Population Coding for Multiple Features in the Human Brain and Convolutional Neural Networks

HARVARD UNIVERSITY
Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Department of Psychology

have examined a dissertation entitled

Population Coding for Multiple Features in the Human Brain and Convolutional Neural Networks

presented by JohnMark Taylor

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature _____

Typed name: Prof. Talia Konkle

Signature _____

Typed name: Prof. Yaoda Xu

Signature _____

Typed name: Prof. George A. Alvarez

Signature _____

Typed name: Prof. Alfonso Caramazza

Signature _____

Typed name: Prof.

Date: May 4, 2021

Population Coding for Multiple Features in the Human Brain and Convolutional Neural Networks

JohnMark Taylor

A dissertation presented to the Department of Psychology in Partial Fulfillment of the Requirement for the Degree of Doctor of Philosophy in the Subject of Psychology

Harvard University
Cambridge, Massachusetts
May 2021

Dissertation Advisors: Yaoda Xu and Talia Konkle JohnMark Taylor

Population Coding for Multiple Features in the Human Brain and Convolutional Neural Networks

Abstract

Entities in the world comprise multiple features: how might an intelligent system, whether biological or synthetic, encode those feature combinations and put them to use in task-relevant processing? While this is a perennial and ubiquitous problem in the cognitive sciences, several recent developments make it timely to examine it anew: the recent characterization of color- and shape-sensitive regions in the primate ventral visual pathway opens a promising window towards examining how color and form are jointly represented in visual processing, the explosive success of artificial neural networks raises the question of how they encode feature combinations to solve the visual tasks at which they excel, and recent developments in computational neuroscience have highlighted the importance of neurons that exhibit nonlinear interaction effects to combinations of stimulus and task variables. In this dissertation, I draw on these recent developments in three projects examining how color and form are encoded in the human ventral visual pathway and convolutional neural networks, and how task and stimulus information are encoded throughout the human visual system. I develop several methods that can characterize the joint coding structure of multiple features. I find that the human ventral visual pathway largely encodes color and form information in an anatomically intermingled but representationally independent manner, that convolutional neural networks code color and form in an increasingly interactive manner throughout processing, and that the superior intraparietal sulcus, but not early visual cortex, encodes stimulus and task information in a nonlinear manner when attentional demands are carefully controlled.

Table of Contents

Chapter 1: Introduction

Preface

A fundamental fact about the world, and arguably of any system that aspires to faithfully and usefully represent the world, is that everything we encounter is composed of multiple *features*. Examples abound: an object consists of visual features like shape, color, texture, size, motion, and orientation, and semantic features like edibility or naturalness. When we encounter a person, that person's face consists of both the unchanging structural aspects of their face that define their identity, and various dynamic features that define their emotional state; additionally, a rich array of knowledge about that person, including their name, occupation, and relationships to others, is linked to that person. Analogous feature lists could be enumerated for nearly any domain of entities that we encounter in the world, or that cognitive scientists study in the lab: places, sounds, words, actions, and so on.

Insofar as such features capture real, useful patterns in the world, then, a basic challenge faced by any adaptively intelligent system, whether biological or synthetic, is how to encode the collections of features associated with different objects. The manner in which they do so will determine the range of operations that a system can perform over those features and objects. One such operation is *invariance*, or picking out the presence of a feature irrespective of the value of other features (for instance, that a shape remains a circle no matter what color it is, what size it is, or where it is). Another such operation is *binding*: determining which features belong to the same object. Additionally, it is often useful to flexibly utilize features across a range of tasks: *grasp* the red object, *remember* the square, *compare* the blue triangle and red circle, *saccade* to the circular object, *grasp* the green apple. Any system that can perform such a variety of tasks, as

the human brain clearly can, must be able to combine information about the features of an object with the demands of the present task.

In this dissertation, I examine the question: what is the nature of the neural architecture, and neural code, that enables such feats of multi-feature processing? I primarily address these questions using the representation of color and shape as a case study, and extend the analysis framework I develop to examine this question in the context of joint coding for visual and task information. While the nature of multi-feature coding is a perennial issue in cognitive science—arguably less an isolated topic of its own than a concern that crosscuts nearly every domain in the field—several exciting advances in the past decade make it timely, and critical, to examine it anew.

First, with respect to the case study of color and shape coding, the underlying neural architecture of color and shape processing has been greatly elucidated in recent years, with several studies in both the macaque and human brain identifying a series of color-sensitive cortical patches in the ventral visual pathway (Chang et al., 2017; Lafer-Sousa et al., 2016; Lafer-Sousa & Conway, 2013). The existence of these regions, which can be easily localized and probed using fMRI, provides fresh opportunities to explore how color and shape information interact (or fail to) throughout the ventral visual pathway.

Second, in the past decade a new "model organism" visual system has arrived on the scene: convolutional neural networks (Serre, 2019; Simonyan & Zisserman, 2015; Storrs & Kriegeskorte, 2019; Yamins & DiCarlo, 2016). These systems have arguably achieved human-level performance on object recognition tasks, and make use of color and shape information in their discriminations, raising the question of how they jointly encode color and shape information in achieving their high rates of performance.

Third, recent work in computational neuroscience has begun to explore in detail the relative advantages and disadvantages of different multi-feature neuronal coding schemes: specifically, it has become clear that it is computationally important to distinguish between 1) *pure* selectivity neurons tuned to a single feature, 2) *linear* mixed selectivity neurons that are tuned to multiple features, but in a purely additive, linear manner, and 3) *nonlinear* mixed selectivity neurons that code for multiple features in a nonlinear (interactive) manner (Fusi et al., 2016; Rigotti et al., 2013). These different neural coding motifs facilitate different operations, making it important to determine which one the brain uses in any particular domain. Of particular relevance to the present work, studies thus far have only been able to examine this issue using invasive single-neuron recording, primarily in macaques, making it desirable to develop methods that can distinguish these coding motifs noninvasively in the human brain.

In this dissertation, I take the following approach. In this introduction, I will review 1) the state of the art on how color and form processing interact in the human brain, 2) the range of proposed neural mechanisms for multifeature coding, including the recent developments in computational neuroscience regarding mixed selectivity neuronal coding, 3) the nature of multi-feature processing in convolutional neural networks, and 4) the range of methods currently available for probing multifeature coding in the human brain. Since all of these topics interconnect, the discussion of each topic will also build upon the discussion of previous topics. In the middle three chapters of this dissertation, I will present a series of studies examining 1) color and form coding in the human brain, 2) color and form coding in CNNs, and 3) mixed selectivity coding for stimulus and task information. Finally, in the concluding chapter, I will synthesize the findings from this series of studies, and discuss what remains to be learned.

Color and Form Coding in the Mind and Brain

Color and form are two of the most salient properties of an object, raising the question of how these features are encoded and linked together to form the colored shapes that populate so many of our visual experiences. This topic has been richly studied in both the psychophysical and neural literatures, although much work remains in order to fully understand how the relevant behavioral phenomena are ultimately grounded in brain activity. Here, I review these literatures, focusing in particular on the question of how color and form information may interact, or fail to interact, throughout various stages of processing. Such interactions can take several forms. First, color might affect the coding of form, or vice versa; for instance, color may be used as a cue in computing form features. Second, certain representations of color and form may be inherently *conjunctive* or *integrated*: that is, the representation encodes a particular *combination* of color and form features. Either possibility is an example of *interactive* color/form coding, as opposed to a more *independent* coding scheme, where processing of either feature proceeds without being modulated by the other feature, or where a representation encodes the presence of a given feature invariantly across values of the other feature (e.g., a representation corresponding to only red squares, versus a representation that responds to all red things irrespective of shape). While vast literatures exist that examine the coding of each feature individually, I here focus on the (still expansive) literature that pertains to the question of how these features are encoded relative to each other.

Psychophysics of Color and Form Processing

The study of how color and form processing interact has a long psychophysical pedigree, and has consistently served as a valuable case study in elucidating the nature of multi-feature

processing more generally. Color is a useful feature in this context, as it can easily be independently varied with respect to shape without affecting the other aspects of a stimulus (such as the retinotopic footprint or the luminance patterns) and is easily parametrized (with every color being being uniquely specified by three coordinates).

Garner (1976) developed a useful experimental paradigm (*Garner Interference)* and theoretical framework for assessing how two visual features might be encoded relative to each other. In this paradigm, participants view a series of stimuli varying with respect to multiple features, and have to report the value of a given feature (e.g., color, shape, or luminance) in each trial, while the values of other features either do or do not vary. The measure of interest is whether judging the value of one feature is affected by variation in the other features. Using this procedure, he drew a distinction between two kinds of features. First, there are *separable* features where changing the value of one does not interfere with performing a task over the other; he found that color and shape appear to be separable in this manner. Second, there are *integral* features where varying the value of one feature slows down processing of the other feature; hue and luminance appear to be two such features meeting this criterion. Minimally, these results suggest that certain features are obligatorily processed along with other features, although they do not by themselves reveal the underlying cognitive mechanism that might underlie these phenomena: for example, two integrated features like hue and luminance could be processed in the same computational units, or they could be processed in distinct units with automatic spreading activation between the units leading to interference. Garner's paradigm thus provides a useful assay for determining how two features interact during processing, and furnished initial evidence suggesting that color and shape processing could proceed independently, at least in the context he studied.

Treisman & Gelade (1980) developed an influential model, *Feature Integration Theory*, that posits several stages of visual feature processing, in which various features are initially processed independently, with the conjunction of these features requiring focused spatial attention. Notably, they brought to bear a variety of converging methods to support their theory. First, visual search for a target defined by one feature (e.g., a red stimulus among blue stimuli or a square among circles) is effortless and equally easy irrespective of the number of distractors, but by contrast, visual search for a target defined by a conjunction of features (e.g., a red square among other red shapes and other squares) is challenging and elicits reaction times that scale linearly with the number of items in the display. Second, textures defined by single features (e.g., a texture defined by vertical bars amidst horizontal bars), but not conjunctions of features, can be used for figure/ground processing. Finally, in brief displays involving search for either a feature or a conjunction of features, correct identification of a feature conjunction nearly always accompanied correct identification of the target's location, whereas finding a single feature did not necessarily require identifying the location. From this collection of findings, the authors formulated the theory that different features are initially encoded on independent feature maps that can be preattentively queried for the presence of a feature, or that can enter into preattentive figure/ground segmentation operations. By contrast, searching for a conjunction of features requires focused, serial spatial attention in order to link features on different maps via their shared location, with a "master map of locations" serving to orchestrate this spatial linking process across maps (Treisman, 1999).

A further line of evidence corroborating feature integration theory is the existence of cases where observers perceive "illusory conjunctions" (Cohen & Rafal, 1991; A. Treisman & Schmidt, 1982), where observers perceive incorrect pairings of the features in a display (e.g.,

seeing a red circle and blue square when in reality a blue circle and red square are present). This can occur either in patients with parietal brain damage, or in healthy individuals in cases of limited presentation time or divided attention.

Thus, both Garner Interference and Feature Integration theory provide evidence that color and form are coded independently, at least for the levels of representation that come into play for the experimental paradigms they used. The findings from these paradigms, and from the scores of studies they inspired, are robustly replicable (and indeed can be confirmed by any curious observer within minutes), and must be explained by any complete cognitive model of color and form processing. That said, several diverse lines of evidence suggest that matters may not be so simple: several experimental demonstrations suggest cases where color and form features appear to be automatically encoded in a conjoined, rather than separate format, or where the processing of one feature can influence the processing of the other feature.

One notable example is the McCollough Effect (McCollough, 1965; Stromeyer, 1969). In this illusion, a participant adapts to two alternating grating patterns: alternating *vertical* black and red gratings, or alternating *horizontal* black and green gratings, then views either alternating black and white vertical gratings, or black and white horizontal gratings (or both at once). The illusion, which is relatively easy to induce, is that the horizontal white test gratings will appear red, and the vertical white test gratings will appear green, suggesting an orientation-specific color aftereffect. That is, these results suggest that the visual system is not merely adapting to given color considered in isolation of orientation (in which case there is no reason that the vertical and horizontal test gratings should show different aftereffects). Moreover, later work suggested that these effects can occur even when the inducing gratings are outside the focus of

attention (Houck & Hoffman, 1986), suggesting that combinations of color and form (at least for a simple form feature, local orientation) can be encoded without attention.

Along similar lines, Holcombe & Cavanagh (2001) presented participants with rapidly alternating grating stimuli, where the pairs of gratings could either be left-tilted-red and right-tilted-green or left-tilted-green and right-tilted-red (such that either condition had the same four features, but differed in how they were conjoined), and had the participants report which color went with which orientation. Participants were able to do so even at high alternation rates (~19hz), consistent with an account where color-orientation combinations are registered rapidly and early in visual processing, rather than requiring a slow and attentionally-demanding binding step.

One intriguing study of parietal lesion patients showing visual extinction (that is, a condition where they can perceive a stimulus contralateral to their lesion when it is presented alone, but not when it is presented along with an ipsilateral stimulus) provides suggestive evidence that color and form conjunctions can be encoded preattentively when they correspond to an object with a familiar, canonical color, such as a red strawberry or yellow banana (Rappaport et al., 2015). Specifically, patients showed higher rates of perceiving and identifying a contralateral stimulus presented with an ipsilateral stimulus when the contralateral stimulus was an object presented in its canonical color. Interestingly, this effect disappeared if the color merely surrounded the object, instead of being on the surface of the object. This suggests that color/form combinations can be encoded preattentively when they correspond to objects with canonical colors.

Even the primary behavioral paradigm marshalled in support of feature integration theory, visual search, has revealed apparent cases of preattentive conjunctive processing of color

and form. Rappaport et al. (2013) found that search for objects in their correct canonical color (e.g., yellow corn), but not in their incorrect color (e.g., blue corn) approached parallel "pop-out" search, suggesting that the binding of color and form can be encoded preattentively for objects with familiar, canonical colors. Such an effect can also be induced via training and is not limited to naturalistic stimuli: Walsh et al. (1998) trained subjects on a search task in which they searched for a green vertical bars among green horizontal and blue vertical bars. With training, subjects approached parallel search on this task. Interestingly, TMS to parietal cortex sharply interfered with performance early in training, but not later in training, suggesting that the attention becomes less necessary for conjunction search for learned color/form conjunctions.

Another class of psychophysical color-shape interactions concerns cases where color can be put to use in shape processing. Notably, even patients with achromatopsia can perceive shapes defined by isoluminant color differences (Barbur et al., 1994; Heywood et al., 1991, 1998); interestingly such color differences can contribute to motion discrimination as well (Cavanagh et al., 1998; Cavanagh & Anstis, 1991). In a similar vein, given a field of colored dots, a set of dots in a single color that form an outline of a circle rapidly "pops out" when observed, but when the same circle is defined by dots of several isoluminant colors, the circle percept is much less visible, suggesting that color information is put to use in the formation of the form percept (Mandelli & Kiper, 2005).

There also exist cases where shape information can seemingly influence color perception. One example is so-called "memory-color" effects, in which a greyscale version of an object with a canonical color (e.g., a red strawberry) is supposedly faintly tinged with the remembered color; however, recent work has suggested that many putative demonstrations of this effect arise from various methodological problems, which when controlled for cause the effect to go away

(Valenti & Firestone, 2019). That said, a recent experiment in which participants viewed various stimuli, such as fruit and faces, in sodium lighting conditions that rendered them objectively monochromatic found that face stimuli, but not other stimulus categories, exhibited an interesting illusion where they appeared green under these monochromatic lighting conditions (Hasantash et al., 2019), suggesting that category-specific color processing (which in this case could only be induced by shape features) may occur in at least one case. Color computations may also draw on the perceived 3D structure of a scene: Bloj et al. (1999) devised a 3D card stimulus whose perceived color flips from pale pink to deep magenta when the subject's percept of the cards switches from convex to concave, suggesting that computations of color appearance may draw on the 3D geometry of a scene, enabling color processing to incorporate knowledge about surface lighting conditions.

Summing up, then, while the framework of Feature Integration Theory explains important patterns of psychophysical data regarding the relationship between color and form processing, other behavioral findings suggest that it is not universally true that these features are processed in isolation of each other until attention glues them together. At least three kinds of exception challenge this principle. First, conjunctions of color and simple form features (notably, orientation) appear to be processed pre-attentively in some cases; second, evidence exists that color and form conjunctions can be processed preattentively when the conjunctions in question are well-learned, either through naturalistic experience or through experimental practice; third, color information can be put to use in form computations, and (in more limited cases) vice-versa.

The Neural Relationship Between Color and Form Coding

In parallel with the expansive psychophysical literature examining how color and form processing relate, a large body of neuroscientific work has examined the neural architecture and coding principles of color and form perception. I here review this literature, focusing on the question of how tuning for color and form information is distributed throughout the brain. Broadly speaking, a variety of outcomes are possible: neurons coding for color and form could be segregated into distinct brain regions, they can be intermingled within the same brain region (which can occur at varying levels of granularity, such as a completely homogeneous mixture, or some mesoscale level of neuronal clustering for neurons coding for different features, etc.), or these features can be encoded by the very same neurons. This last possibility of single neuron mixed tuning can be even further subdivided into neurons that encode both features, but in an independent, additive manner, such that its tuning profiles to color and form merely "stack" on top of each other without affecting the tuning properties of the other feature (i.e., only "main effects" if an ANOVA were to be conducted), or neurons that encode these features in an interactive manner, such that the neuron's tuning across one feature varies across values of the other feature (i.e., literally an "interaction effect"). These different organizations could facilitate different operations: for instance, spatially segregated processing of color and shape could aid in computations where it is valuable to disregard the other feature (for instance, for computation of 3D shape structure it may be useful to disregard color), whereas a more intermingled organization could aid in cases where one feature can be put to use in processing the other feature. The computational abilities of neurons that code interactively for multiple features will be discussed in more detail later in this chapter; in the context of color and form processing, however, such neurons could play a role in detecting particular conjunctions of features, or in

enabling processing of a given feature that varies across values of the other feature (for example, if certain color computations are specific to some shape features or visual categories but not others).

In this section, I summarize what is known about where regions across the ventral visual hierarchy (in both the human brain and in non-human primates) fall with respect to these distinctions. While documenting the neural organization and tuning profile of color- and shape-sensitive neurons is not by itself *sufficient* for understanding all of the psychophysical phenomena discussed in the previous section--for instance, the specific neural dynamics underlying say, conjunction search would remain to be explained--it is arguably an essential first step, and understanding the topographic organization and representational format of color and form representations in the ventral visual pathway will facilitate understanding how these representations interact with processes like attention to give rise to phenomena like visual search and illusory conjunctions.

A potential distinction between color and form processing emerges as early as the retina. The three kinds of cones absorb incoming light, thereby parametrizing an initially infinite-dimensional signal (since the light can be any intensity level at each value of an infinitely divisible range of wavelengths) into a three dimensional signal that emerges from the differing absorption spectra of the three cone types. Inputs from cones then feed into retinal ganglion cells (RGCs), which fall into several varieties (Conway, 2009). Importantly, some RGCs have spatially structured center-surround receptive fields that seem tuned to luminance contrasts (e.g., when the center is either brighter or dimmer than the surround, depending on the exact type of cell), while some are color-tuned (e.g., responding more to red and less to green), but not in a spatially structured way; thus, at the level of RGCs, luminance processing is more spatially fine-

grained than color processing. This distinction appears to hold in the lateral geniculate nucleus as well, with cells exhibiting a similar range of tuning profiles.

The next stage of processing, V1, contains cells with a variety of tuning profiles to color and form information, with some uncertainty persisting regarding the degree of segregation and clustering that these cell types show. Johnson et al. (2001) showed macaques drifting grating stimuli at varying orientations and spatial frequencies, that were defined either by luminance contrasts (that is, the gratings alternated between dark and light) or isoluminant color contrasts (such that the gratings were defined by, for example, alternating red and green zones). They drew a distinction between three kinds of cells: *luminance* cells, that are orientation-specific, lack color tuning, and that are tuned to high spatial frequencies; *color* cells that respond to isoluminant color patterns but not pure luminance-defined patterns, lack orientation tuning, and prefer low spatial frequency (Conway, 2001 found cells with this response profile as well, concluding that they exhibit circularly symmetric center-surround organization); and *color-luminance cells*, which respond equally to luminance and color patterns, and exhibit orientation-tuning and preferences to high spatial frequencies. The latter class of cells offer an example of how color can be put to use in computing form information (i.e., by detecting color-defined boundaries), and offer a potential clue as to how achromatopsia patients can nonetheless perceive color-defined boundaries. As mentioned, the specific topography of V1 neurons with different tuning profiles has been a matter of dispute: an early influential account (Livingstone & Hubel, 1988) posited that color-tuned, orientation-insensitive cells are concentrated into cytoarchitectonically-defined zones called “blobs”, with orientation-specific cells lacking color tuning lying in the “interblob” regions. However, other work has challenged this dichotomy,

suggesting at best limited differences in orientation and color tuning between the blob and interblob zones of V1 (Friedman et al., 2003; Lennie et al., 1990; Leventhal et al., 1995).

An analogous controversy persists in V2 regarding the degree of segregation for color and form tuning: in Livingstone and Hubel's (1988) model, the cytoarchitectonically-defined *thin stripe* zones are posited to contain cells with higher color tuning, but less orientation, tuning, with the *interstripe* zones showing the opposite response profile; however, Gegenfurter (2003) reviewed results from six studies and found that each of these zones contains a high fraction (30-40%) of cells selective for the "non-preferred" feature as well, suggestive of a more distributed, rather than segregated, account of color and form processing in this region.

V3 also has been shown to encode both color and form information (e.g., Seymour et al., 2010), but is often omitted in reviews of color processing (Conway, 2009; Conway et al., 2010), being studied more often for its role in stereoscopic depth perception (e.g., Adams & Zeki, 2001).

V4 neurons have been shown to encode information about both color and midlevel form information. Interestingly, the color representations that it encodes better match the similarity structure of human perception than those in V1 (Brouwer & Heeger, 2009, 2013). V4 neurons exhibit tuning to midlevel form features such as curvature (Gallant et al., 1993; Gallant et al., 2000), texture (Pasupathy et al., 2019; Roe et al., 2012), and convexity (Pasupathy & Connor, 2001). Some evidence suggests that V4 exhibits mesoscale segregation of color and form tuning: Conway et al., (2007) found that color-tuned neurons in V4 tended to be clustered into "globs" (not to be confused with the "blobs" of V1) several millimeters across, with the "interglob" neurons showing dramatically less color tuning and relatively more form tuning.

Moving onto regions higher in the ventral visual hierarchy, various studies have revealed regions in occipitotemporal cortex that are involved in the high-level analysis of shape and color information, with some potential evidence of macroscale segregation for the processing of these features. Specifically, macaque inferotemporal cortex (IT) has been shown to contain neurons tuned to high-level shape features (Bao et al., 2020; James J. DiCarlo et al., 2012; Lehky & Tanaka, 2016; Wang et al., 1996), and human fMRI studies have revealed arguably homologous regions in the human lateral and ventral OTC that respond more to coherent shapes than scrambled stimuli (Denys et al., 2004; Grill-Spector et al., 1998; Kourtzi & Kanwisher, 2001; Malach et al., 1995). Damage to this cortical region can result in loss of form perception, with spared color perception (Benson & Greenberg, 1969; Goodale & Milner, 2004).

Analogously, a series of color-sensitive regions, which can be roughly divided into posterior, central, and anterior zones, have been identified that exhibit color-tuning in the macaque, and that respond more to colored than greyscale stimuli in the human brain (Brewer et al., 2005; Chang et al., 2017; Conway, 2018; Conway et al., 2007; Hadjikhani et al., 1998; Lafer-Sousa et al., 2016; Lafer-Sousa & Conway, 2013). The posterior color-sensitive zones plausibly correspond to the "globs" lying within retinotopically-defined V4; what about the central and anterior zones? Several studies have examined color-sensitive regions lying slightly anterior to V4. One group identified a region they called V8, which exhibited retinotopic topography for the foveal visual field and, unlike V4, responded to color afterimages (Hadjikhani et al., 1998). Another group identified a color sensitive region immediately anterior to retinotopic V4, which they called V4α and argued lacked retinotopic organization (Bartels & Zeki, 2000). Still another group identified two retinotopic regions they called VO1 and VO2 in this vicinity (Brewer et al., 2005). This profusion of naming conventions and methods for defining regions in this vicinity

makes it somewhat challenging to precisely match these regions across studies; nonetheless, this neighborhood of color-sensitive cortex anterior to V4 (but not yet reaching the far anterior temporal lobe) has been shown to exhibit various interesting properties. Several studies suggest that cortex in this region bears an especially close connection with conscious color perception: in addition to the finding that this region responds to color afterimages, one human intracranial study both measured responses from and stimulated tissue in this region, and found that the color eliciting the largest neural response in their measuring site closely matched the illusory evoked color when that site was stimulated (Murphey et al., 2008), and several studies have found that achromatopsia patients often have lesions in this neighborhood (Bouvier & Engel, 2006). Finally, color-sensitive cortex even further anterior has been identified that appears to be especially active for demanding tasks involving color, but not necessarily to the passive viewing of color (Beauchamp et al., 1999); another study found that a region in this overall vicinity has also been found to be active when retrieving color knowledge (Simmons et al., 2007). These three groups of color-sensitive regions, then, can be roughly divided into the posterior zones overlapping with retinotopic V4, zones lying somewhat anterior to V4 that have shown close links with conscious color perception, and even further anterior regions that might be involved in demanding color tasks.

To what extent do these ventral stream regions involved in form or color processing also have information about the other feature—that is, do form-processing regions also encode color information, and vice versa? Several studies of macaque IT have found that regions encoding shape information also contain neurons encoding color information (Komatsu & Ideura, 1993; McMahon & Olson, 2009); similarly, several studies have reported color decoding in area LO of the human brain (Bannert & Bartels, 2013, 2018). In the color-sensitive ventral stream regions,

Chang et al. (2017) found that neurons in all three sectors (posterior, central, and anterior) also appeared to encode form information. However, I do note one potentially ambiguity in the interpretation of these results: their stimuli were not matched in their retinotopic extent, so if the differently-shaped stimuli stimulate different parts of the visual field to varying degrees, then a color-tuned neuron that responds to the "amount" of its preferred color in its receptive field irrespective of shape (e.g., could nonetheless appear to have "shape" tuning. Lafer-Sousa et al. (2016) examined the corresponding color-sensitive regions in the human brain using univariate methods, and found that the anterior color regions, but not the posterior and central color regions, responded more to coherent objects than to scrambled stimuli. However, visual information can be encoded in distributed, fine-grained patterns across cortex that univariate metrics cannot detect (e.g., Haxby et al., 2001), leaving it possible that these regions nonetheless encode form information in the human brain.

Many regions throughout the ventral visual pathway thus appear to encode information about both color and form. A further question to address, as mentioned earlier, is the *format* of these multi-feature representations: are these features encoded in an additive, independent manner, such that the coding of one feature does not affect the encoding of the other, or are they encoded in a more interactive manner? It should be noted that these two possibilities are not mutually exclusive: a neuron, or neural population, can exhibit both additive and interactive effects, in the same way that any quantity affected by multiple variables can exhibit main effects and interaction effects. Additive coding could facilitate invariant representation of either feature, whereas interactive coding could enable rapid detection of particular color/form conjunctions, or processing of one feature that varies across values of the other feature. Various studies have examined this issue in the aforementioned brain regions. Engel (2005), using an fMRI adaptation

paradigm, found that V1 and V2 exhibited adaptation to the color and orientation of oriented gratings that exhibited an interaction effect, with their adaptation to combinations of the two features being greater than would be expected to their adaptation of either feature alone. Similarly, Seymour et al. (2010) employed a design (explained in more detail in Chapter 2) in which they trained an SVM classifier to discriminate between blocks that contained either red clockwise and green counterclockwise spirals, or red counterclockwise and green clockwise spirals, under the logic that only regions containing information about the color/form *combinations* above and beyond the individual features should be able to discriminate these block types. They found that V1, V2, V3, and V4 could successfully discriminate these blocks, consistent with them encoding color and orientation in a conjunctive manner. The finding that some regions encode color and orientation in a conjunctive, interactive manner may provide a potential mechanism for the psychophysical results in the previous section suggestive of automatic, preattentive encoding of color/form conjunctions. Bushnell & Pasupathy (2012) studied the responses of V4 neurons to stimuli of many different artificial shapes in two different colors, and found that many neurons exhibited a similar ordinal shape tuning preference across colors, though some exhibited a multiplicative gain to stimuli of one color relative to the other (one form of an interaction effect). McMahon & Olson (2009) showed stimuli of two different shapes and two different colors to macaques while measuring from IT, and found that neurons tended to encode these two features additively, with no interaction effect. Finally, Chang et al. (2017) found that a large percentage of neurons in the posterior, central, and anterior color-sensitive patches in the macaque ventral visual pathway exhibited a color-shape interaction effect, as documented by an ANOVA within each neuron they recorded; one provocative finding

was that neurons in the most anterior color region coded color differently for faces than for other stimuli, consistent with the existence of face-specific color illusions (Hasantash et al., 2019).

Despite this vast literature, spanning different species, brain regions, stimulus types, and recording modalities, several important questions remain regarding the relationship between color and form processing in the human ventral visual pathway. First, to what extent do higher-level ventral stream regions exhibiting univariate sensitivity to shape or color in the human brain also contain *distributed information* about the other feature, and for regions containing information about both, are there differences in the relative *magnitude* of information about either feature? The fact that different studies have examined different regions with different methods and stimuli makes it difficult to answer this question in a holistic manner. Second, while various studies have found evidence for interactive coding of color and orientation (a relatively simple form feature) in early visual cortex, it remains unknown whether these features are also encoded in this manner in higher-level visual regions. Finally, it remains unknown whether color and form are encoded in an interactive manner when the manipulated form feature is a more complex form feature than orientation; that is, is this interactive coding format a persistent coding motif throughout processing, or is it a relatively rare coding format amidst a largely additive relationship between color and form? Study 1 of this dissertation will endeavor to answer these questions by comprehensively examining the joint coding of color and form across the entire ventral visual hierarchy.

Color and Form Processing in CNNs

The past decade has seen a revolution in artificial intelligence: thanks to advances in computing power and data availability (Deng et al., 2009), artificial neural networks, which had experienced something of a winter for several decades, have made rapid advancements in a variety of domains, from image recognition to game playing. Of relevance to this thesis, this revolution has also catalyzed a paradigm shift in visual neuroscience: artificial neural networks inspired by the primate visual system have approached, or even surpassed, human vision in various benchmarks such as object recognition and image segmentation, raising the question of whether these synthetic visual systems follow similar operating principles to those of biological visual systems. Such a comparison is useful from both a scientific and a technological standpoint: an example of a scientific benefit is that such a comparison enables a better understanding of whether a given functional component (e.g., feedback connections) is essentially to perform a certain task, and an example of a technological benefit is that identifying differences in how brains and artificial neural networks operate facilitates using our knowledge of the brain to improve artificial intelligence algorithms.

A variety of studies have examined how CNNs encode visual information and how their representations might map onto those of the brain. Yamins et al. (2014) found that convolutional neural networks trained to recognize objects explained a high proportion of variance in neural responses, with the higher layers of the network best explaining IT and the middle layers of the network best explaining V4, even though the networks were not trained on neural data, with other studies reaching similar conclusions (Cadieu et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014). fMRI and MEG studies of visual processing in the human brain also suggested a

hierarchical correspondence between CNNs and brains, with corresponding levels of the hierarchy exhibiting more similar representations (Cichy et al., 2016; Eickenberg et al., 2017).

Despite these clear advancements over past models, other work has shed doubt on the correspondence between CNNs and the primate ventral visual pathway. One recent study found limited correspondence between CNN representations and the human ventral pathway, especially when artificial objects were used as stimuli (Xu & Vaziri-Pashkam, 2020). In line with this, CNNs appear to make use of different features in object recognition compared to humans (Linsley et al., 2017; Ullman et al., 2016). CNNs have also been shown to exhibit failures utterly unlike those shown by biological visual systems; specifically, they are vulnerable to "adversarial images", which can cause a CNN to misclassify an image purely by changing a small number of pixels (Serre, 2019). Furthermore, CNNs appear to show a default bias towards classifying images based on texture rather than global shape, although this difference can be mitigated by training them on a dataset that manipulates the images to make texture a noninformative cue (Geirhos et al., 2019).

One obstacle to designing CNNs that better align with biological vision, or that do not exhibit their idiosyncratic errors, is that in some ways they are black boxes: the values of their millions of parameters emerge automatically via training, which enables their remarkable success but also renders the operating principles underlying that very success somewhat obscure.

That said, one thing is unambiguously true about CNNs: they must make use of the features present in the image in some way. For example, it is difficult to imagine how they could correctly categorize objects without *some* sort of representation of object form, though the extent to which they emphasize different form features like texture versus global shape requires further study (Geirhos et al., 2019). Furthermore, though this topic is less studied, they also appear to

make use of color information in their classification decisions. Most CNNs take as input an RGB image, and they can freely make use of this information to improve their classification performance over the course of their training. And they in fact appear to do so: one study found that CNNs trained on ImageNet showed higher classification performance on colored images than their greyscale equivalents, suggesting that they in fact make use of color information (Singh et al., 2020), and one study found that both early and late layers appear to encode color information (Flachot & Gegenfurtner, 2018).

Given that CNNs encode both color and form information, the same question can be asked about them as can be asked in the brain: how are these features processed together throughout processing? Are they encoded in an independent format, such that coding for one is unaffected by coding for the other? Or are they encoded in an interactive format, showing sensitivity to particular color/form *conjunctions*? Only one study has examined this topic, and in a relatively limited way: it examined tuning curves from four sample neurons in a network, and found that they were selective to both color and form information (Rafegas & Vanrell, 2018). However, they did not examine these color/form representations at a population level (i.e., what is the overall geometry of color and form coding across the entire population in each layer), and they only examined a small number of stimuli, in only one network.

The goal of Study 2 of this dissertation is to take a more comprehensive look at the joint coding of color and form information in CNNs. This will advance several important goals.

First, understanding how CNNs encode combinations of features is one window into better understanding their internal operating principles. Recent work has shown that CNNs not only extract information about object identity, but also extract category-orthogonal information about an object's size and location (Hong et al., 2016). However, the representational format of

these multifeature representations is largely unknown; probing this issue in the context of color/form coding will therefore be a useful case study into the principles of multifeature representation in CNNs more generally. Characterizing the internal representations of CNNs in terms of how they jointly encode multiple features may be a useful step towards better characterizing their operating principles in human-interpretable terms.

Second, understanding how CNNs jointly encode multiple features is another axis along which they can be compared with biological visual systems. As described in the previous section, much work has been brought to bear in understanding how color and form are encoded in the brain, so characterizing how CNNs encode multiple features can allow them to be compared with biological visual systems in this respect as well.

