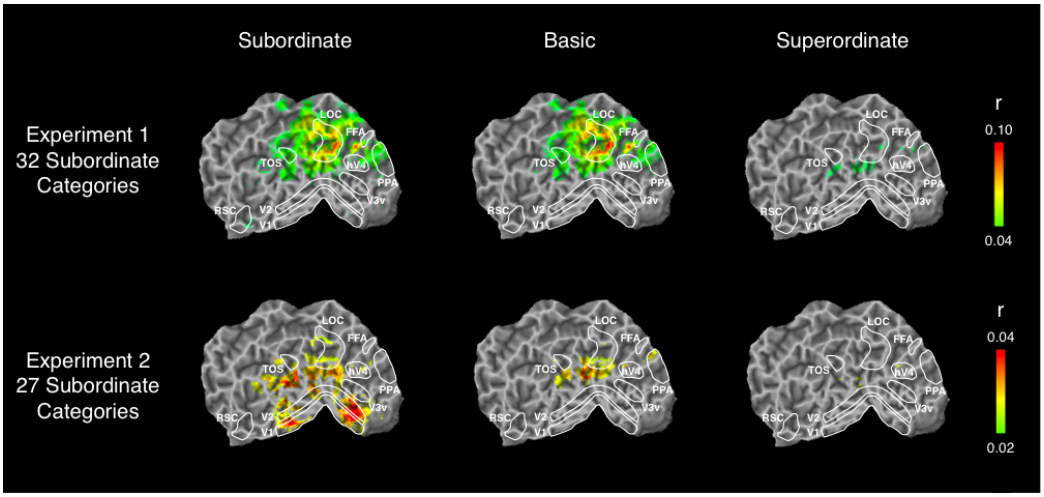
Taken together, the insights we gathered in this dissertation have helped paint a clearer picture of how both the vertical (taxonomy) and horizontal (typicality) dimensions of our category structure are computed and represented by the visual cortex in our brains' quest to understand and portray the world. Nevertheless, many more unanswered questions remain, some of which are borne from our new characterizations of these processes themselves. For example, what components of the neural representation space of object and scene taxonomies are shared and unique across different populations? Does this representational space warp in relationship to differences in native language, subject-level familiarity or expertise with different parts of the category space? Or, even more fundamentally, how much of the structure of this category space is shared across primate species or uniquely human? Future answers to these questions will not only help indicate how stable such principles of categorization are across multiple instantiations of vision solvers (brains), but also to what extent we can potentially use our findings as a guiding principle for the implementation of human-level-performance artificial recognition systems.

Appendix A

Basic Level Category Structure Emerges Gradually Across Human Occipito-Temporal Cortex



Supplementary Fig. A.1. **The relationship among the various taxonomic levels was closer to human behavioral categorization in later regions than in early visual cortex.** To show this, we performed an additional analysis where we computed the correlations between neural representation of categories in Experiment 2 and behavioral similarity (top) and low-level similarity of stimulus images (bottom). We plotted these average across-subject correlations on a flat cortical representation of the group map occipito-temporal region. We observed that low-level features were indeed most correlated with neural representation in early visual cortex, but this correlation decreased as we moved up the ventral visual stream. Conversely, an opposite pattern was observed for the correlation with behavioral similarity: this quantity increased gradually from V1 to LOC, where it reached a peak. This offers further evidence that our results are in agreement with previous studies: category distinctions become more pronounced and more similar to behavioral reports as we move up the ventral visual stream. The authors would like to thank Clara Fannjiang for her assistance in running these additional analyses.



Supplementary Fig. A.2. **Category boundary effect searchlight analysis.** A clear neural basic level advantage may not be clearly evident until we reach higher-level, perhaps amodal representations of visual information. To investigate whether we can find evidence of such a higher-level area, we performed a searchlight analysis [80] in which we tiled the cortical gray matter of each subject with fixed-size spheres (radius = 4 voxels) and computed our category boundary effect measure for the neural activity pattern contained within each such sphere across the entire cortex. This new analysis did not reveal any significant category boundary effect beyond occipito-temporal cortex: we show above average maps of all subjects' occipito-temporal cortex aligned to Talairach atlas. However, we found that the subordinate level is well represented in early visual areas, more strongly than LOC, which is consistent with our ROI analyses. Moreover, the basic level analysis only survives multiple comparisons testing in object-selective cortex. Given the inherent limitations of a searchlight analysis (fixed searchlight size, ignoring putative boundaries between functional selective regions, highly stringent multiple comparison correction thresholds), the results are much more coarse than the ROI-specific analyses, so our null result in this case does not necessarily imply that a high-level area with a clear basic-level advantage does not exist. In fact, the search for such an area and / or representation is an interesting avenue of future study.

Appendix B

Typicality Sharpens Category Representations in Object-Selective Cortex

List of Subordinate Categories (categories in blue selected for fMRI Experiment)

- Natural object
  - Animals
    - Birds
      - Cockatiel
      - Humming bird
      - Vulture
      - Hawk
      - Owl
      - Hen
      - Ostrich
      - Swan
    - Cats
      - Egyptian
      - Angora
      - Manx
      - Abyssinian
      - Tortoiseshell
      - Siamese
      - Persian
      - Sphinx
    - Dogs
      - Malamute
      - Mastiff
      - Pug
      - Schipperke
      - Chihuahua

- Welsh Corgi
- Schnauzer
- Komondor

* Fish
  - Goldfish
  - Clownfish
  - Angelfish
  - Sturgeon
  - Flying fish
  - Pufferfish
  - Needlefish
  - Catfish

– Plants

* Flowers
  - Violet
  - Chrysanthemum
  - Blue daisy
  - Cosmos
  - Ice poppy
  - Orchid
  - Sunflower
  - Toadflax
* Garden Plants
  - Zucchini
  - Broccoli
  - Cabbage
  - Corn
  - Cucumber

        - Jalapeno
        - Soybeans
        - Tomato
    - Herbs
        - Parsley
        - Mint
        - Basil
        - Catnip
        - Chives
        - Cilantro
        - Sage
        - Oregano
    - Trees
        - Aspen tree
        - Coffee tree
        - Conifer tree
        - Apple tree
        - Bonsai
        - Magnolia tree
        - Palm tree
        - Willow

- Man-Made object
    - Transportation
        - Boats
            - Canoe
            - Rowboat
            - Galleon

- Cruise ship
- Battleship
- Icebreaker
- Sailboat
- Aircraft carrier

* Cars
  - Sedan
  - Sports car
  - Minivan
  - Limousine
  - Mini car
  - Racecar
  - Station wagon
  - Antique car
* Planes
  - Airliner
  - Fighter plane
  - Seaplane
  - Glider
  - Delta plane
  - Biplane
  - Stealth plane
  - Gyroplane
* Trains
  - Commuter train
  - Freight train
  - Subway
  - Tram

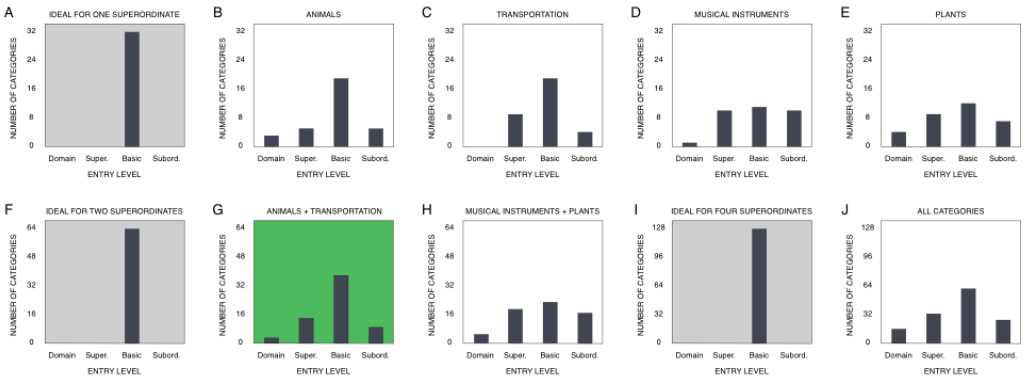
- Monorail
- Bullet train
- Incline railway
- Trolley

– Musical Instruments

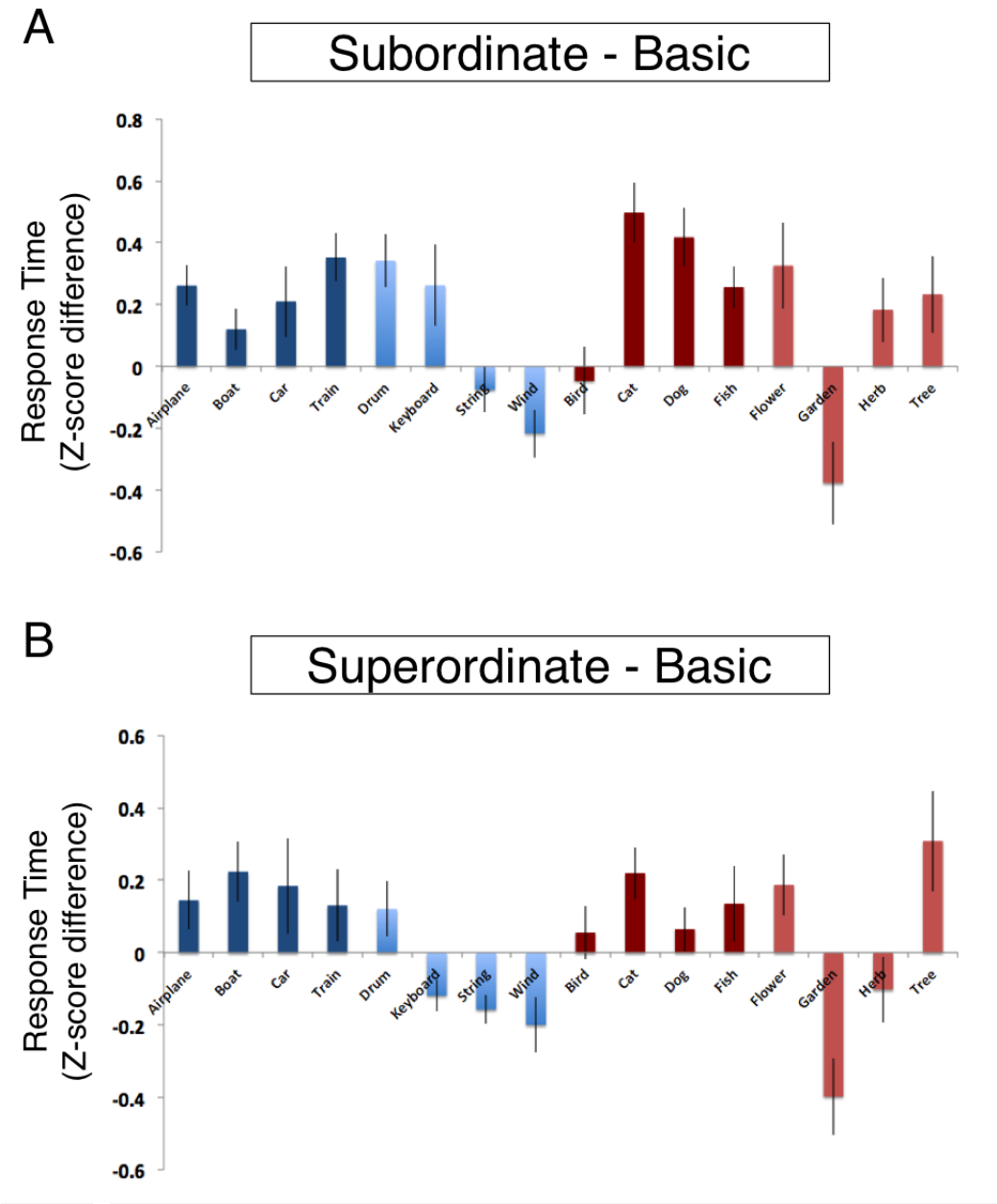
* Drums
  - Bass drum
  - Bongos
  - Timpani
  - Snare drum
  - Steel drum
  - Tenor drum
  - Timbale
  - Bodhran
* Keyboards
  - Upright piano
  - Hammond organ
  - Clavichord
  - Grand piano
  - Harpsichord
  - Mechanical piano
  - Pipe organ
  - Synthesizer
* Strings
  - Violin
  - Sitar
  - Dulcimer
  - Cello