Third, the architecture of a CNN can be fixed by the experimenter; for instance, feedback or recurrent connections can be included or omitted as desired. This makes them a useful testing ground for understanding the *source* of the multifeature representations encoded in any given region: for example, in a purely feedforward network these representations can only have developed via feedforward operations. This is in contrast with the brain, where the origin of a neural representation (feedforward or feedback?) can be less transparent. Thus, studying multifeature representation in CNNs can help clarify what kinds of multifeature representations can emerge in different architectures.

Fourth, CNNs have a well-characterized task that is stipulated by the network's designer. This makes it more tractable to make normative claims about the nature of its representations: that is, if a layer jointly encodes color and form in a particular way, this can only have occurred 1) to improve task performance, 2) as a byproduct of network architecture, or 3) as a byproduct of training. By contrast, if (say) V1 exhibits certain representational hallmarks, their functional

purpose is less clear: presumably the representations in V1 subserve many different downstream operations, making it difficult to pin down why V1 represents information in one way as opposed to another.

Thus, Study 2 of this dissertation examines the joint coding of color and form information in CNNs across a large collection of shapes and colors, across a variety of networks, and across several training regimes in order to better understand how CNNs solve the problem of representing multiple features, adding a new dimension to the burgeoning subfield of "CNN electrophysiology." As will be described later, I take a population-coding approach in order to chart the joint representational coding geometry of these two features.

Nonlinear Mixed Selectivity: A New Principle of Neural Coding

At various points thus far, I have discussed the concept of neurons that exhibit interactive tuning to color/form combinations, potentially enabling rapid detection of particular conjunctions or processing of one feature that varies across values of the other feature. As it happens, work from the past decade has suggested that neurons exhibiting interactive tuning to multiple variables in this way—whether aspects of the stimulus or of the present task—exhibit important computational advantages. These *nonlinear mixed selectivity neurons* have advantages over neurons tuned to just one feature (*pure selectivity neurons*), or neurons tuned to multiple features in a linear, additive manner (*linear mixed selectivity neurons*). Given this, it is important to establish the prevalence of this coding motif in the human brain in different domains, a goal that research thus far has barely even begun to address. In this section, I summarize what is known about the computational properties of these neurons and how they may contribute to flexible, task-driven processing, and lay out a potential test case for examining this coding motif in the

human brain: comparing how stimulus and task information is jointly encoded across early, ventral stream, and dorsal stream brain regions.

A persistent, and perplexing, observation about neural tuning in various brain regions is that neurons appear to be tuned not just to one variable, but to heterogeneous mixtures of stimulus, task, and cognitive variables. For example, Asaad et al. (1998) found that many neurons in macaque prefrontal cortex appeared to respond in a manner that reflected both the stimulus the monkey was viewing, and the response (in this case a left or right saccade) they made in response to that stimulus, often combining this information in a nonlinear manner. Neurons in the macaque lateral intraparietal area (LIP) also appear to be tuned to mixtures of task and stimulus variables (Meister et al., 2013) in this way. One prominent framework for understanding these multitasking neurons was the notion of *adaptive coding* (Duncan, 2001), which posits that many neurons, especially in frontoparietal regions, could dynamically change their response properties to serve the needs of the present task.

In the last decade, however, a computationally rich framework has emerged for understanding the capabilities and functions of these multiplexing neurons. Rigotti et al. (2013) analyzed neural recording data from macaque PFC while the monkey performed one of two tasks over one of four different stimuli. They found that a high percentage of PFC neurons not only contained both task and stimulus information, but combined this information in a nonlinear, interactive manner, a coding hallmark they dubbed *nonlinear mixed selectivity*. While this coding motif had been observed in past studies, this group analyzed the computational properties associated with this motif in detail. In particular, they argued that neurons that encode multiple variables in a nonlinear, but not a linear (additive) manner encoded a higher-dimensional representational space, enabling a larger array of readout operations to be performed over this

space. A simple example of such a function is an "exclusive-or" operation: if neurons encode two variables in a purely linear, additive manner, then it becomes impossible for a linear classifier to selectively respond to combinations of those variables that are either high-low or low-high, limiting the range of readout operations that can be performed over that population. Intriguingly, they found that this higher dimensionality was not just an arbitrary mathematical characterization of the data, but appeared to be causally relevant in behavior: the dimensionality of the neural populations they recorded was lower during error trials than in trials where the monkey made a correct response. As it has become increasingly plausible that neural populations, rather than single neurons, may be the correct unit of analysis in thinking about neural computation (Saxena & Cunningham, 2019), the connection that the authors draw between mixed selectivity neurons and a more flexible population code may be an important ingredient in better understanding how neural populations compute.

A handful of studies in the past decade have drawn upon the framework devised by Rigotti et al. (2013). Similar to the study in Rigotti et al. (2013), a study by Blackman et al., (2016) found that neurons in macaque PFC fire in a manner that reflects mixtures of stimulus, task, and response information. Hardcastle et al. (2017) recorded from neurons in rat entorhinal cortex, a region involved in spatial coding and navigation, and found that the majority of neurons they recorded from encoded mixtures of navigation-relevant variables, such as position, speed, and head direction; another study found that the subiculum subregion of the hippocampus also has this property (Ledergerber et al., 2020). Grunfeld & Likhtik (2018) argued that the medial prefrontal cortex may employ mixed selectivity coding to integrate information about the current situation and the animal's emotional state to assess current levels of threat. Diomedi et al. (2020) recorded from neurons in macaque medial parietal cortex involved in visually guided reaching,

and found that neurons appeared to multiplex information about gaze and reaching variables, suggesting a role for mixed selectivity in this behavior. Interestingly, this coding format may be used not just in brain regions involved in "high-level" operations such as navigation and decision-making, but even in primary somatosensory regions: Nogueira et al. (2021) recorded from neurons in mouse somatosensory cortex corresponding to their whiskers, and found that these neurons nonlinearly multiplexed information about which whiskers were in contact with something, and the displacement angle of the whiskers; through computational modeling, they found that the code they identified could be used to perform a wide range of tasks. Further theoretical work has identified additional computational properties of neurons exhibiting nonlinear mixed selectivity: neural populations with nonlinear mixed selectivity support more reliable readout (Johnston et al., 2019), and simulated neurons with levels of mixed selectivity similar to PFC emerge in a random network that undergoes Hebbian learning (Lindsay et al., 2017), suggesting a mechanism by which neural populations with this property might emerge.

Even in this still-nascent literature, then, various studies have shown that characterizing whether neural populations follow a nonlinear mixed selectivity code is important for understanding their function, and hallmarks of this coding motif have been found in cognitive, sensory, and motor domains. Nonetheless, a notable gap remains: very few studies have probed for this coding motif in the human brain, as most attempts to index it have relied on invasive neural recording techniques. To my knowledge, exactly one group has studied this coding principle in the human brain, using intracranial recording techniques (Zhang et al., 2017, 2020). This group examined how neurons in a patient's posterior parietal cortex encode mixtures of motor variables (what kind of body part, what side of the body, and whether the movement is imagined or attempted), and found that some of these variables were mixed in a linear manner,

with others being mixed in a nonlinear manner. Given that human behavior exhibits a high degree of flexibility and context-sensitivity, which are benefits that nonlinear mixed selectivity is taken to confer (Fusi et al., 2016), understanding the prevalence of this coding scheme in the human brain is an important, yet almost entirely unexamined frontier.

A promising case study for examining nonlinear mixed selectivity in the human brain is the representation of stimulus and task information in the posterior parietal cortex. Similar to the PFC neurons that Rigotti et al. (2013) examined in their study, PPC neurons have been shown to encode information about both stimulus and task information. In the macaque, the lateral intraparietal area (LIP) has been shown to exhibit this multiplexing property (Huk et al., 2017; Meister et al., 2013; Park et al., 2014). A potentially homologous region in the human brain, the superior IPS, has analogously been shown to encode both stimulus and task information (Jeong & Xu, 2016; Vaziri-Pashkam & Xu, 2017, 2019; Xu, 2018a, 2018b). Given the benefits of nonlinear mixed selectivity coding, it is therefore important in its own right to establish whether this region jointly encodes task and stimulus information in a linear versus nonlinear manner. One attractive feature of this parietal region, in contrast with more frontal regions, is that its representations are encoded at a spatial scale visible to fMRI via multivoxel pattern analysis (Jeong & Xu, 2016; Vaziri-Pashkam & Xu, 2017, 2019), unlike neural representations in frontal cortex, which may lack the requisite topographic distribution to be readily visible to fMRI (Bhandari et al., 2018).

Thus, the goal of Study 3 of this dissertation is to examine the coding format of the multiplexed code for stimulus and task information that exists in superior IPS. Besides advancing our understanding of how this region flexibly codes for visual information in a task-sensitive manner, a topic of great importance in its own right, this endeavor would be a first proof-of-

concept that nonlinear mixed selectivity can be noninvasively queried in the human brain, opening the door to studying this coding motif in the human brain across the other processing domains in which it has been shown to be important in other species.

But how?

Methods for Examining Interactive Multifeature Representations

Thus far in this introduction, a recurring topic has been the concept of neural representations in which combinations of variables, whether colors and shapes or combinations of stimulus and task variables, are encoded in an independent versus interactive manner. Even merely framing the problem lays bare its complexity: multiple *neurons* encode multiple kinds of *features*, in which each kind of feature can take on multiple *values*, and the challenge is to understand the nature of this population multifeature code, and how it enables the many readout operations that can be performed over that code. In a sense, the manner in which *single* neurons encode multiple features has been studied since the very dawn of visual neurophysiology, when Hubel and Wiesel identified V1 neurons that are sensitive to both the location and orientation of a stimulus (Hubel & Wiesel, 1962), but as more recent work is moving towards treating neural *populations*, rather than single neurons, as the relevant unit of analysis (Saxena & Cunningham, 2019), it is now vital to develop methods and frameworks that can characterize the manner in which neural populations encode multiple features, specifically with regard to the important question of how these features are encoded relative to each other, whether in an independent or interactive manner. In this last section of the introduction, I briefly summarize some of the methods on offer, along with their limitations.

One method that is commonly used to study the relationship between two features in a neural population, particularly using fMRI, is the method of *cross-decoding*. In this method, a decoder (often a support vector machine), is trained to distinguish between two values of variable X at one value of variable Y, and is then tested on distinguishing those values of variable X for a *different* value of variable Y (e.g., train a classifier to discriminate a red and blue circle and test it on discriminating a red and blue square). If the classifier can cross-decode successfully, the inference is typically that variable X is encoded in a manner that is invariant or tolerant to changes in variable Y. By contrast, if the classifier fails to cross decode, it is taken to imply that these variables are encoded in a more interactive or conjunctive manner. On the face of it, this analysis might appear to answer the question of whether neurons in a population are tuned interactively to combinations of features. However, this is not quite true: successful cross-decoding can coexist with an interactive representation, and a failure to cross-decode can occur even in a neural population which purely exhibits main effects; in other words, these measures are double-dissociable. As an example of the first scenario (successful cross-decoding with an interactive representation), imagine a neural population with exactly one neuron tuned to shapes irrespectively of color, one neuron tuned to colors irrespective of shape, and 98 neurons that respond to heterogeneous mixtures of shapes and colors in an interactive fashion. Here, despite the fact that 98 of the neurons exhibit interactive, nonlinear tuning, cross-decoding would succeed: each of the two neurons exhibiting pure tuning to either feature would effectively provide a dimension that a putative classifier could exploit for cross-decoding (for intuition, imagine a jumbled pile of red shapes, and a differently-jumbled pile of blue shapes far above it; despite the heterogeneous color/shape tuning, a classifier could cross-decode color across different shapes, since there's a dimension along which color unambiguously varies). As an

example of the second scenario (a failure to cross-decode even in a population with no interaction effects), imagine a population with just two neurons that each shows main effects to both shape and color. In this case, their representation of four stimuli with two different shapes and two different colors would constitute a parallelogram in 2D space. Depending on the dimensions of the parallelogram, a plane bisecting two adjacent vertices of the parallelogram wouldn't necessarily fall between the remaining two vertices, and so cross-decoding would fail despite the lack of an interaction effect (this occurs because there are not sufficiently many dimensions for the classifier to exploit in making its classifications in this case). Many more scenarios could be enumerated, but for now, I merely note that the cross-decoding measure is in fact only an indirect measure of whether a neural population encodes multiple features in an interactive format.

A more direct measure suggests itself: why not simply run an ANOVA on each neuron individually and document whether it exhibits an interaction effect? This method indeed usefully captures the "ground truth", and in cases with sufficiently clean data it has proven informative; for example, it has been used to study an interactive relationship between color and shape processing in single neurons, and the relationship between stimulus and task information in single neurons (Chang et al., 2017; Rigotti et al., 2013). However, this univariate approach is much more challenging in the case of fMRI, for several reasons. First, each voxel is noisy, such that it might take vast and unfeasible amounts of data to observe a significant interaction effect for a single voxel even in cases where it exists. Second, interaction effects could be heterogeneous across voxels, such that one voxel might respond preferentially to a particular combination of features, whereas another voxel might respond less to that combination. Consequently, averaging across voxels to improve noise and running an ANOVA on these

averaged responses would hide heterogeneous tuning of this nature. Third, interaction effects might be genuinely present in a subset of voxels, but not others: thus, running a univariate ANOVA on every voxel and attempting to statistically aggregate the responses might drown the true interaction effects in a sea of noise.

Ideally, then, a method capable of testing for the presence of interactive tuning in the case of fMRI should be able to 1) aggregate small effects across voxels, such that results that might be insignificant for a single voxel become significant when evidence is pooled across voxels, 2) take into account heterogeneous interaction effects across voxels, and 3) be sensitive even when only a subset of voxels exhibit interaction effects. To my knowledge, only one study has proposed a solution to this problem. Allefeld & Haynes (2014) suggest that the MANOVA could be used for this purpose, since no decoding approaches (as of then) existed that could test for interaction effects. However, the MANOVA relies on many assumptions that are unfeasible to verify in fMRI data: for example, it requires that the number of trials in each condition be greater than the number of dependent variables (voxels in this case; this condition sharply limits the number of voxels that can be used and raises the problem of how to select these voxels in the first place), that the dependent variables be relatively uncorrelated (unfeasible for adjacent voxels), and that the data follow a multivariate normal distribution (hard to verify for fMRI data).

How then, to proceed? In two studies of this dissertation, I employ a novel method for testing sensitively for an interaction effect in fMRI data (although this method is perfectly general and can be applied to any recording modality, and indeed to any data whatsoever where detecting distributed, heterogeneous interaction effects is of interest). In short, this method makes use of support vector machines to measure subtle, heterogeneous interaction effects in

fMRI data. SVM decoders are abundantly used to sensitively test for whether a brain region contains information about a given binary classification, even when any given voxel might only have weak (and by itself statistically insignificant) information about the classification, when voxels exhibit heterogeneous effects for that classification (e.g., some voxels might prefer class A, and some might prefer class B), and when only a subset of voxels contain information about that classification. Note that these are analogous to the challenges I mentioned earlier for devising a method that can sensitively test for interaction effects; if SVM could be repurposed to test for interaction effects, it could overcome the challenges I mentioned. This is exactly the approach I take. While I describe the method in more detail in the corresponding chapters, the overall approach is that I compute *difference vectors* between the patterns associated with conditions that vary on one feature while keeping the other feature constant (e.g., the difference vectors for red square - blue square, and red circle - blue circle), and train an SVM decoder to discriminate not the original patterns, but these *difference vectors*; in other words, the SVM tests whether pattern differences across one condition are identical across changes in another condition, or whether a "difference of difference" (another way of describing an interaction term) exists in aggregate in the population. I therefore call this method *pattern difference decoding*. Unlike the MANOVA approach, SVM decoding also does not rely on strict assumptions about the data, advantages that this new decoding method inherits.

This method is designed to overcome the noisiness inherent in fMRI data to test for the presence of interaction effects, but obviously the mere presence of an interaction effect is not the only question of interest, since interaction effects can take many forms. As mentioned earlier, Study 2 of this dissertation examines how color and form are jointly represented in convolutional neural networks, which have the useful property of having no noise, either in their operation

(each unit's activation is a deterministic function of network input) or in their measurement (the experimenter can freely observe the activation of any unit without the use of a noisy proxy like the BOLD signal). Thus, in Study 2 of this dissertation, I use CNNs as a testing ground to further develop methods for understanding how multiple features are encoded in a neural population. Again, while I describe the methods in more detail in the corresponding section, the basic idea is that I employ a version of representational similarity analysis (Kriegeskorte & Kievit, 2013) that measures the extent to which the coding geometry of one feature varies across values of the other feature. This approach makes a step towards the important goal of understanding how multiple features are represented together within a neural population.

Thesis Outline

To briefly sum up the thesis thus far: we do not yet know how exactly neural populations encode multiple features, either with each other or in conjunction with task information. Color and shape are a useful case study, with reasonably well-charted neural correlates, but we do not yet understand the degree of anatomical segregation for the processing of these features (for instance, do regions exhibiting univariate sensitivity to one feature also have information about the other), nor do we understand how these features tend to be jointly represented in the same brain regions (independently or interactively). CNNs also encode color and form information, but as with the brain, the joint coding format of these features remains unknown. Finally, interactive coding for stimulus/task information has been shown to be an important motif of flexible neural computation, but the extent of this coding motif in the human brain remains completely unknown, and there do not yet exist methods for testing for this coding motif noninvasively.

In the studies that follow, I address these questions. In Study 1, I comprehensively examine the coding of color and form across the entire human ventral visual pathway, charting the topography and coding format of the coding for these features. In Study 2, I examine how color and form information are jointly encoded in CNNs. In Study 3, I examine how stimulus and task information are jointly encoded across the human visual system, contrasting parietal, ventral, and early visual regions. Throughout these studies, I develop methods for understanding the joint coding of multiple features in neural populations.

Chapter 2: Representation of Color, Form, and their Conjunction Across the Human Ventral Visual Pathway

Introduction

Research over the past several decades has provided us with a wealth of knowledge regarding the representation of color and form information in the primate brain. Both color and form information have been shown to be represented and transformed across multiple levels of processing, with the relevant neural processes spanning the entire visual processing hierarchy, from the retina to higher-level ventral stream regions. Notably, human fMRI studies have identified form-processing regions in lateral and ventral occipito-temporal cortex (OTC) (Malach et al., 1995; Grill-Spector et al., 1998; Kourtzi & Kanwisher, 2001; Orban et al., 2004), and both monkey neurophysiology and human fMRI studies have reported color-processing regions in ventral OTC (Hadjikhani et al., 1998; Brewer et al., 2005; Conway et al., 2007; Lafer-Sousa & Conway, 2013; Lafer-Sousa et al., 2016; Chang et al., 2017; Conway, 2018). How can we reconcile the presence of these regions showing univariate sensitivity to color or form with the fact that color and form information have been reported throughout much of the ventral visual pathway? A further question, for regions encoding information about both color and form, is how these features are represented together within a brain region: is each feature encoded completely invariantly to changes in the other feature in an orthogonal manner, or are these features encoded in a more interactive manner, where coding for one feature influences the representation of the other feature (or some combination of these two coding motifs)? Using fMRI and multi-voxel pattern analysis (MVPA), by replicating and extending a previous study conducted by Seymour

et al. (2010), we aim to address these questions and provide an up-to-date documentation of the representation of color, form, and their conjunction across the human ventral visual pathway.

Color and Form Processing Across the Visual Hierarchy

Past work has demonstrated that both color and form information is successively transformed across a series of processing stages, spanning from early visual cortex to anterior temporal lobe regions. V1 and V2 have been shown to contain cells with a range of different tuning profiles to color and form information, with some degree of mesoscale segregation of neurons specialized for these features (Livingstone & Hubel, 1988; Gegenfurter et al., 1996; Conway, 2001; Johnson et al., 2001; Ts'o et al., 2001; Johnson et al., 2008; Conway, 2010; Shapley & Hawken, 2011). V3 also contains both color and form information (e.g., Seymour et al., 2010), although some theoretical accounts of the primate color processing hierarchy omit it entirely (e.g., Conway, 2009; Conway et al., 2010). Area V4 is an important hub for both color processing (Brewer et al., 2005; Conway et al., 2007; Brouwer & Heeger, 2009; Brouwer & Heeger, 2013; Bannert & Bartels, 2018), and the coding of mid-level form features such as curvature (Gallant et al., 1993; Gallant et al., 2000), convexity (Pasupathy & Connor, 2001), and texture (Roe et al., 2012; Pasupathy et al., 2019). Color- and shape-tuned V4 neurons have been reported to show some segregation in macaques (Conway et al., 2007).

While color and form information clearly coexists within each early visual area, for higher ventral regions beyond V4, at least some evidence suggests that coding for these features may exhibit more anatomical separation. Specifically, macaque inferotemporal cortex (IT) has been shown to contain neurons tuned to high-level shape features (e.g., Tanaka, 1996; DiCarlo et al., 2012; Lehky & Tanaka, 2016; Bao et al., 2020), and human fMRI studies have revealed

arguably homologous regions in the human lateral and ventral OTC that respond more to coherent shapes than scrambled stimuli (Malach et al., 1995; Grill-Spector et al., 1998; Kourtzi & Kanwisher, 2001; Orban et al., 2004). Damage to this cortical region can result in loss of form perception, with spared color perception (Benson & Greenberg, 1969; Goodale & Milner, 2004). Analogously, a series of color-sensitive regions, which can be roughly divided into posterior, central, and anterior zones, have been identified that exhibit color-tuning in the macaque, and that respond more to colored than greyscale stimuli in the human brain (Hadjikhani et al., 1998; Brewer et al., 2005; Conway et al., 2007; Lafer-Sousa & Conway, 2013; Lafer-Sousa et al., 2016; Chang et al., 2017; Conway et al., 2018). Damage to the central and anterior color regions has been linked to neuropsychological deficits in color knowledge or color naming (reviewed in Siuda-Krzywicka & Bartolomeo, 2019) and impaired color processing with largely spared form processing (Bouvier & Engel, 2006).

The existence of regions reliably showing sensitivity to color and form in univariate contrasts is consistent with a modular view of feature representation in high-level vision, with different features encoded by anatomically distinct neural populations. However, color and form information may be encoded in distributed, fine-grained activation patterns that univariate methods cannot detect (e.g., Haxby et al., 2001). Indeed, using fMRI MVPA, several human studies have found that area LO can decode color information (Bannert and Bartels, 2013; Bannert and Bartels, 2018). This is consistent with neurophysiological findings showing that macaque IT and color regions contain both color and form information (Komatsu & Ideura, 1993; McMahon & Olson, 2009; Chang et al., 2017).

Although past studies have provided us with a wealth of information regarding how color and form are processed in the primate brain, individual studies have often examined different

brain regions with different methods or stimuli, or focused on just one feature or the other, making it difficult to construct an overarching view of how these features are coded relative to each other within a brain region and across different regions along the primate ventral processing pathway. A particularly important theoretical concern is reconciling the existence of regions showing univariate sensitivity for color or form with the evidence suggesting that tuning for these features might be broadly distributed throughout the ventral visual pathway. It is presently unknown whether different features are represented equally strongly in these regions at the multivariate level, or whether there remains a multivariate feature coding bias consistent with a region's univariate feature preference. A primary goal of the present study is thus to systematically document the coding of color and form information throughout the human ventral visual processing pathway, how the relative coding strength of these two types of feature may change across brain regions, and whether there is a close correspondence in a region's univariate and multivariate selectivity for a particular feature, using sensitive multivariate measures and well-controlled stimuli varying in their complexity.

Independent or Interactive Coding of Color and Form?

The mesoscale segregation of neurons specialized for processing color and form features in early visual areas, the existence of higher visual regions showing univariate response preference to color or form, and the behavioral deficits associated with damage to these regions are consistent with independent coding of color and form in the primate brain. Available behavioral evidence also supports this view. For example, visual search for single features is fast, but search for feature conjunctions is slow, an observation that led Treisman and Gelade (1980) to posit *feature integration theory*: different visual features are initially encoded on their own

distinct feature maps, and focused attention then spatially links the different features associated with the same object to form conjunctions of features. Consistent with this framework, "illusory conjunctions" can be induced, where the features of different objects are mismatched in conscious perception; this can occur in patients with parietal lesions, or in normal participants under conditions of divided attention (Treisman & Schmidt, 1982; Cohen & Rafal, 1991; Friedman-Hill et al., 1995). It is worth noting that while independent access of color and form features may be supported by distinctive neuronal populations either from separate brain regions or commingled within the same region, it may also be supported by neurons coding both features in an additive/orthogonal manner, where the coding of each feature is unaffected by coding for the other, thereby enabling independent feature readout.

Meanwhile evidence exists showing that color and form may be automatically coded in an interactive fashion. For instance, achromatopsia patients can still perceive shapes defined by isoluminant color boundaries (Victor et al., 1989; Heywood et al., 1991; Barbur et al., 1994; Heywood et al., 1998), demonstrating that one feature can contribute to the processing of the other feature. In another example, human observers are able to correctly perceive color/orientation bindings even when stimuli with different colors and orientations are rapidly alternating, suggesting that at least for certain stimuli, color and form features are automatically encoded in a conjoined format without requiring a separate, laborious attention-driven binding step (Holcombe & Cavanagh, 2001; see also evidence from Stromeyer, 1969; Cavanagh, 1991; Mandelli & Kiper, 2005). At the level of neural coding, one signature of interactive feature coding of color and form is the presence of non-additive neural responses to different feature combinations, where the coding for each feature depends on the value of the other; such tuning can also coexist with an additive component of the neural tuning function, in the same way that a

main effect and an interaction can coexist in any quantity influenced by multiple variables. Non-additive tuning has been found in human early visual areas in fMRI studies (Engel, 2005; Seymour et al., 2010; see more details of the latter study below). In macaques, such non-additive feature coding has been reported in V4 and color regions, is largely absent in IT, and was not explicitly tested in V1 and V2 despite the presence of neurons exhibiting tuning for both color and form (Friedman et al. 2003, McMahon & Olson, 2009, Bushnell & Pasupathy, 2012, and Chang et al., 2017). Notably, the studies that have demonstrated interactive color/form coding in the human brain thus far have involved relatively simple form features, such as orientation, leaving it unknown whether this processing format is used for the conjunction of color with more complex form features as well.

A second goal of the present study is thus to examine the prevalence of non-additive color and form coding in ventral visual cortex, whether it is present for both the conjunction of color and simple form features and that of color and more complex form features, and whether it can be found in lower as well as higher ventral regions in the human brain. Due to the "combinatorial explosion" involved in directly encoding every possible combination of color and form features, it is possible that interactive coding may only apply for some form features but not others, making it important to determine how broadly it applies.

Present Study

To answer the outstanding questions raised above, we replicated and extended a previous human fMRI multivoxel pattern analysis (MVPA) study by Seymour et al. (2010) that examined color and orientation coding in early visual areas. In this study, spiral stimuli were shown that were either clockwise or counterclockwise, and either red or green (see Figure 2.2 for an

illustration). In the single conjunction condition, spiral stimuli for each orientation and color combination were shown in different blocks of trials, with the phase of each spiral alternating over the course of the block to ensure that any form decoding was not a confound from differing retinotopic footprints of the stimuli. Decoding fMRI response patterns from these blocks revealed that color and orientation were both present in V1, V2, V3, and V4. In the double conjunction condition, *pairs* of stimuli with both features differing (e.g., either Red-Clockwise and Green-Counterclockwise, or Red-Counterclockwise and Green-Clockwise) were shown alternating throughout a block of trials, such that the two kinds of block had the same individual features, but differed in how they were conjoined. Decoding fMRI response patterns between these blocks revealed interactive coding of color and orientation throughout V1 to V4; in other words, these regions appeared to encode not just color and orientation, but also the specific way they were combined.

While Seymour et al. (2010) was elegantly designed and theoretically informative, because only the coding of color and orientation in early visual areas was examined, it provides us with an incomplete picture regarding the coding of these features in higher visual regions and the coding of color and more complex form feature in ventral visual cortex. Additionally, the finding that color and form features are interactively encoded in early visual cortex is important to replicate, as this result rules out neural models that posit purely separate processing of these features in early processing.

In this study, we extended Seymour et al. (2010)'s paradigm in two important ways. First, in addition to human early visual areas, we examined higher-level ventral stream regions exhibiting univariate sensitivity to either color or shape information. Second, besides examining the coding of color and spiral stimuli (a low-level form feature), we documented the coding of

color and a mid-level form feature, curvature, across the ventral visual regions. Together, our approach allowed us to comprehensively probe every region in the ventral visual pathway with the same stimuli to document the coding strength of color and form features (both simple and complex) within a brain region as well as across different brain regions and examine whether there is a close correspondence in a region's univariate and multivariate selectivity for a particular feature. Our study additionally allowed us to replicate the interactive coding for color/orientation conjunctions in early visual areas as reported by Seymour et al. (2010) and test whether such a coding scheme is specific to simple form features in early visual areas, or is a broader motif of color-form processing in human ventral cortex. As an additional extension of their study, we devised a new analysis technique, *pattern difference MVPA*, that we used as a further method to probe for interactive color-form tuning; this method can also be used to look for subtle interaction effects in any fMRI data beyond the present study.

With the exception of the central color-sensitive region, we found that information about color and form was always co-localized in the same brain regions throughout the ventral visual cortex, even for color and shape regions defined based on their univariate sensitivity to one feature. Nevertheless, preference of color and form information varied broadly across the regions, with preference obtained by univariate and multivariate measures by and large agreeing with each other. Color and form features are thus represented in the human brain in a biased distributed manner. Furthermore, color and form were coded together in a manner such that coding of each feature was tolerant to changes in the other feature throughout the ventral visual cortex. Meanwhile, evidence also exists for interactive color-form coding in several regions, most reliably for the coding of color with simple form features in early visual regions.

Results

Using fMRI MVPA, in the two experiments of this study, we examined the representation of simple and complex form features, color, and their conjunction in human early visual areas (V1 to V4) and higher-level ventral regions showing univariate sensitivity to shape (LOT and VOT) and color (posterior and middle color regions) (see Figure 2.1 for examples of these regions). This study served to both replicate the results of a study from Seymour et al. (2010), and extend their results from early visual cortex to higher-level ventral visual regions and from orientation to more complex form features. We aimed to understand the coding strength of these two types of features within a given brain region and across different brain regions along the ventral visual cortex, whether there is a close correspondence in a region's univariate and multivariate selectivity for a particular feature, and whether these two types of features are represented in a predominantly independent/orthogonal or an interactive manner when representations of both features are found within the same brain region. We examined the coding of color and orientation in Experiment 1 by showing clockwise and counterclockwise spirals appearing in red and green colors, and the coding of color and curvature in Experiment 2 by showing spiky and curvy tessellations appearing in red and green colors (Figure 2.2). The phase of all stimuli alternated once per second, equating the overall stimulation across the visual field (and ruling out the possibility that any "form" decoding could merely be due to differences in the spatial envelope of the stimuli). In some of the runs, only a single stimulus type was present in each block. FMRI pattern decoding from these runs were used to determine which brain regions contain color and/or form information and how the relative coding strength of color and form may change across the ventral visual pathway. FMRI response patterns from these runs were also used in a novel measure to test the presence of interactive coding of color and form in a brain

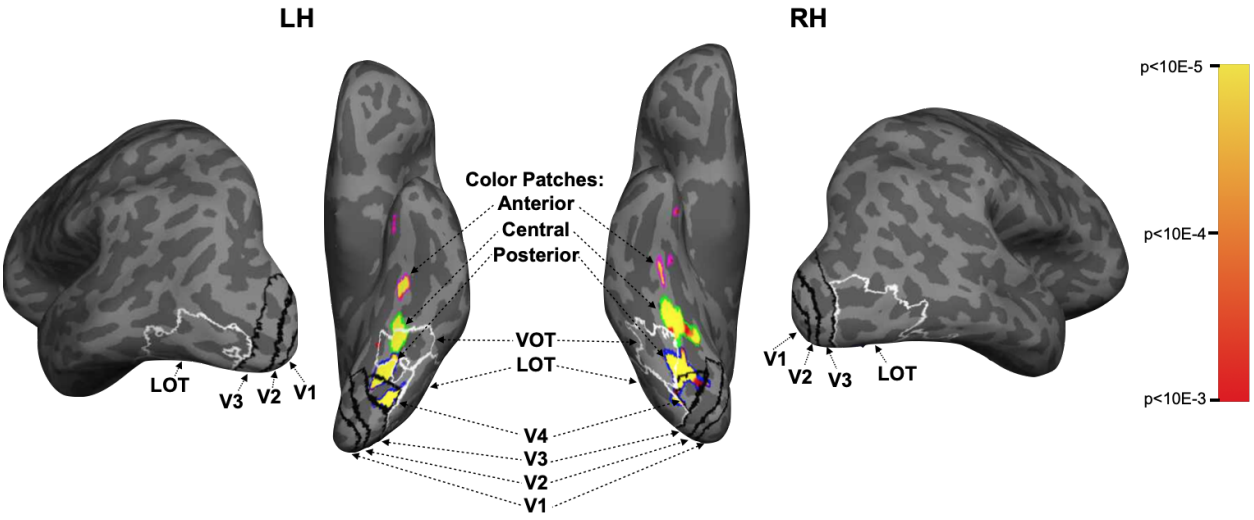


Figure 2.1 Lateral and ventral views of left and right hemisphere from an example participant, showing all regions of interest used in the study. Retinotopically defined areas V1, V2, V3, and V4 shown with black outlines; object-sensitive regions LOT and VOT shown with white outlines; posterior, central, and anterior color-sensitive regions shown with blue, green, and magenta outlines, respectively, along with their activation maps from the color versus greyscale localizer used to define them.

region. In the other runs, these stimuli were presented in blocks where stimuli of different forms and colors were alternated, which we analyzed using a method adapted from Seymour et al. (2010) as another metric to test for the presence of interactive coding.

ROI Overlap

Since retinotopic V4, the posterior color region, and area VOT overlap to some degree, we quantified this overlap for each pair of these ROIs. Across all the participants from both Experiments 1 and 2, V4 and the posterior color region overlapped by 40.7% +/- 2.4% (mean +/- s.e.). VOT and the posterior color region overlapped by 16.4% +/- 2.7%. VOT and V4 overlapped by 17.5% +/- 3.5%. There is thus a sizable overlap between V4 and the posterior color region, with both also overlapping slightly with VOT. Despite these overlaps, as described below, there were significant differences in how color and form were represented in these brain

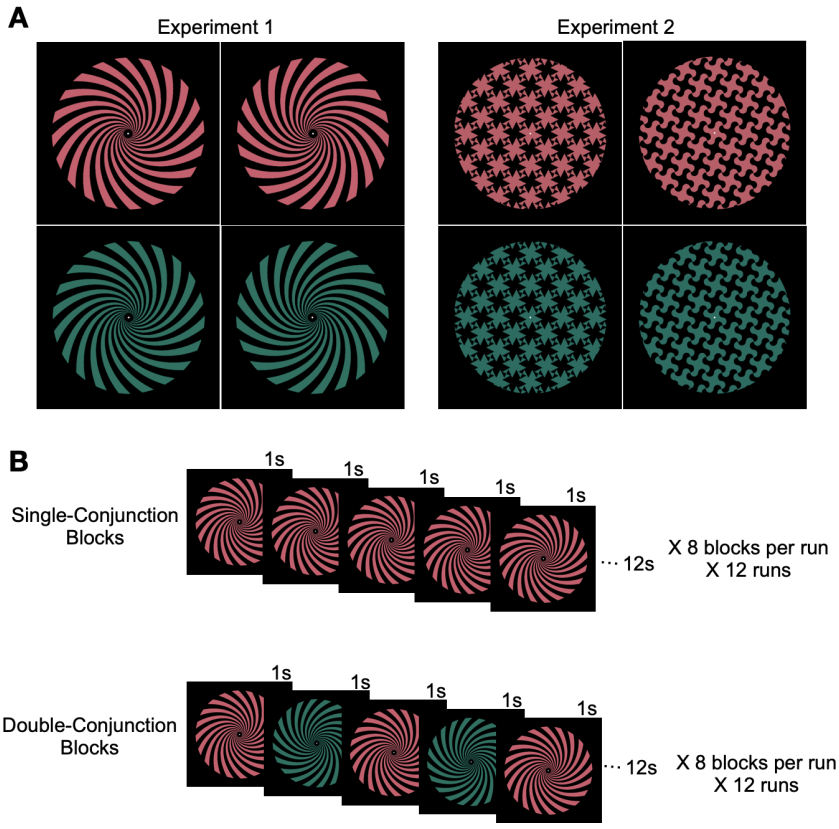


Figure 2.2. Stimuli and experimental design. **A.** In Experiment 1 (left), logarithmic spiral stimuli (adapted from Seymour et al., 2010) were shown that could be oriented clockwise or counterclockwise, and colored red or green. These spirals have the property that their arms are a fixed angle from radius at all points, ensuring that gross radial biases in cortical retinotopic maps could not drive decoder performance. In Experiment 2 (right), spiky and curvy tessellation stimuli were used, with the same colors as Experiment 1. The stimuli alternated phase once per second, such that black shapes within the circular aperture became colored, and vice versa. **B.** The two kinds of blocks present in both experiments. Stimuli were either presented in single-conjunction blocks, where a single stimulus type (e.g., Red-CW spiral) was presented for the entire block with its phase alternating once per second, or in double-conjunction blocks, where stimuli varying with respect to both features alternated once per second within a block. Thus, in Experiment 1, the two kinds of double-conjunction block were Red-CW/Green-CCW and Red-CCW/Green-CW; in Experiment 2, the two kinds of double-conjunction block were Red-Spiky/Green-Pinwheel and Red-Pinwheel/Green-Spiky.

regions that could not be predicted by the amount of anatomical overlap. Consequently, we grouped brain regions in a later analysis by their overall functional response profile, rather than by the amount of anatomical overlap.

Feature decoding

To document whether color and form information were present in a brain region, we compared color and form decoding accuracy in each region against chance level performance (Figure 2.3). Here decoding was performed between fMRI response patterns differing in one feature dimension while allowing these patterns to take on either value of the other feature dimension (e.g., color decoding in Experiment 1 was performed by contrasting the red clockwise and red counterclockwise conditions against the green clockwise and green counterclockwise conditions). Except for the central color region, which showed no significant form decoding in either experiment ($ts < 1.14$, $ps > .18$), both color and form were decodable significantly above chance in both experiments in every brain region examined, including V1 through V4, the two shape regions LOT and VOT, and the posterior color region ($ts > 2.27$, $ps < .03$, one-tailed as only values above chance-level performance are meaningful here. Corrected for multiple comparisons using the Benjamini-Hochberg procedure across the set of tests done within each ROI; this applies to all t-tests performed in this study, see Methods for more details. Figure 2.3 depicts these results with the significance level of each t-test for above-chance decoding labeled with asterisks at the top of each bar). Color and form information is thus prevalent throughout the ventral visual cortex.

To characterize the coding strength of color and the two types of form features (i.e., orientation and curvature) within a given region, we next conducted detailed comparisons within

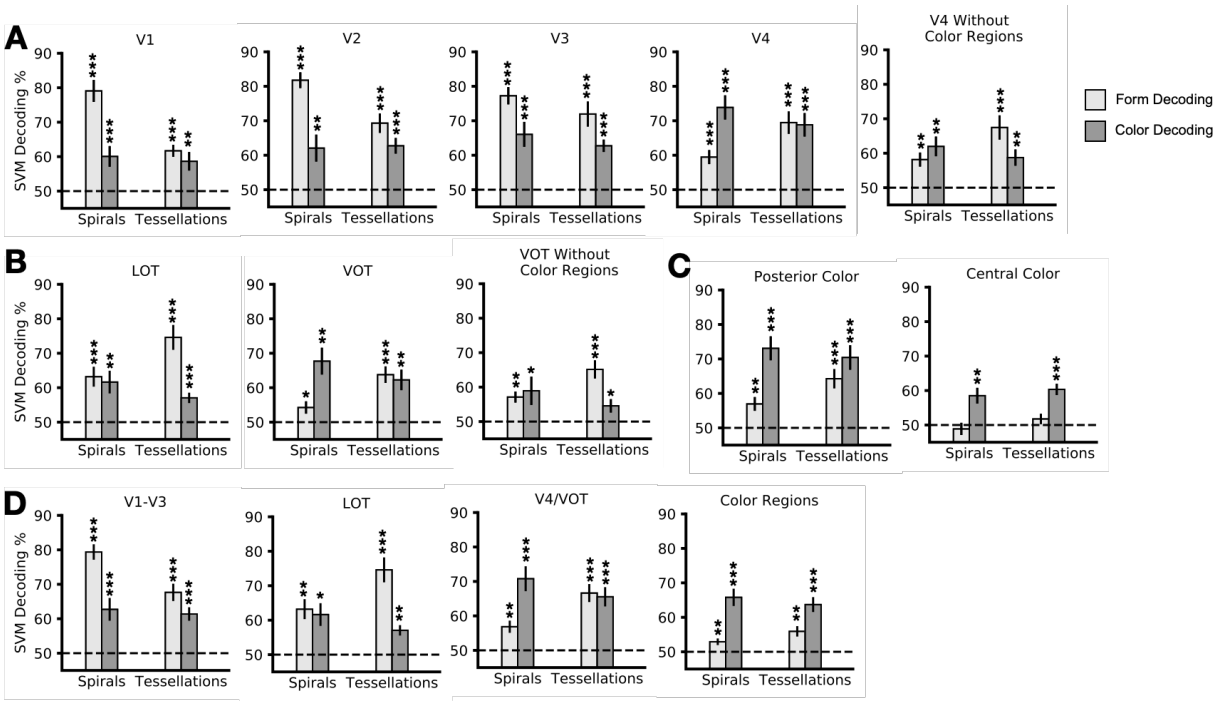


Figure 2.3. Results of color and form decoding in both experiments for **(A)** early visual areas, **(B)** shape regions, **(C)** color regions, and **(D)** sectors, which were formed by averaging the decoding of brain regions showing similar response profiles. Overall, V1 and V2 show a preference for orientation over curvature and color. V3 shows an equal preference to orientation and curvature over color. VOT and V4 showed equal preference to color and curvature over orientation; the overlap of V4 and VOT with the color regions partially, but not entirely, drove color decoding in these regions. Removing the color region overlap resulted in VOT showing a preference for curvature over orientation and color. LOT showed a preference for curvature over color and orientation. Lastly, the posterior color region showed a preference for color over orientation but not over curvature, while the central color region showed a preference for color over both form features. . * $p < .05$; ** $p < .01$; *** $p < .001$ for t-tests testing for above chance (> 50%) decoding (one-sample t-tests, one-tailed).

and across the two experiments (all statistical results are reported in Table 2.1). We noted that color coding did not vary between the two experiments in any of the brain regions examined even though only a subset of the participants completed both experiments. Because color stimulation was comparable between the two experiments (as color covered half of the stimuli in both experiments), this indicates that participant performance was comparable and fairly stable across the two experiments. This enabled us to directly compare orientation and curvature coding between the two experiments and evaluate how the processing of these two form features may

differ within a brain region. To account for the fact that partially overlapping participants partook in both experiments, a linear mixed effects analysis was performed to determine the influence of experiment, feature type, and their interaction on decoding in each region and between regions, and a partially-overlapping t-test (Derrick et al., 2017) was performed to compare between pairs of conditions across the two experiments. Within each experiment, within-subjects t-tests were used to compare color and form decoding. All t-tests were corrected within each kind of comparison for a given ROI (i.e., across the two comparisons of color versus form decoding within each of two experiments, and across the two comparisons of feature decoding between the two experiments).

As shown in Figure 2.3 and Table 2.1, overall, in early visual areas, V1 and V2 showed a main effect of higher form than color decoding, with decoding further being higher for orientation than for either curvature or color. V3 also showed a main effect of higher form than color decoding, but with similar decoding for both form features. V4, on the other hand, showed a main effect of higher color than form decoding, with decoding further being higher for color and curvature than for orientation. In the two form-selective regions, VOT, like V4, showed a main effect of higher color than form decoding, with decoding further being higher for color and curvature than for orientation. LOT, on the other hand, showed no main effect of feature decoding, but higher decoding for curvature than for either color or orientation, consistent with its role in object shape processing. Both color regions showed a main effect of higher color than form decoding. While the posterior color region showed higher decoding for color and curvature than for orientation with no significant difference between decoding for color and curvature, the middle color region showed higher decoding for color than for either kind of form feature.

Since V4 and VOT overlapped somewhat with the posterior color region, we performed additional analyses examining decoding in these regions when their overlap with the color-sensitive regions were removed. The same feature decoding analyses were performed in these regions as in the other regions (Figure 2.3). Mixed-effects analyses were performed for each feature to compare decoding in these regions with or without the parts of these regions that overlapped with the color regions. For form decoding, V4 showed no difference when the overlap with color-sensitive regions was removed ($Z = .58$, $p = .57$, two-tailed), but VOT showed a slight trend towards an increase in form decoding ($Z = 1.66$, $p = .096$, two-tailed). However, in both ROIs, color decoding significantly decreased when the posterior color region was removed (Zs > 3.4, ps < .01, two-tailed), though color decoding remained significantly above chance (ts > 2.26, ps < .03, one-tailed). Removing the overlapping color region from V4 and VOT also changed the relative coding strength of color and form in these regions (see the detailed stats reported in Table 2.1). Both regions no longer showed an overall main effect of higher color than form decoding, with VOT now showing a greater sensitivity to curvature than to color or orientation changes. The latter is consistent with VOT's role in object shape processing. Thus, removing the color-sensitive voxels from VOT and V4 removed their apparent feature preference for color.

Based on the overall similarity of their response profiles and their anatomical proximity, ROIs were grouped into sectors to allow us to directly compare the feature coding characteristics between the different sectors: early visual areas V1-V3, lateral visual area LOT, ventral visual areas V4/VOT, and Color Regions (including the posterior and central color regions). Decoding accuracies were averaged within each sector across the component brain regions. The decoding profiles within each sector are reported in Figure 2.3 and Table 2.1, and they are overall

Table 2.1 Summary of statistical comparisons within each ROI for color and form decoding results. Mixed-effects analyses were conducted to test the effect of experiment, feature type, and their interaction. Within-subject t tests were conducted to test the difference between color and form decoding within each experiment (stats reported were two-tailed and corrected for multiple comparisons across the two experiments). Partially-overlapping t-tests were conducted to compare the decoding of each feature across experiments (stats reported were two tailed and corrected for multiple comparisons across the two comparisons). † p < .10; * p <.05; ** p <.01; *** p <.001.

	Main Effects and Interaction			Form vs. Color Within Experiment		Spirals vs. Tessellations Within Feature	
ROI	**Experiment**	**Feature**	**Interaction**	**Spirals**	**Tessellations**	**Form**	**Color**
V1	z = 1.06 p = .29	z = 5.80 p < .001 ***	z = 3.51 p < .001 ***	$t(11)$ = 4.81 p = .001 **	$t(12)$ = 1.10 p = .30	$t(16.6)$ = 6.21 p < .001 ***	$t(16.6)$ = .97 p = .34
V2	z = .029 p = .98	z = 5.50 p < .001 ***	z = 2.56 p = .01 *	$t(11)$ = 5.13 p < .001 ***	$t(12)$ = 2.05 p = .06 †	$t(16.6)$ = 4.12 p = .002 **	$t(16.6)$ = .20 p = .84
V3	z = .85 p = .39	z = 2.61 p = .009 **	z = .082 p = .93	$t(11)$ = 2.21 p = .048 *	$t(12)$ = 2.71 p = .037 *	$t(16.6)$ = 1.14 p = .27	$t(16.6)$ = 1.20 p = .27
V4	z = .74 p = .46	z = 3.87 p < .001 ***	z = 2.91 p = .004 **	$t(11)$ = -3.88 p = .005 **	$t(12)$ = .18 p = .85	$t(16.6)$ = -2.59 p = .036 *	$t(16.6)$ = 1.17 p = .25
V4 w/out Color	z = .69 p = .49	z = .97 p = .33	z = 2.31 p = .02 *	$t(11)$ = -.15 p = .27	$t(12)$ = 2.01, p = .13	$t(16.6)$ = -2.38 p = .06 †	$t(16.6)$ = .97 p = .34
LOT	z = 1.11 p = .26	z = .42 p = .67	z = 3.1 p = .002 **	$t(11)$ = .40 p = .70	$t(12)$ = 4.93 p < .001 ***	$t(16.6)$ = -2.52 p = .04 *	$t(16.6)$ = 1.24 p = .23
VOT	z = 1.36 p = .18	z = 4.06 p < .001 ***	z = 3.26 p = .001 **	$t(11)$ = -3.19 p = .017 *	$t(12)$ = .58 p = .57	$t(16.6)$ = -2.66 p = .033 *	$t(16.6)$ = 1.18 p = .25
VOT w/out Color	z = .88 p = .37	z = .55 p = .58	z = 2.70 p = .007 **	$t(11)$ = -.44 p = .66	$t(12)$ = 3.96, p = .004 **	$t(16.6)$ = -2.59 p = .038 *	$t(16.6)$ = .92 p = .37
Posterior Color	z = .62 p = .54	z = 3.80 p < .001 ***	z = 1.69 p = .091 †	$t(11)$ = -3.33 p = .013 *	$t(12)$ = 1.59 p = .14	$t(16.6)$ = -2.23 p = .079 †	$t(16.6)$ = .61 p = .55
Central Color	z = .72 p = .47	z = 4.19 p < .001 ***	z = .33 p = .74	$t(11)$ = -3.27 p = .007 **	$t(12)$ = -5.71 p < .001 ***	$t(16.6)$ = -1.44 p = .34	$t(16.6)$ = -.65 p = .52
V1-V3	z = .75 p = .46	z = 5.51 p < .001 ***	z = 2.40 p = .017 *	$t(11)$ = 4.86, p = .002 **	$t(12)$ = 2.64, p = .02 *	$t(16.6)$ = 4.23 p = .001 **	$t(16.6)$ = .88 p = .39
V4/VOT	z = 1.09 p = .28	z = 4.50 p < .001 ***	z = 3.52 p = .001 **	$t(11)$ = -3.77 p = .006 **	$t(12)$ = .41 p = .69	$t(16.6)$ = -2.89 p = .02 *	$t(16.6)$ = 1.33 p = .20
Color Regions	z = .78 p = .43	z = 5.33 p < .001 ***	z = 1.50 p = .13	$t(11)$ = -4.30 p = .002 **	$t(12)$ = -3.81 p = .002 **	$t(16.6)$ = -1.69 p = .22	$t(16.6)$ = .73 p = .47

consistent with the profile of the individual regions comprising the sector. Three-way mixed-effects models (sector x feature x experiment) performed on each pair of sectors reveal significant or trending two-way and/or three-way interactions involving sector for each pair, verifying that each of these sectors indeed exhibits a distinct feature encoding profile from each of the others (significant or trending effects included: for Color Regions vs. LOT, sector x feature and 3-way interaction; for Color Regions vs. V1-V3, sector x feature and 3-way interaction; for Color regions vs. V4/VOT, 3-way interaction; for LOT vs. V1-V3, sector x feature and 3-way interaction; for LOT vs. V4/VOT, sector x feature; for V1-V3 vs. V4/VOT, sector x feature and 3-way interaction; all $Zs > 1.8$, $ps < .07$).

We found only scattered and limited evidence for hemispheric differences in color or form coding. In Experiment 1, V1 showed higher form decoding in the right hemisphere, and V3 showed higher color decoding in the right hemisphere ($ts > 2.36$, $ps < .05$; both two-tailed and uncorrected), but these effects were not present in Experiment 2 ($ts < .60$, $ps > .56$; two tailed and uncorrected), and no other ROIs exhibited a hemispheric difference for decoding of either feature ($ts < 1.7$, $ps > .12$; two tailed and uncorrected).

Overall, with the exception of the central color region, all other regions examined showed significant decoding for both color and form, even for shape and color regions showing univariate selectivity for color or form. This shows that a distributed color and form representation is more prevalent throughout the human ventral visual cortex than a segregated, modular organization. Meanwhile, significant coding bias also exists in every region examined, indicating the processing of color and form is not the same in the different brain regions: even early visual areas show some feature coding preference, and in higher visual regions, such a

preference appears to be largely consistent between multivariate decoding and the univariate feature preferences that define the regions.

Feature Cross-Decoding

To understand how color and form are coded together in a brain region, we next examined the extent to which each feature is encoded in a manner that is tolerant to changes in the other feature. To do so, we performed cross-decoding and trained an SVM classifier on one feature (e.g., form) within one value of the other feature (e.g., red), and tested the classifier in the other value of the other feature (e.g., green). Additionally, to obtain a baseline measure of feature decoding with an equal amount of data for comparison purposes, we also performed within-feature decoding, and trained and tested a classifier in one feature within the same value of the other feature. Figure 2.4 depicts the results of these analyses. Every region that showed successful decoding of a given feature in the previous analysis also exhibited significant cross-decoding of that feature ($ts > 1.92$, $ps < .05$; one-tailed t-test, corrected for multiple comparisons with the eight comparisons performed for each ROI). Meanwhile, several ROIs also exhibited a significant or trending drop in decoding when performing cross-feature rather than within-feature decoding (shown in Figure 2.4): specifically, V1 showed a significant or trending cross-decoding drop for both color and orientation in Experiment 1 ($ts > 2.00$, $ps < .08$; one-tailed and corrected for multiple comparisons across the four cross-decoding drops tested in V1), and V2 exhibited a significant or trending cross-decoding drop for color in Experiment 1, and for both color and curvature in Experiment 2 ($ts > 1.61$, $ps < .09$; one-tailed and corrected across the four cross-decoding drops tested in V2).

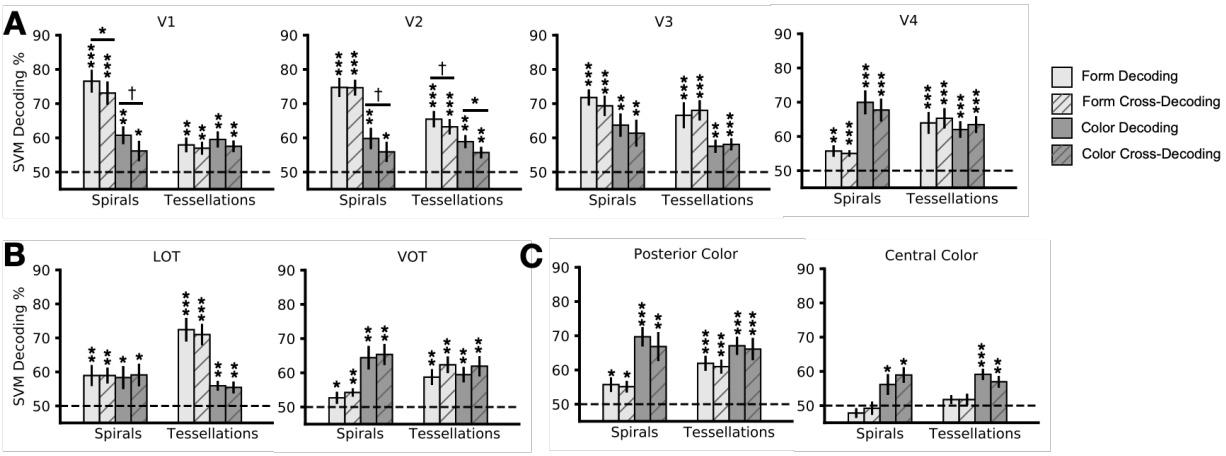


Figure 2.4. Results of feature cross-decoding analysis for **(A)** early visual areas, **(B)** shape regions, and **(C)** color regions. Solid bars show decoding accuracy for features trained and tested within the same value of the other feature (e.g., train on RedCW vs. RedCCW, test on RedCW vs. RedCCW); striped bars show decoding where training and testing for a feature is done across values of the other feature (e.g., train on RedCW vs. RedCCW, test on GreenCW vs. GreenCCW). Every region exhibiting successful decoding of a feature also exhibits significant cross-decoding; that said, V1 and V2 show a significant or trending drop in cross-decoding in several cases. † p <.10, * *p* <.05, ** p <.01, *** p <.001 for t-tests testing for above chance (> 50%) decoding (one-sample t-tests, one-tailed) and for t-tests comparing within-feature decoding to be above cross-decoding (within-subjects t-tests, one-tailed).

As the presence of a cross-decoding drop is an informative index of an interactive, rather than a completely orthogonal, relationship between color and form coding, to examine this effect in detail, we performed a set of further analyses. To increase power, we combined the effect from both color and form decoding (since a drop in either is suggestive of interactive coding between the features) and tested the amount of decoding drop in each ROI using one-tailed t-tests. Figure 2.5 shows the results of this analysis for the main voxel set used throughout this study (i.e., 300 most active voxels in each ROI). To examine how the results may depend upon the number of voxels included in each ROI, we also conducted this analysis separately for the top 100, 200, 300, 400, or 500 most active voxels in each ROI. Given that SVM is sensitive to both power and noise (such that including too few voxels may exclude some of the informative voxels and thus provide insufficient power whereas including too many voxels may add noise),

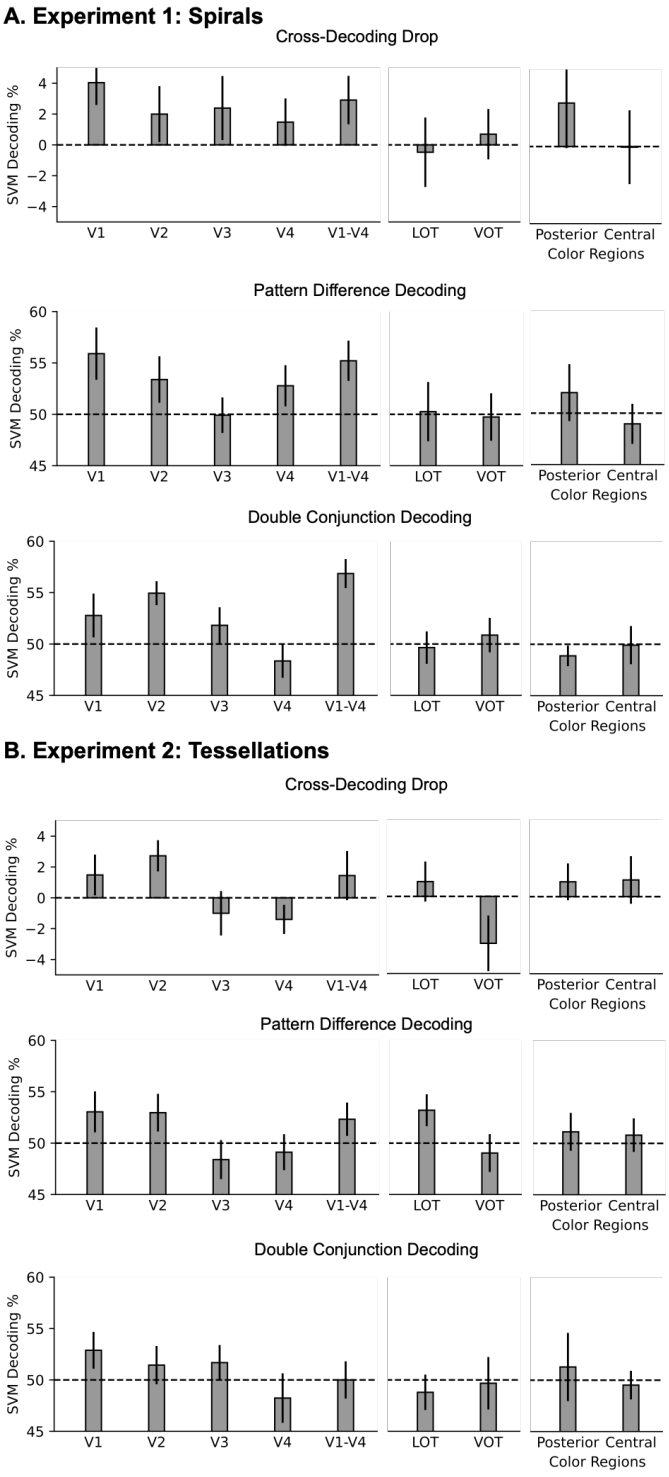


Figure 2.5. Results of the three analyses testing for interactive color/form coding — cross-decoding drop, pattern difference decoding, and double conjunction coding — for the Experiment 1 (A) and Experiment 2 (B), when the most active 300 voxels from each ROI were used for the analysis.

testing the effect at a range of voxel sizes may provide us with a more sensitive way to document the effect. Tables 2.2 and 2.3 (leftmost set of columns) show the results of this analysis for Experiment 1 (spirals) and Experiment 2 (tessellations), respectively. No correction for multiple comparison was applied here to allow our results to be more comparable to those of Seymour et al. (2010) (since correction for multiple comparisons was not mentioned in Seymour et al., we assume the results were uncorrected; corrected p-values can be easily derived from the uncorrected p-values reported using the Benjamini-Hochberg procedure). In Experiment 1, V1, V2, and a macro-ROI composed of V1 through V4 exhibit a significant drop in cross-decoding across multiple voxel selection conditions, with V3 showing trends. In Experiment 2, V2 and the posterior color region showed significant drops in cross-decoding, but only for one voxel selection condition each with a trend in one other voxel selection condition; V1 and LOT each exhibited a trend for one voxel selection condition. Thus, the strongest evidence for interactive coding based on the cross-decoding drop metric is for orientation-color conjunctions in early visual regions, with only scattered evidence for such coding in higher-level visual regions, or for curvature-color conjunctions. Other than these cases, however, color and form exhibit no significant drop in cross-decoding across the ventral visual pathway.

Pattern Difference Analysis

Despite the presence of some cross-decoding drop, all ROIs examined showed above chance cross-decoding for both color and form. Successful cross-decoding merely requires that the test patterns lie on the same side of the SVM classification boundary as the corresponding training patterns. As long as the difference between the patterns is large enough, an interactive and non-orthogonal color and form representations may show successful cross-decoding with no

Table 2.2. Statistical results from Experiment 1 (spirals) for the three types of analyses that measure interactive coding for color and form: cross-decoding drop, pattern difference decoding, and double conjunction decoding. Analyses were performed separately for the top 100 to 500 most active voxels in each ROI. All results were from one-sample, one-tailed t-tests examining whether the effects were significantly above chance. No correction for multiple comparisons were applied to make these results comparable to those of Seymour et al. (2010). Corrected p values may be derived using the Benjamini-Hochberg procedure. † p < .10, * p <.05, ** p <.01, and *** p <.001.

	Cross-Decoding Drop					Pattern Difference Decoding					Double Conjunction Decoding				
ROI	**Top100**	**Top200**	**Top300**	**Top400**	**Top500**	**Top100**	**Top200**	**Top300**	**Top400**	**Top500**	**Top100**	**Top200**	**Top300**	**Top400**	**Top500**
V1	$t(11) = 1.78$ $p = .051$ †	$t(11) = 2.29$ $p = .02$ *	$t(11) = 2.67$ $p = .01$ *	$t(11) = 2.00$ $p = .04$ *	$t(11) =1.48$ $p = .08$ †	$t(11) = 1.62$ $p = .07$ †	$t(11) = 2.01$ $p = .03$ *	$t(11) = 2.22$ $p = .02$ *	$t(11) = 1.23$ $p = .12$	$t(11) = 2.13$ $p = .03$ *	$t(11) = 3.06$ $p = .005$ **	$t(11) = 1.60$ $p = .07$ †	$t(11) = 1.25$ $p = .12$	$t(11) = 1.87$ $p = .04$ *	$t(11) = 2.10$ $p = .03$ *
V2	$t(11) = 3.18$ $p = .004$ **	$t(11) = 2.87$ $p = .008$ **	$t(11) = 1.05$ $p = .16$	$t(11) = 1.54$ $p = .08$ †	$t(11) = .85$ $p = .21$	$t(11) = 1.38$ $p = .097$ †	$t(11) = 3.28$ $p = .004$ **	$t(11) = 1.43$ $p = .09$ †	$t(11) = 2.37$ $p = .02$ *	$t(11) = 2.07$ $p = .03$ *	$t(11) = 1.02$ $p = .17$	$t(11) = 3.1$ $p = .005$ **	$t(11) = 4.07$ $p < .001$ ***	$t(11) = 3.68$ $p = .002$ **	$t(11) = 3.79$ $p = .002$ **
V3	$t(11) = .31$ $p = .38$	$t(11) = 1.25$ $p = .12$	$t(11) = 1.10$ $p = .15$	$t(11) = 1.40$ $p = .09$ †	$t(11) = 1.73$ $p = .06$ †	$t(11) =.57$ $p = .29$	$t(11) = .62$ $p = .27$	$t(11) = -.05$ $p = .51$	$t(11) = 1.25$ $p = .11$	$t(11) = 1.01$ $p = .17$	$t(11) = .82$ $p = .21$	$t(11) = 1.63$ $p = .07$ †	$t(11) = .99$ $p = .17$	$t(11) = 1.92$ $p = .04$ *	$t(11) = 1.55$ $p = .08$ †
V4	$t(11) = -.82$ $p = .78$	$t(11) = 0.0$ $p = .5$	$t(11) = .92$ $p = .19$	$t(11) = .46$ $p = .33$	$t(11) = .49$ $p = .32$	$t(11) = .75$ $p = .23$	$t(11) = 1.16$ $p = .14$	$t(11) = 1.33$ $p = .11$	$t(11) = .64$ $p = .27$	$t(11) = .70$ $p = .25$	$t(11) = 1.80$ $p = .049$ *	$t(11) = .15$ $p = .44$	$t(11) = -.96$ $p = .82$	$t(11) = .044$ $p = .48$	$t(11) = .32$ $p = .38$
V1-V4	$t(11) = 2.70$ $p = .01$ *	$t(11) = 1.48$ $p = .08$ †	$t(11) = 1.77$ $p = .052$ †	$t(11) = 2.24$ $p = .02$ *	$t(11) = 1.86$ $p = .045$ *	$t(11) = 1.43$ $p = .09$ †	$t(11) = .76$ $p = .23$	$t(11) = 2.55$ $p = .014$ *	$t(11) = 2.84$ $p = .008$ **	$t(11) = 2.75$ $p = .009$ **	$t(11) = 2.39$ $p = .018$ *	$t(11) = 4.16$ $p < .001$ ***	$t(11) = 4.64$ $p < .001$ ***	$t(11) = 2.32$ $p = .02$ *	$t(11) = 3.17$ $p = .004$ **
LOT	$t(11) = -.129$ $p = .55$	$t(11) = -.22$ $p = .58$	$t(11) = -.20$ $p = .58$	$t(11) = .016$ $p = .49$	$t(11) = .32$ $p = .38$	$t(11) = -.64$ $p = .73$	$t(11) = .30$ $p = .38$	$t(11) = .08$ $p = .47$	$t(11) = .10$ $p = .46$	$t(11) = .17$ $p = .43$	$t(11) = .19$ $p = .43$	$t(11) = -.04$ $p = .51$	$t(11) = -.21$ $p = .58$	$t(11) = .07$ $p = .47$	$t(11) = .23$ $p = .41$
VOT	$t(11) = -.14$ $p = .55$	$t(11) = .07$ $p = .47$	$t(11) = .41$ $p = .35$	$t(11) = -.02$ $p = .51$	$t(11) = -.19$ $p = .57$	$t(11) = -1.50$ $p = .92$	$t(11) = -.52$ $p = .69$	$t(11) = -.11$ $p = .54$	$t(11) = -.94$ $p = .92$	$t(11) = -.70$ $p = .75$	$t(11) = 1.59$ $p = .07$ †	$t(11) = 1.01$ $p = .17$	$t(11) = .50$ $p = .32$	$t(11) = -.76$ $p = .77$	$t(11) = .48$ $p = .32$
Posterior Color	$t(11) = .18$ $p = .43$	$t(11) = .60$ $p = .28$	$t(11) = .92$ $p = .19$	$t(11) = .87$ $p = .20$	$t(11) = .94$ $p = .18$	$t(11) = .53$ $p = .31$	$t(11) = .95$ $p = .18$	$t(11) = .69$ $p = .25$	$t(11) = .54$ $p = .30$	$t(11) = .13$ $p = .44$	$t(11) = 1.09$ $p = .15$	$t(11) = .80$ $p = .22$	$t(11) = -1.08$ $p = .85$	$t(11) = -.15$ $p = .56$	$t(11) = -.83$ $p = .79$
Central Color	$t(11) = .36$ $p = .36$	$t(11) = -.62$ $p = .73$	$t(11) = -.02$ $p = .51$	$t(11) = -.44$ $p = .67$	$t(11) = -.23$ $p = .59$	$t(11) = -.33$ $p = .63$	$t(11) = -1.3$ $p = .90$	$t(11) = -.51$ $p = .69$	$t(11) = -.39$ $p = .65$	$t(11) = -.19$ $p = .57$	$t(11) = .62$ $p = .27$	$t(11) = -.42$ $p = .66$	$t(11) = -.05$ $p = .52$	$t(11) = -.41$ $p = .65$	$t(11) = -.19$ $p = .57$

Table 2.3. Statistical results from Experiment 2 (tessellations) for the three types of analyses that measure interactive coding for color and form: cross-decoding drop, pattern difference decoding, and double conjunction decoding. Analyses were performed separately for the top 100 to 500 most active voxels in each ROI. All results were from one-sample, one-tailed t-tests examining whether the effects were significantly above chance. No correction for multiple comparisons were applied to make these results comparable to those of Seymour et al. (2010). Corrected p values may be derived using the Benjamini-Hochberg procedure. † p < .10, * p <.05, ** p <.01, and *** p <.001.