- Bass
- Banjo
- Viola
- Mandolin

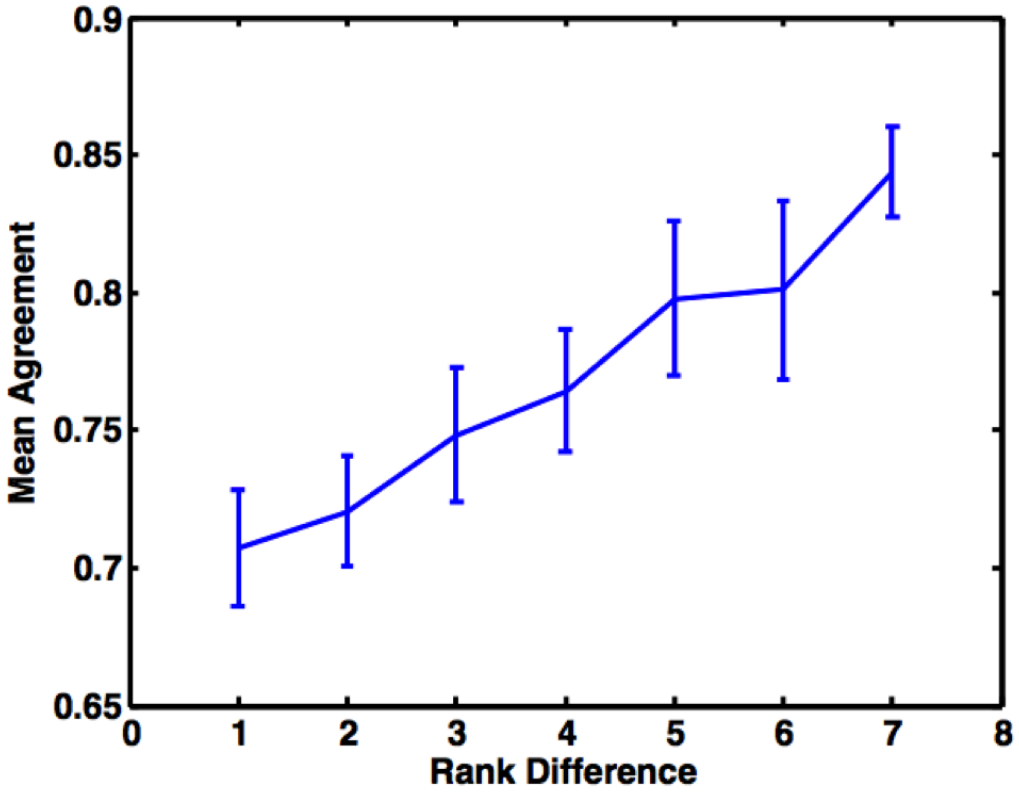
* Wind Instruments
  - Clarinet
  - Oboe
  - Flute
  - Pennywhistle
  - Trombone
  - Trumpet
  - Saxophone
  - Tuba



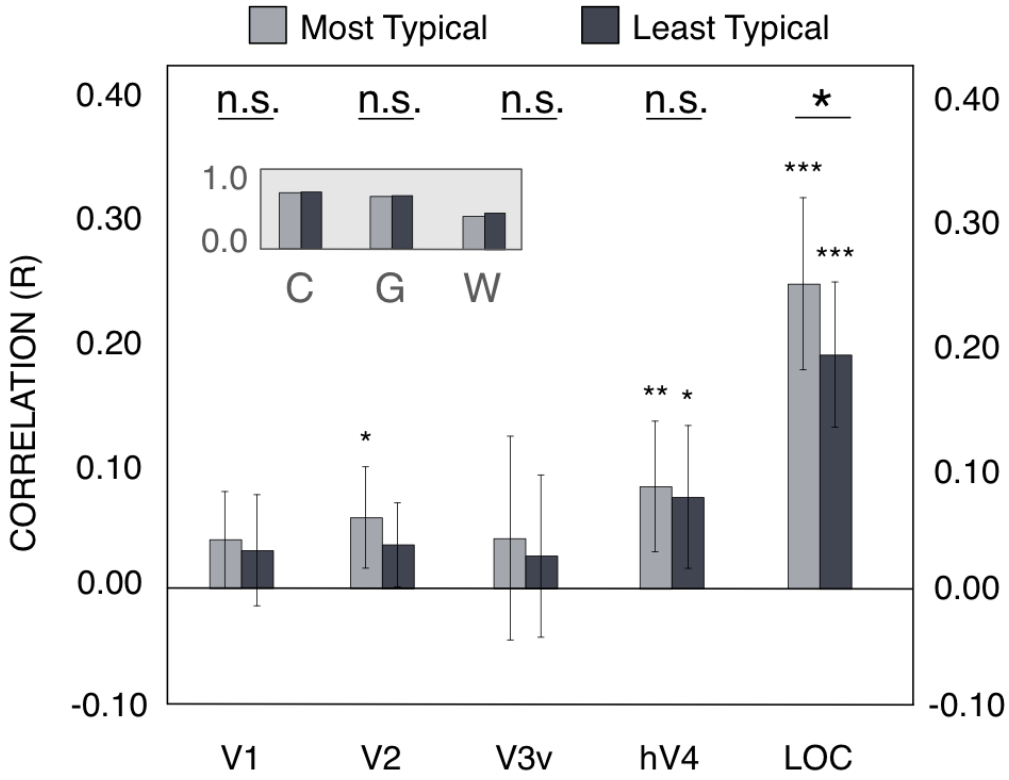
Supplementary Fig. B.1. **Distribution of entry-levels for the 128 subordinate category dataset.** The X-axis denotes the number of categories behaviorally verified to have the corresponding entry level on the Y-axis. Ideally, we would like entry levels for all categories to lie at the basic level (A, F, I). (B-E) Entry-level distribution for each subordinate category in a particular superordinate category. (G-H, J) Distribution of entry levels collapsed across superordinate categories: (G) animals and transportation categories; (H) plants and musical instruments; (J) all superordinate categories. Green highlight indicates category set distribution chosen for the fMRI experiment data analysis.



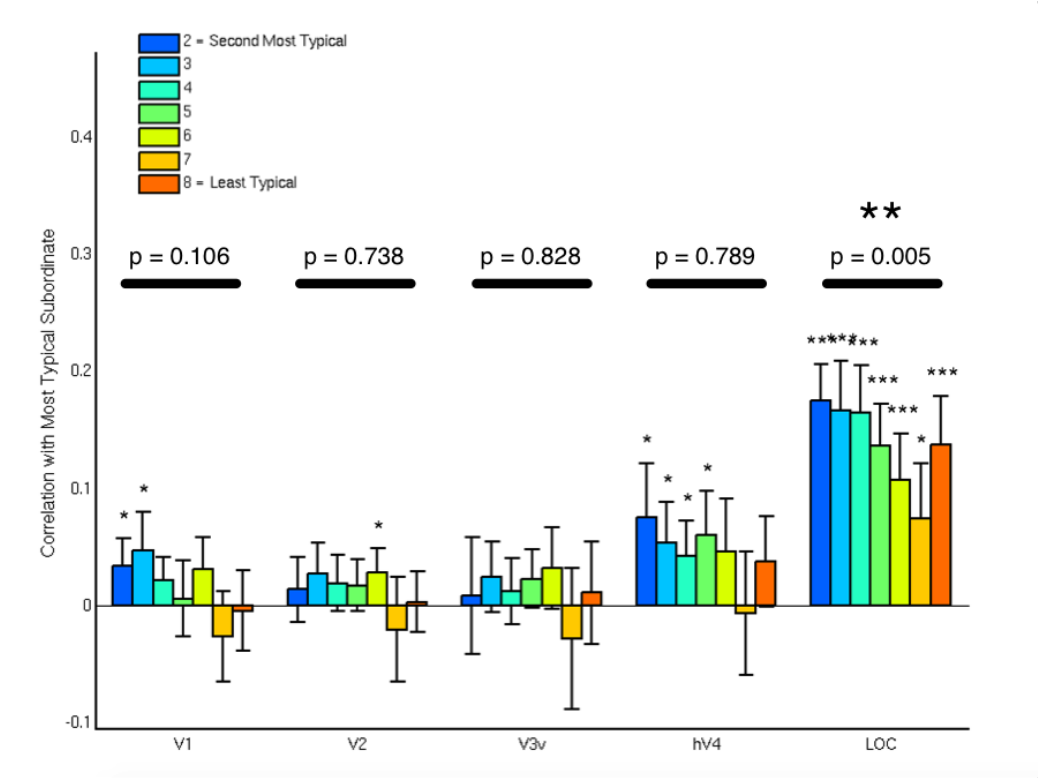
Supplementary Fig. B.2. **Response time z-score difference between putative categorization levels.** (A) Advantage of basic-level classification versus subordinate level classification (positive values indicate strong basic-level effects). (B) Advantage of basic-level classification versus superordinate level classification (positive values indicate strong basic-level effects). Error bars: standard error of the mean.



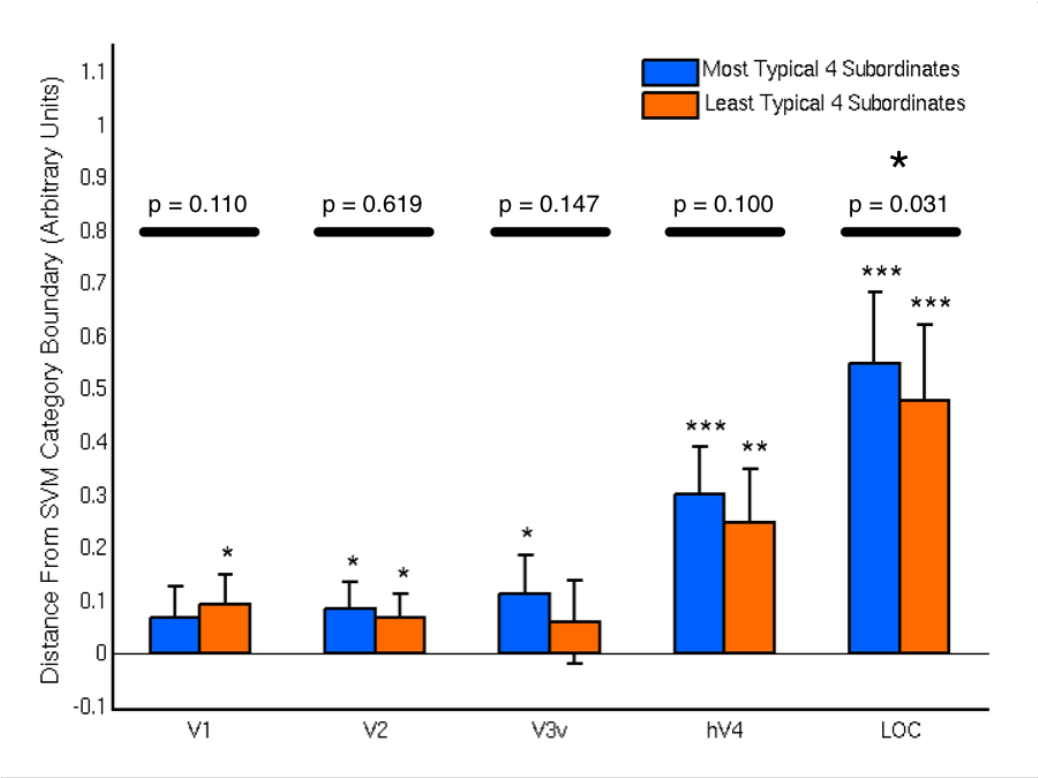
Supplementary Fig. B.3. **Inter-Subject Reliability for Typicality Rankings.** The X-axis denotes the absolute value difference in typicality rank between categories, where rank 1 indicates the most typical subordinate and rank 8 indicates the least typical subordinate in a given basic category. The Y-axis denotes the level of agreement between subjects, where chance is 0.5 (typicality rankings were obtained through judgment of pairs of two categories). We observed a mean agreement of 75% ± 2%, mean ± s.e.m. across all rank differences.



Supplementary Fig. B.4. **Correlation with central category tendency when omitting the most and least typical exemplar in the computation of the central tendency.** Correlation between central category tendency and most typical exemplar in each category (light gray) or least typical exemplar in each category (dark gray), averaged across all 8 basic level categories. In object-selective cortex (LOC), typical categories are more similar to the average category representation than less typical exemplars and this effect is not present in early visual areas. We performed a similar analysis using the image-level features from our stimulus set: LAB color histograms, GIST features, and multi-scale Gabor wavelet features. All features show similar values for both highly typical and less typical exemplar correlations and all features show an opposite trend to our LOC results (higher correlation for less typical exemplars). $***$ $p < .001$, $**$ $p < .01$, $*$ $p < .05$, n.s. - not significant. Error bars: 95% confidence interval.



Supplementary Fig. B.5. **Similarity of Each Subordinate with Most Typical Subordinate Category.** Each bar represents similarity between a subordinate category at a particular typicality rank (2–8) and the most typical subordinate (rank 1) in its corresponding basic category, averaged across all basic categories. Using Friedman non-parametric tests, we found that similarity decreases significantly with typicality rank in object-selective cortex (LOC), but not in early visual regions (V1, V2, V3v, hV4). $***$ $p < .001$, $**$ $p < .01$, $*$ $p < .05$. Error bars: 95% confidence interval.



Supplementary Fig. B.6. **Distances from SVM Category Boundary.** We hypothesized that if more typical subordinates are better separated form other basic categories, then they should exhibit larger distances (compared to less typical subordinates) to a putative category boundary separating their basic category from others. To test this hypothesis, we trained a series of linear support vector machine (SVM) classifiers and computed, for each of our 8 typicality ranks, the average distance per subject per ROI between all subordinates of that rank and their corresponding category boundary. Graph shows average distance from linear support vector machine (SVM) category boundaries averaged over most typical four subordinates and least typical four subordinates. We found that typical subordinates usually lie farther from their respective category boundary than less typical subordinates in LOC, but not in early visual areas (V1, V2, V3v, hV4). $***$ $p < .001$, $**$ $p < .01$, $*$ $p < .05$. Error bars: 95% confidence interval.

Supplementary Fig. B.7. **Stimuli for "Pug" Subordinate Category.** We selected 16 images from each subordinate category varying greatly in pose, color, shape, and shape-occlusion. Shown above are the 16 images presented for the "pug" category.

Appendix C

Category Boundaries and Typicality Warp the Neural Representation Space of Real-World Objects in Human Ventral Visual Cortex

Experiment 1: Full Names and Typicality Ratings for Initial Set of 48 Subordinate Categories. Typicality ratings are computed as proportion chosen in a pairwise 2AFC task (0 = never chosen, low typicality; 1 = always chosen, high typicality). Color highlights indicate subordinates that were eventually selected as stimuli for the scanning phase of Experiment 1 (green = high typicality; orange = middle typicality; red = low typicality; black = not chosen for scanning).

- Dogs
  - Golden retriever (0.83)
  - Beagle (0.76)
  - Saint Bernard (0.71)
  - Mastiff (0.67)
  - Collie (0.65)
  - Basset hound (0.62)
  - Elk hound (0.60)
  - Welsh corgi (0.58)
  - Malamute (0.58)
  - Doberman (0.55)
  - Pug (0.51)
  - Bloodhound (0.50)
  - Schnauzer (0.49)
  - Terrier (0.48)
  - Sheepdog (0.46)
  - Schipperke (0.40)
  - Pomeranian (0.36)
  - Chow-chow (0.35)
  - Airedale (0.32)

  - Dinmont (0.30)
  - Poodle (0.27)
  - Chihuahua (0.23)
  - Afghan hound (0.17)
  - Komondor (0.11)

- Cars
  - Ford Mustang (0.84)
  - Chevrolet Crossfire (0.79)
  - BMW Z4 (0.77)
  - Rolls Royce (0.71)
  - Lamborghini Diablo (0.71)
  - Kia Rio (0.70)
  - Volvo Hatchback (0.69)
  - Volkswagen Cabrio (0.68)
  - Toyota Prius (0.66)
  - Lotus Elise (0.62)
  - Caddillac (0.58)
  - Mini Cooper (0.50)
  - Mitsubishi Miev (0.42)
  - Land Rover (0.39)
  - Nissan Cube (0.36)
  - Isuzu Vehicross (0.34)
  - Honda Element (0.30)
  - Antique car (0.28)

  - Racecar (0.28)
  - Minivan (0.27)
  - Jeep Wrangler (0.23)
  - Ford Ranger (0.16)
  - Limousine (0.13)
  - Hummer (0.10)

Full Names and Typicality Ratings for the 64 Subordinate Categories Used in Experiment 2. Typicality ratings are computed as proportion chosen in a pairwise 2AFC task (0 = never chosen, low typicality; 1 = always chosen, high typicality). Color highlights indicate typicality tier (green = high typicality; red = low typicality).

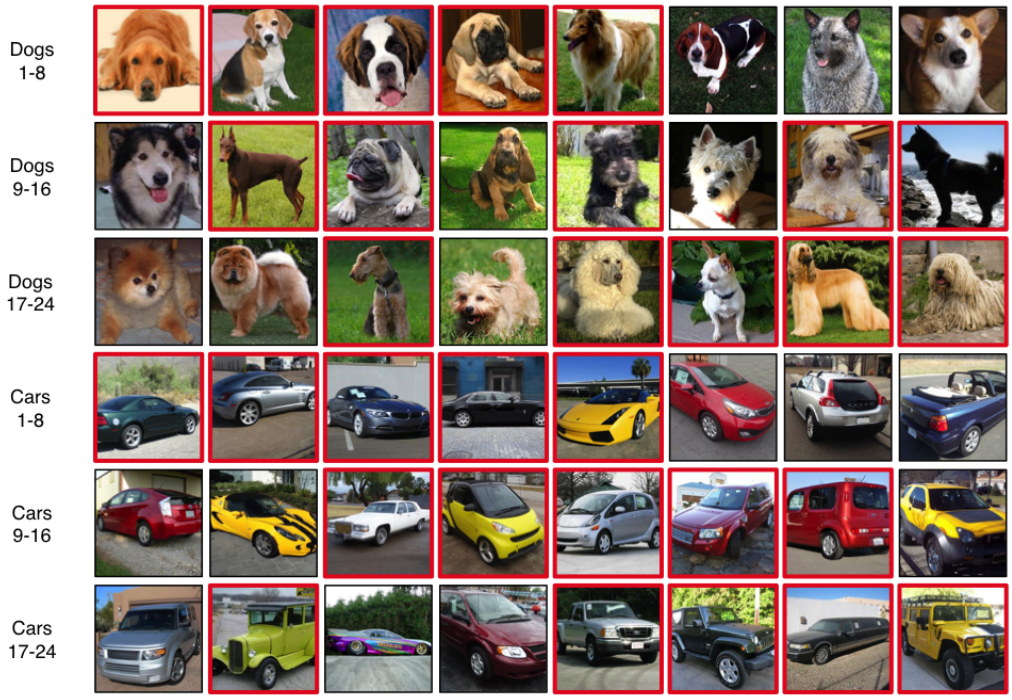
- Natural objects / Animals
  - Birds
    - Cockatiel (0.78)
    - Humming bird (0.75)
    - Vulture (0.64)
    - Hawk (0.57)
    - Owl (0.53)
    - Hen (0.42)
    - Ostrich (0.25)
    - Swan (0.22)
  - Cats
    - Egyptian (0.70)
    - Angora (0.64)
    - Manx (0.62)

        * Abyssinian (0.57)
        * Tortoiseshell (0.56)
        * Siamese (0.54)
        * Persian (0.41)
        * Sphinx (0.13)
    - Dogs
        * Malamute (0.70)
        * Mastiff (0.63)
        * Pug (0.53)
        * Schipperke (0.53)
        * Chihuahua (0.40)
        * Welsh Corgi (0.32)
        * Schnauzer (0.31)
        * Komondor (0.24)
    - Fish
        * Goldfish (0.76)
        * Clownfish (0.74)
        * Angelfish (0.66)
        * Sturgeon (0.62)
        * Flying fish (0.43)
        * Pufferfish (0.34)
        * Needlefish (0.34)
        * Catfish (0.31)

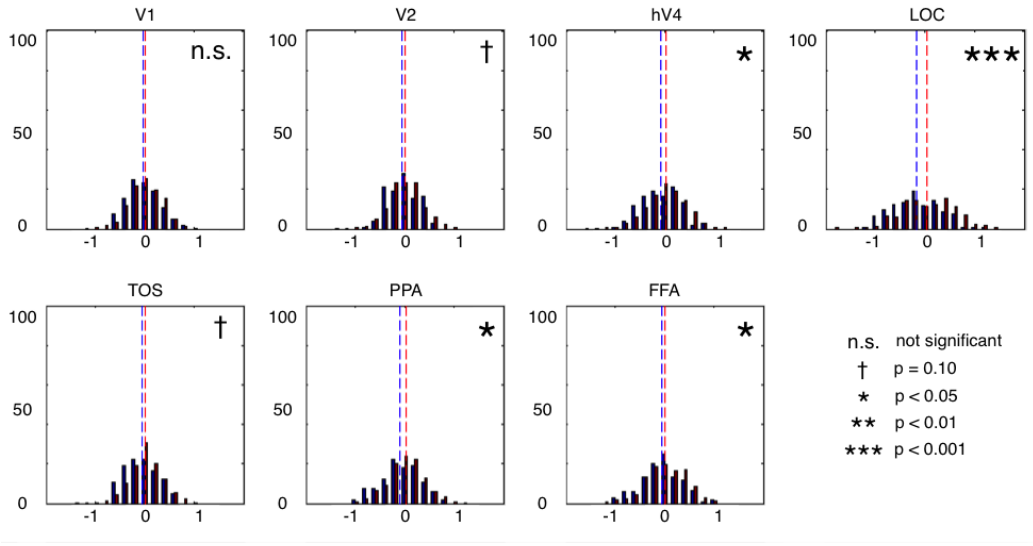
- Man-Made objects / Vehicles
    - Boats
        * Canoe (0.67)

    - Rowboat (0.53)
    - Galleon (0.49)
    - Cruise ship (0.48)
    - Battleship (0.45)
    - Icebreaker (0.41)
    - Sailboat (0.27)
    - Aircraft carrier (0.26)
  - Cars
    - Sedan (0.67)
    - Sports car (0.61)
    - Minivan (0.60)
    - Limousine (0.51)
    - Mini car (0.37)
    - Race car (0.36)
    - Station wagon (0.33)
    - Antique car (0.25)
  - Planes
    - Airliner (0.88)
    - Fighter plane (0.58)
    - Seaplane (0.55)
    - Glider (0.50)
    - Delta plane (0.47)
    - Biplane (0.31)
    - Stealth plane (0.28)
    - Gyroplane (0.28)
  - Trains
    - Commuter train (0.71)

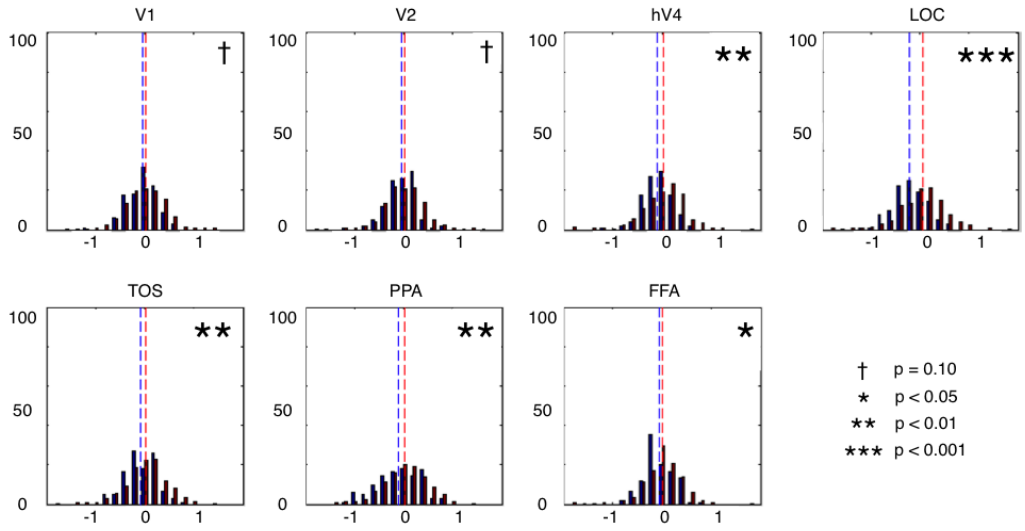
* Freight train (0.70)
* Subway (0.65)
* Tram (0.57)
* Monorail (0.44)
* Bullet train (0.41)
* Incline railway (0.40)
* Trolley (0.32)