	Cross-Decoding Drop					Pattern Difference Decoding					Double Conjunction Decoding				
ROI	**Top100**	**Top200**	**Top300**	**Top400**	**Top500**	**Top100**	**Top200**	**Top300**	**Top400**	**Top500**	**Top100**	**Top200**	**Top300**	**Top400**	**Top500**
V1	$t(12) = -.03$ $p = .51$	$t(12) = 1.62$ $p = .07$ †	$t(12) = 1.08$ $p = .15$	$t(12) = .58$ $p = .28$	$t(12) = .42$ $p = .34$	$t(12) = .40$ $p = .35$	$t(12) = 1.02$ $p = .16$	$t(12) = 1.47$ $p = .08$ †	$t(12) = 2.57$ $p = .01$ *	$t(12) = 1.44$ $p = .09$ †	$t(12) = .31$ $p = .38$	$t(12) = 1.34$ $p = .10$	$t(12) = 1.55$ $p = .07$ †	$t(12) = .79$ $p = .22$	$t(12) = .50$ $p = .31$
V2	$t(12) = 1.04$ $p = .16$	$t(12) = 1.25$ $p = .12$	$t(12) = 2.59$ $p = .01$ *	$t(12) = 1.67$ $p = .06$ †	$t(12) = .64$ $p = .27$	$t(12) = 2.19$ $p = .03$ *	$t(12) = 1.75$ $p = .053$ †	$t(12) = 1.56$ $p = .07$ †	$t(12) = .97$ $p = .17$	$t(12) = 1.08$ $p = .15$	$t(12) = .42$ $p = .34$	$t(12) = 1.86$ $p = .04$ *	$t(12) = .74$ $p = .24$	$t(12) = .66$ $p = .26$	$t(12) = .41$ $p = .34$
V3	$t(12) = .45$ $p = .33$	$t(12) = -.28$ $p = .61$	$t(12) = -.67$ $p = .74$	$t(12) = -1.29$ $p = .88$	$t(12) = -.85$ $p = .80$	$t(12) = .09$ $p = .47$	$t(12) = -1.70$ $p = .95$	$t(12) = -.81$ $p = .78$	$t(12) = -1.53$ $p = .92$	$t(12) = -1.16$ $p = .87$	$t(12) = .48$ $p = .32$	$t(12) = -.41$ $p = .65$	$t(12) = .94$ $p = .18$	$t(12) = -.11$ $p = .54$	$t(12) = .41$ $p = .35$
V4	$t(12) = .14$ $p = .44$	$t(12) = -1.21$ $p = .88$	$t(12) = -1.41$ $p = .91$	$t(12) = -.48$ $p = .68$	$t(12) = -.39$ $p = .64$	$t(12) = .072$ $p = .47$	$t(12) = .21$ $p = .42$	$t(12) = -.48$ $p = .68$	$t(12) = -.69$ $p = .75$	$t(12) = -1.20$ $p = .87$	$t(12) = -.49$ $p = .68$	$t(12) = -.33$ $p = .63$	$t(12) = -.70$ $p = .75$	$t(12) = -.24$ $p = .59$	$t(12) = -.73$ $p = .76$
V1-V4	$t(12) = .23$ $p = .41$	$t(12) = -.17$ $p = .57$	$t(12) = .87$ $p = .20$	$t(12) = .78$ $p = .22$	$t(12) = .93$ $p = .19$	$t(12) = -.18$ $p = .57$	$t(12) = 1.78$ $p = .05$ *	$t(12) = 1.38$ $p = .097$ †	$t(12) = 2.10$ $p = .03$ *	$t(12) = 1.73$ $p = .055$ †	$t(12) = .88$ $p = .20$	$t(12) = 1.3$ $p = .10$	$t(12) = 0$ $p = .5$	$t(12) = .97$ $p = .18$	$t(12) = .99$ $p = .17$
LOT	$t(12) = -1.8$ $p = .95$	$t(12) = 1.72$ $p = .06$ †	$t(12) = .71$ $p = .24$	$t(12) = .78$ $p = .23$	$t(12) = .14$ $p = .44$	$t(12) = -.05$ $p = .52$	$t(12) = .40$ $p = .35$	$t(12) = 1.99$ $p = .04$ *	$t(12) = .20$ $p = .42$	$t(12) = -.07$ $p = .53$	$t(12) = -.37$ $p = .64$	$t(12) = -.04$ $p = .52$	$t(12) = -.67$ $p = .74$	$t(12) = .58$ $p = .29$	$t(12) = .17$ $p = .43$
VOT	$t(12) = -1.15$ $p = .86$	$t(12) = -.76$ $p = .77$	$t(12) = -1.6$ $p = .93$	$t(12) = -.68$ $p = .75$	$t(12) = -.08$ $p = .53$	$t(12) = .21$ $p = .42$	$t(12) = .41$ $p = .35$	$t(12) = -.50$ $p = .69$	$t(12) = -.58$ $p = .71$	$t(12) = -.24$ $p = .59$	$t(12) = .36$ $p = .36$	$t(12) = 1.11$ $p = .14$	$t(12) = -.12$ $p = .55$	$t(12) = -.23$ $p = .59$	$t(12) = -.13$ $p = .55$
Posterior Color	$t(12) = 2.95$ $p = .006$ **	$t(12) = 1.67$ $p = .06$ †	$t(12) = .77$ $p = .22$	$t(12) = .68$ $p = .25$	$t(12) = -.19$ $p = .57$	$t(12) = 1.00$ $p = .17$	$t(12) = 1.39$ $p = .095$ †	$t(12) = .59$ $p = .29$	$t(12) = .54$ $p = .30$	$t(12) = -.23$ $p = .59$	$t(12) = .26$ $p = .40$	$t(12) = .27$ $p = .40$	$t(12) = .37$ $p = .36$	$t(12) = .35$ $p = .37$	$t(12) = .20$ $p = .42$
Central Color	$t(12) = -.88$ $p = .80$	$t(12) = .52$ $p = .31$	$t(12) = .67$ $p = .26$	$t(12) = .17$ $p = .43$	$t(12) = .28$ $p = .39$	$t(12) = -.14$ $p = .91$	$t(12) = .84$ $p = .21$	$t(12) = .47$ $p = .32$	$t(12) = .43$ $p = .34$	$t(12) = -.18$ $p = .57$	$t(12) = 2.30$ $p = .02$ *	$t(12) = 1.31$ $p = .11$	$t(12) = -.33$ $p = .63$	$t(12) = .75$ $p = .23$	$t(12) = .60$ $p = .28$

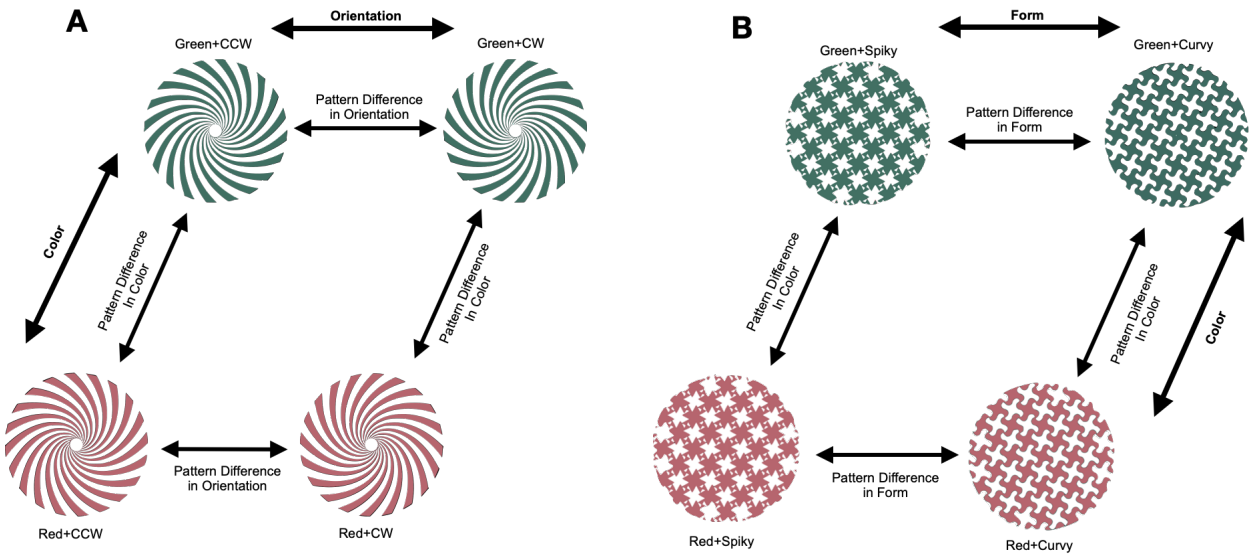


Figure 2.6. Logic of the pattern difference MVPA analysis for **A,** the color and orientation spiral stimuli in Experiment 1, and **B,** the color and curvature tessellation stimuli in Experiment 2. In this analysis, we examined which ROIs might code features in a manner that depends on the value of the other feature. From each ROI, we extracted and z-normalized the patterns associated with pairs of conditions matched on one feature but varying on the other, and took the difference between these patterns (e.g., GreenCCW - RedCCW). We did the same for the other value of the constant feature (e.g., GreenCW - RedCW). We then used SVM to determine whether these difference patterns were distinguishable from each other. This was done both possible ways — discriminating pattern differences in form across colors, and distinguishing pattern differences in color across the two values of each form feature — and the decoding accuracies were averaged. of interactive color and form coding in a brain region.

decoding drop. In other words, our cross-decoding drop analysis may underestimate the presence of interactive color and form coding in a brain region.

To remedy this, we performed a novel *pattern difference MVPA* analysis to specifically focus on the interactive effects that may be present in the response patterns in a brain region. Specifically, we extracted two *difference vectors,* each between two stimuli that differed on the same feature dimension (e.g., one difference vector could be RedCW minus GreenCW, while the other could be RedCCW minus GreenCCW). We then tested whether these two difference vectors could be discriminated using SVM. We did this separately for both color and form and

then averaged the results (see Figure 2.6 for a detailed illustration of this approach). If the encoding of one feature is completely independent and orthogonal to values of the other feature, then chance-level decoding is expected; by contrast, if the encoding of one feature changes based on the other feature, then above chance-level decoding is expected. This analysis essentially examines whether there is any interactive color and form coding in an ROI, with the SVM classification step serving to aggregate small interaction effects across voxels.

To test sensitively and exhaustively for the presence of interactive coding using this analysis method, for each ROI we performed the analysis separately for the top 100, 200, 300, 400, or 500 most active voxels. Figure 2.5 depicts the results of this analysis for the top 300 most active voxels (since this was the main voxel set used throughout this study); Tables 2.2 and 2.3 show the results for all voxel sets. In Experiment 1 (spirals), we found above-chance pattern difference decoding in a number of voxel sets from V1, V2, and the macro-ROI composed of V1-V4 (one-sample, one-tailed t-tests; no correction for multiple comparisons). In Experiment 2 (tessellations), we found above-chance pattern difference decoding in at least one voxel set from V1, V2, the V1-V4 macro-ROI, and LOT.

Double Conjunction Decoding

As another way to test for the presence of interactive coding of color and form, in an independent set of data, following Seymour et al. (2010), we examined which ROIs are able to discriminate between two pairs of stimuli, where each pair has the same set of four individual features, but conjoined in different ways. Specifically, we trained a classifier to discriminate between two kinds of blocks, each consisting of alternating pairs of stimuli with different form and color features, such that the same set of four features is present in each kind of block, but

combined in different ways (e.g., one kind of block alternated between RedCW spirals and GreenCCW spirals, and the other alternated between RedCCW and GreenCW spirals). If a region encodes these features in a completely additive, orthogonal manner, such that tuning to a feature does not depend on the value of the other feature, then patterns of activity in this region should not be able to distinguish these two kinds of block; by contrast, if there is any interactive coding of features, such that some voxels are sensitive to particular *pairings* of color and form features, then an SVM classifier should be able to distinguish these two kinds of blocks.

As in the decoding drop and pattern difference analyses, we performed this analysis separately on the top 100, 200, 300, 400, and 500 most active voxels from each ROI. Figure 2.5 shows the results of this analysis for the set of 300 voxels, and Tables 2.2 and 2.3 show the results for all voxel sets (one-sample, one-tailed t-test, no correction for multiple comparisons applied). In Experiment 1 (spirals), we found above-chance double conjunction decoding for V1, V2, V3, V4, and the V1-V4 macro ROI, with some variation in the robustness of the effect across ROIs; for example, the effect was significant in V2 for four of the five voxel sets, but only significant in V4 for one voxel set. VOT also exhibited a trend in the set of 100 voxels. By contrast, in Experiment 2 (tessellations), there was only a trend for one voxel set in V1, and significant decoding for one voxel set in V2 and the central color region.

In order to compare our results more directly with those of Seymour et al. (2010), we also re-ran the analysis with two changes to the pipeline to match the analysis of Seymour et al. First, we included all voxels falling under $p < .01$ in a task versus rest contrast, instead of using the top 300 voxels in such a contrast. Second, instead of z-normalizing the beta values going into the analysis across voxels within each trial, we normalized the beta values of each *voxel* across all its trials. When we used the $p < .01$ activation threshold for voxel selection, we found no significant

conjunction decoding in any individual ROI, or in the V1-V4 macro-ROI, with either within-voxel normalization (*ts* < 1.07, *ps* > .15) or across-voxel normalization (*ts* < 1.17, *ps* > .13), with the exception of a trend in V1 (*t(12)* = 1.56, *p* = .07). When we selected the most active 300 voxels (as we primarily used in our study), but used the within-voxel normalization method used by Seymour et al. (2010), we found significant conjunction coding in V2, V3, and the V1-V4 macro-ROI (*ts* > 2.62, *ps* < .02), along with a trend in V1 (*t(12)* = 1.59, *p* = .07) but no significant or trending decoding in V4 (*t(11)* = -.14, *p* = .55). All in all, then, we replicate their finding of conjunction coding for V1, V2, and V3 when we apply the normalization method their study used, but not in V4.

Discussion

Using fMRI pattern decoding and examining color and orientation representation in Experiment 1 and color and curvature coding in Experiment 2, the present study provides a comprehensive and updated documentation of the coding of color and form information across the ventral visual processing pathway in the human brain.

Broadly, we found that color and form information is nearly always anatomically commingled in the human ventral visual pathway. This includes early visual areas V1 to V4, and higher ventral visual regions defined based on their univariate selectivity for color or shape, including the posterior color region, LOT and VOT; this is especially striking in the case of LOT, since it is nowhere in the anatomical vicinity of the color-sensitive regions. The only exception to this pattern is the central color region which showed significant color decoding, but no form decoding, in both experiments, making it unique among the regions we examined. We were unable to reliably localize the anterior color region in every participant here due to its

location near the MRI signal dropout zone (at a rate similar to Lafer-Sousa et al., 2016). Overall, across the human ventral visual processing pathway, we found a largely distributed representation of color and form features, even in higher visual regions defined by their univariate selectivity for one feature or the other.

That said, coding preference for either feature, quantified using MVPA, varied across regions, and depended on the specific form feature tested. V1 and V2 were most sensitive to orientation changes, and less so to either curvature and color changes, thus showing a preference for orientation over curvature and color. V3 showed higher sensitivity to either form feature than to color. VOT and V4, which greatly overlapped, showed equally strong sensitivity to color and curvature changes, but a decreased sensitivity to orientation changes. The latter could potentially be due to the mirror symmetry of the clockwise and counterclockwise spirals used, since some evidence suggests that responses in VOT may be invariant to mirror-symmetric transformations (Dilks et al., 2011). The overlap of V4 and VOT with the color regions partially, but not entirely, drove color decoding in these regions: removing the color region overlap significantly decreased color decoding in these regions, but it remained above chance. Interestingly, removing the color region overlap also resulted in VOT showing a preference for curvature over orientation and color, consistent with this region's univariate selectivity for complex object shapes. LOT showed roughly equal sensitivity to color and orientation changes, but far greater sensitivity to curvature changes. LOT thus showed a preference to curvature over color and orientation, consistent with its univariate selectivity for complex object shapes. Finally, the posterior color region showed greater sensitivity to color than orientation, but an equal sensitivity to color and curvature, while the central color region showed a greater sensitivity to color than to either form feature. Thus, despite an overall distributed representation of color and form features, even early visual areas

show a feature preference, and in higher visual regions, their feature preferences are largely consistent between univariate and multivariate measures. Overall these results show that color and form features are represented in the human brain in a biased distributed manner.

Decoding for each feature depends on the amount of variation we introduced within each feature. For example, by reducing or increasing the difference between the two colors we examined, we could change the magnitude of color decoding and shift the relative encoding strength of color and form in a brain region. Because similarity within a feature changes across brain regions (e.g., two similar colors in one region may become dissimilar in another region), it would not have been possible to equate color and form variations for all the brain regions examined. Thus we have chosen what we believe to be reasonably large variations within each feature, including choosing two spirals with opposite directions, two tessellation stimuli with either all straight or all curved contours (thereby greatly varying an important midlevel form feature, curvature), and two hues that are maximally distinctive. These feature variations allow us to make a reasonable evaluation of the relative coding strength of color and form in each brain region, and more importantly, how the feature coding bias may change across visual regions. Although it could be argued that perhaps a wider array of colors and form features could have been sampled, by using a small number of stimuli chosen to greatly differ with respect to a chosen dimension (hue, orientation, curvature), we were able to maximize our power, giving us more confidence that any null results were not due to an inadequate number of trials. Furthermore, the logic of the double-conjunction design we used in one of the analyses requires two pairs of stimuli that differ with respect to two features.

We note that the color and form information encoded in different regions may play different roles in visual information processing. For example, only feature information in some

regions may be directly available to conscious perception, whereas the same information in other regions could be put to other uses. This can be seen in achromatopsia patients who can perceive isoluminant, color-defined shapes (e.g., a red square on a green background), even if they cannot report the colors that define the shape (Victor et al., 1989; Heywood et al., 1991; Barbur et al., 1994; Heywood et al., 1998). Perhaps this could be one functional role of color information in form-processing regions such as LOT. Differences in how the encoded feature information is utilized by the visual system may explain the human lesion results and visual search results that on the surface support a modular view of feature processing in the human brain, even though the underlying neural representation may follow a biased distributed organization as shown in this study.

To understand how color and form may be represented together in regions that code for both features, we performed several additional analyses. Broadly, these analyses examined the extent to which color and form are encoded in an *orthogonal* manner (with coding for each feature unaffected by the value of the other feature), an *interactive* manner (where coding for each feature depends on the value of the other feature), or some mixture of these motifs. In order to exhaustively test for the presence of interactive tuning and examine whether the results depend upon the set of voxels examined in each ROI, we performed each of these analyses on the 100, 200, 300, 400 and 500 most active voxels in each region.

Using a cross-decoding approach, we found most regions encode color and form information in a manner that is tolerant to changes in the other feature, indicating some independence in representation between these two features in each region. At the circuit level, such independence could be achieved by either intermingled but specialized neurons tuned to each feature, or by neurons tuned to both features but responding in a linearly additive manner.

However, several regions exhibited a drop in cross-decoding, suggesting an additional interactive component of their tuning. This was found in early visual cortex, where several regions showed a significant cross-decoding drop, especially for the spiral stimuli in Experiment 1 that varied based on their orientation. The effect was fairly consistent even when different numbers of voxels were included in the analysis. Some regions also exhibited a cross-decoding drop in Experiment 2, though these effects were not as robust across changes in the number of voxels included.

As a further test for interactive coding of color and form, we devised a novel analysis method, *pattern difference MVPA*, that sensitively tests for the presence of multivariate interaction effects in voxel populations. While related to the cross-decoding method for testing for the tolerance of feature representations relative to each other; it is conceptually distinct: testing for a cross-decoding drop examines whether the representations for two values of the target feature remain on the same sides of a classification boundary when the value of another feature changes, but such a method may overlook small interaction effects that leave the representations on the correct side of the boundary. By contrast, pattern difference MVPA focuses on the presence of *any* interactive effects and tests whether the pattern differences based on one feature are *exactly* preserved when another feature changes. Using this method, for the spiral stimuli in Experiment 1 that varied in orientation, we found evidence for interactive color-form coding in early visual cortex, and the effect was present across changes in the number of voxels included in the analysis. Meanwhile, for the tessellation stimuli in Experiment 2 that varied in curvature, only scattered evidence for interactive coding in early visual cortex and LOT was found for some of the voxel sets.

As a final test of interactive coding, in a separate data set, using the double conjunction methodology developed from Seymour et al. (2010), we also found evidence for interactive coding of color and orientation in early visual cortex across changes in voxel numbers, largely replicating their results. Again, evidence for interactive coding of color and curvature was scarce and was only found in a few voxel sets.

Thus, across three different analysis techniques and two independent datasets, we found evidence for interactive coding for color and orientation in early visual cortex, which remained fairly consistent even when varying numbers of voxels were included in the analyses. On the other hand, evidence for interactive coding for color and curvature was scarce and we could not find an effect for a brain region with one voxel set that was significant across all three analysis techniques. Overall, these results show that replicable evidence exists for interactive coding of color and form in early visual cortex and for simple form features, but less so in higher-level visual regions or for more complex form features, where color and form appear to be encoded more orthogonally. It should be noted that even in early visual cortex we obtained much stronger decoding results for single features than for feature conjunctions and that cross-decoding accuracy was above chance. This suggests that, despite the presence of interactive color and orientation coding in early visual areas, color and form representations still exhibit a high degree of independence in all regions examined; put another way, color and form representation in these regions consists mostly of “main effects”, with any interaction effects being small and mostly confined to simple form features in early visual regions, with somewhat weaker evidence for interactive coding of color with more complex form features.

Treisman and colleagues have famously argued that independently coded features can be conjoined via their shared location (Treisman & Gelade, 1980). One proposed neural mechanism

for achieving this has been long-range synchronized firings between neurons corresponding to different features of the same object at the same spatial location (Singer, 1999), with the posterior parietal cortex (PPC) serving a critical role in mediating this process (Robertson, 2003) as damage to PPC can result in feature binding deficits (Cohen & Rafal, 1991; Friedman-Hill et al., 1995). However, it is unclear how such a code would be generated and read out, and the wiring patterns and temporal firing precision of neurons between brain regions may be insufficient to implement this code (Shadlen & Movshon, 1999). Nevertheless, binding through a shared location via a neural mechanism other than synchrony is still possible. Every region we examined was either defined through retinotopic mapping or plausibly overlaps with a region that exhibits retinotopy (e.g., the posterior color patches overlap with V4, and the central color patches potentially overlap with retinotopic regions VO1 and/or VO2; see Brewer et al., 2005; Larsson & Heeger, 2006; Wandell & Winawer, 2011). The co-existence of color and form representation within most ROIs we examined, together with the co-existence of a detailed spatial map, could facilitate a binding by location mechanism at the local level without evoking long-range couplings between brain regions through neural synchrony, thereby serving as a potential binding mechanism (see also Di Lollo, 2012). Our results do not directly bear on the role of parietal cortex in binding: past accounts posit that it plays a purely spatial role in linking different features (e.g., Cohen & Rafal, 1991; Friedman-Hill et al. 1995), whereas more recent accounts emphasize its role in encoding and maintaining task-relevant visual information (Bettencourt & Xu, 2016; Vaziri-Pashkam & Xu, 2017; Xu, 2017; Xu; 2018a; Xu, 2018b). Our findings are consistent with an account in which the commingling of color and form information on spatially organized ventral stream cortical maps at least implicitly defines the binding of features, but parietal cortex may still be necessary to explicitly extract these bindings for

conscious perception and task-relevant processing. At minimum, the present study charts the anatomical layout and coding scheme of the ventral stream feature representations over which any putative parietal mechanism involved in feature binding might operate.

One potential limitation of this study was that the stimuli were non-naturalistic and arguably "texture-like". It is possible that this fact may have contributed to several of the null results, such as the failure to find form coding in the central color patches (which Chang et al., 2017 found in the macaque), and the limited scope of conjunction coding that the study identified. However, one key advantage of the stimuli used was that it allowed the whole central visual field to be equally stimulated, increasing the odds of identifying conjunction decoding anywhere in the central visual field. Moreover, past work has found that object ensembles containing repeated shapes activate high-level object shape regions just like single objects, supporting the use of such stimuli to drive these regions (Cant & Xu, 2012). Although stimuli were not scaled for eccentricity, this would only account for the null interactive coding findings if this coding motif only occurs over specific spatial scales, which would imply that it plays a rather specific rather than general role in visual processing.

One confound in comparing MVPA decoding across different experiments is that decoding accuracy can be affected not just by the strength of the underlying neural tuning, but also by factors like different analysis parameters, different levels of noise, and differences in data quality. For the present two experiments, however, the analysis pipelines were completely identical, removing analysis-related confounds. Furthermore, color decoding was statistically indistinguishable between experiments for every ROI, providing a common metric that suggests that levels of noise and data quality did not substantially differ, lending validity to the between-experiment comparisons. One important confound in fMRI decoding approaches is that two

experimental conditions can be discriminated by a linear classifier purely on the basis of differences in their noise covariance across voxels, even if their pattern centroids are the same (Hebart & Baker, 2018). Since our pattern difference analysis was novel, we therefore performed a control analysis in which we subtracted the mean pattern centroid of the training data within each condition (that is, within each of the two sets of difference vectors being compared) to equate the pattern centroids between conditions while maintaining differences in covariance structure, and then fed these transformed patterns into a support vector machine. As a test case, we examined the macro-ROI consisting of the most active 300 voxels across V1-V4 in Experiment 1 (spirals), since this is where we found the most robust evidence for interactive color-form coding. Performing this transformation of this data caused decoding to drop to chance (mean decoding accuracy 50.09%; $t(11) = .21$, $p = .42$ for one-sample, one-tailed t-test comparing against chance), suggesting that this confound does not account for the results of the pattern difference analysis.

As an experimental method, fMRI depends on the heterogeneity of neuronal tuning across voxels at the probed spatial resolution. Our results thus should be understood within the limitations of this method, like all other fMRI studies. That said, the spatial scale measured by fMRI often reasonably tracks the documented spatial heterogeneity of neuronal feature tuning in several of the ROIs that were examined. For example, V1 orientation columns are organized at a scale visible to fMRI and plausibly contribute to fMRI MVPA decoding (Yacoub et al., 2008; Pratte et al., 2016), V2 neurons are organized into "stripe" patterns, approximately 1-2m wide, with different kinds of stripes exhibiting different feature tuning (Ts'o et al., 2001), and monkey IT neurons are often organized into clusters .5mm in diameter containing neurons with similar tuning (Wang et al., 1996; Tsunoda et al., 2001). As such, organization at the mesoscale visible

to fMRI is not arbitrary or meaningless, but well-suited to capture the spatial tuning heterogeneity across neurons in many cases. This has enabled the representations visible to fMRI to be linked to the underlying neural computations, with fMRI decoding strength from human ventral and dorsal visual regions being tightly correlated with behavioral performance. For example, color decoding in V4, but not V1, reflected perceptual color space (Brouwer & Heeger, 2009), orientation decoding in early visual areas during the delay period of a visual working memory task tracked behavioral change detection performance (Bettencourt & Xu, 2016), and both object exemplar decoding and object category decoding in ventral and dorsal regions reflected perceived object similarity as measured by behavioral visual search and similarity judgement tasks (Mur et al., 2013; Charest et al., 2014; Cohen et al., 2017; Xu & Vaziri-Pashkam, 2019). Thus, the mesoscale neuronal organization visible to fMRI can be used to probe the underlying neural computations. That being said, methods like neurophysiological recordings are needed to verify the present set of findings.

In the present study, we found significant decoding of color and form much more reliably and broadly than we found evidence of interactive coding for these features, raising the question of what underlying patterns of neuronal tuning may account for these results. As discussed above, our method depends upon neural tuning being heterogeneously clustered across voxels in a way that can be detected by the spatial resolution of fMRI. The null results for interactive coding that we find are therefore consistent with a scenario in which neurons exhibiting interactive color/form tuning exist in higher-level ventral regions, but are not clustered in a sufficiently heterogeneous manner across voxels to be visible to fMRI MVPA. However, even if this is the case, it is interesting that such heterogeneity would be present for form- and color-coding neurons in higher-level ventral regions so as to enable decoding of individual features,

and present for conjunction-coding neurons in early visual cortex so as to enable conjunction decoding in *these* regions, but absent for conjunction-coding neurons in higher-level ventral regions. At the very least, if these neuronal populations do exist, we can conclude that they are distributed very differently from the other neuronal populations involved in color and form coding in the ventral visual cortex. It is also possible such neuronal populations simply do not exist, thereby avoiding the potential combinatorial explosion involved in having dedicated neurons for encoding the combination of every form and every color.

To conclude, our comprehensive approach illuminates the overall architecture of color and form processing in the human brain. Color and form information was not anatomically segregated into distinct anatomical regions defined by their univariate sensitivity to either feature, but instead was generally co-localized in the same brain regions in a biased distributed manner throughout the ventral visual processing pathway, with decoding from color and shape regions largely consistent with their univariate preferences. This challenges a strictly modular view of color and form processing. Convergent evidence from several analyses suggests that the joint coding of color and form within a region tends to be additive, with an additional interactive component present in a subset of cases, most reliably for the joint coding of color with simple form features in early visual cortex. Thus, the predominant relationship between color and form processing in the human ventral visual hierarchy appears to be one of anatomical coexistence but mostly representational independence.

Materials and Methods

Participants

Experiment 1 included 12 healthy, right-handed adults (7 females, between 25 and 34 years old, average age 30.6 years old) with normal color vision and normal or corrected to normal visual acuity. Experiment 2 included 13 healthy adults (7 females, between 25 and 34 years old, average age 28.7 years old). Four participants partook in both experiments. Participants were members of the Harvard community with prior scanning experience. All participants gave informed consent prior to the experiments and received payment. The experiments were approved by the Committee on the Use of Human Subjects at Harvard University.

Stimuli

Experiment 1: Colored spirals

Stimulus design and experimental design for Experiment 1 were largely adapted from Seymour et al. (2010), with identical stimuli and tasks but some differences in the number and timing of the blocks. Participants viewed colored spiral stimuli that varied by color—red or green—and orientation—clockwise (CW) or counterclockwise (CCW) —resulting in four different kinds of spirals (Figure 2.2A). Spirals were presented on a black background.

The spirals used were *logarithmic spirals*, defined by the formula $r=ae^{b\theta}$, which have the property that the angle between the radius of the spiral and an arm of the spiral at any point is fixed, in this case at 45 degrees. This property ensures that there is a constant relationship between the location of an edge of a spiral arm in visual space and the radial component of its angle, as would not be the case if oriented gratings were used (for example, a horizontal oriented

grating would have a maximal radial component along the horizontal midline, and minimal radial component along the vertical midline). This constraint accounts for the known *radial bias* in early visual cortex, in which radial orientations are preferentially represented in early visual topographic maps (e.g., zones of cortex corresponding to the top of the visual field have an over-representation of vertically oriented angles), ensuring that successful decoding of orientation could not simply be due to activation of different sub-regions of topographic maps (Sasaki et al., 2006; Mannion et al., 2009; Seymour et al., 2010). Stimuli were generated by first drawing 40 spiral lines at evenly spaced angles from the origin according to the above formula and filling in alternating regions of the spiral with the stimulus color and the background color, black, resulting in 20 spiral arms. The spiral subtended a circular region covering 9.7 degrees of visual angle, with an internal aperture in the middle, within which a white fixation dot was displayed. As mentioned earlier, the spiral arms could be oriented either clockwise or counterclockwise. Additionally, depending on which of the spiral arms were colored and which were black, each spiral could be presented in one of two phases.

The exact spiral colors used in the experiment were generated using the following procedure. To generate initially isoluminant shades of red and green, each participant performed a flicker-adjustment procedure inside the scanner (Kaiser, 1991), in which a flickering checkerboard with the two colors being adjusted flashed at 30 hz, and participants adjusted the colors until the flickering sensation was minimal. Specifically, the two colors had RGB values of the form red-hue = [178, 178 - X, 89] and green-hue = [0, X, 89], where participants adjusted the "X" parameter until isoluminance was achieved. This procedure guarantees that the two colors are isoluminant and sum to neutral gray, thereby equally stimulating all chromatic channels.

Participants performed ten trials of this procedure, and the average "X" value was used to produce the initial colors. However, since this procedure might theoretically have some associated imprecision, each color was presented at either +/-10% of its initially calibrated luminance value on any given run of the experiment, where the number of high-luminance and low-luminance runs was balanced across the red and green colors. This manipulation ensures that any residual between-hue luminance differences will be far smaller than the within-hue luminance differences, reducing the likelihood that luminance, rather than hue, could drive MVPA classification during analysis. The luminance adjustment procedures were identical to those of Seymour et al. (2010), with the minor difference that their study varied the luminance settings of a given color within a run, whereas we varied it between runs.

Experiment 2: Colored tessellation patterns

For this experiment, we constructed two different tessellation stimuli, consisting either of a curvy or a spiky pattern within a circular aperture (Figure 2.2A). These stimuli were deliberately designed so as not to resemble any real-world entities, and we decided upon a curvy versus spiky contrast because curvature is a salient mid-level visual feature, in contrast with orientation, which can be considered a lower-level visual feature (Gallant et al., 1993; Srihasam et al., 2014; Yue et al., 2014). The "phase" of the tessellation stimuli could also vary, based on whether a given region of the stimulus was currently colored or black. Exactly the same procedure as Experiment 1 was used to calibrate the colors of the two stimuli, and the stimuli subtended the same visual angle (9.7°) as in Experiment 1.

Procedure

Experiment 1: Colored Spirals

Participants viewed 12s blocks of the stimuli and had to detect a 30% luminance increment or decrement using a button press (index finger for increase, middle finger for decrease). On any given block, two 500ms luminance changes were presented, one in the first half and one in the second half of the block, and never in the first or last two stimuli of the block. The number and timing of the increments and decrements within the blocks was balanced across the whole experiment, and across all stimulus conditions described below. There were 9s fixation blocks between the stimulus blocks and at the end of the run, with a 12s fixation block at the beginning of the run. This allowed us to better separate fMRI responses from adjacent blocks (note that Seymour et al., 2010 included no fixation blocks between stimulus blocks in their design).

The experiment included two kinds of runs (Figure 2.2B). In the *single-conjunction runs*, only a single kind of spiral (RedCW, RedCCW, GreenCW, or GreenCCW) was presented for a given block, with its phase alternating once per second. This phase alternation ensures that all conditions were equated in their retinotopic footprint over the course of each block, removing this as a possible confound in form decoding. Since two starting phases were possible, each of the four spiral types could begin on either starting phase, resulting in 8 different block types for this condition. Each run contained one instance of each of the 8 types of block, totaling 180s per run. Participants completed 12 such runs, thus viewing a total of 24 blocks of each of the four spiral types over the whole session.