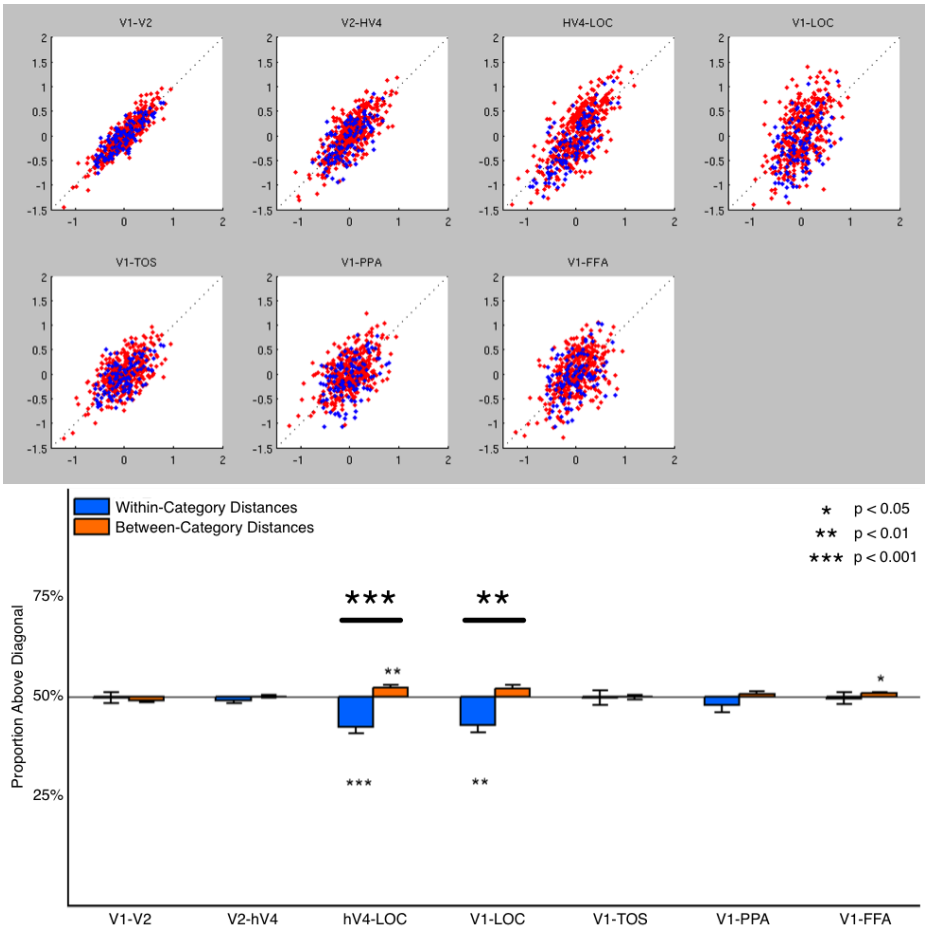
Supplementary Fig. C.1. **Representative Images from the Initial 48 Subordinate Categories in Experiment 1.** We selected twenty-found subordinates from each of two basic level categories known from prior work to be well differentiable in their elicited patterns of activity in occipito-temporal cortex (see e.g. [68]). Full names of the subordinate categories ordered by their typicality are given in the first list above. Red borders indicate subordinates that were ultimately chosen as stimuli for the scanning phase of Experiment 1.



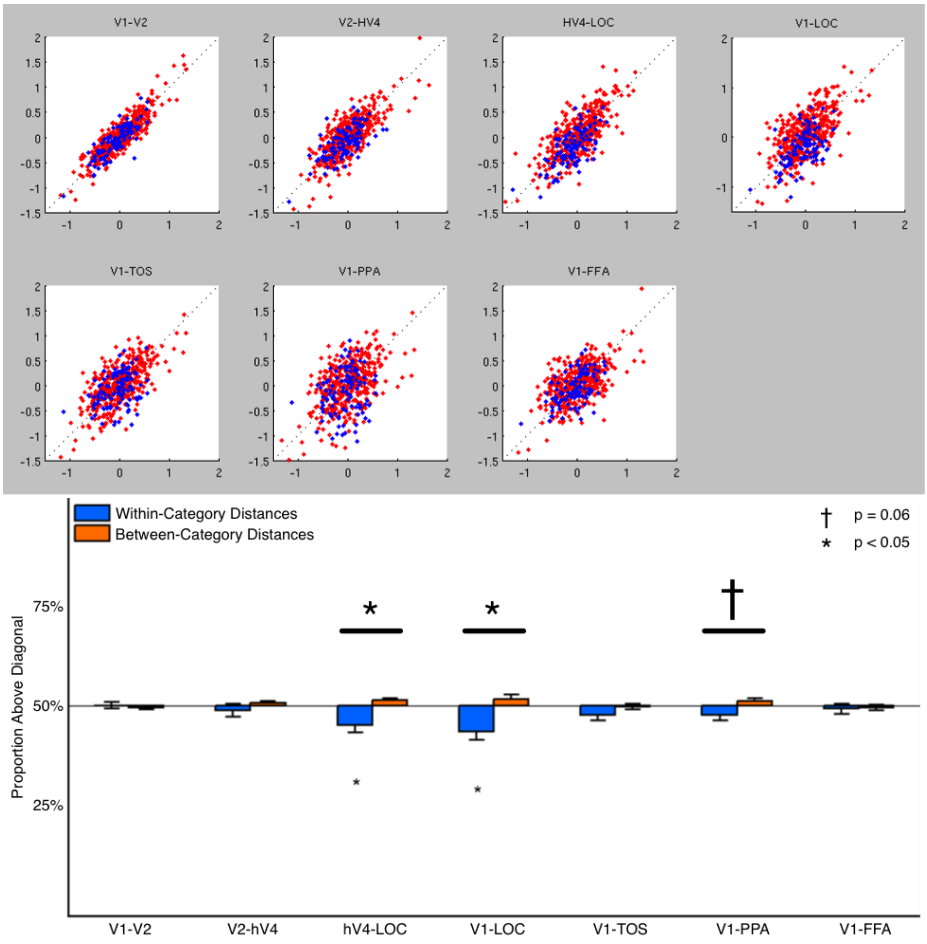
Supplementary Fig. C.2. **Category Distance Histograms for Natural / Animals Superordinate Category Subset of Experiment 2.** Graphs show z-scored Pearson correlation distance histograms for within-category distances (blue) and between-category distances (red) for early visual (V1, V2, hV4), object-selective (LOC), scene-selective (PPA, TOS), face-selective (FFA) regions. The natural basic categories "dog" and "car" are reasonably separable in virtually all brain regions except V1, with the highest distinction arising in LOC (top right, grey). The natural basic categories are reasonably separable in virtually all brain regions (with the exception of V1), with the highest distinction arising in LOC (top right). This suggests that a sharp qualitative change in the structure of the feature space arises between hV4 and LOC, which may be mirrored in other stimulus selective regions of occipito-temporal cortex.



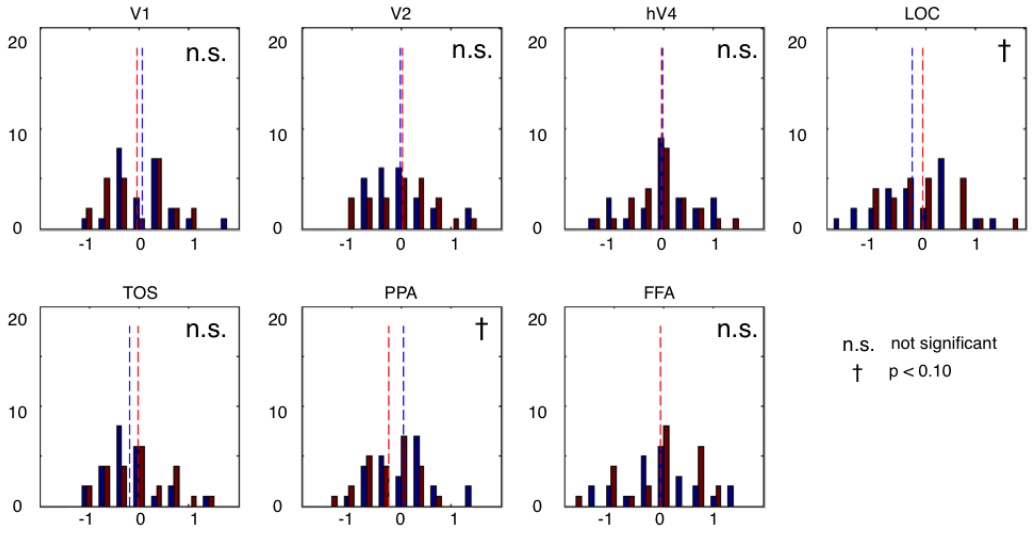
Supplementary Fig. C.3. **Category Distance Histograms for Man-Made / Vehicles Superordinate Category Subset of Experiment 2.** Graphs show z-scored Pearson correlation distance histograms for within-category distances (blue) and between-category distances (red) for early visual (V1, V2, hV4), object-selective (LOC), scene-selective (PPA, TOS), face-selective (FFA) regions. The man-made basic categories are reasonably separable in virtually all brain regions, with the highest distinction arising in LOC (top right). This suggests that a sharp qualitative change in the structure of the feature space arises between hV4 and LOC, which may be mirrored in other stimulus selective regions of occipito-temporal cortex.



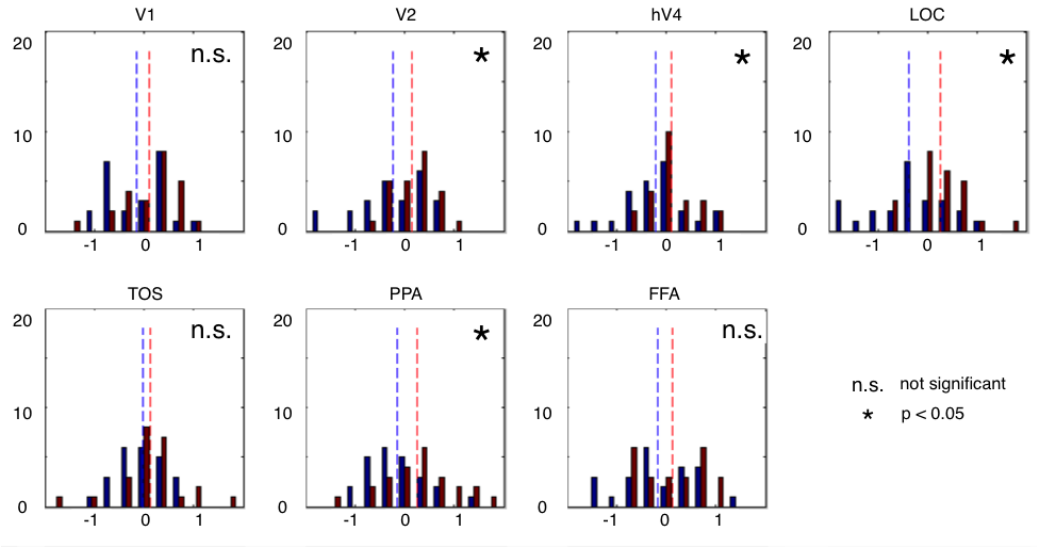
Supplementary Fig. C.4. **Category Warping for Natural / Animals Superordinate Category Subset of Experiment 2.** (Top, Middle) Graphs show how representations of distances corresponding to subordinate category pairs change as we move up the ventral visual stream. Representations are relatively stable between early visual regions (V1, V2, hV4), as well as between early visual cortex and scene- and face-selective regions (PPA, TOS, FFA). However, we see a striking shift in the quality of the representation as we move between hV4 and LOC. Here, within-category distance pairs lie below the diagonal, while between-category distance pairs sit above the diagonal, which indicates that the feature space of LOC shrinks relative distances within categories and expands relative distances between categories, compared to the feature space of V1.



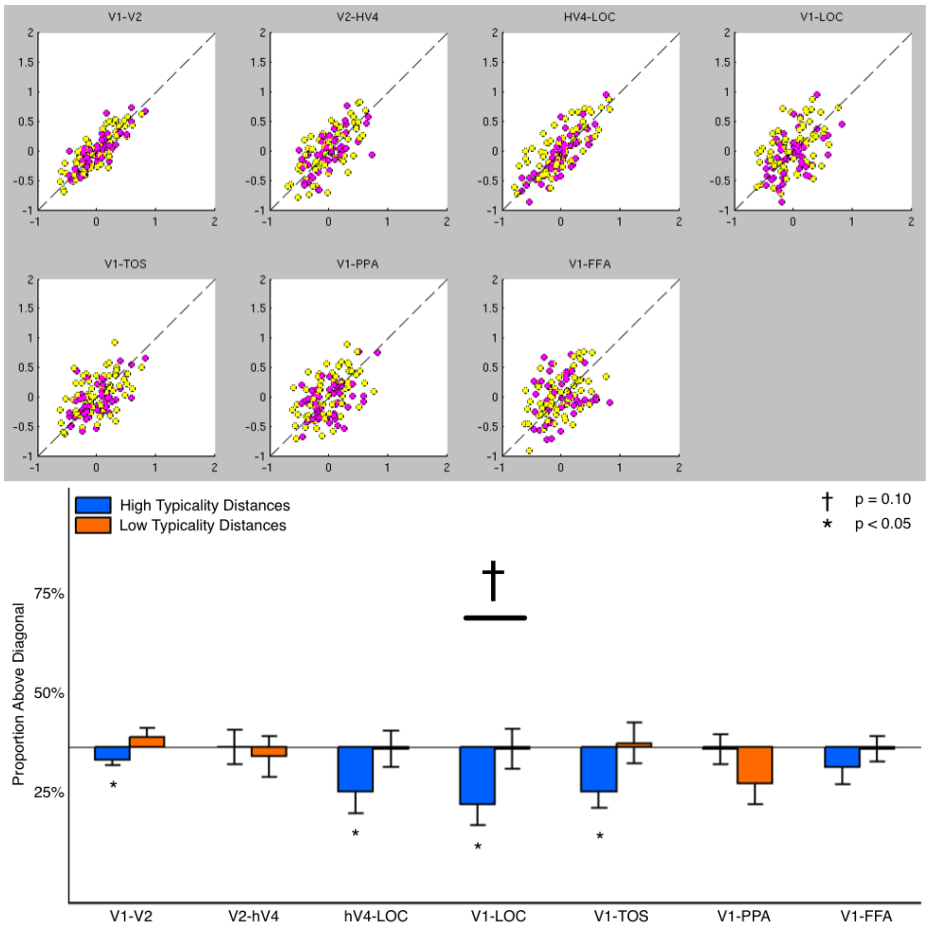
Supplementary Fig. C.5. **Category Warping for Man-Made / Vehicles Superordinate Category Subset of Experiment 2.** (Top, Middle) Graphs show how representations of distances corresponding to subordinate category pairs change as we move up the ventral visual stream. Representations are relatively stable between early visual regions (V1, V2, hV4), as well as between early visual cortex and face- and dorsal scene-selective regions (TOS, FFA). However, we see a striking shift in the quality of the representation as we move between hV4 and LOC. Here, within-category distance pairs lie below the diagonal, while between-category distance pairs sit above the diagonal, which indicates that the feature space of LOC shrinks relative distances within categories and expands relative distances between categories, compared to the feature space of V1. A weak category warping effect is also observed between early visual cortex and PPA, which may be due to this region's predilection for processing and representing contextual effects [85].



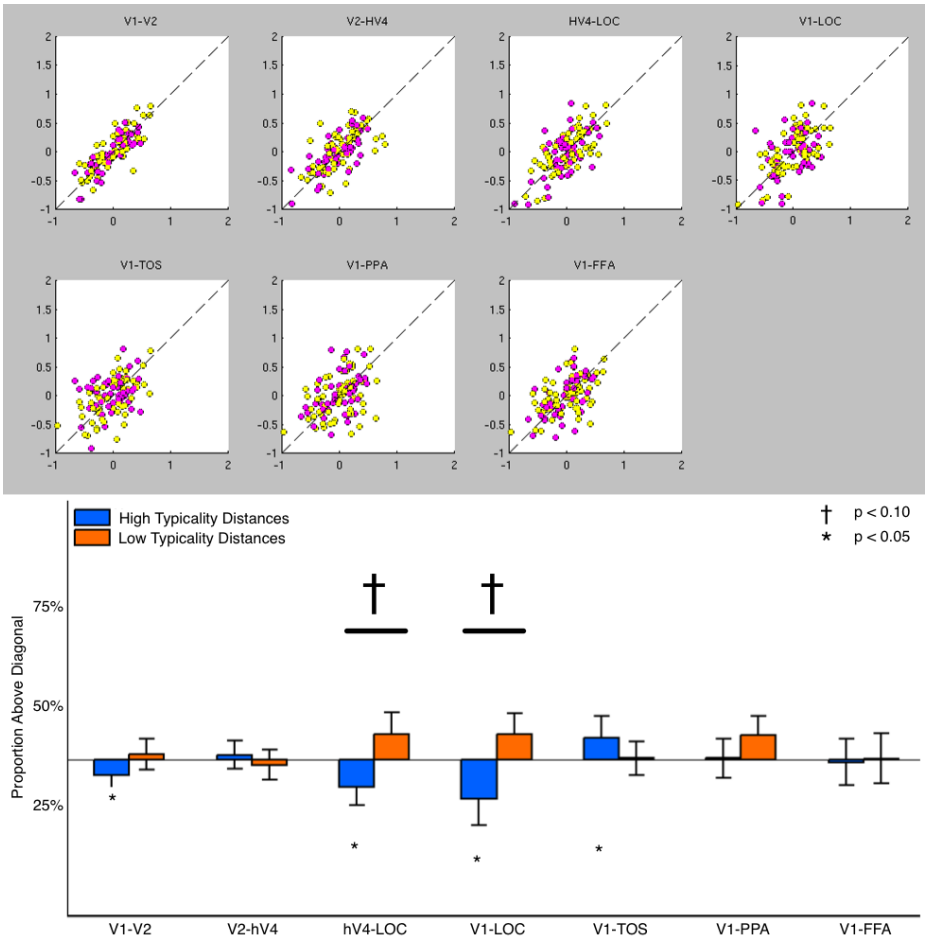
Supplementary Fig. C.6. **Typicality Distance Histograms for Natural / Animals Superordinate Category Subset of Experiment 2.** Graphs show z-scored Pearson correlation distance histograms for within-highly-typical-subordinates distances (blue) and within-less-typical-subordinates distances (red) for early visual (V1, V2, hV4), object-selective (LOC), scene-selective (PPA, TOS), face-selective (FFA) regions. In early visual regions, face- and scene-selective regions, typicality does not significantly modulate the representation of real-world objects. By contrast, we observed a trend for typical and less typical subordinates to be more strongly separable in LOC (top right), which suggests a qualitative change in the structure of the feature space may arise between hV4 and LOC, which is not mirrored in other stimulus selective regions of occipito-temporal cortex.



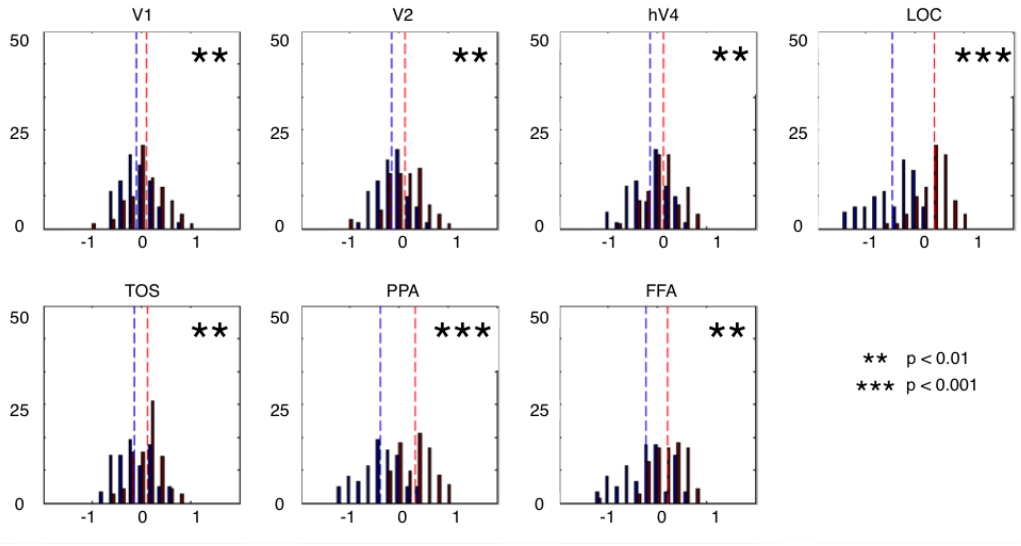
Supplementary Fig. C.7. **Typicality Distance Histograms for Man-Made / Vehicles Superordinate Category Subset of Experiment 2.** Graphs show z-scored Pearson correlation distance histograms for within-highly-typical-subordinates distances (blue) and within-less-typical-subordinates distances (red) for early visual (V1, V2, hV4), object-selective (LOC), scene-selective (PPA, TOS), face-selective (FFA) regions. In contrast to the natural superordinate category, here we observed that typicality indeed modulated the representation of object categories beginning in intermediate visual regions (V2, hV4), strongest in object-selective regions (LOC), and to a lesser degree in the most anterior scene-selective region (PPA). By contrast, dorsal scene-selective regions and face-selective cortex were not affected by typicality (TOS, FFA), perhaps due to the lack of explicit face information in the man-made stimuli.



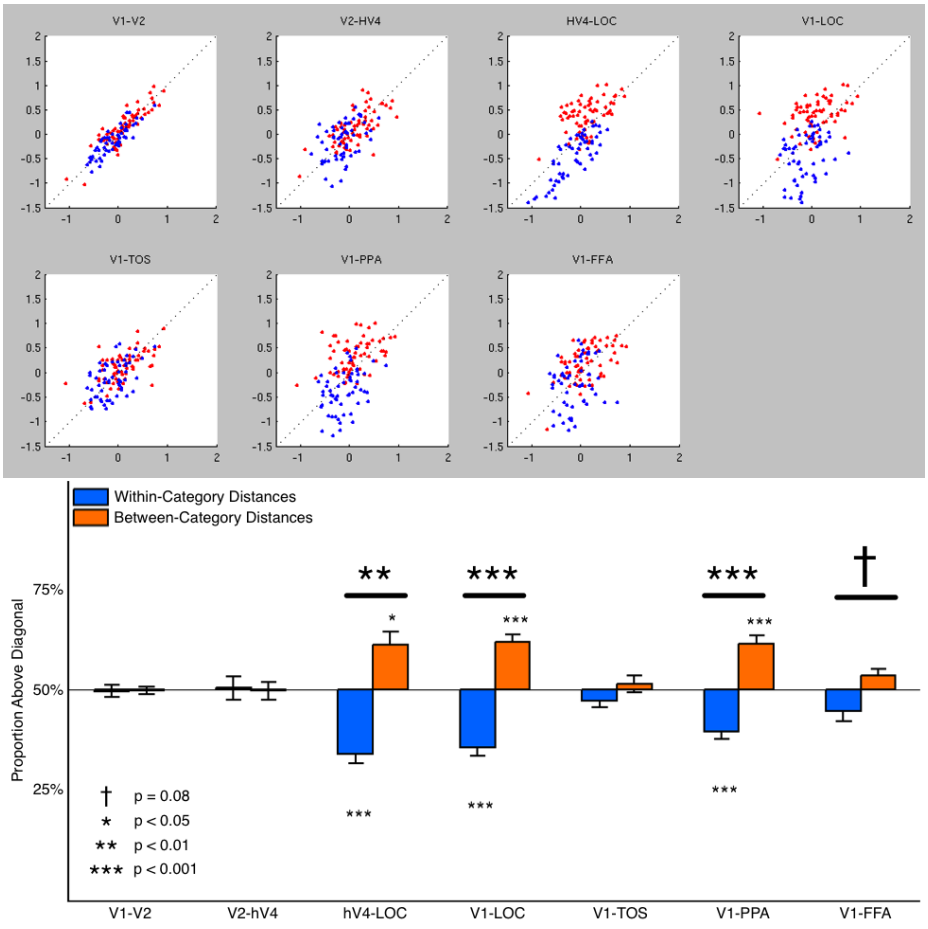
Supplementary Fig. C.8. **Typicality Warping for Natural / Animals Superordinate Category Subset of Experiment 2.** (Top, Middle) Graphs show how representations of distances corresponding to subordinate category pairs of high (purple) and low (yellow) typicality change as we move up the ventral visual stream. Representations are relatively stable between early visual regions (V1, V2, hV4), as well as between early visual cortex and scene- and face-selective regions (PPA, TOS, FFA). However, we see a trend for the representation to shift as we move between hV4 and LOC. Here, high typicality subordinate category pairs exhibit a tendency to lie below the diagonal, which indicates that the feature space of LOC shrinks relative distances between typical exemplars within a category compared to the feature space of V1.