In the *double-conjunction runs*, there were two block conditions: a block could either alternate between RedCW and GreenCCW, or between RedCCW and GreenCW, with the phase

of each spiral type alternating at each presentation. Since each block condition could begin on either one of the two spirals in one of the two phases, there were therefore four different block types for each block condition. Due to how the spirals were constructed and how the stimuli alternated phase within each block type, every pixel took on values of red, green, and black an equal number of times both over the course of any given block and at any given timepoint for the four block types within each block condition. This ensured that pixel-level information could not drive decoding during the MVPA analysis. The stimulus timing, number of blocks, and task for these runs was otherwise identical to that of the single-conjunction runs. Participants completed 12 double-conjunction runs, and thus viewed each kind of double conjunction block 48 times. The single-conjunction runs and double-conjunction runs alternated in sets of three (e.g., three double-conjunction runs, then three single-conjunction runs), with the type of the initial run set counterbalanced across participants. Note that while Seymour et al. (2010) interleaved single-conjunction blocks and double-conjunction blocks within the same run, we separated them into different runs. This allowed us to form two completely independent datasets to more rigorously validate results showing interactive coding of color and form.

Experiment 2: Colored Tessellation Patterns

Exactly the same task and experimental design was used in Experiment 2 as in Experiment 1, with only the stimuli varying. Due to how the tessellation stimuli were constructed and the manner in which they alternate phase within the double conjunction blocks, they shared with the spirals the property that each pixel takes on values of red, green, and black an equal number of times over the course of the block, and at corresponding timepoints for the two block conditions across the four block types within each block condition.

Localizer Experiments

As regions of interest in both experiments, we included retinotopically-defined regions V1, V2, V3, and V4 in early visual cortex, and functionally-defined shape and color regions in occipitotemporal visual cortex.

To localize topographic visual field maps, we followed standard retinotopic mapping techniques (Sereno et al., 1995). A 72° polar angle wedge swept either clockwise or counterclockwise (alternating each run) across the entire screen, with a sweeping period of 36.4s and 10 cycles per run. The entire display subtended 23.4 × 17.6° of visual angle. The wedge contained a colored checkerboard pattern that flashed at 4 Hz. Participants were asked to detect a dimming in the polar angle wedge. Each participant completed 4–6 runs, each lasting 364s.

We localized two shape regions in lateral occipitotemporal (LOT) and ventral occipitotemporal (VOT) cortex, following the procedure described by Kourtzi & Kanwisher (2001), and subsequently used in several of our own lab's studies (Vaziri-Pashkam & Xu, 2017; Vaziri-Pashkam et al., 2019). LOT and VOT approximately correspond to the locations of LO and pFs (Malach et al., 1995; Grill-Spector et al.,1998; Kourtzi & Kanwisher, 2000) but extend further into the temporal cortex in order to include as many form-selective voxels as possible in occipitotemporal regions. Specifically, in a separate scanning session from the main experiment (usually the same one as the retinotopic mapping session), participants viewed black-and-white pictures of faces, places, common objects, arrays of four objects, phase-scrambled noise, and white noise in a block design paradigm, and responded with a button press whenever the stimulus underwent a slight spatial jitter, which occurred randomly twice per block. Each block contained 20 images from the same category, and each image was presented for 750ms each, followed by a 50ms blank display, totaling 16s per block, with four blocks per stimulus category.

Each run also contained a 12s fixation block at the beginning, and an 8s fixation block in the middle and end. Images subtended 9.5° of visual angle. Participants performed either two or three runs, each lasting 364s.

We also localized a series of color-sensitive regions in ventral temporal cortex, using a procedure similar to Lafer-Sousa et al. (2016). Two runs of a color localizer were presented during the main scan session, one at the middle and one at the end of the session. In these runs, participants viewed 16s blocks consisting of either colorful, highly saturated natural scene images selected from the online Places scene database (Zhou, Lapedriza, Khosla, Oliva, & Torralba, 2018) or greyscale versions of these images. Participants responded when an image jittered back and forth, which occurred twice per block. Images subtended 9.5° of visual angle. Each run contained 16 blocks, 8 for each of the two stimulus types, for a total run duration of 292s including an initial 20s fixation block, and an 8s fixation block in the middle and the end of the run.

MRI Methods

MRI data were collected using a Siemens PRISMA 3T scanner, with a 32-channel receiver array headcoil. Participants lay on their backs inside the scanner and viewed the back-projected display through an angled mirror mounted inside the headcoil. The display was projected using an LCD projector at a refresh rate of 60 Hz and a spatial resolution of 1280x1024. An Apple Macbook Pro laptop was used to create the stimuli and collect the motor responses. Stimuli were created using Matlab and Psychtoolbox (Brainard 1997).

A high-resolution T1-weighted structural image (1.0 x 1.0 x 1.3 mm) was obtained from each participant for surface reconstruction. All Blood-oxygen-level-dependent (BOLD) data

were collected via a T_2*-weighted echo-planar imaging (EPI) pulse sequence that employed multiband RF pulses and Simultaneous Multi-Slice (SMS) acquisition. For the two main experiments, including the color localizer runs, 69 axial slices tilted 25° towards coronal from the AC-PC line (2mm isotropic) were collected covering the whole brain (TR = 1.5s, TE = 30ms, flip angle = 75°, FOV = 208m, matrix = 104x104, SMS factor = 5). For the retinotopic mapping and LOC localizer sessions, 64 axial slices tilted 25° towards coronal from the AC-PC line (2.3mm isotropic) were collected covering the whole brain (TR = 0.65s, TE = 34.8ms, flip angle = 52°, matrix = 90x90, SMS factor = 8). Different slice prescriptions were used here for the different localizers to be consistent with the parameters used in our previous studies. Because the localizer data were projected into the volume view and then onto individual participants' flattened cortical surface, the exact slice prescriptions used had minimal impact on the final results.

Data Analysis

FMRI data were analyzed using FreeSurfer (surfer.nmr.mgh.harvard.edu), FsFast (Dale, Fischl, & Sereno, 1999) and in-house Python scripts. The exact same analysis pipeline was used for the two experiments, except that any analyses comparing clockwise versus counterclockwise spirals in Experiment 1 instead compared the spiky and curvy tessellation patterns in Experiment 2, due to the differing stimuli used. Preprocessing was performed using FsFast. All functional data was motion-corrected to the first image of the run of the experiment. Slice-timing correction was applied, but smoothing was not. A generalized linear model (GLM) with a boxcar function convolved with the canonical HRF was used to model the response of each trial, with the three motion parameters and a linear and quadratic trend used as covariates in the analysis. The first

eight TRs of each run (prior to the presentation of the first stimulus) were included as nuisance regressors to remove them from further analysis. A beta value reflecting the brain response was extracted for each trial block in each voxel. ROIs were defined on the cortical surface and then projected back to native functional space for further analysis.

ROI Definitions

Using independent localizers, we defined ROIs in early visual areas and in higher visual regions showing univariate selectivity to shapes or colors. Figure 2.1A depicts all ROIs for an example participant.

V1 to V4. Areas V1 through V4 were localized on each participant's cortical surface by manually tracing the borders of these visual maps activated by the vertical meridian of visual stimulation (identified by locating the phase reversals in the phase-encoded mapping), following the procedure outlined in Sereno et al. (1995).

LOT and VOT. Following the procedure described by Kourtzi & Kanwisher (2001), LOT and VOT were defined as the clusters of voxels in lateral and ventral occipitotemporal cortex, respectively, that respond more to photos of real-world objects than to phase-scrambled versions of the same objects ($p <$. 001 uncorrected). These regions correspond to the location of LO and pFs (Malach et al., 1995; Grill-Spector et al.,1998; Kourtzi & Kanwisher, 2000), but extend further into the temporal cortex in our effort to include as many object-selective voxels as possible in occipito-temporal regions.

Ventral Stream Color Regions. Following Lafer-Sousa et al. (2016), several color regions were identified in ventral temporal cortex as clusters of voxels responding more to colored images than to greyscale versions of the same images ($p <$ .001, uncorrected). Since participants

had varying numbers of such regions, we divided the regions in each hemisphere into anterior, central, and posterior color regions, following Lafer-Sousa et al. (2016). We were able to identify posterior and central color regions in every hemisphere of every participant in both experiments. In Experiment 1, we were able to localize the anterior color region in both hemispheres of 7/12 participants, one hemisphere of 3/12 participants, and neither hemisphere of 2/12 participants. In Experiment 2, we were able to localize the anterior color region in both hemispheres of 8/13 participants, one hemisphere of 3/13 participants, and neither hemisphere of 2/13 participants. The inconsistency in localizing this color region was possibly due to its location being close to the ear canals where large MRI susceptibility effects and signal dropoff could occur. We note that our rate of localizing this color region was similar to that of Lafer-Sousa et al. (2016), who reported that this region was found in both hemispheres of 6/13 participants, one hemisphere of 4/13 participants, and neither hemisphere of 3/13 participants. These anterior regions were generally relatively small (mean 49 voxels, std 46 voxels, min 4 voxels, max 163 voxels), precluding us from conducting meaningful decoding analyses in these regions. We thus omit them from further analysis.

V4 and VOT *with Color Regions Removed.* We observed that the color regions overlapped with areas V4 and VOT in some cases. To document the extent to which color and form decoding in V4 and VOT might be affected by the color regions within them, we also ran several of the analyses on versions of V4 and VOT with the color-sensitive regions removed.

ROI Overlap Analysis

As noted just previously, we observed that areas V4 (defined retinotopically), the posterior color region (defined using a color versus greyscale localizer), and area VOT (defined

using an object versus scrambled localizer) overlapped to some degree. To quantify this overlap, we computed the pairwise percent overlap between each of these ROIs, where percent overlap was defined as the percentage of the number of overlap voxels over the averaged number of voxels for the two ROIs as we did in a previous study (Cant & Xu, 2012; see also Kung et al., 2007).

Multivoxel Pattern Analysis

In order to equate the number of voxels used in each ROI, the top 300 most active voxels in a stimulus-versus-rest GLM contrast across all the runs were selected. In addition to the ROIs described above, we also constructed an ROI for each participant consisting of the 300 most active voxels from the entire V1-V4 sector defined by the union of V1-V4, in order to test more sensitively for potentially subtle effects in several analyses. For several of the analyses (noted in each section below that describes the analysis), we analyzed subsets of the 100, 200, 300, 400, and 500 most active voxels per ROI, to determine the extent to which the presence of an effect depended on the number of voxels selected. A beta value was extracted from each voxel of each ROI for every trial block. To remove response amplitude differences across stimulus conditions, trial blocks and ROIs, beta values were z-normalized across all the voxels for each trial block in each ROI. For each of the contrasts of interest (described below), these beta values were used to train and test a linear support vector machine (SVM) classifier (with regularization parameter c = 1), using leave-one-run-out cross-validation. T-tests were performed to compare the decoding accuracy of the various comparisons to chance (one-sample, one-tailed t-test; one-tailed was used because below-chance decoding is not conceptually meaningful). To account for the fact that four participants partook in both experiments with the other participants being different

between the two experiments, in cases where decoding was compared between pairs of conditions between the two experiments, a *partially-overlapping t-test* (Derrick et al., 2017) was performed. Likewise, to examine the influence of experiment, feature type, and their interaction on decoding in each region and between regions, a linear mixed effects analysis was performed (since this analysis, unlike the classical ANOVA, is able to explicitly account for subject-specific variance when only a subset of participants complete both experiments). Correction for multiple comparisons was applied using the Benjamini–Hochberg procedure with false discovery rate controlled at $q < 0.05$, with the details of this correction described for each analysis below (Benjamini and Hochberg, 1995). Specific details for each analysis were as follows.

Feature Decoding. To assess the extent to which regions carried information about single features, in the single-conjunction blocks we trained and tested the classifier on color (red vs. green) and form (CW vs. CCW spirals in Experiment 1 and curvy vs. spiky tessellations in Experiment 2), where *both* values of the other feature were fed into each bin of the classifier (e.g., for color decoding, RedCW and RedCCW versus GreenCW and GreenCCW). Each condition was compared to chance (one sample t-test, one-tailed), decoding for each feature was also compared between experiments (partially-overlapping t-test, two-tailed), and decoding for the two features was also compared within each experiment (within-subjects t-test, two-tailed). Correction for multiple comparisons was performed within the set of comparisons done for each ROI (i.e., four tests for feature decoding and two tests for comparing feature decoding between experiments). We additionally performed mixed-effects analyses in each ROI to examine how decoding accuracy changes based on feature (color and form), the form features used in the two experiments (orientation and curvature), and their interaction. To test for broad trends in feature coding across the visual hierarchy, we also averaged the decoding accuracy of ROIs showing

qualitatively similar response profiles via their proximity and their ordinal pattern of their feature decoding strengths over the two experiments, and the same analyses were performed for these sectors as were performed in the individual ROIs. Further linear mixed-effects analyses were used to verify that the decoding profiles in these sectors in fact varied from one another.

Additionally, to document whether there exist any hemispheric differences in color and form coding, within each experiment we ran a within-subjects t-test between the left and right hemisphere for both color and form coding. Since this analysis was exploratory, no corrections for multiple comparisons were performed.

Finally, to examine the extent to which feature decoding results for V4 and VOT are driven by their overlap with the color regions, we constructed ROIs consisting of V4 and VOT minus their overlap with the color regions. The same feature decoding analyses were run for these ROIs as for the other ROIs. Additionally, two-way mixed-effects analyses with ROI and experiment as factors were run to examine whether decoding for either color or form significantly decreased in either region when the color-sensitive patches were removed (analyses conducted separately for each feature).

Feature Cross-Decoding. To assess whether the two features were represented independently in each ROI (i.e., whether the representation of one feature was invariant to changes in the other feature) and whether there was any evidence of interactive feature coding, in the single conjunction blocks we performed cross-feature decoding in which we trained a classifier to discriminate two values of a relevant feature while the irrelevant feature was held at one value, and tested the classifier's performance on the relevant feature when the irrelevant feature changed to the other value (e.g., train an orientation classifier on RedCW vs. RedCCW, and test orientation decoding on GreenCW vs. GreenCCW, or vice versa, with the results from

the two directions averaged together). We did this for both features serving as the relevant feature. For comparison purposes, we also performed within-feature decoding, where we held the irrelevant feature constant between training and testing. This allowed us to compare the cross- and within-feature decoding using a matched number of trials. Decoding of each condition was compared to chance (one-sample t-test, one-tailed). Additionally, within-feature and cross-feature decoding were compared (within-subjects t-test, one-tailed; one-tailed was performed because only a decrease, not an increase, in performance from cross-decoding is interpretable) within each feature and experiment to determine whether coding for each feature is tolerant to changes in the other. Correction for multiple comparisons was performed within the set of comparisons done for each ROI (i.e., eight comparisons for comparing each condition to chance; four comparisons for comparing within-feature decoding to cross-decoding for each condition).

Since both kinds of cross-decoding drop — a drop in color decoding across form features, or a drop in form decoding across colors — are conceptually similar in that they both reflect a more interactive feature representation, a one-tailed t-test was performed within each experiment to take both effects into account to test for an overall main effect of lower decoding in the cross-feature versus within-feature decoding conditions (note that this is the same as assessing a main effect of decoding difference between the within-feature and cross-feature decoding conditions across the two types of features using an ANOVA test, but looking at this main effect in a particular direction). Since this comparison provides critical evidence regarding whether or not interactive color and form coding may exist in a brain region, to perform an exhaustive search, we ran this particular analysis separately for the top 100, 200, 300, 400, and 500 most active voxels in each ROI. Given that SVM is sensitive to both power and noise (such that including too few voxels may exclude some of the informative voxels and thus provide

insufficient power, whereas including too many voxels may add noise), testing the effect at a range of voxel sizes allowed us to assess the stability of any positive results obtained and how it may be affected by the number of voxels included in the analysis. For this comparative analysis, we report p-values uncorrected for multiple comparisons to make our results comparable to the conjunction decoding results reported by Seymour et al. (2010). Since correction for multiple comparisons was not mentioned in Seymour et al., we assume the results were uncorrected. Corrected p- values can be easily derived from the uncorrected p-values reported using the Benjamini-Hochberg procedure.

Pattern Difference MVPA. To probe for the presence of interactive color and form representation in an ROI, we ran a novel analysis to examine whether the encoding of one feature (form or color) depends on the value of the other feature. Specifically, we first took the difference between the z-normalized beta values associated with RedCW and RedCCW, and between GreenCW and GreenCCW (Figure 2.6). We then trained and tested an SVM (leave one run out cross-validation) on these difference vectors to examine whether the pattern differences associated with the two orientations change based on the color of the stimulus. We also performed the opposite analysis, comparing the beta value differences for the two different orientations (RedCW — GreenCW versus RedCCW — GreenCCW). The mean classification accuracies of these two directions of the analysis were then averaged. If the encoding of one feature is invariant to values of the other feature, SVM should discriminate these vectors at chance (50%); by contrast, if the encoding of one feature changes based on the other feature, the classification should be above chance. Effectively, this analysis examines whether the voxels in an ROI exhibit an interaction effect in their tuning for color and form, with the SVM classification step serving to aggregate small and potentially heterogeneous interaction effects

across voxels. The same analysis was performed for the tessellation stimuli in Experiment 2, replacing CW and CCW with the spiky and curvy stimulus conditions. One sample, one-tailed t-tests were performed for each ROI to determine if decoding of the pattern differences was above chance (one-tailed t-tests were used because below-chance decoding is not conceptually meaningful).

As in the cross-decoding drop analysis, we also ran this analysis separately on the top 100, 200, 300, 400, and 500 most active voxels in each ROI, so as to test exhaustively for the presence of interactive color-form coding in each ROI, and determine the extent to which the results depend upon the number of voxels selected. The results of these tests are reported without correction for multiple comparisons so that our results may be comparable to those of Seymour et al. (2010). Corrected p-values can be easily derived from the uncorrected p-values reported using the Benjamini-Hochberg procedure.

We note that the information captured by this analysis is distinct from the information conveyed by feature cross-decoding. Feature cross-decoding would succeed so long as the patterns being cross-decoded end up on the correct side of the SVM decision boundary, even if the differences between the respective patterns were distinct (i.e., if main effects in feature coding far exceeded any interaction effects). By contrast, this method provides a more direct and sensitive test regarding the existence of interactive coding in the representational space.

Double Conjunction Decoding. As another way of examining which regions may contain interactive coding of color and form, we trained and tested the classifier on the two kinds of double conjunction blocks in each experiment (e.g., RedCW/GreenCCW and RedCCW/GreenCW). These blocks contained color and form features alternating once per second. Due to the sluggishness of the hemodynamic response, the pattern of BOLD activity

present in each region would roughly constitute a superposition of the patterns associated with the two kinds of stimuli in each block. Since these two kinds of blocks both contained the two color and two form features used (e.g., red, green, clockwise, and counterclockwise), but differ in how they were conjoined, only regions encoding color and form in an interactive manner should be able to decode the two kinds of blocks from each other. The results of this analysis were compared against chance (50% decoding) using a one-sample, one-tailed t-test (one-tailed t-tests were used because below-chance decoding is not conceptually meaningful).

As in the cross-decoding drop and pattern difference analyses, we performed this analysis separately on the top 100, 200, 300, 400, and 500 most active voxels in each ROI; p-values are reported without correction for multiple comparisons so that our results may be comparable to those of Seymour et al. (2010). Corrected p-values can be easily derived from the uncorrected p-values reported using the Benjamini-Hochberg procedure.

Chapter 3: Joint Representation of Color and Form in Convolutional Neural Networks: A Stimulus-Rich Network Perspective

Abstract

To interact with real-world objects, any effective visual system must jointly code the unique features defining each object. Despite decades of neuroscience research, we still lack a firm grasp on how the primate brain binds visual features. Here we apply a novel network-based stimulus-rich representational similarity approach to study color and form binding in five convolutional neural networks (CNNs) with varying architecture, depth, and presence/absence of recurrent processing. All CNNs showed near-orthogonal color and form processing in early layers, but increasingly interactive feature coding in higher layers, with this effect being much stronger for networks trained for object classification than untrained networks. These results characterize for the first time how multiple basic visual features are coded together in CNNs. The approach developed here can be easily implemented to characterize whether a similar coding scheme may serve as a viable solution to the binding problem in the primate brain.

Author Summary

Visual experience consists of different features, like form and color; how might an intelligent visual system encode the combinations of these features that compose whole objects? This question has a long history as the "binding problem" in visual neuroscience, but with the recent development of artificial neural networks showing human-level object recognition performance, it arises anew; while successful, it is largely unknown how these systems internally process features to recognize objects. Here, we examine how these networks encode

combinations of color and form, and find a consistent coding principle across all networks we examined: while form and color are initially processed independently, as processing proceeds in these networks they are processed in an increasingly interactive manner.

Introduction

Natural visual experience comprises a juxtaposition of different visual features, such as an object's color, position, size, and form, with the form features including both simple form features such as local orientations and contours, and the complex form features including global shape and texture, which often define an object's identity. To recognize an object under different viewing conditions, our visual system must successively reformat and "untangle" the different features to make object identity information explicitly available to a linear readout process in a manner that is tolerant to variations in other features, an ability that has been hailed as the hallmark of primate high-level vision (DiCarlo & Cox, 2007; Hong et al., 2016).

Meanwhile, our interaction with the world often involves objects with uniquely defined features, such as grabbing the blue pen on the desk. How would an object representation that sheds all its identity-irrelevant features support our ability to interact with specific objects? One possibility is that different visual features are initially processed separately and are bound together via attention (i.e., Feature Integration Theory, Treisman & Gelade, 1980). Despite decades of neuroscience research, the coding mechanism for such a binding process remains unknown, with existing proposals facing various challenges. For example, Singer (1999) proposed that neurons coding for different features of the same object could engage in synchronous oscillations, serving as a binding signal, but it is unclear how such a signal would be generated and read out (Shadlen & Movshon, 1999). Alternatively, there might exist neurons

that encode particular feature conjunctions; however, this view collides with the problem of "combinatorial explosion": there are more possible feature conjunctions than neurons in the brain. Given that the tuning of neurons is affected by a diverse set of mechanisms, including feedforward activation, lateral inhibition, and feedback connections, a step towards understanding how neurons encode feature combinations would be to understand how each of these factors independently contribute to the conjunctive coding of visual features, but this is challenging due to the complexity of the primate visual system.

Recently, convolutional neural networks (CNNs) have achieved human-level object recognition performance (Kriegeskorte, 2015; Yamins & Dicarlo, 2016; Rajalingham, et al., 2018; Serre, 2019). Specifically, these CNNs have been trained to disregard identity-irrelevant object features to correctly identify objects across different viewing conditions, thereby forming transformation-tolerant visual object representations much like those in high-level primate vision. In both human fMRI and monkey neurophysiological studies, representations formed in lower and higher layers of the CNNs have been shown to track those of the primate lower and higher visual processing regions, respectively (Cadieu et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014; Güçlü & van Gerven, 2015, Cichy et al., 2016, and Eickenberg et al., 2017).

Although CNNs are fully computable and accessible, they are extremely complex models with thousands or even millions of free parameters. Consequently, despite their success in object recognition, the general operating principles at the algorithmic level (see Marr, 1980) that enable CNNs' success remain poorly understood (e.g., Kay, 2018). This includes understanding how different types of visual features are represented together in CNNs during the course of visual processing. Several studies have examined how individual features are encoded

in CNNs, with some finding that coding for object identity-irrelevant features increases in higher CNN layers (Hong et al., 2016). Additional approaches to understanding internal CNN representations (summarized in Rafegas et al., 2020, and Serre, 2019) include synthesizing images that maximally drive individual CNN units (e.g., Zeiler & Fergus, 2014), ablating sets of units and examining how this impairs network performance (e.g., Zhou et al., 2018), and using principal components analysis to visualize how different features are encoded in a given layer (e.g., Aubry & Russell, 2015). Of relevance to the present work, several studies have reported the color encoding characteristics of CNN units (Aubry & Russell, 2015; Flachot & Gegenfurtner, 2018; Rafegas & Vanrell, 2018; Rafegas et al., 2020). Of these four studies, Aubry and Russell (2015), Flachot and Gegenfurter (2018), and Rafegas et al. (2020) examine color coding, but do not examine how color is jointly coded with form. Rafegas and Vanrell (2018) examine this issue, but only do so for a small number of units in a single CNN. No study to date to our knowledge has examined how combinations of these features are jointly encoded across entire populations of CNN units.

Because CNNs are not trained to interact with specific objects but simply to produce the correct object labels at the end of its processing, it is possible that different features are initially encoded in an entangled, intermingled fashion, and are gradually separated, with object identity information gradually made more explicit and independent of other visual features over the course of processing (DiCarlo & Cox, 2007). Alternatively, CNN architecture and training for object recognition may automatically give rise to *interactive,* rather than independent, coding of object features in later stages of processing, rendering unnecessary a separate binding operation to encode the relationship between independent features. This could constitute a novel binding mechanism that has not been considered before in neuroscience research. Thus, studying how

CNNs jointly encode different object features during the course of visual information processing is not only timely in its own right, but also provides us with a unique opportunity to gain insight into the potential computational algorithm that a successful object recognition system may use to code different object features together. Moreover, the wiring of most CNNs is restricted to feedforward connections, making it tractable to isolate how various aspects of conjunctive tuning may arise from feed-forward processing alone. Equally importantly, given that the internal representations of CNNs are fully image computable and freely inspectable, CNNs provide ideal testing grounds for developing analysis methods to study feature coding across an entire processing hierarchy and with a large number of objects, and generating hypotheses that can be tested in biological visual systems. Finally, CNNs are trained for a well-defined task (typically object recognition), making it possible to examine how task demands shape the joint representation of multiple features.

In this study, we examined how an object's color and form may be coded together during visual processing in CNNs. We employ a network-based, stimulus-rich approach in which we characterize the joint representations of these two features not for a few pairs of objects at a few processing stages as is traditionally done in neuroscience and vision research, but rather, across a large number of objects and across the entire processing hierarchy of a CNN. We devise a new metric that captures the extent to which color and form are encoded in an interactive manner—that is, the extent to which the coding for one feature varies across values of the other. Broadly, our method involves examining whether the similarity structure of different colors varies across different object form features. With this approach, we found that coding for color and form becomes increasingly interactive throughout CNN processing. These results thus characterize, within a novel population-coding framework, how multiple visual features are encoded together

in CNNs. The approach developed here can be easily implemented to characterize whether the primate brain may use a similar coding scheme to solve the binding problem.

Results

In this study, we examined in detail how color and naturalistic object form features may be represented together in five CNNs trained for object recognition using ImageNet (Deng et al., 2009) images. These CNNs, chosen for their high object recognition performance, architectural diversity, and prevalence in the literature, included AlexNet (Krizhevsky, Sutskever, & Hinton, 2012), VGG19 (Simonyan & Zisserman, 2015), GoogLeNet (Szegedy et al., 2015), ResNet-50 (He, Zhang, Ren, & Sun, 2015), and CORNet-S (Kubilius et al., 2018). Specifically, AlexNet was included for its high object recognition performance, relative simplicity, and prevalence in the literature. VGG19, GoogLeNet and ResNet-50 were chosen based on their high object recognition performance and architectural diversity. Both AlexNet and VGG19 have a shallower network structure, whereas GoogLeNet and ResNet-50 have a deeper network structure. CORNet-S is a shallow recurrent CNN designed to approximate the structure of the primate ventral visual pathway, and exhibits high correlation with neural and behavioral metrics. This CNN has recently been argued to be the current best model of the primate ventral visual regions (Kar et al., 2019). These networks differ in both their architecture, and in some cases, their training regimes; specifically, for training data augmentation AlexNet, VGG19 and ResNet-50 use cropping, horizontal flips, and RGB adjustments (simulating changes in lighting), whereas CORNet-S only uses cropping and horizontal flips, and GoogLeNet only employs cropping. We sampled between 6 to 9 layers in each of these CNNs (Table 3.1).

Table 3.1. The five CNNs included in the present study and the layers sampled in each CNN.

Network (Layers Used/Total Layers)	Layers Used
AlexNet (8/25)	Conv1, Conv2, Conv3, Conv4, Con5, FC1, FC2, FC3
CORnet-S (7/42)	Conv1, Relu2, Relu5, Relu8, Relu11, AvgPool1, FC1
GoogLeNet (6/144)	Conv1, MaxPool2, MaxPool5, MaxPool11, AvgPool1, FC1
ResNet-50 (6/177)	Conv1, Relu4, Relu8, Relu14, AvgPool1, FC1
VGG19 (9/47)	Conv1, MaxPool1, MaxPool2, MaxPool3, MaxPool4, MaxPool5, FC1, FC2, FC3

We used representational similarity analysis (RSA, Kriegeskorte & Kievit, 2013) to characterize how color and form information is represented together in these networks. For most analyses, we studied a set of 50 objects (chosen from a set created by Brady et al., 2013), each colored in 12 colors that were calibrated to equate their mean luminance and saturation in the CIELUV color space before being converted back to RGB as inputs to each network (Figure 3.1a-b). Equating luminance was necessary to ensure that any results were truly driven by color-related processing rather than luminance contrast mechanisms, and equating saturation was important for equating the hue variability within each object. Two versions of each object were shown: a textured version, with internal object detail preserved, and a silhouette version with all non-white pixels set to a uniform color, thus comprising a global form contour without internal details (Figure 3.1b). Examining both the normally-textured objects and the silhouettes allowed us to make more precise inferences about how color and form information interact: while the textured objects differ both by their outline and texture, the silhouettes vary with respect to their outline alone, and so to the extent that any results also hold for the silhouettes, it would demonstrate that these results do not depend on texture differences. Additionally, removing

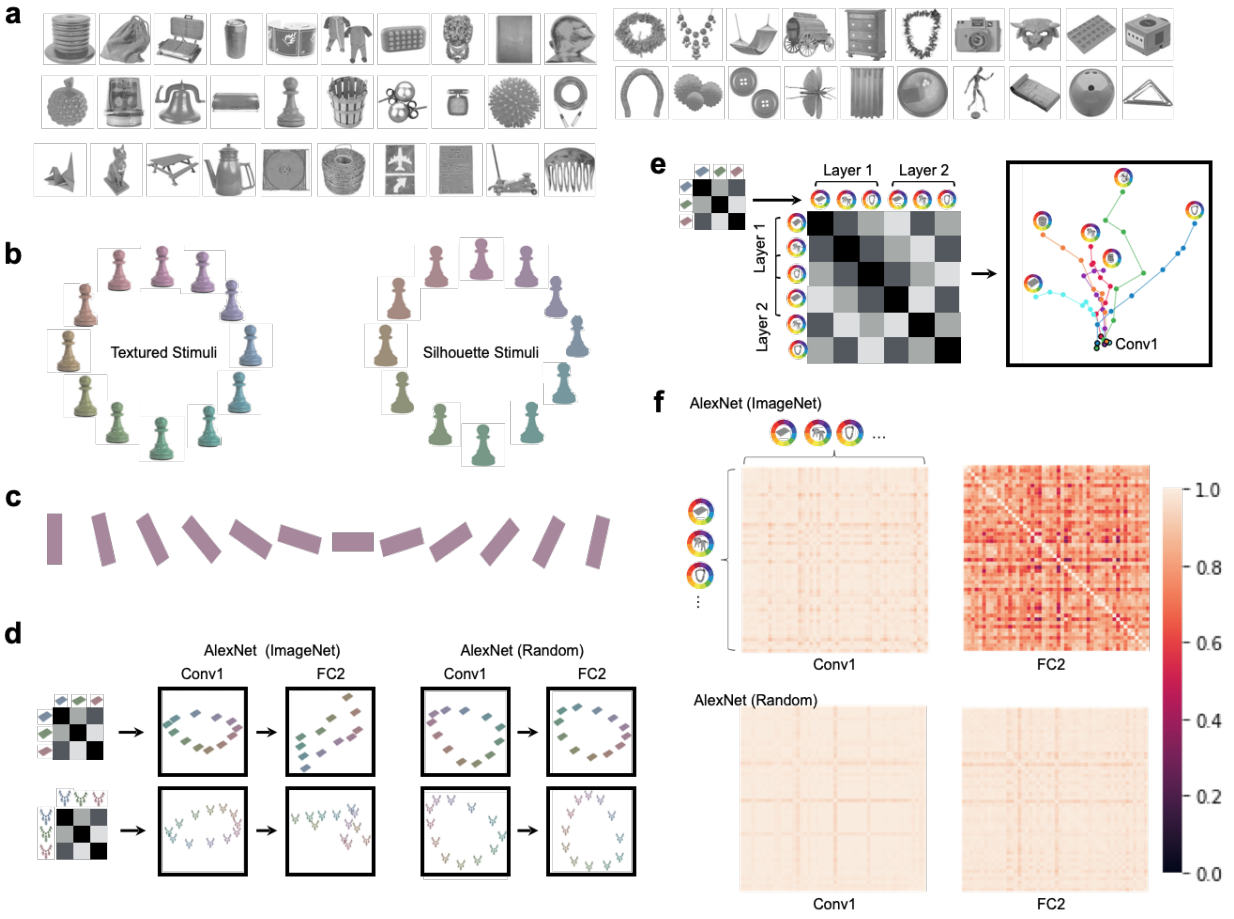


Figure 3.1. Stimuli used and example color space characterization using RSA and MDS. a. The 50 objects included in the main stimulus set, chosen from an initial set of 500 objects to maximize their mean pairwise pattern dissimilarity in AlexNet FC2. b. The 12 isoluminant and iso-saturated colors (based on the CIELUV color space) and the two versions of the object shapes used in the main analysis. Objects appeared either with their original textures preserved ("Textured Stimuli"), or as uniformly shaded silhouette stimuli ("Silhouette Stimuli"). c. The 12 oriented bar stimuli used in a control analysis. d. An illustrative color similarity matrix for a given object (left-most column) and actual MDS plots showing the representational structure of two example objects each in the 12 colors calibrated in CIELUV color space from Conv1 and FC2 of AlexNet trained with ImageNet (2nd and 3rd columns from left). Color space correlations were first obtained from each object in the 12 colors (3 colors were illustrated here) in a given layer to construct a color similarity matrix for that layer. The first two dimensions of this similarity matrix were then projected onto the 2D space using MDS. While the similarity spaces of these objects have a similar elliptical pattern at the beginning of the trained AlexNet, by the end of processing the color spaces of these objects are substantially different both from each other and from those at the beginning of the processing. By contrast, in a version of AlexNet with randomized weights (two right columns), the color spaces of both objects remain roughly similar at both the beginning and end of processing, as shown by the similar arrangement of the colors of each object. e. An illustrative color space similarity matrix (left) and an actual MDS plot showing the color spaces of six example objects over the course of

processing in AlexNet (right). Color spaces were computed separately for each of these objects in each sampled layer of AlexNet (illustrative color space depicted by the small matrix on the left), and the resulting color spaces (only three objects and two layers illustrated here) were correlated with one another to construct a color space similarity matrix (note this is a second order correlation matrix, different from the color similarity matrix illustrated in d). The first two dimensions of this similarity matrix were then projected onto the 2D space using MDS. Each dot in the MDS plot represents the color space of a given object at a given layer, where the distance between two dots reflects the similarity between two color spaces. Each trajectory traces the color space of a given object. The dot corresponding to the initial layer has a black outline, and the dot corresponding to the final layer is marked by a picture of the object for that trajectory. While the color spaces of different objects are initially very similar, by the end of processing they have substantially diverged. f. Actual representational similarity matrices showing the pairwise color space similarity for each pair of objects in both the first and penultimate layers of AlexNet, in both its trained and untrained variants.

texture details removes much category-diagnostic information from the stimuli, clarifying whether category-specific effects may be driving any results.