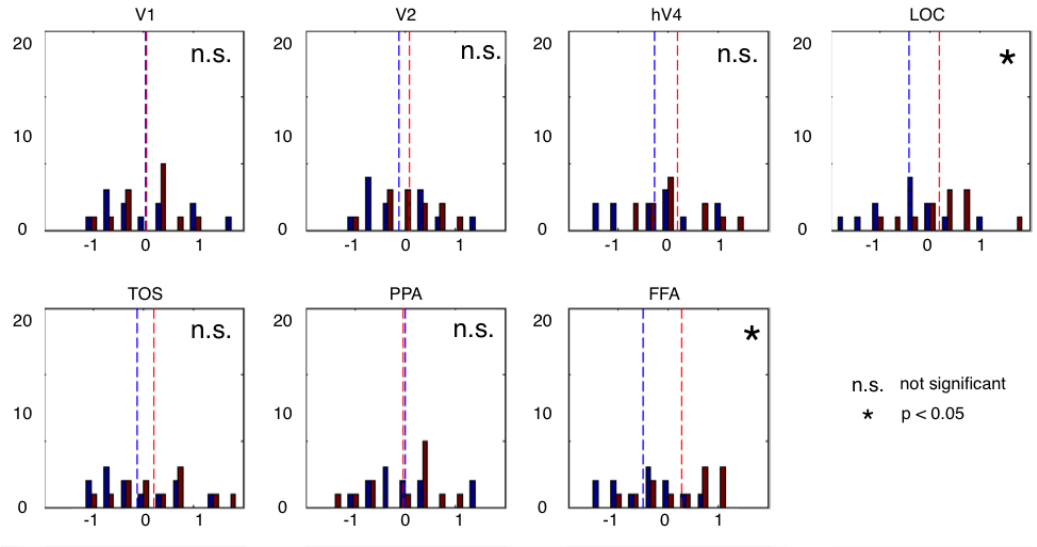
Supplementary Fig. C.9. **Typicality Warping for Man-Made / Vehicles Superordinate Category Subset of Experiment 2.** (Top, Middle) Graphs show how representations of distances corresponding to subordinate category pairs of high (purple) and low (yellow) typicality change as we move up the ventral visual stream. Representations are relatively stable between early visual regions (V1, V2, hV4), as well as between early visual cortex and scene- and face-selective regions (PPA, TOS, FFA). However, we see a trend for the representation to shift as we move between hV4 and LOC. Here, high typicality subordinate category pairs exhibit a tendency to lie below the diagonal, which indicates that the feature space of LOC shrinks relative distances between typical exemplars within a category compared to the feature space of V1.



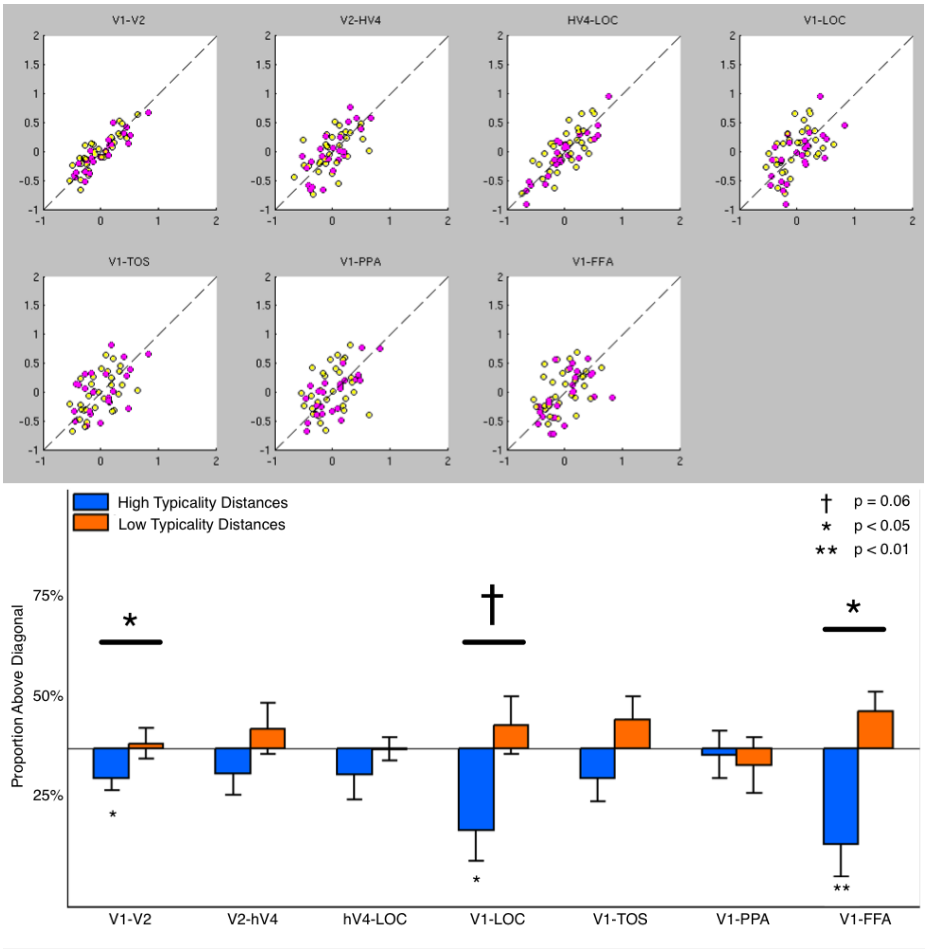
Supplementary Fig. C.10. **Category Distance Histograms for "Dog" and "Car" Subset of Experiment 2.** Graphs show z-scored Pearson correlation distance histograms for within-category distances (blue) and between-category distances (red) for early visual (V1, V2, hV4), object-selective (LOC), scene-selective (PPA, TOS), face-selective (FFA) regions. The natural basic categories "dog" and "car" are reasonably separable in virtually all brain regions, with the highest distinction arising in LOC (top right, grey). The natural basic categories are reasonably separable in virtually all brain regions (with the exception of V1), with the highest distinction arising in LOC (top right). This suggests that a sharp qualitative change in the structure of the feature space arises between hV4 and LOC, which may be mirrored in other stimulus selective regions of occipito-temporal cortex.



Supplementary Fig. C.11. **Category Warping for "Dog" and "Car" Subset of Experiment 2.** (Top, Middle) Graphs show how representations of distances corresponding to subordinate category pairs change as we move up the ventral visual stream. Representations are relatively stable between early visual regions (V1, V2, hV4), as well as between early visual cortex and dorsal scene- and face-selective regions (TOS, FFA). However, we see a striking shift in the quality of the representation as we move between hV4 and LOC. Here, within-category distance pairs lie below the diagonal, while between-category distance pairs sit above the diagonal, which indicates that the feature space of LOC shrinks relative distances within categories and expands relative distances between categories, compared to the feature space of V1. This effect is also observed in the ventral scene-selective region PPA, which may be due to this region's predilection for processing and representing contextual effects [85].



Supplementary Fig. C.12. **Typicality Distance Histograms for "Dog" and "Car" Subset of Experiment 2.** Graphs show z-scored Pearson correlation distance histograms for within-highly-typical-subordinates distances (blue) and within-less-typical-subordinates distances (red) for early visual (V1, V2, hV4), object-selective (LOC), scene-selective (PPA, TOS), face-selective (FFA) regions. In early visual regions and scene-selective regions, typicality does not significantly modulate the representation of real-world objects. By contrast, we observed a trend for typical and less typical subordinates to be more strongly separable in LOC (top right), which suggests a qualitative change in the structure of the feature space may arise between hV4 and LOC, and to a lesser degree in face-selective cortex (FFA).



Supplementary Fig. C.13. **Typicality Warping for "Dog" and "Car" Subset of Experiment 2.** (Top, Middle) Graphs show how representations of distances corresponding to subordinate category pairs of high (purple) and low (yellow) typicality change as we move up the ventral visual stream. We observe that typicality modulates the representation of object categories across the ventral visual stream, with discrete step changes between V1 - V2 and hV4 - LOC. Here, high typicality subordinate category pairs exhibit a tendency to lie below the diagonal, which indicates that the feature space of LOC shrinks relative distances between typical exemplars within a category compared to the feature space of V1. This trend is also mirrored between early visual cortex and face-selective regions (FFA), perhaps due to the presence of face stimuli in our "dog" basic category.

Bibliography

[1] J. T. Abbott, J. L. Austerweil, and T. L. Griffiths. "Constructing a hypothesis space from the Web for large-scale Bayesian word learning". *Proceedings of the 34th Annual Conference of the Cognitive Science Society*. Ed. by N. Miyake, D. Peebles, and R. P. Cooper. Austin, TX: Cognitive Science Society, 2012, pp. 54–59.

[2] H. J. Aizenstein, A. W. MacDonald, V. A. Stenger, R. D. Nebes, J. K. Larson, S. Ursu, and C. S. Carter. "Complementary category learning systems identified using event-related functional MRI". *Journal of Cognitive Neuroscience* 12 (2000), pp. 977–987.

[3] K. Amano, B. A. Wandell, and S. O. Dumoulin. "Visual field maps, population receptive field sizes, and visual field coverage in the human MT complex". *Journal of Neurophysiology* 102 (2009), pp. 2704–2718.

[4] A. Amedi, G. Jacobson, T. Hendler, R. Malach, and E. Zohary. "Convergence of visual and tactile shape processing in the human lateral occipital complex". *Cerebral Cortex* 12 (2002), pp. 1202–1212.

[5] A. Amedi, R. Malach, T. Hendler, S. Peled, and E. Zohary. "Visuo-haptic object-related activation in the ventral visual pathway". *Nature Neuroscience* 4 (2001), pp. 324–330.

[6] J. M. Anglin. *Word, object, and conceptual development*. New York, NY: Norton, 1977.

[7] J Arizpe, D. Kravitz, E. Bilger, and C. Baker. "Increasing extent of category selectivity with increasing power". *Journal of Vision* 14 (2014), p. 117.

[8] F. G. Ashby and W. T. Maddox. "Human category learning". *Annual Review of Psychology* 56 (2005), pp. 149–178.

[9] F. G. Ashby and W. T. Maddox. "Relations between prototype, exemplar, and decision bound models of categorization". *Journal of Mathematical Psychology* 37 (1993), pp. 372–400.

[10] B. Berlin. *Ethnobiological classification: Principles of categorization of plants and animals in traditional societies*. Princeton, NJ: Princeton University Press, 1992.

[11] I. Biederman. "Recognition-by-components: A theory of human image understanding". *Psychological Review* 94 (1987), pp. 115–147.

[12] D. H. Brainard. "The psychophysics toolbox". *Spatial Vision* 10 (1997), pp. 433–436.

[13] K. Braunlich and C. A. Seger. "Categorical evidence, confidence, and urgency during probabilistic categorization". *Neuroimage* 125 (2016), pp. 941–952.

[14] R. Brown. "How shall a thing be called". *Psychology Review* 65 (1958), pp. 14–21.

[15] R. Bruffaerts, P. Dupont, R. Peeters, S. D. Peyne, G. Storms, and R. Vandenberghe. "Similarity of fMRI activity patterns in left perirhinal cortex reflects semantic similarity between words". *Journal of Neuroscience* 33 (2013), pp. 18597–18607.

[16] C. F. Cadieu, H. Hong, D. L. K. Yamins, N. Pinto, D. Ardila, E. A. Solomon, N. J. Majaj, and J. J. DiCarlo. "Deep neural networks rival the representation of primate IT cortex for core visual object recognition". *PLOS Computational Biology* 10 (2014), e1003963.

[17] T. A. Carlson, R. A. Simmons, N. Kriegeskorte, and L. R. Slevc. "The emergence of semantic meaning in the ventral temporal pathway". *Journal of Cognitive Neuroscience* 26 (2013), pp. 120–131.