Broadly, our analyses examined how color and form information are jointly encoded over the course of processing in CNNs. To examine this question within a population coding framework, we define a color space as the similarity profile of a set of colors for a given stimulus in a given representational context (e.g., for a given object and CNN layer). To the extent that color spaces are invariant across objects, color and form information can be said to be independent or orthogonal; to the extent that they differ, they can be said to be interactive or entangled. Several analyses were performed on these color spaces. The specific analyses involved (1) comparing the color spaces across objects within each layer, to determine whether color and form are encoded independently versus interactively in that layer, and examining how this is affected by variations in stimuli, analysis parameters, and network training regime; (2) comparing the color space for each object across layers, to determine how color information for each object is transformed over the course of processing; (3) examining whether color space differences across objects are preserved across layers, and (4) whether the form similarity of two

objects predicts their color space similarity. While we use MDS plots for visualization purposes and to build intuition, all conclusions are grounded in subsequent statistical analyses. To our knowledge, these analyses provide the first in-depth and comprehensive network-based description of how colors and form are coded together in CNNs.

Visualizing Color Space Representation Across Objects and CNN Layers

As our primary analysis, we applied RSA to examine the extent to which coding for color varies across objects, and the extent to which the magnitude of this variability changes across CNN layers. As an initial exploratory analysis, we visualized how the color spaces of two example objects may differ at the beginning and end of processing in AlexNet, examining both a trained and untrained version of the network (Figure 3.1d). Specifically, we extracted the activation patterns for the 12 colors of these two objects (textured versions) from the first and the penultimate layers of AlexNet (Conv1 and FC2). Within each layer, we performed all pairwise Pearson correlation among the 12 patterns to create a representational similarity matrix (RSM). After subtracting each correlation from 1 to convert the similarities to dissimilarities, we used multidimensional scaling (MDS) to visualize the resulting representational dissimilarity space projected onto 2D space, with a closer distance between a pair of colored objects indicating more similar representations (Figure 3.1d). In the version of AlexNet trained on ImageNet, color appeared to be coded similarly at the beginning of the network for these two objects, as reflected by the similar elliptical shape of the color spaces; by the end of processing, however, the color spaces of these two objects seem to differ substantially, both from each other and from their color spaces at the beginning of processing. By contrast, in an untrained version of AlexNet the two objects had similar, ellipse-shaped color spaces at both the beginning and end of processing.

To generalize from these two objects and examine how the color space for different objects might diverge over layers, we visualized the evolution of the color spaces of six example objects over the course of processing in the trained version of AlexNet (Figure 3.1e). To do this, for each object (textured versions) and for each sampled layer of AlexNet, we first constructed a "color space" RSM by performing all pairwise Pearson correlations of the patterns associated with the 12 different colors of that object. We vectorized the off-diagonal values of this RSM to create a "color space" vector. Next, we performed all pairwise correlations of these "color space" vectors across objects and layers to form a "color space similarity" RSM that quantifies how similarly color is coded in different objects and layers; finally, each value was subtracted from 1 to convert the matrix to a dissimilarity matrix. We then used 2D MDS to visualize the resulting representational similarity space, where each dot represents the color space of a given object at a given layer (i.e., the similarity profile among the different colors of the object at that layer), and the distance between two dots reflects how similarly color is coded in those two spaces (Figure 3.1e). In these objects, color appeared to be initially coded in a very similar manner (as reflected by the dense clustering of the bold-outlined dots representing the different color spaces in the initial layers of processing), but color coding increasingly diverged as processing proceeded in the network (as reflected by the separation of the dots at the end of processing, indicated by the object icons next to the dots). In other words, over the course of processing, color coding for each object both increasingly differed from the color coding for other objects, and from the color coding of that object at the beginning of processing. Figure 3.1f shows the full RSM of the color space similarities among every pair of objects for both a trained and an untrained version of Alexnet. As an initial observation, the color spaces of all objects appear to be very similar to each other early in processing for both trained and untrained AlexNet, but by the end of

processing the color spaces associated with the different objects have diverged for the trained, but not the untrained, version of AlexNet.

Quantifying Color Space Differences Across Objects within a CNN Layer

Next, we quantified and further explored the divergence in object color spaces over the course of processing that we qualitatively observed in the previous section. To quantify the color space divergence among different objects within a layer and over the course of processing, we computed the averaged pairwise color space vector correlations for the 50 objects in each layer of each CNN, and for both the textured and silhouette stimuli. This between-object color space similarity measure is a correlation-of-correlations metric: it captures the extent to which color coding varies across form features, with the first order correlation capturing the structure of the color space for each individual object and the second-order correlation capturing how the color space structure varies across the different objects. The lower this second-order correlation, the more it can be inferred that color and form are encoded in an interactive versus an independent manner.

As an additional measure to ensure that there were no lurking differences in the "baseline" between-object color space correlations for different models and training regimes, for the trained version of AlexNet, the untrained version of AlexNet, and the trained version of GoogLeNet, we performed a resampling procedure in which we randomly shuffled the labels of the colors within each object (independently for each object) and computed the mean pairwise color space correlations among all pairs of objects; this was done one hundred times, and the resulting color space correlations were averaged. The resulting mean correlations were near zero in every single layer of all three tested models. We thus found no evidence that spurious

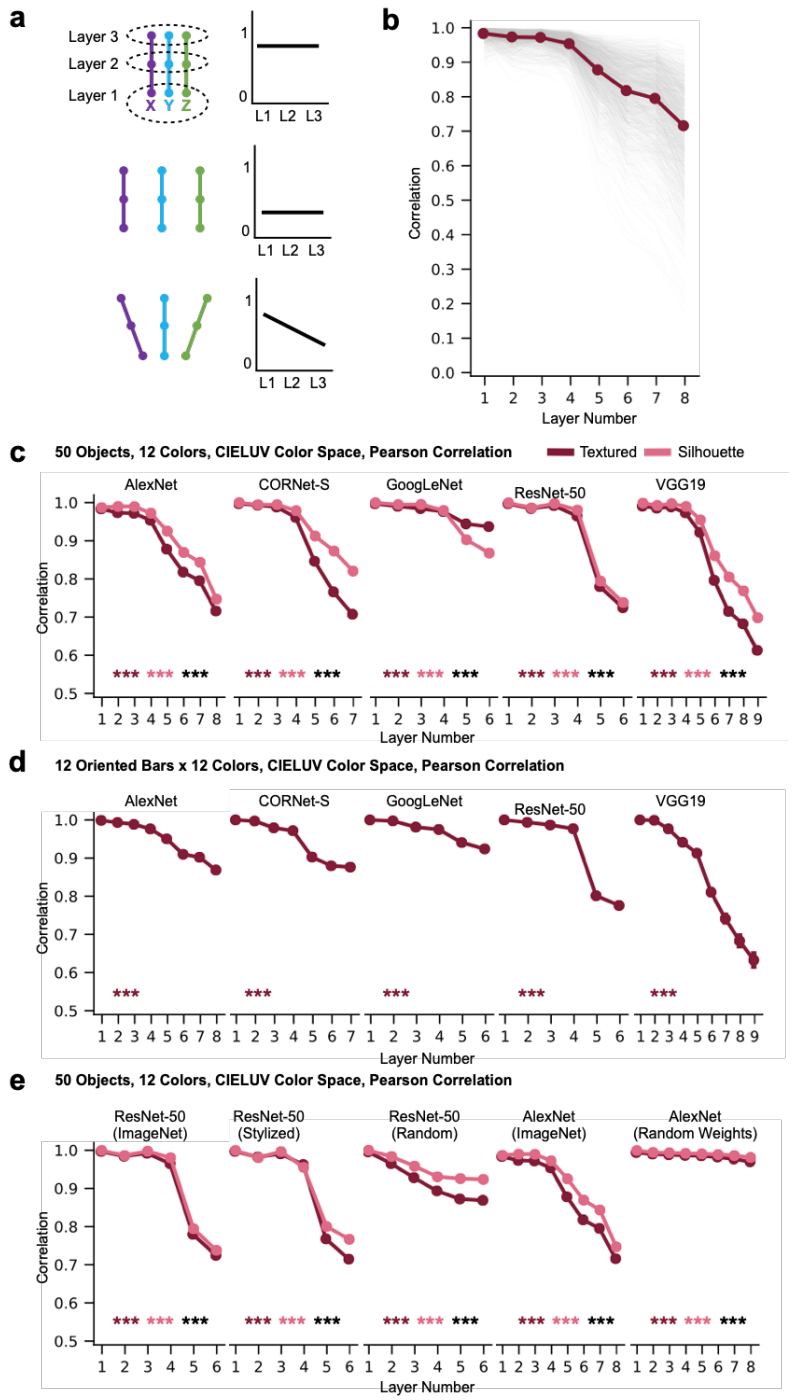


Figure 3.2. Color space representation across objects within a CNN layer. a. A schematic illustration of three possible scenarios. In each scenario, the left figure illustrates the color space transformation of three objects in three hypothetical CNN layers, with each colored dot depicting a color space structure of an object at a CNN layer and each trajectory depicting an object. The right figure in each scenario illustrates how the mean pairwise correlation of all object color spaces for a given layer changes across layers. In the first scenario, the color spaces of the three objects remain relatively similar within each layer throughout processing. In the second scenario,

they are dissimilar within each layer throughout processing. In the third scenario, they are similar in the first layer, but become dissimilar in later layers. b. The pairwise color space similarity for every pair of textured objects in each layer of Alexnet for the full set of 50 objects in the 12 colors calibrated in CIELUV color space. The color space structure of each object is measured with Pearson correlation (this also applies to c-e below). Each thin grey line is the color space similarity for a single pair of objects, with the bold line showing the mean across all pairs. c. The mean pairwise color space similarity for each sampled layer of each of the five CNNs, for the full set of 50 objects in the 12 colors calibrated in CIELUV color space. Results are shown for both the textured objects (in maroon) and the silhouette objects (in pink). Linear regression was used to measure the downward trend of the correlation values across layers for each object pair. The mean of the resulting slopes (one slope per object pair) were tested against zero for each of the two versions of the objects, and the difference between the two sets of slopes was also tested (with significance levels marked by maroon and pink asterisks, respectively, for each of the two versions against zero, and by black asterisks for the differences between the two versions). In all cases, mean pairwise color space similarity decreases over the course of processing, with this decline being greater for the textured than for the silhouette objects, with the exception of GoogLeNet. d. Mean color space similarity across the 12 oriented bar stimuli in the 12 colors calibrated in CIELUV color space. Even in these minimally simple form stimuli, mean pairwise color space similarity decreases over the course of processing. e. Mean color space similarity across the 50 objects in the 12 colors calibrated in CIELUV color space in CNNs with different training regimes. Comparisons are made among ResNet-50 trained with the original ImageNet images, trained with stylized ImageNet images, and with 100 untrained random-weight initializations of the network. Comparisons are also made between AlexNet trained with the original ImageNet images, and with 100 untrained random-weight initializations of the network. Averaged results are shown from the 100 untrained versions of each network. The untrained networks exhibit a much smaller decline in their mean pairwise colorspace correlation across objects than the trained networks. *** $p < .001$.

differences among the different model architectures, layer types, or training regimes could have driven our main results in the absence of genuine color space correlations among objects.

We further quantified, using regression analysis, whether this mean between-object color space similarity significantly declines over the course of processing, and whether this decline varies significantly between the textured and silhouette versions of the stimuli (see Methods for analysis details). For the entire set of 50 objects, several patterns of results, shown in Figure 3.2a, could be possible: the color spaces of different objects might be highly similar in every layer,

they might be highly dissimilar in every layer, or they might begin similar to each other, but diverge over the course of processing, similar to the pattern of the six example objects in Figure 3.1e.

Figure 3.2b shows the color space correlation between every pair of objects, as well as the mean of these correlations, for the textured version of the objects run through AlexNet. We observe a reliable decrease in the color space correlation of different objects as processing proceeds. To generalize these findings across networks and quantify statistical significance, Figure 3.2c depicts the mean between-object color space correlation within every layer of every network as well as how the mean changes across layers, for both the textured and silhouette versions of the objects. In all CNNs, and in both stimulus conditions (textured and silhouette), the mean color space correlations were high in lower layers but then significantly decreased from mid to high CNN layers.

To quantify whether there was indeed a decrease in the correlation values over layers, we tested the presence of a negative slope. We note that our conclusions do not require a purely monotonic progression of the correlation values. Nevertheless, the presence of a negative slope would still signal any presence of a negative linear trend. We found an overall significantly negative slope across all the layers for each of the networks tested (see the asterisks marking the significance level of the slope at the lower part of each plot). Coding of color thus remained relatively similar across objects in lower layers but then became increasingly different from mid to high CNN layers, reflecting what is depicted in Figure 3.1e and Figure 3.2a bottom panel. Since this increase in interactive tuning occurred even for the silhouette stimuli, it did not depend on the internal texture features of the stimuli, and can occur with respect to global form features alone. That being said, for most of the networks, the textured stimuli did exhibit a greater drop in

their color space similarity over the course of processing than the silhouette stimuli, with the exception of GoogLeNet, suggesting the existence of greater interactive coding for texture features above and beyond global form features alone.

To ensure that these results did not arise due to the particular similarity metric used in constructing the RSMs (i.e., Pearson correlation), we repeated the same analysis using Euclidean Distance to measure the similarity between the different colors of each object; Pearson correlation was still used to measure the second-order similarity *between* the color spaces (Supplementary Figure 3.1a). Additionally, to ensure that our results did not depend on the human-based color space we used (i.e., CIELUV color space), we repeated the same analysis, but using stimuli where the saturation and luminance were equated according to a new color space we constructed, "synthetic HSV", which was not based on human psychophysical measurements, but was constructed to parametrize the concepts of luminance and saturation for CNNs (Supplementary Figure 3.1b; see also Methods). As with the stimuli tuned in CIELUV color space, images were converted back to RGB space prior to running them through the networks. For both manipulations, the results remained qualitatively similar: color space correlations among different objects began high in early layers, and dropped significantly in later layers in all conditions.

To what extent do these results depend on this specific stimulus set? For example, it is possible these results have arisen due to the objects subtending different areas of space. Some stimuli, like the top hat, covered large areas, while other stimuli, like the necklace, covered small areas (Figure 3.1a). This could have activated different numbers of units in CNN layers and affected how colors are coded for each object. Additionally, it is theoretically possible that results could have been driven by differences in object category rather than form per se, perhaps

due to the fact that all networks were trained to recognize objects. To investigate this possibility and to examine whether our results hold even for minimally simple stimuli, we repeated the same analysis on 12 oriented bars presented in the same 12 colors equated in CIELUV color space as used earlier (Figure 3.1c). We found the same overall result: as processing proceeds in each network, color coding increasingly differs across different form features (Figure 3.2d). Thus our results hold for objects equated in their spatial coverage. Moreover, results obtained from complex natural objects can be generalized to simple form stimuli.

Overall, across all conditions we examined, we found a consistent pattern: all CNNs showed near-orthogonal color and form processing in early layers, but increasingly interactive feature coding in higher layers.

The Effect of Training on CNN Color Space Representation

The ImageNet images used to train the CNNs studied so far contain real-world objects with natural color-form covariation (e.g., bananas are yellow). Could the interactive color and form coding observed so far in CNNs be driven by such covariation in the training images? To address this question, we compared results from the version of ResNet-50 trained on the original ImageNet images and the version trained on stylized ImageNet images in which the original texture and color of every single image was replaced by the style of a randomly chosen painting, removing the real-world color-form covariation in the natural objects (Geirhos et al., 2019). Interestingly, the version of ResNet-50 trained on stylized images still exhibited a significant, steep decrease in their color space correlation over the course of processing (Figure 3.2e), almost as steep as that observed in the version of ResNet-50 trained on the original ImageNet images. This suggests that the interactive color and form coding observed in CNNs does not rely on the

presence of consistent color and form pairing naturally occurring in the training images. That said, the slopes were slightly, though significantly, steeper for the version of ResNet-50 trained on the original than the styled images of ImageNet in both stimulus conditions ($ts > 2.89$, $ps < .004$). Thus training on naturalistic images does appear to increase the degree of interactive color and form coding in this CNN, although the effect is fairly small.

To understand the extent to which the effects we observe may arise due to the intrinsic architecture of the networks versus being a result of object classification training, we examined 100 random-weight initializations of AlexNet and 100 random-weight initializations of ResNet-50, and compared the results with those from the ImageNet image-trained AlexNet and ResNet-50 and the stylized ImageNet image-trained ResNet-50. Results for the 100 random initializations of each network were computed independently, and then averaged together at the final stage. As shown in Figure 3.2e, while the random networks still exhibited a significant decline in their mean pairwise colorspace correlation across objects, this decline was small, and much smaller than in the corresponding trained version of each network (matched-pairs t-tests; $ps < .001$).

Overall, these results show that the intrinsic CNN architecture is not sufficient to give rise to the large interactive color and form coding observed so far. Training on the object classification task, even with inconsistent pairings of color and form in the object stimuli, appear to play a more significant role in creating such coding.

Transformation of Color Space Representations Across CNN Layers and Architectures

Instead of focusing on color space differences across objects within a layer, here we took an orthogonal approach and tested how the color space of a given object may change across layers by correlating the color space vector of a given object between layers. Color coding for a given object may remain similar across layers, resulting in closely clustered color spaces across layers (Figure 3.3a, right), or it may transform substantially over the course of processing, leading to dispersed color spaces (Figure 3.3a, left). To quantify such transformations, for each object, we correlated the color space vector from each layer with the color space vector from the first and penultimate processing layers of the network (Figure 3.3b). We then used regression to examine whether the correlation significantly decreases with an increasing number of intervening layers from the reference (first or penultimate) layer. This was done both for the networks trained on ImageNet, 100 random initializations of AlexNet, and 100 random initializations of ResNet-50 (with correlation values averaged across random networks). Across the main set of 50 objects in 12 colors calibrated in CIELUV color space, in all cases, there was a significant and steady decrease in correlation with the target layer with increasing number of intervening layers (see the asterisks marking the significance level of the slope at the lower part of each plot). Even for the first few layers, although color space correlations within a layer were fairly high among the different objects (see Figure 3.2c), the color spaces of each object still differed *across* layers. For the random networks, while there was a statistically significant decrease in correlation with an increasing number of intervening layers, this effect was much smaller than in the trained networks, and in the case of the random version of AlexNet nearly nonexistent, suggesting that the color space transformations we observe are in large part induced by training the networks to recognize objects. Overall, color space was successively and

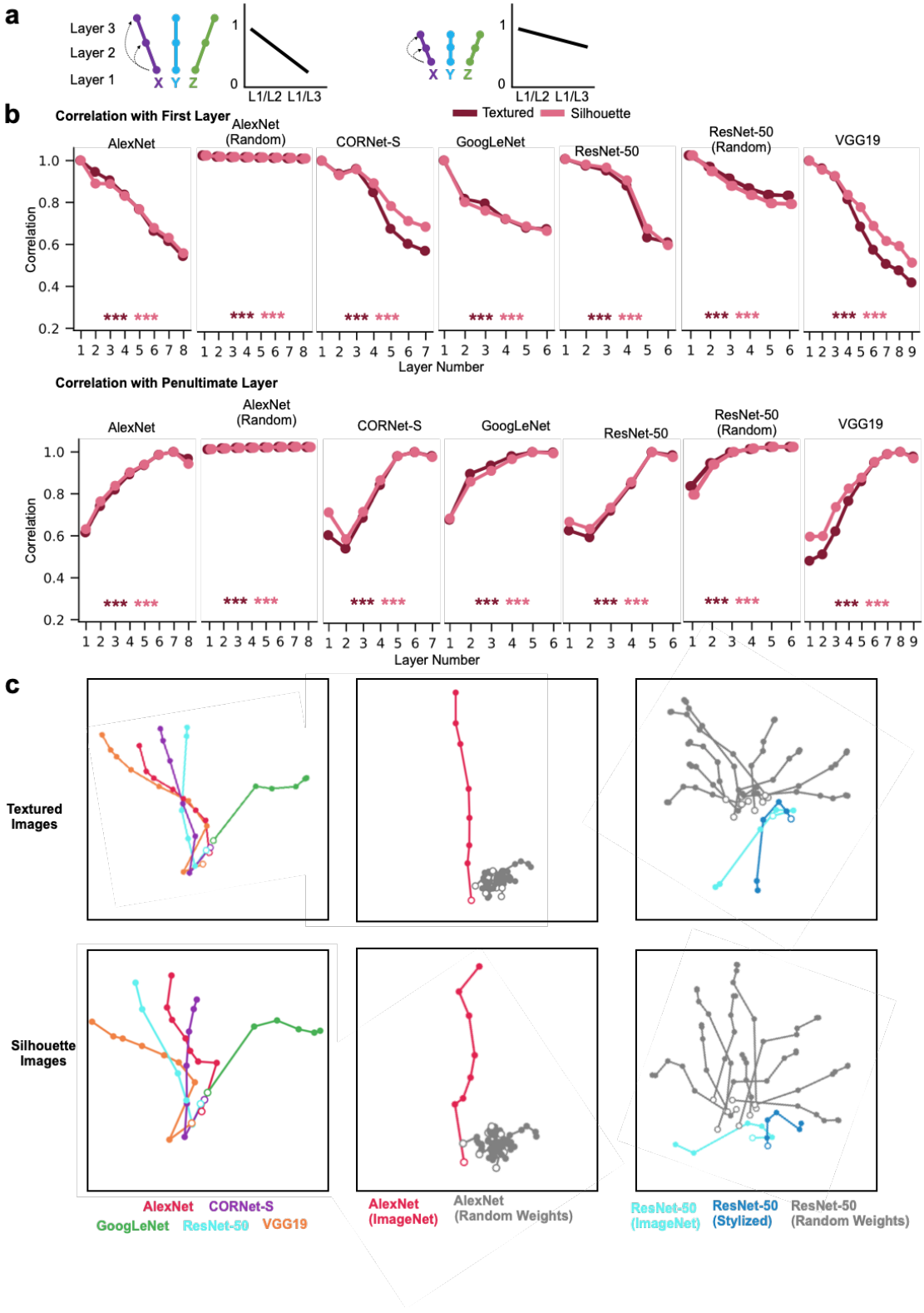


Figure 3.3. Color space representation across different CNN layers and different CNN architectures. a. A schematic illustration of two possible scenarios of color space correlation

across layers within a CNN, using the same notations as those in Figure 3.2a. In this analysis, within each object, the color space structure from the first layer is correlated with each of the other layers, as shown on the left of each scenario. The averaged correlation over all objects for each layer is plotted in a line graph on the right of each scenario. In the first scenario, the color space structure within each object differs substantially across processing, resulting in a large decrease in correlation across layers. In the second scenario, the color space structure for each object remains relatively stable across processing, resulting in a relatively small decrease in correlation across layers. b. Mean within-object across-layer color space correlations for each network for the full set of 50 objects in the 12 colors calibrated in CIELUV color space for both the textured and silhouette versions of the objects. Top row shows the correlations with the first layer of each network, bottom row shows correlations with the penultimate layer of each network. Results for random networks are averaged across 100 random initializations of the network. Linear regression was used to measure the downward or upward trend of the correlations for each object across layers. The resulting slopes were tested against zero for each of the two versions of the objects (with significance levels marked by maroon and pink asterisks, respectively). For all trained networks, the color space similarity within an object significantly decreases with more intervening layers, and correlations between early and late layers were fairly modest. This trend is far smaller for the versions of the networks with random weights c. MDS plots depicting color space correlation across different CNN layers and architectures. This was done by constructing a color space similarity matrix for each object, including its color space correlation across all sampled layers of all CNNs. The resulting correlation matrix was then averaged across objects and visualized using MDS. This was performed for the five trained networks (left column), AlexNet trained with ImageNet images and with 10 random-weight initializations (middle column), ResNet-50 trained with ImageNet images, trained with stylized ImageNet images, and with 10 random-weight initializations (right column), and for both the textured (top row) and silhouette images (bottom row). The hollow dots denote the first layer of each network. In the 5 trained CNNs (left column), color spaces are almost identical in the first layer and then gradually fan out during the course of processing, though in a similar overall direction. Color spaces in the untrained networks, however, differ substantially from the trained ones (middle and right columns). ** $p < .01$, *** $p < .001$.

substantially transformed over the course of processing for the trained networks, with the correlations between the color spaces at the beginning and end of processing being quite modest.

To understand how the color space of an object may be encoded differently among the different CNNs, for each of the 50 objects, we also correlated its color space vector across all CNNs and layers. We then visualized the resulting correlations, averaged over all objects, using MDS plots (after subtracting each correlation from 1 to convert similarities to dissimilarities). As

shown in Figure 3.3c (leftmost column), across the 5 CNNs, for both the textured and silhouette objects, while the color spaces evolved substantially from their initial state over the course of processing (consistent with the quantitative analyses above), the color representations nonetheless evolved in a relatively similar way across networks, with the representations being almost identical in the first layer for all 5 CNNs and then gradually fanning out in a roughly similar direction during the course of processing, with GoogLeNet showing a greater divergence compared to the other CNNs. Of particular note, the penultimate layer of each network appears to encode color in a more similar manner compared to the penultimate layers of other networks than it does to the first layer of that network, suggesting that each network substantially transforms its color representations over the course of processing, but in a similar manner to other networks (with the exception of GoogLeNet). Supplementary Figure 3.2 shows the exact between-network correlation values for both the initial and penultimate layers of each network, which in conjunction with Figure 3.3b corroborates the qualitative observations from the MDS plots regarding the color space differences among different models and layers.

To further understand how training on object classification may affect the color space of an object, we repeated the above analysis and correlated the color space vector of the same object across the layers of various instantiations of the same network with different training regimes: (1) across AlexNet trained with ImageNet images and 10 random initializations of AlexNet; and (2) across ResNet-50 trained with the original ImageNet images, ResNet-50 trained with stylized ImageNet images, and 10 random initializations of ResNet-50 (Figure 3.3c, middle and right columns respectively, and Supplementary Figures 3.3 and 3.4). In both cases, while color was initially encoded in a similar manner between the random and trained versions of the networks, over the course of processing, the color spaces for objects in the trained

networks substantially diverged from those in the random networks. Interestingly, while the color spaces of the different random initializations of AlexNet tended to cluster together over the course of processing and did not diverge as the trained network did (consistent with the possibility that that the observed transformation in the color space was induced by training the network), those of the different random initializations of ResNet-50 diverged substantially but in different directions as those of the trained networks. On average, in the penultimate layers, color spaces for the two trained versions of ResNet-50 tended to be more correlated with each other than they are with the random initialization of the network; this was more so for the textured than the silhouette version of the objects. Additionally, the version of ResNet-50 trained on ImageNet was approximately as similar in its penultimate layer to the version of ResNet-50 trained on stylized ImageNet as it was to the other networks trained on ImageNet (Supplemental Figures 3.2 and 3.4), suggesting that both a network's architecture and its training can contribute to the transformation of color information in a network.

Overall, these results demonstrate that, within a given network, color representations for each object transform dramatically over the course of processing. Across the different networks, color spaces evolved in a roughly similar manner across the trained networks. This transformation of color space was not a mere byproduct of a network's intrinsic architecture, as the color spaces of untrained networks evolved substantially differently from the trained networks.

Transformation of Color Space Similarity Across CNN Layers

To understand how color space similarity may change across objects over different CNN layers, instead of testing the color space of a single object, here we asked: is the profile of color

space similarities among different objects preserved across processing (Figure 3.4a top), or does the pattern of color space similarities among objects change throughout processing (Figure 3.4a bottom)? To test this, for the main set of 50 objects in the 12 colors calibrated in CIELUV color space, for each CNN, we took either the first or penultimate layer as our target layer and first generated its color space similarity RSM by performing all pairwise correlations of color space vectors between objects. We then vectorized the off-diagonal elements of this RSM to form a "color space similarity" vector and correlated this vector with those from all other layers of the CNN. This was done for both the networks trained on ImageNet, and on 100 random initializations of AlexNet and ResNet-50 with untrained weights. We found that for all networks, correlations decreased as we moved away from the reference layer, with the first and last layers being only moderately correlated (Figure 3.4b). This transformation occurred in the untrained networks as well, though to a lesser extent than in the trained networks. Thus, if two objects had a highly similar color space at the beginning of a CNN, they did not necessarily have a highly similar color space at the end of the CNN. Patterns of color space differences among objects appear to dramatically change throughout the course of processing.

The Effect of Object Form Similarity on Color Space Similarity

It is possible that color space similarity covaries with object form similarity, such that a small change in form features leads to a small change in the associated representational geometry for color. However, given that each feature can vary relatively independently of the other feature, it is also possible that color coding does not closely follow form coding. To arbitrate between these two possibilities, and better understand what factors might be driving the divergence in color space across objects, for the main set of 50 objects in the 12 colors calibrated in CIELUV

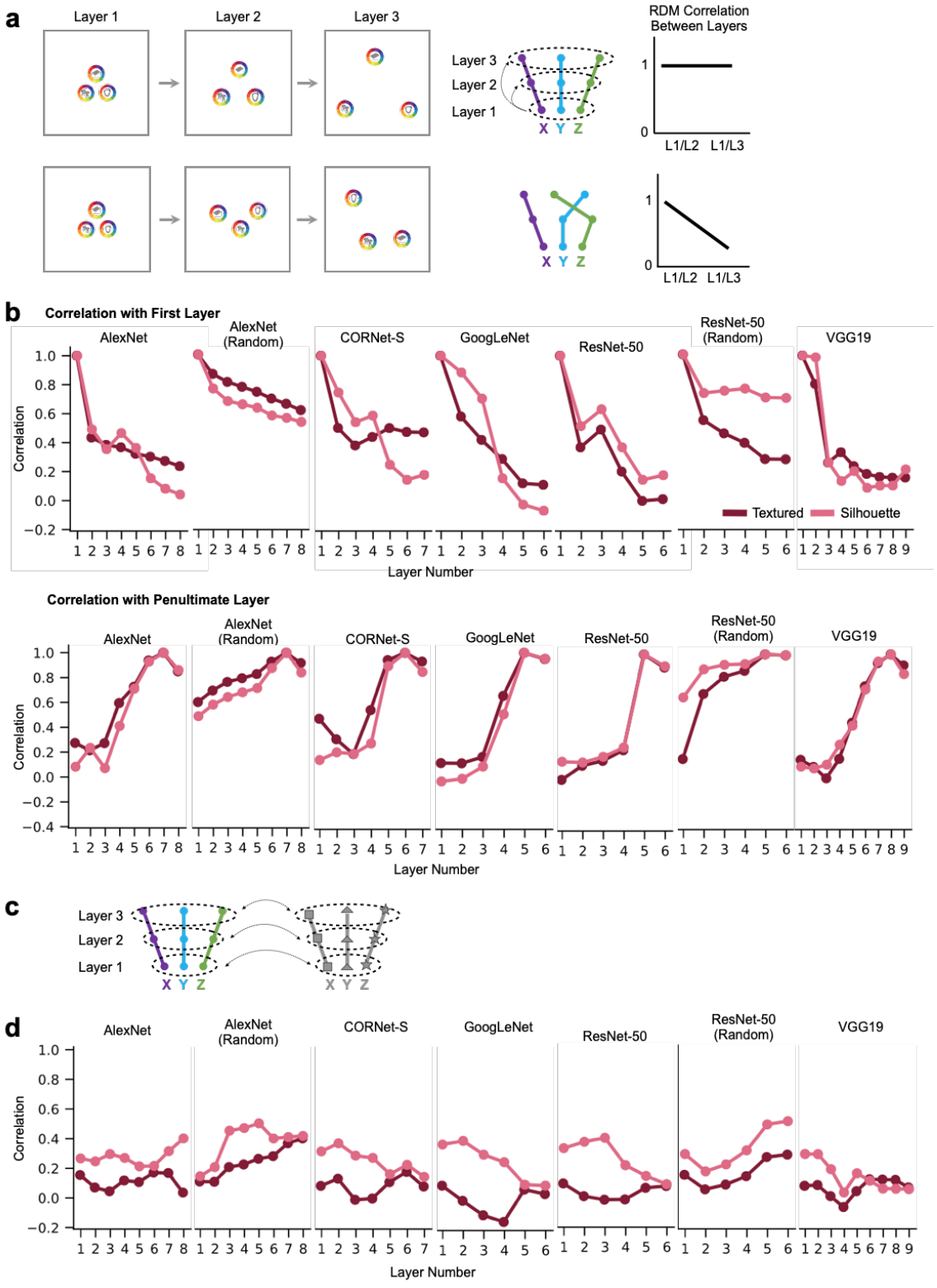


Figure 3.4. The evolution of color space similarity among objects across CNN layers and the dependence of color space similarity on object form similarity. a. A schematic illustration of two

possible scenarios of the evolution of color space similarity among objects across CNN layers, using the same notations as those in Figure 3.2a. In this analysis, we examine whether or not patterns of color space similarity among objects (as shown on the left) are preserved across layers by correlating the color space similarity matrix (i.e., the second-order RSM quantifying the similarity among the color spaces of different objects) from the first layer with each of the other layers, as shown in the middle. These correlations are then plotted in a line graph on the right. In the first scenario (top row), the relative color space similarity among the different objects is preserved in the different CNN layers (i.e., the configuration of the three color spaces stays the same across the different layers), even as the absolute similarity among color spaces decreases. In the second scenario, the relative color space similarity is not preserved in different CNN layers (i.e., the configuration changes across the different layers). b. The correlations of the color space similarity across different CNN layers for the full set of 50 objects in the 12 colors calibrated in CIELUV color space for both versions of the objects. Top row shows the correlations with the first layer of each network, bottom row shows correlations with the penultimate layer of each network. In most cases, correlations between the early and late layers are fairly modest. Results for random networks are averaged across 100 random initializations of the network. c. A schematic illustration of comparing color space similarity and object form similarity, using the same notations as those in Figure 2a. In this analysis, the achromatic object form similarity matrix is extracted for each CNN layer and then correlated with the corresponding color space similarity matrix of that layer. d. Correlations between the form similarity and color space similarity for each layer of each network for the full set of 50 objects in the 12 colors calibrated in CIELUV color space for both versions of the objects. Results for random networks are averaged across 100 random initializations. No reliable trends were evident, but in general correlations were modest.

color space, for each CNN layer, we first performed all possible pairwise correlations of the CNN layer output for the grayscale versions of the object forms to form an object form similarity RSM and vectorized the off-diagonal elements of this RSM to form an object form similarity vector. We then correlated this object form similarity vector with the color space similarity vector from the same CNN layer (Figure 3.4c). If similar object forms had similar color space structure, we expected to obtain a high correlation between the two. We performed the analysis for both the networks trained on ImageNet, and 100 random initializations of AlexNet and ResNet-50. As shown in Figure 3.4d, correlations varied across layers and networks, showing no consistent pattern; correlations tended to be higher for the silhouette than for the textured stimuli, but tended to be modest. Interestingly correlations existed for the random networks as well,

suggesting that a network's intrinsic architecture can induce a correlation between the shape similarities of different objects and the color spaces similarities of those objects. Overall, color space similarity does not closely track object form similarity, suggesting some separation between the two, and that the difference in color space similarity between two objects does not scale linearly with the difference in their form features.