[18] T. Çukur, S. Nishimoto, A. G. Huth, and J. L. Gallant. "Attention during natural vision warps semantic representation across the human brain". *Nature Neuroscience* 16 (2013), pp. 763–770.

[19] L. L. Chao, J. V. Haxby, and A. Martin. "Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects". *Nature Neuroscience* 2 (1999), pp. 913–919.

[20] R. M. Cichy, Y. Chen, and J. D. Haynes. "Encoding the identity and location of objects in human LOC". *Neuroimage* 54 (2011), pp. 2297–2307.

[21] A. Clarke, P. J. Pell, C. Ranganath, and L. K. Tyler. "Learning warps object representations in the ventral temporal cortex". *Journal of Cognitive Neuroscience* (in press).

[22] A. Clarke and L. K. Tyler. "Object-specific semantic coding in human perirhinal cortex". *Journal of Neuroscience* 34 (2014), pp. 4766–4775.

[23] A. C. Connolly, J. S. Guntupalli, J. Gors, M. Hanke, Y. O. Halchenko, Y. C. Wu, H. Abdi, and J. V. Haxby. "The representation of biological classes in the human brain". *Journal of Neuroscience* 32 (2012), pp. 2608–2618.

[24] B. R. Conroy, B. D. Singer, J. S. Guntupalli, P. J. Ramadge, and J. V. Haxby. "Inter-subject alignment of human cortical anatomy using functional connectivity". *NeuroImage* 81 (2013), 400—411.

[25] D. D. Cox and R. L. Savoy. "Functional magnetic resonance imaging (fMRI) brain reading: detecting and classifying distributed patterns of fMRI activity in human visual cortex". *Neuroimage* 19 (2003), pp. 261–270.

[26] R. W. Cox. " AFNI: software for analysis and visualization of functional magnetic resonance neuroimages". *Comput. Biomed. Res.* 29 (1996), 162—173.

[27] N. Dalal and B. Triggs. "Histograms of Oriented Gradients for Human Detection". *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 1. 2005, pp. 886–893.

[28] T. Davis, K. F. LaRocque, J. A. Mumford, K. A. Norman, A. D. Wagner, and R. A. Poldrack. "What do differences between multi-voxel and univariate analysis mean? How subject-, voxel-, and trial-level variance impact fMRI analysis". *Neuroimage* 97 (2014), pp. 271–283.

[29] T. Davis, B. C. Love, and A. R. Preston. "Learning the exception to the rule: Model-based fMRI reveals specialized representations for surprising category members". *Cerebral Cortex* 22 (2012), pp. 260–273.

[30] T. Davis, B. C. Love, and A. R. Preston. "Striatal and hippocampal entropy and recognition signals in category learning: Simultaneous processes revealed by model-based fMRI". *Journal of Experimental Psychology: Learning, Memory, and Cognition* 38 (2012), pp. 821–839.

[31] T. Davis and R. A. Poldrack. "Quantifying the internal structure of categories using a neural typicality measure". *Cerebral Cortex* 26 (2014), pp. 1–18.

[32] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei. "ImageNet: A large-scale hierarchical image database". *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 1. 2009, pp. 248–255.

[33] J. J. DiCarlo and D. D. Cox. "Untangling invariant object recognition". *Trends in Cognitive Sciences* 11 (2007), pp. 333–341.

[34] D. D. Dilks, J. B. J. dan A. M. Paunov, and N. Kanwisher. "The occipital place area is causally and selectively involved in scene perception". *Journal of Neuroscience* 33 (2013), pp. 1331–1336.

[35] P. E. Downing, Y. Jiang, M. Shuman, and N. Kanwisher. "A cortical area selective for visual processing of the human body". *Science* 293 (2001), pp. 2470–2473.

[36] O. Duchenne, A. Joulin, and J. Ponce. "A graph-matching kernel for object categorization". *IEEE International Conference on Computer Vision*. IEEE, 2011, pp. 1792–1799.

[37] O. Duchenne. "Image Alignment Techniques for Object Recognition and Detection". PhD thesis. École Normale Supérieure Paris, 2012.

[38] E. Eger, J. Ashburner, J. D. Haynes, R. J. Dolan, and G. Rees. "fMRI activity patterns in human LOC carry information about object exemplars within category". *Journal of Cognitive Neuroscience* 20 (2008), pp. 356–370.

[39] E. Eger, C. A. Kell, and A. Kleinschmidt. "Graded size sensitivity of object-exemplar-evoked activity patterns within human LOC subregions". *Journal of Neurophysiology* 100 (2008), pp. 2038–2047.

[40] R. A. Epstein and N. Kanwisher. "A cortical representation of the local visual environment". *Nature* 392 (1998), pp. 598–601.

[41] G. Erdogan, Q. Chen, F. Garcea, B. Z. Mahon, and R. A. Jacobs. "Multisensory part-based representations of objects in human lateral occipital cortex". *Journal of Cognitive Neuroscience* (in press).

[42] S. L. Fairhall, S. Anzellotti, P. E. Pastas, and A. Caramazza. "Concordance between perceptual and categorical repetition effects in the ventral visual stream". *Journal of Neurophysiology* 106 (2011), pp. 398–408.

[43] S. L. Fairhall and A. Caramazza. "Brain regions that represent amodal conceptual knowledge". *Journal of Neuroscience* 33 (2013), pp. 10552–10558.

[44] D. J. Felleman and D. C. VanEssen. "Distributed hierarchical processing in the primate cerebral cortex". *Cerebral Cortex* 1 (1991), pp. 1–47.

[45] C. Firestone and B. J. Scholl. ""Moral perception" reflects neither morality nor perception". *Trends in Cognitive Sciences* 20 (2016), pp. 75–76.

[46] B. Fischl. " FreeSurfer." *NeuroImage* 62 (2012), 774–781.

[47] J. Fodor. *Modularity of Mind*. Cambridge, MA: MIT Press, 1983.

[48] J. A. Fodor. *The modularity of mind: an essay on faculty psychology*. Cambridge, MA: MIT Press, 1993.

[49] D. J. Freedman and E. K. Miller. "Neural mechanisms of visual categorization: Insights from neurophysiology". *Neuroscience and Biobehavioral Reviews* 32 (2008), pp. 311–329.

[50] A. Gilbert, T. Regier, P. Kay, and R. Ivry. "Whorn hypothesis is supported in the right visual field but not the left". *Proceedings of the National Academy of Sciences USA* 103 (2006), pp. 489–494.

[51] G. Golarai, D. G. Ghahremani, S. Whitfield-Gabrieli, A. Reiss, J. L. Eberhardt, J. D. E. Gabrieli, and K. Grill-Spector. "Differential development of high-level visual cortex correlates with category-specific recognition memory". *Nature Neuroscience* 10 (2007), pp. 512–522.

[52] J. Gonzales-Castillo, Z. S. Saad, D. A. Handwerker, S. J. Inati, N. Brenowitz, and P. A. Bandettini. "Whole-brain, time-locked activation with simple tasks revealed using massive averaging and model-free analysis". *Proceedings of the National Academy of Sciences USA* 109 (2012), pp. 5487–5492.

[53] K. Grill-Spector. "The neural basis of object perception". *Current Opinion in Neurobiology* 13 (2003), pp. 159–166.

[54] K. Grill-Spector, Z. Kourtzi, and N. Kanwisher. "The lateral occipital complex and its role in object recognition". *Vision Research* 41 (2001), pp. 1409–1422.

[55] K Grill-Spector, T Kushnir, T Hendler, S Edelman, Y Itzchak, and R Malach. "A sequence of object-processing stages revealed by fMRI in the human occipital lobe". *Human Brain Mapping* 6 (1998), pp. 316–328.

[56] M. Guggenmos, V. Thoma, R. M. Cichy, J. D. Haynes, P. Sterzer, and A. Richardson-Klavehn. "Non-holistic coding of objects in lateral occipital complex with and without attention". *Neuroimage* 46 (2015), pp. 4024–4031.

[57] A. Harel, D. J. Kravitz, and C. I. Baker. "Task context impacts visual object processing differentially across the cortex". *Proceedings of the National Academy of Sciences USA* 111 (2014), pp. 962–971.

[58] J. V. Haxby, M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, and P. Petrini. "Distributed and overlapping representations of faces and objects in ventral temporal cortex". *Science* 293 (2001), pp. 2425–2430.

[59] J. V. Haxby, J. S. Guntupalli, A. C. Connolly, Y. O. Halchenko, B. R. Conroy, M. I. Gobbini, M. Hanke, and P. J. Ramadge. "A common, high-dimensional model of the representational space in human ventral temporal cortex". *Neuron* 72 (2011), pp. 404–416.

[60] J. D. Haynes and G. Rees. "Predicting the orientation of invisible stimuli from activity in human primary visual cortex". *Nature Neuroscience* 8 (2005), pp. 686–691.

[61] J. D. Haynes and G. Rees. "Predicting the stream of consciousness from activity in human visual cortex". *Current Biology* 15 (2005), pp. 1301–1307.

[62] H. R. Heekeren, S. Marrett, D. A. Ruff, P. A. Bandettini, and L. G. Ungerleider. "Involvement of human left dorsolateral prefrontal cortex in perceptual decision making is independent of response modality". *Proceedings of the National Academy of Sciences USA* 103 (2006), pp. 10023–10028.

[63] H. Hong, D. L. K. Yamins, N. J. Majaj, and J. J. DiCarlo. "Explicit information for category-orthogonal object properties increases along the ventral stream". *Nature Neuroscience* 19 (2016), pp. 613–622.

[64] T. Horikawa, M. Tamaki, Y. Miyawaki, and Y. Kamitani. "Neural decoding of visual imagery during sleep". *Science* 80 (2013), pp. 639–642.

[65] M. S. Horton and E. M. Markman. "Developmental differences in the acquisition of basic and superordinate categories". *Child Development* 51 (1980), pp. 708–719.

[66] A. G. Huth, S. Nishimoto, A. T. Vu, and J. L. Gallant. "A continuous semantic space describes the representation of thousands of object and action categories across the human brain". *Neuron* 76 (2012), pp. 1210–1224.

[67] M. C. Iordan, M. R. Greene, D. M. Beck, and L. Fei-Fei. "Basic-level category structure emerges gradually across human ventral visual cortex". *Journal of Cognitive Neuroscience* 27 (2015), pp. 1427–1446.

[68] M. C. Iordan, M. R. Greene, D. M. Beck, and L. Fei-Fei. "Typicality sharpens category representations in object-selective cortex". *Neuroimage* 134 (2016), pp. 170–179.

[69] M. C. Iordan, A. Joulin, D. M. Beck, and L. Fei-Fei. "Inter-subject alignment of functional cortical regions". *Proceedings of the Machine Learning and Interpretation in Neuroimaging (MLINI) Workshop, Advances in Neural Information Processing Systems (NIPS)*. 2015.

[70] J. D. Johnson and M. D. Rugg. "Recollection and the reinstatement of encoding-related cortical activity". *Cerebral Cortex* 17 (2007), pp. 2507–2515.

[71] P. Jolicoeur, M. A. Gluck, and S. M. Kosslyn. "Pictures and names: making the connection". *Cognitive Pshychology* 16 (1984), pp. 243–275.

[72] N. Kanwisher, J. McDermott, and M. M. Chun. "The fusiform face area: a module in human extrastriate cortex specialized for face perception". *Journal of Neuroscience* 17 (1997), pp. 4302–4311.

[73] K. N. Kay, T. Naselaris, R. J. Preener, and J. L. Gallant. "Identifying natural images from human brain activity". *Nature* 452 (2008), pp. 352–355.

[74] M. L. Kellenbach, A. A. Wijers, and G. Mulder. "Visual semantic features are activated during the processing of concrete words: event-related potential evidence for perceptual semantic priming". *Cognitive Brain Research* 10 (2000), pp. 67–75.

[75] C. S. Konen and S. Kastner. "Two hierarchically organized neural systems for object information in human visual cortex". *Nature Neuroscience* 11 (2008), pp. 224–231.

[76] T. Konkle and A. Caramazza. "Tripartite organization of the ventral stream by animacy and object size". *Journal of Neuroscience* 33 (2013), pp. 10235–10242.

[77] T. Konkle and A. Oliva. "A real-world size organization of object responses in occipito-temporal cortex". *Neuron* 74 (2012), pp. 1114–1124.

[78] Z. Kourtzi and N. Kanwisher. "Representation of perceived object shape by the human lateral occipital complex". *Science* 293 (2001), pp. 1506–1509.