Discussion

Despite decades of neuroscience research, we still lack a full understanding on how feature conjunctions are represented in the primate brain. In this study, we took advantage of the recent development in CNNs trained to perform object classification and examined how such an information processing system jointly represents different object features across the entire processing hierarchy of a CNN. We did this through using a variation of RSA to examine how color coding varies across different objects, which provides an index that reflects the extent to which color and form are encoded in an interactive, as opposed to independent, manner. Our investigation not only allowed us to gain insight into the internal representations of CNNs, but also enabled us to develop a novel network-based stimulus-rich approach to study feature binding across the entire network and a large stimulus set, which can be easily implemented to study feature binding in biological visual systems. Although we tested the joint coding of color and form here, our approach can be applied to study the joint coding of any pair of features: the variation in coding for feature X across values of feature Y can be computed by first computing the similarity space for feature X *within* each value of feature Y, then correlating these similarity spaces *across* each value of feature Y.

With this approach, we found that color coding increasingly varies across different real-world objects in higher levels of each CNN. This held true for both the naturally textured stimuli and the uniformly colored "silhouette" stimuli, suggesting that interactive coding of color and form in higher CNN layers exists for global form features alone (which are preserved in the silhouettes), even in the absence of texture and much of the category-diagnostic visual information. The textured interior of an object form, however, did further increase the amount of interactive coding between colors and forms, likely due to the presence of additional form features in these textured objects. This interactive coding appears to be quite general, as it was present not only for complex real-world object forms, but also for minimally simple oriented bar features with stimulus size equated. Finally, this effect did not depend on the distance metric that was used to compute the representational geometry of the features, nor did it depend on our particular selection of color space used to calibrate the luminance and saturation of the images.

All sampled networks contained both convolutional and fully connected layers, including the final category output layer. The increasing degree of interactive coding we observed throughout processing was found both in the fully connected layers prior to the category readout layer, and the category readout layer itself. A priori, one would assume that the final fully connected layer encodes object category orthogonally to color, since it is trained to output category labels. However, our results show that there is a greater amount of color and form interaction in the final compared to the first sampled layer, with the amount of interaction steadily increasing during the course of visual processing. Additionally, while the networks we examine vary broadly with respect to parameters such as the number of units and layers, we observe the same increasing entangling of color and form information as processing proceeds in

each network. This suggests that it is a truly general property of CNN information processing rather than a quirk arising from the architecture of any particular network or layer type.

The interaction between color and form coding was not a mere byproduct of the intrinsic architecture of a CNN, as the interaction effect was profoundly attenuated in untrained CNNs with random weights. That said, the increasing color-form interaction over the course of processing was still significant for the random networks, suggesting that a small part of the effect arises from intrinsic aspects of the CNN's architecture. The interaction between color and form coding did not appear to depend on the existence of natural covariation between form and color in the training set, as the magnitude of interactive tuning is nearly as large in a CNN trained on objects stripped of their naturalistic form-color pairings. Thus training for object recognition is needed to produce the interactive coding of color and form, even when no consistent color and form pairing is present during training. This suggests that the interactive coding of color and form is not intrinsically tied to the CNN architecture and that object recognition training automatically gives rise to increasingly tangled color and form representations in higher levels of processing, even when color is not informative to object recognition after training.

To our knowledge, only one study has examined the joint encoding of color and form features in CNNs: Rafegas and Vanrell (2018) examined the responses of several individual color-selective CNN units in both the first and last convolutional layer and found that they were also sensitive to both the color and the orientation of an image patch. Based on this result they suggest that color and form are "entangled" at all stages of processing. At first glance, this seems to conflict somewhat with our finding that color and form are first encoded orthogonally but are increasingly encoded interactively with further processing. However, this difference can be readily explained by the difference in analysis approach. These authors examine interactive

color-form tuning in single units, whereas our method involves examining changes in the population-level representational geometry of color coding across form changes. Interactive color-form coding in single units could coexist with a stable population-level representational geometry for color coding across forms if color and form interact in the same way across different colors. For example, this would occur if for every unit selective for vertical red-green edges there exists a unit selective for vertical blue-yellow edges. This distinction highlights an important point in studying neural networks whether biological or artificial: the way that a single unit jointly encodes two features does not transparently reveal how these two features are jointly encoded at a population level.

In additional analyses, we found that an object's color space greatly diverged from its initial color space over the course of processing and that two objects with a similar color space at the beginning of processing did not necessarily have a similar color space at the end of processing. Thus the color space representation for a given object as well as the relative similarity of color spaces between objects dynamically changed over the course of processing**.** These color space transformations were greatly reduced in the untrained networks, suggesting that training networks for object recognition induces a transformation in how color is encoded over processing. Moreover, the color space of an object tended to transform in similar ways across the trained networks, but differently in the untrained networks. This relative consistency across the trained, but not the untrained, networks with vastly varying architectures suggests that this resculpting of color space may be of adaptive value for the network's object classification task. Interestingly, the achromatic form similarity of two objects only weakly predicted the similarity of their respective color spaces. This demonstrates that, in general, color space similarity does not closely track object form similarity, suggesting some separation between the

two. The untrained networks also exhibited a correlation between shape similarity and color space similarity, suggesting that these correlations may be a byproduct of a network's intrinsic architecture.

Overall, these results show that colors are not represented similarly across different objects in an orthogonal manner in CNNs. Rather, colors are encoded increasingly differently across objects in an interactive and object-specific manner during the course of CNN processing. This is more consistent with a late feature integration account (but without needing an additional binding operation), rather than color and form being represented in an initially entangled and intermingled fashion and only being represented separately and explicitly in later layers. To our knowledge, these results provide the first detailed and comprehensive documentation of how color and form may be jointly coded in CNNs, unveiling important details regarding the algorithms employed by CNN for visual processing, which up to now have remained largely hidden. While some studies have examined the coding of complex, high-level form features in CNNs, such as those involved in face recognition (Grossman et al., 2019), this study is the first to document the joint population coding of two completely *different* visual features, color and form. Moreover, the present study examines this interaction both for complex, naturalistic form features and simple form features, in the case of the oriented bar stimuli.

It should be noted that interactive tuning does not imply that there exist units tuned exclusively to a single color/form conjunction (a "grandmother unit"); units could plausibly be tuned to heterogeneous combinations of color and form combinations in a "mixed selectivity" coding scheme. Such a coding scheme has been reported in the macaque prefrontal cortex for the coding of stimulus identity and task and has been shown to vastly increase the neural representational capacity of that brain region (Rigotti et al., 2013). The present results suggest

that an interactive coding scheme along these lines may be more prevalent and can automatically emerge in a complex information processing system even though, compared to a biological brain, CNNs have a relatively simple structure, consisting only of a single feed-forward sweep and lacking mechanisms such as feedback connections (except for the recurrent network, CORNet-S, we included here) and oscillatory synchrony. Such a coding scheme may well be used by sensory regions in the primate brain to support the flexible encoding of a wide range of sensory feature combinations. Indeed, although initial evidence from visual search (Treisman and Gelade, 1980) and neuropsychology studies (Zeki, 1990) suggests that color and form might be initially encoded independently, and only combined in a late binding operation, other strands of evidence suggest that color and form might be encoded in an interactive manner early on during processing (Rentzeperis et al., 2014). For example, Seymour et al. (2009) and our own work (Taylor & Xu, 2020) have shown that nonlinear tuning for color/orientation combinations might emerge as early as V1, V2, V3, and V4. Consistent with the present observation, interactive coding of color and form has also recently been observed in the color selective neurons of macaque color patches (Chang et al., 2017).

Despite its significance in visual cognition, how feature conjunctions are coded in the human brain remains unresolved. A population code that instantiates interactive tuning for feature combinations, as we observe here, is a candidate mechanism that should be explored in more detail, and analogous analyses should be applied in monkey neurophysiology and human fMRI studies to see if similar response profiles exist in the primate brain. Indeed, given that a number of previous studies have shown that the representations formed in lower CNN layers better correlate with lower than higher primate ventral visual regions and, conversely, the representations formed in higher CNN layers better correlate with higher than lower primate

ventral visual regions (Cadieu et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014; Güçlü & van Gerven, 2015, Cichy et al., 2016, and Eickenberg et al., 2017; Xu & Vaziri-Pashkam, 2020), our results may be used to directly predict and compare with responses from corresponding primate cortical regions.

To summarize, despite the success of CNNs in object recognition tasks, presently we know very little about how visual information is processed in these systems. The present study provides the first detailed and comprehensive documentation of how color and form may be jointly coded in CNNs. Our development of a novel network-based stimulus-rich approach to study feature binding in CNNs can be easily implemented to study neural mechanisms supporting feature binding in the primate brain. Equally importantly, the discovery of the interactive coding scheme used by CNNs to encode feature conjunctions could be a viable coding scheme that the primate brain may employ to solve the binding problem.

Methods

CNN Selection

We chose five CNNs in our analyses: AlexNet, CORNet-S, GoogLeNet, ResNet-50, and VGG19. These CNNs were selected based on several different criteria. AlexNet (Krizhevsky, Sutskever, & Hinton, 2012) was included for its high object recognition performance, relative simplicity, and prevalence in the literature. VGG19 (Simonyan & Zisserman, 2015), GoogLeNet (Szegedy et al., 2015), and ResNet-50 (He, Zhang, Ren, & Sun, 2015) were chosen based on their high object recognition performance and architectural diversity. Additionally, both AlexNet and VGG19 have a shallower network structure, whereas GoogLeNet and ResNet-50 have a deeper network structure. Finally, CORNet-S (Kubilius et al., 2018), a shallow recurrent CNN

designed to approximate the structure of the primate ventral visual pathway, was included for its high correlation with neural and behavioral metrics. This CNN has recently been argued to be the current best model of the primate ventral visual regions (Kar et al., 2019). For most analyses, we used pre-trained implementations of these CNNs optimized for object recognition using ImageNet (Deng et al., 2009). The exact training method differed somewhat among networks; specifically, AlexNet, VGG19 and ResNet-50 use cropping, horizontal flips, and RGB adjustments (simulating changes in lighting) for data augmentation, whereas CORNet-S only uses cropping and horizontal flips, and GoogLeNet only employs cropping. To understand how the specific training images would impact CNN representations, we also examined responses from an alternative version of ResNet-50 that was trained on stylized ImageNet images in which the original texture of every single image was replaced with the style of a randomly chosen painting. This biased the model towards representing holistic form information rather than texture information (Geirhos et al., 2019). Finally, in order to determine to what extent the architectural parameters of a network (number of layers, kernel size, etc.), independent of any training, affects the results, we also examined multiple initializations of AlexNet and ResNet-50 with randomly assigned weights and no training. The PyTorch implementations of all models were used, and custom scripts designed to interface with PyTorch were used for all analyses (Paszke et al., 2017).

Since these CNNs have varying, and often large, numbers of layers, we performed analyses over a subset of layers in each model, in order to simplify analysis and roughly equate the number of layers analyzed in each model. The first layer, several intermediate layers, the penultimate layer (i.e., the last layer before the object category label outputs), and the final layer (i.e., the object category label output layer) were used for each model. Selection of intermediate

layers varied based on the model, but since all CNNs we examined tended to be structured into architecturally significant "segments" (e.g., VGG19 has repeated "segments" composed of alternating conv and relu layers followed by a pooling layer, CORNet-S has "segments" meant to correspond to different visual brain areas etc.), we included the intermediate layer corresponding to the boundaries between "segments". The specific layers we included are listed in Table 3.1. For this study, we adopt the convention of labeling layers by the kind of layer, followed by the number of times that kind of layer was used up to that point in processing (e.g., the third convolutional layer is conv3).

In cases where we wished to compare coding at the beginning and end of processing in the network, the first and *penultimate* layers were used; this is because the final layer is the category output layer, and thus the penultimate layer can be seen as the last "feature-coding" layer.

In order to compare our results across different CNNs, for some analyses we coded a variable, layer_fraction, that reflects what fraction of a network's layers have been traversed up to a given layer in the course of a CNN's processing hierarchy (all layer types were included). For example, the first layer in a ten-layer network would have a value of .1 for this variable, and the final layer would have a value of 1.0.

Stimulus Selection

We used a set of real-world object stimuli from Brady et al. (2013) as our main object stimuli. This stimulus set consists of images (400 x 400 jpegs) of 540 different objects, where the colored portions of these objects are all initially of the same hue (so as to facilitate manipulating the colors of the objects in a consistent way). To derive the stimuli used in our analyses, we

selected a smaller subset of these objects (as detailed below), and then manipulated the color and texture of these objects.

Main Object Stimuli in CIELUV Colorspace

For our main stimulus set, we chose 50 objects intended to be maximally dissimilar with respect to their high-level visual features. To do this, we converted the initial 540 objects to grayscale, ran them through AlexNet, and extracted their activations from AlexNet's penultimate (last pre-classification) layer, FC2. We then constructed a representational similarity matrix (RSM) by computing all pairwise correlations of the CNN output from layer FC2 for each object with each other, and used this RSM to select a set of 50 objects whose mean pairwise correlation was minimally low. With this procedure, the mean pairwise similarity went from $r = .25$ (min $r = -.02$, max $r = .92$) for the original set of 540 objects, to a mean pairwise similarity of $r = .13$ (min $r = -.02$, max $r = .78$). The resulting set of 50 objects are shown in Figure 3.1a; visual inspection confirms that the resulting object set spans a wide range of different form and internal texture features.

We then recolorized each of the 50 objects, roughly following the procedure outlined in Chang, Bao, and Tsao (2017). This procedure guaranteed that all stimuli had the same mean luminance and saturation over the non-background portions of the image. Equating mean luminance was necessary in order to equate each image's contrast with the background, and ensure that any results were not mere byproducts of contrast-sensitive mechanisms rather than color processing as such. Clearly, even a color-blind network would encode lingering contrast differences, so equating contrast across colors is necessary to ensure that any results are genuinely driven by color-sensitive mechanisms. Equating mean saturation was necessary to

ensure the validity of some of our analyses; in particular, since we examine whether color coding varies based on form, failing to equate saturation could introduce spurious results, since a relatively unsaturated object image would by definition have less hue variation (this is especially problematic since the RGB values fed into the network are integers, so reducing the hue variation could hide important structure by introducing a loss of precision due to rounding). Additionally, without equating saturation, it would remain a possibility that variability in certain units could be driven by the degree of saturation, rather than hue per se. Since luminance and saturation are not explicitly encoded in the RGB color space, we adopted the approach of first tuning the hue, saturation, and luminance of every image (every color of every object) in LUV color space before converting the image back to RGB to feed into the network. Each object stimulus was colored in each of 12 different hues, evenly spaced around the colorwheel. Specifically, we converted all images to the CIELUV colorspace, which is constructed such that equal distances in the space correspond to roughly equal psychophysical differences; this was done so that we could use the same stimuli on a future study comparing our results to those of human observers. Next, we computed the mean luminance (L) of each stimulus over the non-background portion of each stimulus, and did the same for the saturation (computed as $\sqrt{u^2 + v^2}$, where u and v are the two chromaticity coordinates in the LUV color space). For each image, a constant value was then added to the luminance and saturation of the non-background pixels so as to bring the mean luminance and saturation of that image to target values that were equated across all objects. This procedure sometimes resulted in pixels whose values overflowed past the range of permissible LUV luminance and saturation values; in cases where this occurred, the variance of the luminance or saturation about the mean was shrunk until all pixel values fell within the permissible boundaries. Once the luminance and saturation for each pixel were set in this

manner, each object was colored in 12 different hues by rotating the U and V (hue) coordinates in each pixel to 12 equally spaced angles. Additionally, a grayscale version of each stimulus was created by setting U and V to zero, while keeping the luminance channel the same. This procedure preserves relative saturation and luminance patterns that were present across each original image, while manipulating hue and equating mean saturation and luminance. The target mean saturation and luminance were derived through successive adjustments, until applying the above transformations to each image did not take any pixel's LUV values outside of their allowable range. Mean saturation was required to be relatively low, due to the nonlinearities of the LUV color space; specifically, a high saturation value may be possible for one luminance/hue combination, but not others, so saturation had to be kept within relatively narrow boundaries. Stimuli were converted from LUV back to RGB color space prior to being run through the networks, as the networks were trained on RGB images. Since internal texture is preserved for these stimuli, we henceforth refer to them as the "Textured" stimuli.

In addition to the above method, which preserved the internal texture of the stimuli, we also constructed a version of each stimulus that consisted of a uniformly-colored "silhouette" of the image (henceforth "silhouette" stimuli), thereby removing internal texture information while sparing global form information. These were constructed by replacing every non-white pixel in the object with a pixel of a uniform color. Twelve such silhouette images were created from each object, using the same 12 hues and the same mean luminance and saturation values as were used for the Textured stimuli. We performed this manipulation because some evidence suggests that CNNs may prioritize texture over global form features (Geirhos et al., 2019), so employing silhouette stimuli allowed us to examine whether our results hold in the absence of texture. Example stimuli from these two methods, in the 12 possible colors, are shown in Figure 3.1b.

Main Object Stimuli in Synthetic HSV Colorspace

The CIELUV color space, while widely used, implements a specific parametrization of luminance and saturation that is based on human psychophysical judgments, which may not be applicable to how CNNs represent visual information. To ensure that our results do not depend on these idiosyncrasies of the CIELUV color space, we constructed a stimulus set whose colors were calibrated in a novel color space that we designed not to align with human data, but with the RGB input that CNNs receive. Specifically, we used a variation of the common hue/saturation/luminance parametrization calculated over RGB values, which we call Synthetic HSV. Some such parametrizations require taking maxima and minima of the RGB channels, an operation which is not available to convolutional kernels. Thus, luminance was stipulated to be the mean of R, G, and B; this definition assigns equal weights to the three channels (unlike some HSV parametrizations, which weight the three channels differently to account for human psychophysical performance), and uses an operation (taking the mean) that is easily implemented by convolutional kernels. Intuitively, saturation reflects the dissimilarity of a color from neutral grey; thus, saturation was stipulated to be the Euclidean distance in RGB coordinates from a color to neutral grey of the same luminance. Once these values are fixed, the range of possible RGB values forms a circle in the 3D RGB space. This occurs because restricting the RGB values to have a fixed average - and therefore a fixed sum - constrains the range of possible RGB values to fall on a single plane, and further restricting the RGB values to have a fixed Euclidean distance from the neutral gray point on that plane (where R = G = B) selects a circle of RGB values on that plane. Within this circle, we stipulated that the RGB triplet with the highest R channel corresponds to a hue value of 0°. To keep the overall level of luminance and saturation in a similar range to the stimuli calibrated in the CIELUV color space, we set the mean

luminance and saturation (in Synthetic HSV) for these stimuli to the values of these parameters for one of the colors calibrated in CIELUV space (specifically, the one with the highest red channel in RGB space). We then constructed stimuli with these chosen values as their mean luminance and saturation, with 12 hue values evenly spaced around 360°. After calibrating the stimuli in synthetic HSV, stimuli were converted back into RGB to feed into each network. Stimulus construction was otherwise identical to the procedure for the stimuli colored based on CIELUV.

Oriented Bars in CIELUV Colorspace

To examine the extent to which our results may hold for stimuli equated in their spatial coverage and for simpler stimuli than the naturalistic object stimuli that were used in the study, we constructed a set of oriented bar stimuli (Figure 3.1c). Twelve orientations, ranging in even increments from 0° to 180°, were used, and each was uniformly colored in the same twelve isoluminant and isosaturated colors that were used for the object stimuli (using CIELUV space).

Analysis Methods

For all analyses, images were fed into each network. Next, unit activations were extracted from each sampled layer and flattened into 1D vectors in cases where the layer was 3D (e.g., if it was a convolutional layer).

Visualizing Color Space Representation Across Objects and CNN Layers

We used representational similarity analysis (RSA) to measure conjunctive tuning for color and object form. Specifically, we examined the extent to which the representational

structure for color changes across the different object forms. To the extent that the representational structure of color varies across the object form, it would provide evidence that CNNs encode color and object form not independently, but interactively.

As an initial analysis, we visualized how the color spaces for two example objects differ at the beginning and end of a CNN. Specifically, we extracted the patterns for all 12 colors of two example objects from the first and the penultimate layers of AlexNet (which are Conv1 and FC2). Within each layer, we performed all pairwise Pearson correlation among the 12 patterns for each object to create a representational similarity matrix (RSM, with the value for each cell being the Pearson correlation coefficient); each value was then subtracted from 1 to convert it to a dissimilarity matrix. Using multidimensional scaling (MDS), we visualized the resulting dissimilarity space (Figure 3.1d).

Next, we visualized how the color spaces of six representative objects might diverge over the course of processing in AlexNet (Figure 3.1e). To do this, for each object and for each sampled layer of AlexNet, we first constructed a "color space" RSM by performing all pairwise Pearson correlations of the patterns associated with the 12 different colors of that object (with the value for each cell of the matrix being the Pearson correlation coefficient). We vectorized the off-diagonal value of this RSM to create a "color space" vector. Next, we performed all pairwise correlations of these "color space" vectors across objects and layers to form a "color space similarity" RSM that quantifies how similarly color is coded in different objects and layers. After converting the matrix to a dissimilarity matrix by subtracting each value from 1, we then used 2D MDS to visualize the similarity of the different color spaces across different objects and CNN layers. Figure 3.1f shows the full color space similarity matrix among all 50 objects for Conv1 and FC2 of both the trained and untrained version of AlexNet.

Following these qualitative observations, to provide a comprehensive and quantitative description of color representation across different objects and CNN layers, we performed a series of analyses. Specifically, we quantified (1) within each layer, how color is coded differently across objects (Figure 3.2a), (2) within each object, how color is coded across different layers of a CNN and different CNNs (Figure 3.3a), and (3) whether or not color space similarity among the different objects within one layer is preserved across CNN layers (Figure 3.4a). We also quantified how color space similarity between two objects may be determined by their form similarity at a given CNN layer (Figure 3.4c). These four analyses are described in detail below. All the analyses were performed for both the Textured and Silhouette stimuli.

Quantifying Color Space Differences Across Objects within a CNN Layer

To understand how color is coded across objects in each CNN layer, we first created a "color space" vector for each object in each layer of each CNN as described above for our main stimulus set of 50 objects and 12 colors calibrated in CIELUV color space. We then performed all pairwise correlations of these "color space" vectors for all the objects for a given CNN layer. We next averaged these correlation values within each layer and used a line plot to visualize how the mean colorspace correlation changes over layers of a given CNN (Figure 3.2). Figure 3.2b shows the color space correlations between every pair of objects in each layer of AlexNet, for the colored object stimuli calibrated in CIELUV color space. Figure 3.2c shows the mean pairwise color space correlations among every pair of objects for the five trained networks we examined. To assess the statistical significance of any change across layers, for each pair of objects, the correlation values between the color spaces of those two objects were Fisher Z-transformed and regressed onto the position of that layer in the CNN (using the layer_fraction variable described

in the CNN Selection section). The resulting slope from this regression reflects the degree to which the color space similarity for these two objects increases or decreases over the course of the network. A positive slope would mean that the color spaces for these two object forms become progressively more similar over the course of processing. A one-sample t-test was used to test the slopes from all possible object form pairs against zero to assess whether the average slope was significantly different from zero (that is, whether the mean color space similarity between objects tends to change over the course of processing). Additionally, a matched-pairs t-test was used to assess whether the slopes significantly differed between the textured and silhouette images. Note that these regression analyses do not require a strictly monotonic decrease in color space correlations as processing proceeds, but only whether there tends to be an increase or decrease on average.

To examine whether our effects hold not only for complex, real-world objects but also for simple artificial forms, we repeated the same analysis on simple oriented bar stimuli, where the different object "forms" were simply different orientations of a centrally placed bar stimulus (Figure 3.2d). This allowed us to examine whether the results would hold for stimuli equated in their overall spatial coverage and whether results obtained from complex nature objects hold for simple form stimuli.

We also performed the same analysis in a number of control conditions, to examine whether specific choices regarding the stimulus set and analysis method affect the results. As our first control, to test how the particular similarity measure we used may impact the results, we repeated our analysis, but used Euclidean distance as our initial similarity metric instead of Pearson correlation (Supplementary Figure 3.1a); this was done because Euclidean distance, unlike Pearson correlation, is an unbounded metric, and we sought to ensure that the choice of

the specific similarity measure did not affect the results (Pearson correlation was still used as the second-order similarity metric to quantify the differences *between* the color spaces of different objects). As our second control, to examine whether our results depended on our particular choice of color space, we repeated the same analysis on the same set of objects whose colors were calibrated in the synthetic HSV space described above, instead of the LUV space used in our main stimulus set (Supplementary Figure 3.1b).

The Effect of Training on CNN Color Space Representation

In order to assess whether the naturally occurring consistent color and form conjunctions present in the training images were necessary to produce the results we observed, we compared models trained on naturally textured stimuli, versus unnaturally textured "stylized" stimuli (Geirhos et al., 2019). Specifically, we compared performance between ResNet-50 trained on ImageNet, and ResNet-50 trained on stylized images. In order to assess the extent to which the results are driven by the intrinsic architecture of the networks, versus being a consequence of training them for object recognition, we also performed this same analysis on 100 initializations of AlexNet and ResNet-50 with random weights and no training of any kind. The same analysis pipeline was applied to each random initialization independently, and the final results (mean color space correlation across objects within a layer) were averaged to obtain the final result (Figure 3.2e). Targeted matched-pairs t-tests were used to compare the slopes of these differently-trained networks with the corresponding networks trained on object recognition.

Transformation of Color Space Representations Across CNN Layers and Architectures

To understand how the color space of an object may evolve over the course of processing and whether colors are coded similarly for an object across different layers of a network, for the main original set of 50 objects and 12 colors calibrated in CIELUV color space, we correlated the color space vector for each object in either the first or penultimate layer of each network with its color space vector from each other layer of the network (Figure 3.3a). These correlation values were then averaged across all objects and plotted in Figure 3.3b. To test for statistical significance, we performed a regression analysis to examine whether correlations with the first and penultimate layers of the network decrease in layers that are further apart. To do this, for each object, we applied Fisher's Z transformation to the correlation values, and regressed them onto the positions of the CNN layers (using layer_fraction) of all layers up to, but not including, the comparison layer (first or penultimate layer). A one-sample t-test was then used to test the slopes from all the objects against zero to assess whether the average slope was significantly different from zero. This was done for the networks trained on ImageNet, on 100 initializations of AlexNet with random weights, and 100 initializations of ResNet-50 with random weights.

To examine whether color coding within an object transforms in a similar manner across different networks, for each object we correlated its color space vector from each sampled layer of each network with every other layer (Figure 3.3c, left column). The resulting similarity space was averaged across objects and visualized using MDS (after subtracting each correlation from 1 to convert to dissimilarity), and the mean pairwise similarities in the first and penultimate layers of each network were reported in Supplementary Figure 3.2.

To examine whether the color space of an object evolves in a similar way in trained networks and in randomly initialized networks, we performed the same analysis as described

above comparing the version of AlexNet trained on object recognition with 10 random initializations of AlexNet (Figure 3.3c, middle column). We also performed the same analysis for ResNet-50, including the ImageNet-trained version, the version trained on stylized images, and ten random initializations (Figure 3.3c, right column). The mean pairwise similarities in the first and penultimate layers for these comparisons were reported in Supplementary Figures 3.3 and 3.4.

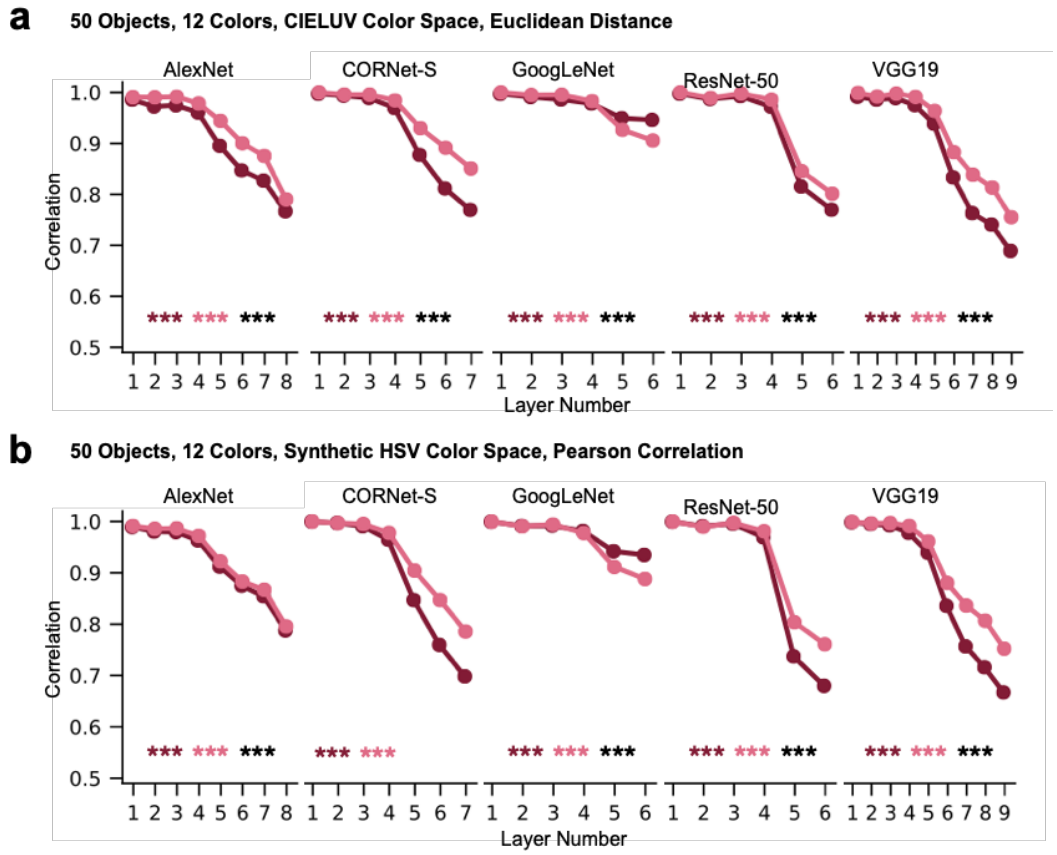
Transformation of Color Space Similarity Across CNN Layers

To understand how patterns of color space similarities among objects change over the course of processing (i.e., whether the similarity profile among the color spaces of different objects remains consistent over the course of processing), for the main original set of 50 objects and 12 colors calibrated in CIELUV color space, we correlated the color space vector of all objects within each layer with one another to construct a color space similarity RSM for that layer. From the similarity matrix formed, we used the off-diagonal values to define a *color space similarity vector.* The resulting color space similarity vector from the first and penultimate layers were then correlated with those from each of the other layers (Figure 3.4b).

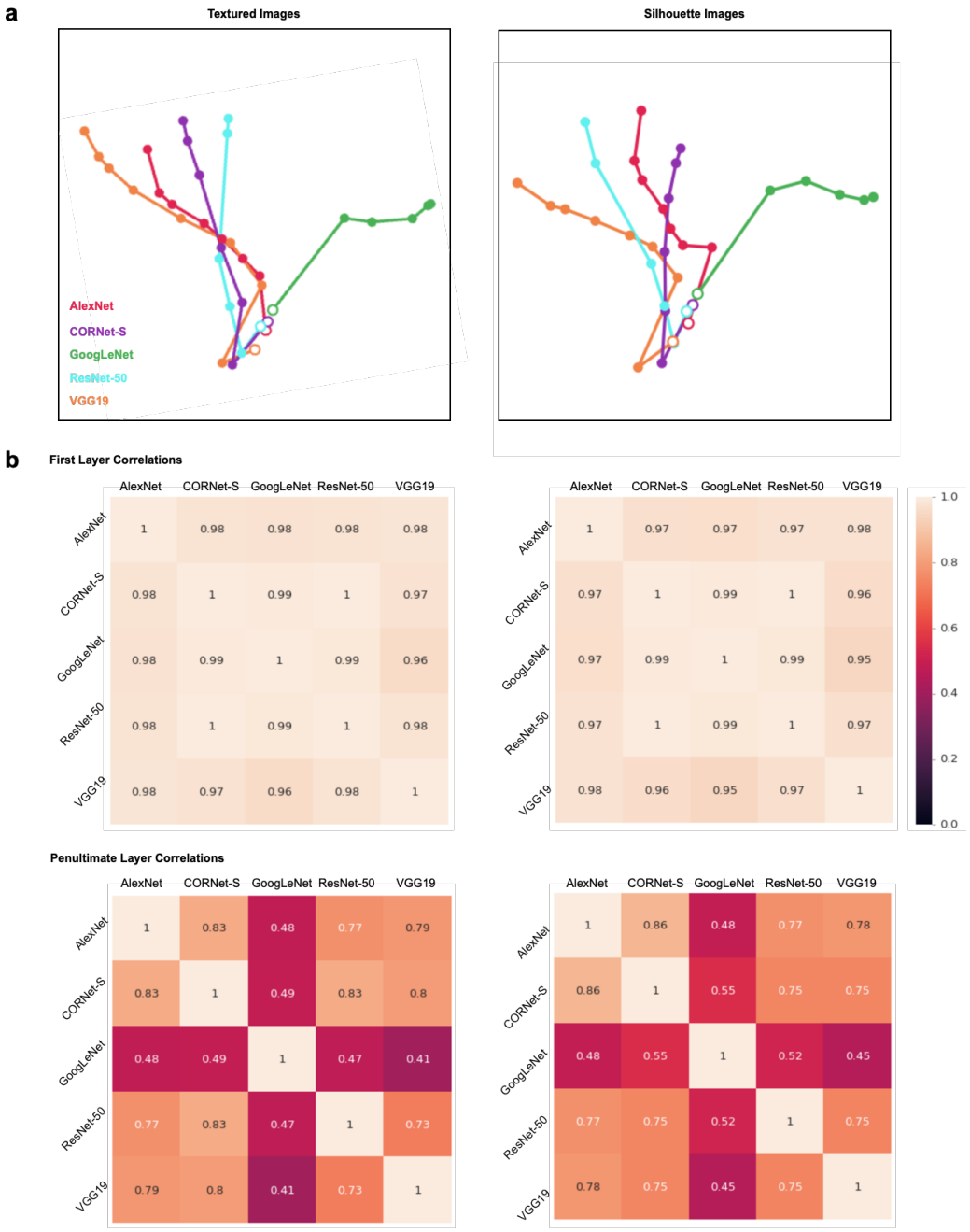
The Effect of Object Form Similarity on Color Space Similarity

To understand how color space similarity of two objects is determined by the form similarity of these two objects (Figure 3.4c) and if two objects with similar forms would also have similar color spaces, for the main set of 50 objects and 12 colors calibrated in CIELUV color space, we first measured the overall object form similarity in each CNN layer for the original set of 50 objects. This was done by calculating all the pairwise Pearson correlations of

the CNN layer output to grayscale versions of all the objects. From the similarity matrix formed, we used the off-diagonal values to define an *object form similarity vector*. We then correlated the object form similarity vector with the corresponding color space similarity vector for that CNN layer. The resulting correlation value from each CNN layer was plotted together in a line graph (Figure 3.4d).