[79] W. Koustaal, A. D. Wagner, M. Rotte, A. Maril, R. L. Buckner, and D. L. Shacter. "Perceptual specificity in visual object priming: functional magnetic imaging evidence for a laterality difference in fusiform cortex". *Neuropsychologia* 39 (2001), pp. 184–199.

[80] N. Kriegeskorte, R. Goebel, and P. A. Bandettini. "Information-based functional brain mapping". *Proceedings of the National Academy of Sciences USA* 103 (2006), pp. 3863–3868.

[81] N. Kriegeskorte, M. Mur, D. A. Ruff, R. Kiani, J. Bodurka, H. Esteky, K. Tanaka, and P. A. Bandettini. "Matching categorical object representations in inferior temporal cortex of man and monkey". *Neuron* 60 (2008), pp. 1126–1141.

[82] G. Langs, A. Sweet, D. Lashkari, Y. Tie, L. Rigolo, A. J. Golby, and P. Golland. "Decoupling function and anatomy in atlases of functional connectivity patterns: Language mapping in tumor patients". *Neuroimage* 103 (2014), pp. 462–475.

[83] D. A. Leopold, I. V. Bondar, and M. A. Giese. "Norm-based face encoding by single neurons in the monkey inferotemporal cortex". *Nature* 442 (2006), pp. 572–575.

[84] D. A. Leopold, A. J. O'Toole, T. Vetter, and V. Blanz. "Prototype-referenced shape encoding revealed by high-level aftereffects". *Nature Neuroscience* 4 (2001), pp. 89–94.

[85] T. Livne and M. Bar. "Cortical integration of contextual information across objects". *Journal of Cognitive Neuroscience* (in press).

[86] B. C. Love. "Environment and goals jointly direct category acquisition". *Current Directions in Psychological Science* 14 (2005), pp. 195–199.

[87] D. G. Lowe. "Distinctive image features from scale-invariant keypoints". *International Journal of Computer Vision* 60 (2004), pp. 91–110.

[88] G. Lupyan, S. L. Thompson-Schill, and D. Swingley. "Conceptual penetration of visual processing". *Psychological Science* 21 (2010), pp. 682–691.

[89] M. J. M. Macé, O. R. Joubert, J. Nespoulous, and M. Fabre-Thorpe. "The time-course of visual categorizations: You spot the animal faster than the bird". *PLOS One* 4 (2009), e5927.

[90] M. L. Mack, C. A. N. Wong, I. Gauthier, J. W. Tanaka, and T. J. Palmeri. "Time course of visual object categorization: Fastest does not necessarily mean first". *Vision Research* 49 (2009), pp. 1961–1968.

[91] R. Malach, J. B. Reppas, R. R. Benson, K. K. Kwong, H. Jiang, W. A. Kennedy, P. J. Ledden, T. J. Brady, B. R. Rosen, and R. B. Tootell. "Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex". *Proceedings of the National Academy of Sciences USA* 92 (1995), pp. 8135–8139.

[92] D. Marr. *Vision: A computational investigation into the human representation and processing of visual information*. New York, NY: Henry Holt and Co., Inc., 1982.

[93] J. L. McClelland and T. T. Rogers. "The parallel distributed processing approach to semantic cognition". *Nature Neuroscience* 4 (2003), pp. 310–322.

[94] C. B. Mervis and M. A. Crisafi. "Order of acquisition of subordinate-, basic-, and superordinate level categories". *Child Development* 53 (1982), pp. 258–266.

[95] E. K. Miller, D. J. Freedman, and J. D. Wallis. "The prefrontal cortex: categories, concepts, and cognition". *Philosophical Transactions of the Royal Society B: Biological Sciences* 357 (2002), pp. 1123–1136.

[96] J. P. Minda and J. D. Smith. "Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation". *Journal of Experimental Psychology: Learning, Memory, and Cognition* 28 (2002), pp. 275–292.

[97] M. Mur, D. A. Ruff, P. DeWeerd, P. A. Bandettini, and N. Kriegeskorte. "Categorical, yet graded - single-image activation profiles of human category-selective cortical regions". *Journal of Neuroscience* 32 (2012), pp. 8649–8662.

[98] G. L. Murphy and H. H. Brownell. "Category differentiation in object recognition: Typicality constraints on the basic category advantage". *Journal of Experimental Psychology: Learning, Memory, and Cognition* 11 (1985), pp. 70–84.

[99] G. L. Murphy and E. J. Wisniewski. "Categorizing objects in isolation and in scenes: what a superordinate is good for". *Journal of Experimental Psychology: Learning, Memory, and Cognition* 15 (1989), pp. 572–586.

[100] R. M. Nosofsky. "Attention, similarity, and the identification-categorization relationship". *Journal of Experimental Psychology: General* 115 (1986), pp. 39–61.

[101] R. M. Nosofsky. "Similarity scaling and cognitive process models". *Annual Review of Psychology* 43 (1992), pp. 25–53.

[102] R. M. Nosofsky. "Typicality in logically defined categories: exemplar-similarity versus rule instantiation". *Memory and Cognition* 19 (1991), pp. 131–150.

[103] A. Oliva and A. Torralba. "Modeling the shape of the scene: a holistic representation of the spatial envelope". *International Journal of Computer Vision* 42 (2001), pp. 145–175.

[104] S. Park, T. Konkle, and A. Oliva. "Parametric Coding of the Size and Clutter of Natural Scenes in the Human Brain". *Cerebral Cortex* 25 (2015), pp. 1792–1805.

[105] M. V. Peelen, C. He, A. Caramazza, and Y. Bi. "Nonvisual and visual object shape representations in occipitotemporal cortex: Evidence from congenitally blind and sighted adults". *Journal of Neuroscience* 34 (2014), pp. 163–170.

[106] D. G. Pelli. "The VideoToolbox software for visual psychophysics: Transforming numbers into movies". *Spatial Vision* 10 (1997), pp. 437–442.

[107] J. Peters, I. Daum, E. Gizewski, M. Forsting, and B. Suchan. "Associations evoked during memory encoding recruit the context-network". *Hippocampus* 19 (2009), pp. 141–151.

[108] Z. Phylyshyn. "Is vision continuous with cognition? The case for cognitive impenetrability of visual perception". *Behavioral and Brain Sciences* 22 (1999), pp. 341–365.

[109] M. Posner and S. Keele. "On the genesis of abstract ideas". *Journal of Experimental Psychology* 77 (1968), pp. 353–363.

[110] M. Riesenhuber and T. Poggio. "Modes of object recognition". *Nature Neuroscience* 3 (2000), pp. 1199–1204.

[111] J. B. Ritchie, D. A. Tovar, and T. A. Carlson. "Emerging object representations in the visual system predict reaction times for categorization". *PLOS Computational Biology* 11 (2015), e1004316.

[112] E. Rosch. "On the internal structure of perceptual and semantic categories". *Cognitive Development and the Acquisition of Language*. Ed. by T. E. Moore. Oxford, UK: Academic Press, 1973.

[113] E. Rosch. "Principles of categorization". *Cognition and categorization*. Ed. by E. Rosch and B. B. Lloyd. Hillsdale, NJ: Erlbaum, 1978, pp. 189–206.

[114] E. Rosch. "Universals and cultural specifics in human categorization". *Cross-cultural perspective on learning*. Ed. by R. Breslin, W. Lonner, and S. Bochner. London, UK: Sage Press, 1974.

[115] E. Rosch and C. B. Mervis. "Family resemblances: Studies in the internal structure of categories". *Cognitive Psychology* 7 (1975), pp. 573–605.

[116] E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson, and P. Boyes-Braem. "Basic objects in natural categories". *Cognitive Pshychology* 8 (1976), pp. 382–439.

[117] R. Rustamov and L. Guibas. "Hyperalignment of multi-subject fMRI data by synchronized projections". *Proceedings of the Machine Learning and Interpretation in Neuroimaging (MLINI) Workshop, Advances in Neural Information Processing Systems (NIPS)*. 2013.

[118] M. Sabuncu, B. D. Singer, B. Conroy, R. E. Bryan, P. J. Ramadge, and J. V. Haxby. "Function-based intersubject alignment of human cortical anatomy". *Cerebral Cortex* 20 (2010), pp. 130–140.

[119] R. Sayres and K. Grill-Spector. "Relating retinotopic and object-selective responses in human lateral occipital complex". *Journal of Neurophysiology* 100 (2008), pp. 249–267.

[120] N. Sigala, F. Gabbiani, and N. K. Logothetis. "Visual categorization and object representation in monkeys and humans". *Journal of Cognitive Neuroscience* 14 (2002), pp. 187–198.

[121] N. Sigala and N. K. Logothetis. "Visual categorization shapes feature selectivity in the primate temporal cortex". *Nature* 415 (2002), pp. 318–320.

[122] M. A. Silver and S. Kastner. "Topographic maps in human frontal and parietal cortex". *Trends in Cognitive Sciences* 13 (2009), pp. 488–495.

[123] F. Smith, G. Balzano, and J. Walker. "Nominal, perceptual and semantic codes in picture categorization". *Semantic factors in cognition*. Ed. by J. W. Cotton and R. L. Klutzy. Hillsdale, NJ: Erlbaum, 1978, pp. 137–168.

[124] J Talairach and P Tournoux. "Co-planar stereotaxic atlas of the human brain". *Thieme Publishing Group* (1988).

[125] J. W. Tanaka and M. Taylor. "Object categories and expertise: Is the basic level in the eye of the beholder?" *Cognitive Psychology* 23 (1991), pp. 457–482.

[126] K. Tanaka. "Inferotemporal cortex and object vision". *Annual Review of Neuroscience* 19 (1996), pp. 109–139.

[127] K. I. Taylor, B. J. Devereux, K. Acres, B. Randall, and L. K. Tyler. "Contrasting effects of feature-based statistics on the categorization and basic-level identification of visual objects". *Cognition* 122 (2012), pp. 363–374.

[128] A. Tosoni, G. Galati, G. L. Romani, and M. Corbetta. "Sensory-motor mechanisms in human parietal cortex underlie arbitrary visual decisions". *Nature Neuroscience* 11 (2008), pp. 1446–1453.

[129] M. Vaziri-Pashkam and Y. Xu. "Object representations in human parietal and occipito-temporal cortices: Similarities and differences". *Journal of Vision* 15 (2015), p. 374.

[130] K. L. Vilberg and M. D. Rugg. "Functional significance of retrieval-related activity in lateral parietal cortex: evidence from fMRI and ERPs". *Human Brain Mapping* 30 (2009), pp. 1490–1501.

[131] K. L. Vilberg and M. D. Rugg. "The neural correlates of recollection: Transient versus sustained fMRI effects". *Journal of Neuroscience* 32 (2012), pp. 15679–15687.

[132] J. Vinberg and K. Grill-Spector. "Representation of shapes, edges, and surfaces across multiple cues in the human visual cortex". *Journal of Neurophysiology* 99 (2008), pp. 1380–1393.

[133] D. B. Walther, E. Caddigan, F.-F. Li, and D. M. Beck. "Natural scene categories revealed in distributed patterns of activity in the human brain". *Journal of Neuroscience* 29 (2009), pp. 10573–10581.

[134] M. F. Wurm and A. Lingnau. "Decoding actions at different levels of abstraction". *Journal of Neuroscience* 35 (2015), pp. 7727–7735.

[135] D. L. K. Yamins and J. J. DiCarlo. "Using goal-driven deep learning models to understand sensory cortex". *Nature Neuroscience* 19 (2016), pp. 356–365.

[136] D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo. "Performance-optimized hierarchical models predict neural responses in higher visual cortex". *Proceedings of the National Academy of Sciences USA* 111 (2014), pp. 8619–8624.

[137] T. Yeo, F. M. Krienen, J. Sepulcre, M. R. Sabuncu, D. Lashkari, M. Hollinshead, J. L. Roffman, J. W. Smoller, L. Zollei, J. R. Polimeni, B. Fischl, H. Liu, and R. L. Buckner. "The organization of the human cerebral cortex estimated by intrinsic functional connectivity". *Journal of Neurophysiology* 106 (2011), 1125–1165.

[138] D. Zeithamova, W. T. Maddox, and D. M. Schnyer. "Dissociable prototype learning systems: evidence from brain imaging and behavior". *Journal of Neuroscience* 28 (2008), pp. 13194–13201.