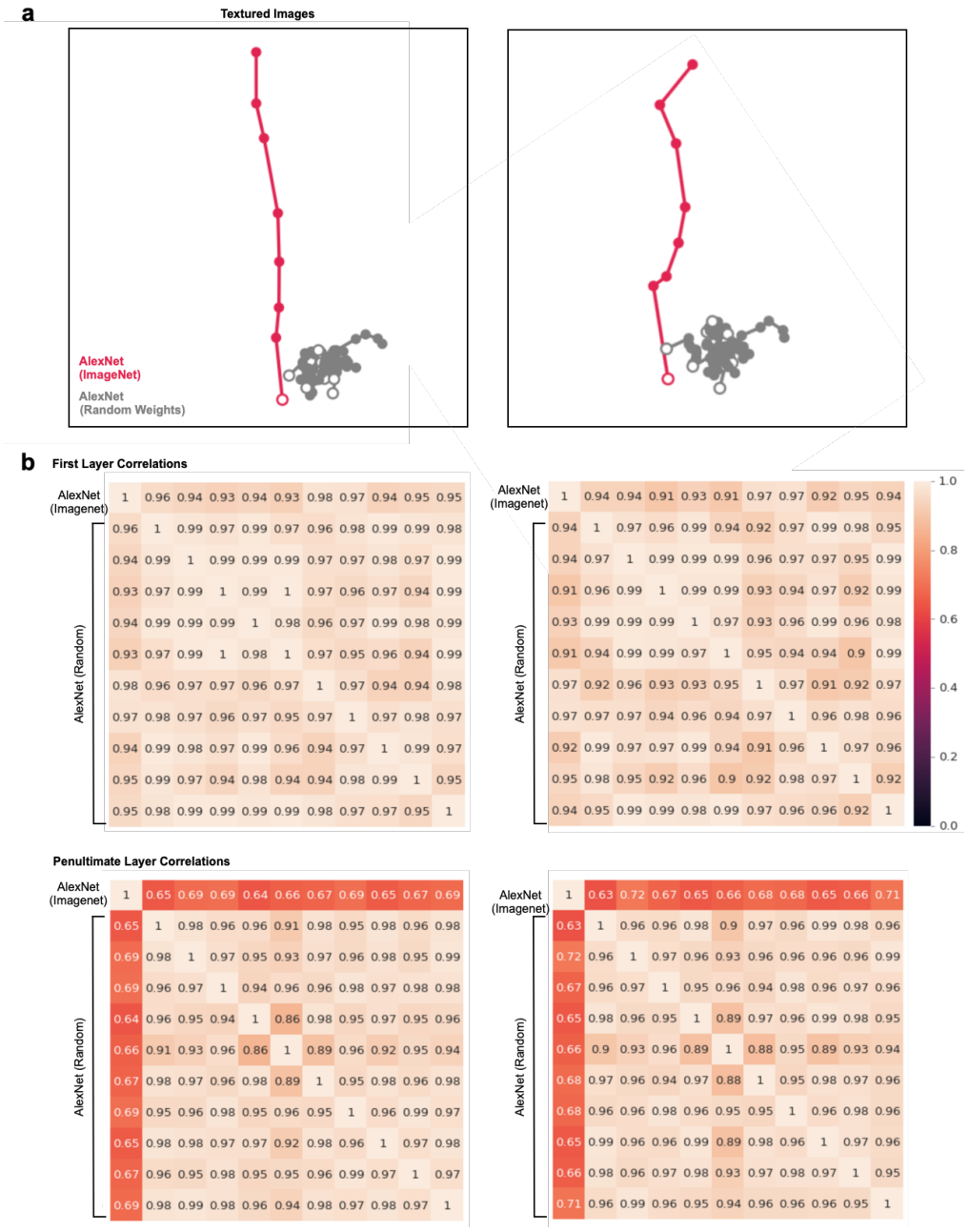


Supplementary Figure 3.1. Mean between-object color space correlations for each layer and network, using two additional measures. a. Same as Figure 2c, but with color space structure of each object measured with Euclidean distance instead of Pearson correlation (similarity between color spaces was still calculated with Pearson correlation). Results remain qualitatively similar as those in Figure 2c. b. Same as Figure 2c, but using colors calibrated in an artificial HSV color space that is not based on human psychophysical judgments. Results again remain qualitatively similar as those in Figure 2c. *** $p < .001$.



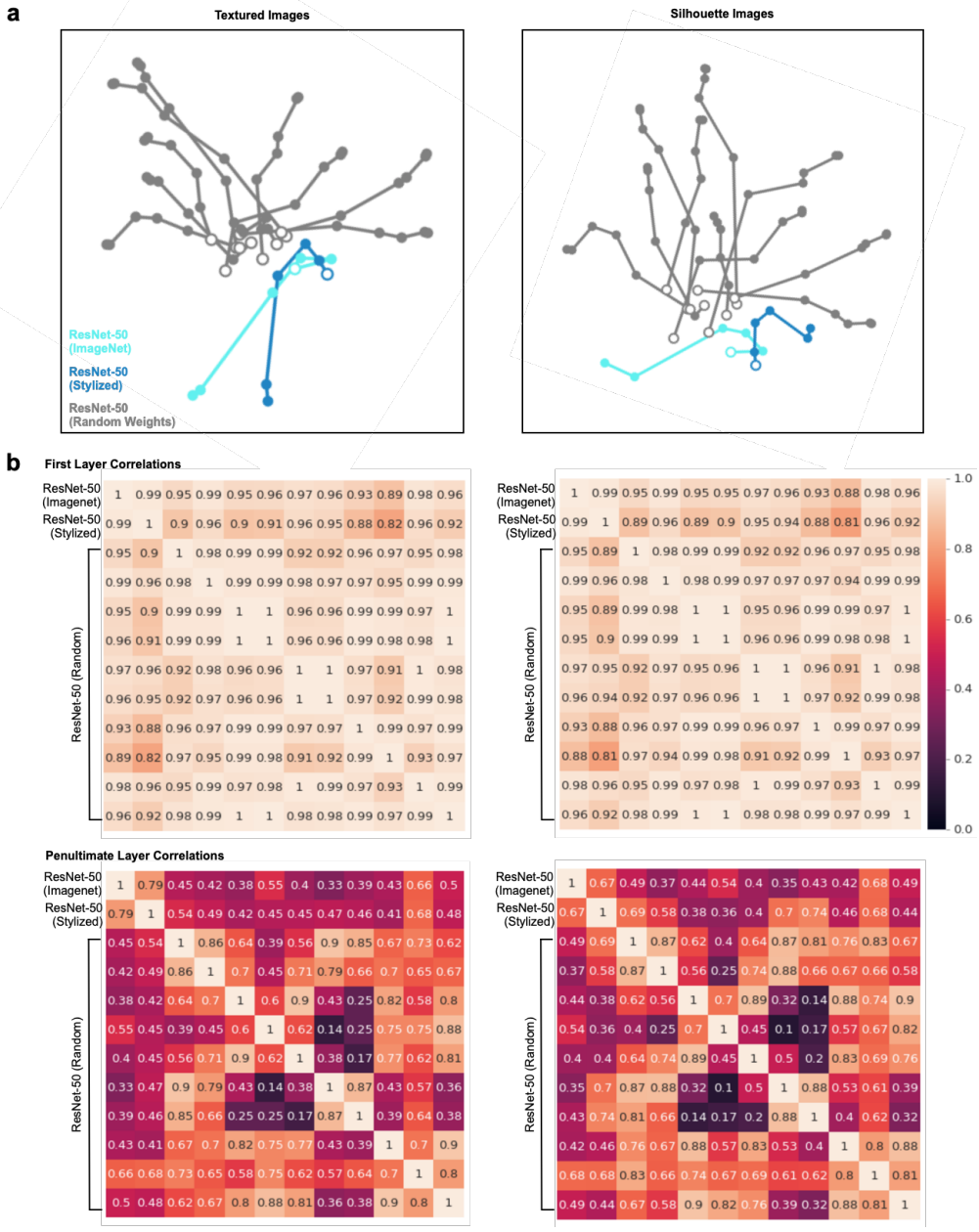
Supplementary Figure 3.2. Color space comparisons for the five trained networks. a. MDS plots depicting color space correlation across different CNN layers and architectures, plotted separately for the two versions of the objects (copied from Figure 3c for convenience). This was done by constructing a color space correlation matrix for each object including its color space correlation across all sampled layers of all CNNs. The resulting correlation matrix was then averaged across objects and visualized using MDS. Each trajectory is a different network, each dot is a different layer, and the hollow dots denote the first layer of each network. b. Exact correlation values for the between network correlations in the first layer of each network (top) and the penultimate layer of each network (bottom) for both versions of the objects (textured

objects left, silhouette objects right). Correlations are very similar across networks in the first layer, but diverge by the end of processing.



Supplementary Figure 3.3. Color space comparisons for AlexNet trained with ImageNet images and with 10 different random-weight initializations. a. MDS plots depicting color space correlation across different layers for ImageNet trained AlexNet and 10 instances of AlexNet with random-weight initializations, plotted separately for the two versions of the objects (copied from Figure 3c for convenience). b. Exact correlation values for the between network

correlations in the first layer of each network (top) and the penultimate layer of each network (bottom) for both ImageNet image trained AlexNet and the 10 instances of AlexNet with random-weight initializations, done separately for the two versions of the objects. Other details are the same as Supplementary Figure 2. Correlations are very similar across the trained and untrained networks in the first layer and remain similar across all the untrained networks, but differ substantially between the trained and untrained networks by the end of processing.



Supplementary Figure 3.4. Color space comparisons for ResNet-50 trained with ImageNet images, trained with stylized ImageNet images, and with 10 random-weight initializations. a. MDS plots depicting color space correlation across different layers for original ImageNet trained

ResNet-50, stylized ImageNet trained ResNet-50, and 10 instances of ResNet-50 with random-weight initializations, plotted separately for the two versions of the objects (copied from Figure 3c for convenience). b. Exact correlation values for the between network correlations in the first layer of each network (top) and the penultimate layer of each network (bottom) for original ImageNet image trained ResNet-50, stylized ImageNet image trained ResNet-50 and the 10 instances of ResNet-50 with random-weight initializations, done separately for the two versions of the objects. Other details are the same as Supplementary Figure 2. Correlations are relatively similar across the trained and untrained networks in the first layer, but diverge substantially between the trained and untrained networks, and between the 10 different instances of the untrained networks by the end of processing.

Chapter 4: Nonlinear Mixed Selectivity Coding for Stimulus and Task Across the Human Visual System

Humans can perform a seemingly unbounded variety of tasks over a similarly limitless range of stimuli, both in daily life and in the psychology lab, which raises the question of what neural code undergirds this flexibility. Recent work has demonstrated that neurons exhibiting *nonlinear mixed selectivity*—that is, neurons whose tuning to multiple variables shows interaction effects—may play an important computational role in implementing flexible behavior by increasing the dimensionality of the coding space of a neural population, which allows a far broader range of readout functions to be performed over the population's activity (Fusi et al., 2016; Rigotti et al., 2013). However, there currently exists no method for noninvasively identifying this coding motif. As such, most studies of nonlinear mixed selectivity coding have studied it in macaques (Dang et al., 2020; Parthasarathy et al., 2017) or rodents (Hardcastle et al., 2017; Nogueira et al., 2021), with only one group examining it in the human brain, via intracranial recording (Zhang et al., 2017, 2020). Furthermore, none of the existing methods permit recording from many brain regions simultaneously, complicating efforts to compare the prevalence of this coding scheme across regions.

In this work, we devised a novel method, *pattern difference decoding*, for measuring nonlinear mixed selectivity tuning noninvasively in the human using fMRI. Our approach exploits the sensitivity of multivoxel pattern analysis (MVPA) decoding approaches to test for whether neural activity in a region exhibits interaction effects, which may be subtle and heterogeneous across a region, making it infeasible to identify them using univariate methods. We applied this method to data from four experiments (Figure 4.1) to compare the extent to

which different visual regions encode task and stimulus information in a nonlinear manner. Specifically, we examine early visual cortex, the lateral occipital complex, and the superior IPS, a region that has been shown to encode both stimulus and task information (Jeong & Xu, 2016; Vaziri-Pashkam & Xu, 2017; Xu, 2018). We found that while all three regions exhibit nonlinear mixed selectivity in some cases, only the superior IPS reliably shows it across all tasks we examined, even when spatial, object, and feature-based attentional demands were equated across the pair of tasks being examined. This approach thus makes it feasible to study nonlinear mixed selectivity coding in the human brain, and reveals important differences in neural coding principles across the human visual system.

All four experiments involved viewing stimuli from eight different categories, and performing two different tasks, with the pair of tasks varying across experiments. In the first three experiments (each with n = 7), participants either performed a oneback on stimulus exemplar, or a oneback on color. In the Colored Background experiment this color covered both the stimulus and the background, in the Colored Dots experiment the stimulus and background were grey but colored dots were superimposed on the stimulus, and in the Colored Object experiment the object itself was colored. experimentFinally, the Oddball/Oneback experiment, participants either performed an oddball task where they responded if the category of a stimulus did not match that of the surrounding block, or a oneback task on object exemplar. Thus, the pair of tasks in each experiment differed with respect to how they varied the participant's attentional demands: in the Colored Background experiment the participant either attended to the object itself or to the background (i.e., locus of spatial attention differed), in the Colored Dots task the participant either attended to the object itself or to dots superimposed on the object (equating the locus of spatial attention while varying the attended item), in the Colored Object task the

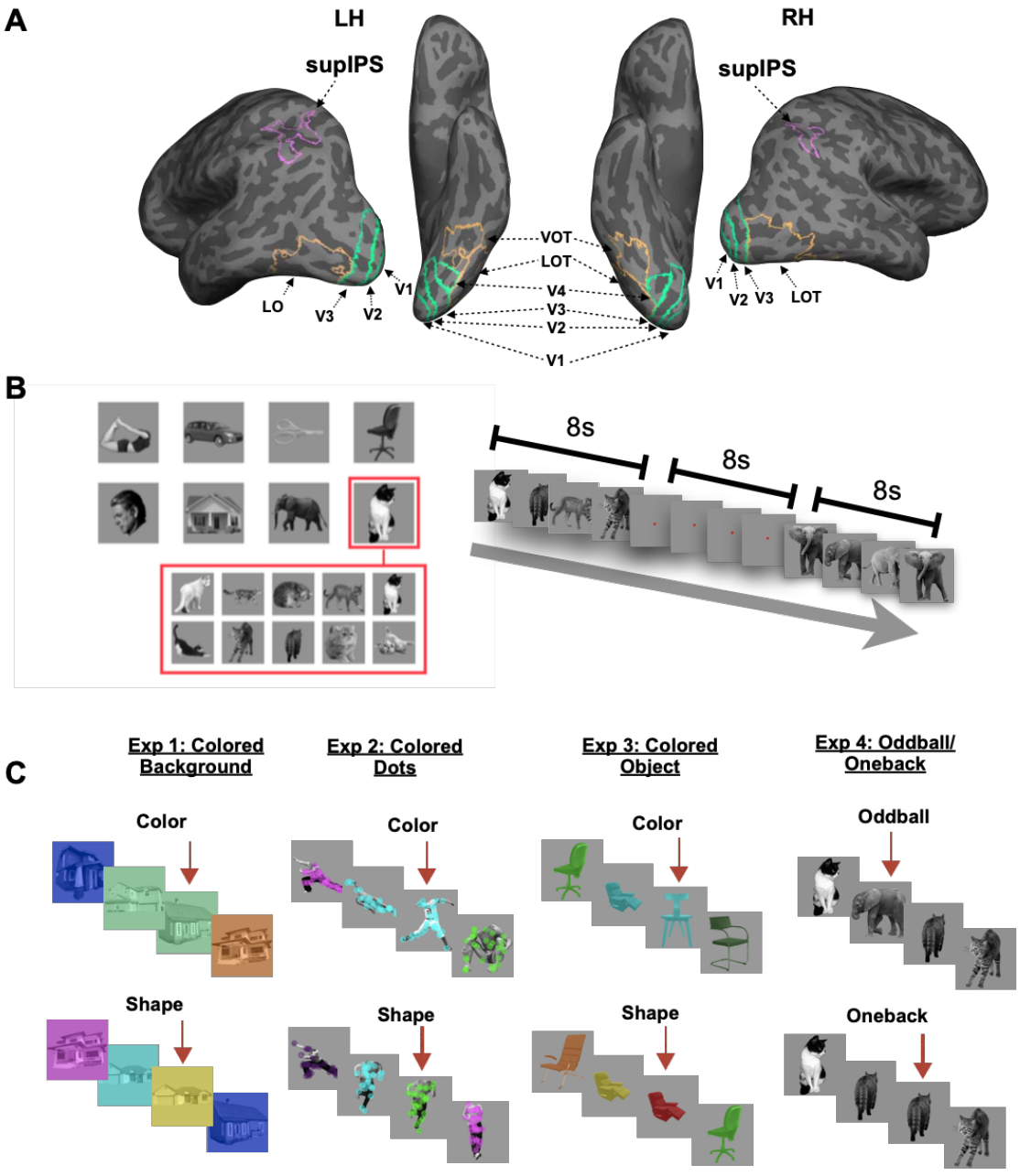


Figure 4.1. Regions examined and experimental design. **A.** Regions of interest analyzed in the experiment included retinotopic early visual regions V1, V2, V3, and V4 (lime), object-sensitive ventral stream regions LO and pFs (orange), and the superior IPS (magenta), a region implicated in task-relevant visual processing, defined using a short term memory localizer. **B.** All four experiments involved presenting stimuli from 8 different categories in 8s blocks, where multiple exemplars of each category were presented in each block. **C.** The pair of tasks used in each of the four experiments. Experiments 1-3 involved either a oneback on color (with the placement of the color varying across experiments) or a oneback on stimulus exemplar; Experiment 4 involved either a category oddball or a oneback on stimulus exemplar.

participant always attended to the same object but varied which feature (color or shape) they attended to, and in the Oddball/Oneback task the participant attended to the same location, object, and features, but differed in the task they performed over those features. Manipulating attention in this way enabled us to better understand the conditions in which nonlinear mixed selectivity emerges in different visual regions.

To examine the extent to which each brain region encodes task and stimulus information in an interactive manner (indicative of nonlinear mixed selectivity tuning), we developed a method we call *pattern difference decoding* (Figure 4.2). Specifically, we extracted the patterns of BOLD activity associated with each stimulus/task combination, z-normalized the patterns from each trial to equate their mean and variance, and computed the BOLD pattern *difference vectors* between pairs of conditions with the same task, but different category (e.g., Elephant/Oddball and Cat/Oddball). We then used these category difference vectors to train a support vector machine (SVM) to determine whether these category difference vectors were discriminable across task—for instance, whether Elephant/Oddball - Cat/Oddball is discriminable from Elephant/Oneback - Cat/Oneback. The opposite analysis (decoding task differences across task) was also performed, with both directions of the analysis being performed for every pair of categories, and the resulting decoding accuracies were averaged. Above-chance decoding implies that at least some voxels in the population exhibit an interaction effect (i.e., a difference-of-differences) in their tuning profile to different task/category populations; using an SVM allows these differences to be detected in distributed patterns of activity even when they are small and heterogeneous in individual voxels, in the same way that the ordinary use of an SVM in MVPA decoding allows potentially small and heterogeneous main effects to be detected in a brain region. In order to equate the number of voxels in each ROI and choose the most

discriminative voxels, the 300 voxels showing the highest F-value in a two-way ANOVA (task X category) in the training data for each iteration of the classifier were used in the analysis. T-tests (one-sample, one-tailed) were used to compare decoding in each ROI and experiment to chance, and partially-overlapping t-tests (Derrick et al., 2017) were used to compare decoding between experiments within each ROI.

Figure 4.2 depicts the results of the analysis. In order to simplify comparison, the results from regions V1-V4 were averaged within each participant, as a two-way ANOVA revealed no significant interaction between experiment and ROI ($F(3, 36) = .65$, $p = .69$), suggesting similar patterns of results; similarly, LO and pFs were averaged ($F(2, 12) = .52$, $p = .61$ for experiment-by-ROI interaction). V1-V4 exhibited significant pattern difference decoding in the Colored Background and Colored Dots experiments ($ts > 3.6$, $ps < .01$), but not in the Colored Object or Oddball/Oneback experiments ($ts < .60$, $ps > .28$). LOC showed significant pattern difference decoding in every experiment except the Oddball/Oneback experiment ($ts > 3.26$, $ps < .01$ for all except Oddball/Oneback; $t(12) = .39$, $p = .35$ for Odball/Oneback). However, supIPS showed above-chance pattern difference decoding in all four experiments ($ts > 2.37$, $ps < .03$). Within each ROI, pairwise t-tests indicated that decoding for the experiments exhibiting above-chance decoding was significantly greater than decoding for the experiments that did not achieve significance ($ts > 2.56$, $ps < .03$). Since the null decoding result for the Oddball/Oneback experiment in LOC interestingly distinguishes it from supIPS, we performed a power analysis to estimate whether the sample size was adequate in this experiment; we found that assuming a true effect size for LOC in the Oddball/Oneback experiment equal to that of the Colored Object experiment, the power was 99.7%, suggesting that this null result was not due to inadequate power.

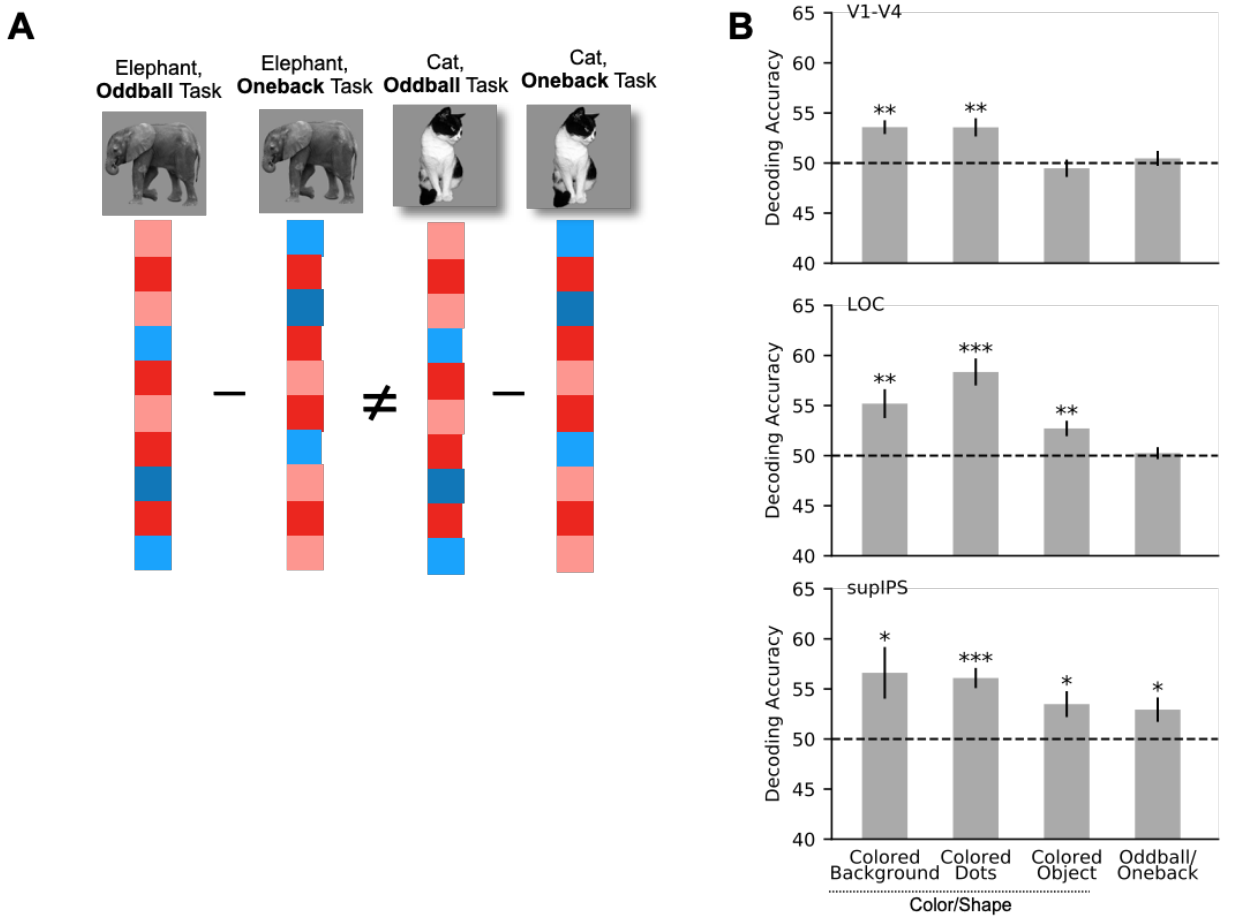


Figure 4.2. Analysis approach and results. **A.** Illustration of pattern difference decoding analysis. BOLD patterns associated with each category/task combination were extracted from each ROI, and the pattern differences associated with pairs of conditions with the same category but different tasks were computed. A support vector machine was then trained to distinguish these task difference vectors across different object categories (for every possible pair of categories) to assess whether there exists evidence for interactive coding of category and task. The "reverse" analysis (discriminating category differences across tasks; e.g., Elephant/Oddball — Cat/Oddball vs. Elephant/Oneback — Cat/Oneback) was also performed, and the results were averaged. **B.** Results of pattern difference decoding for each ROI and experiment. *** $p < .001$, ** $p < .01$, * $p < .05$, one-tailed t-tests against chance decoding.

Thus, these three sectors of the human visual system exhibit differing profiles of nonlinear mixed selectivity to task/stimulus combinations. All three regions show it for the Colored Background and Colored Dots experiments, where the two tasks vary the locus of spatial or object-based attention; only supIPS and LOC show it in the Colored Object

experiment, which equates spatial and object-based attention while varying feature-based attention; and only supIPS shows it in the Oddball/Oneback experiment, where spatial, object-based, and feature-based attention were equated, with only the task differing. These results are consistent with a model in which nonlinear mixed selectivity can arise for different reasons in different regions. For example, if a neural population only encodes the details of a stimulus under the right attentional conditions (whether spatial, object-based, or feature-based), this would manifest itself as an interaction effect when the relevant form of attention is manipulated. Other regions, however, may exhibit nonlinear mixed selectivity even when attentional demands are equated, suggesting that these regions may genuinely be encoding task-sensitive encoding of visual information, above and beyond mere attentional gating of certain visual details; this may offer a clue as to the neural coding scheme used by supIPS to encode task-sensitive visual representations. More broadly, the *pattern difference decoding* analysis developed here can be used to probe for subtle, heterogeneous, distributed interaction effects not just in other fMRI studies, but for any neural recording modality yielding multivariate data, and indeed in any scientific data where identifying distributed interaction effects is of interest.

Supplemental Methods

Experiments 1-3 have been reported in detail in Vaziri-Pashkam and Xu (2017); throughout this section, the methods are reproduced for the reader's convenience.

Participants

All participants had normal color vision and normal or corrected-to-normal visual acuity, and were between 18 and 35 years old. In experiments 1-3, a total of 7 healthy adults (4 females) participated in all three studies. In experiment 4, 13 healthy adults (7 females) participated. Four participants partook in all four studies. The experiments were approved by the Committee on the Use of Human Subjects at Harvard University.

Experimental Design and Procedures

Experiment 1: Colored Background

In this experiment, we used gray-scaled object images from 8 object categories (faces, bodies, houses, cats, elephants, cars, chairs, and scissors). These categories were chosen as they covered a good range of natural object categories encountered in our everyday visual environment and were the typical categories used in previous investigations of object category representations in ventral visual cortex (e.g., Haxby et al., 2001; Kriegeskorte et al., 2008). For each object category, 10 unique exemplar objects were selected. These exemplars varied in identity, pose (for cats and elephants), expression (for faces), and viewing angle to reduce the likelihood that object category decoding would be driven by the decoding of any particular exemplar. Objects were placed on a light gray background and covered with a semitransparent colored square subtending 9.24° of visual angle. Thus, both the object and the background

surrounding the object were colored. On each trial, the color of the square was selected from a list of 10 different colors (blue, red, light green, yellow, cyan, magenta, orange, dark green, purple, and brown). Participants were instructed to view the images while fixating at a centrally presented red dot subtending 0.46° of visual angle.

In a block design paradigm, participants performed a one-back repetition detection task when the exact same object exemplar repeated back to back or the exact same color repeated back to back. In each block, 10 colored exemplars from the same object category were presented sequentially, each for 200 ms followed by a 600 ms fixation period between the images. In half of the runs, participants attended to the object shapes and ignored the colors, and pressed a response button whenever the same object repeated back to back. Two of the objects in each block were randomly selected to repeat. In the other half of the runs, participants attended to the colors and ignored object shapes and detected a one-back repetition of the colors, which occurred twice in each block.

Each experimental run consisted of 1 practice block at the beginning of the run and 8 experimental blocks with 1 for each of the 8 object categories. The stimuli from the practice block were chosen randomly from 1 of the 8 categories, and data from the practice block were removed from further analysis. Each block lasted 8 s. There was a 2 s fixation period at the beginning of the run and an 8 s fixation period after each stimulus block. The presentation order of the object categories was counterbalanced across runs for each task. To balance for the presentation order of the two tasks, task changed every other run with the order reversed halfway through the session so that, for each participant, one task was not presented on average earlier than the other task. Each participant completed one session of 32 runs, with 16 runs for each of the two tasks. Each run lasted 2 min 26 s.

Experiment 2: Colored Dots

The stimuli and paradigm used in this experiment were similar to those of Experiment 1, except that, instead of both the object and the background being colored, a set of 30 semitransparent colored dots, each with a diameter subtending 0.93° of visual angle, were placed on top of the object, covering the same spatial extent as the object. This ensured that participants attended to approximately the same spatial envelope whether or not they attended the object shape or color in the two tasks. Other details of the experiment were identical to those of Experiment 1.

Experiment 3: Colored Object

The stimuli and paradigm used in this experiment were similar to those of Experiment 1, except that only the objects were colored, making color more integrated with shape in this experiment than in the previous two. Participants thus attended to different features of the same object when doing the two tasks. Other details of the experiment were identical to those of Experiment 1.

Experiment 4: Oddball/Oneback

Like the previous three experiments, this experiment involved presenting stimuli from each of 8 categories in a block design. The categories and stimulus exemplars used were exactly the same, except that no colors were associated with the stimuli as in the previous experiments. In this experiment, participants performed one of two tasks, which was announced at the beginning of each run. In the *Oddball Task*, participants responded if a given stimulus exemplar did not match the category of the surrounding block (e.g., if there was a cat in a block of

elephants). The oddball stimulus never occurred earlier than the third trial, so as to allow the participant to determine what the "true" category of the block was. In the *Oneback task*, participants responded if there was an exact repeat of stimulus exemplar (e.g., the same face twice in a row). Each run contained ten blocks. Eight of these blocks did not contain the task, and two did; however, the participant did not know in advance which would contain the task, such that they had to remain prepared to respond during the whole run. The two blocks with the task were discarded during analysis to remove motor contamination of the neural responses. The eight blocks without the task spanned all eight stimulus categories. The two blocks with the task each contained a randomly chosen category, subject to the constraint that the block category was different from the categories of the two blocks before and after that block.

Stimulus timing was otherwise identical to experiments 1-3, with the exception that there was a 12s fixation (instead of 2s fixation) to begin each run, for a total run duration of 2:52. Participants completed two practice blocks prior to the first TR being collected to remind them of the task for that run, while the scanner reached equilibrium.

Participants performed 10 runs of each task, with the task alternating each run with the exception of the tenth and eleventh runs, where a task repeated twice in a row; this ensured that no task was presented earlier or later on average than the other task, which might have introduced confounds due to scanner drift or participant fatigue. The two possible task orders were counterbalanced across participants. To determine the block order in each run, a 10x10 balanced Latin square (i.e., a square with different numbers in every row and column, in which no sequence of two values occurs more than once in the square) was randomly generated for each task for a given subject, where numbers 1-8 (one for each category) corresponded to blocks

without the task and 9-10 corresponded to task blocks. This balancing scheme ensured that all categories and task blocks occurred equally often in any given part of the run.

Localizer experiments

All the localizer experiments conducted here used previously established protocols, and the details of these protocols are reproduced here for the reader's convenience.

To localize topographic visual field maps, we followed standard topographic mapping techniques (Sereno et al., 1995; Swisher et al., 2007) and optimized our parameters to reveal the maps in parietal cortex (Swisher et al., 2007). A 72° polar angle wedge swept across the entire screen, with a sweeping period of 55.467 s and 12 cycles per run. The entire display subtended 23.4 × 17.6° of visual angle. The wedge contained a colored checkerboard pattern that flashed at 4 Hz. Participants were asked to detect a dimming in the polar angle wedge. Each participant completed 4–6 runs, each lasting 11 min and 5.6 s.

To identify superior IPS, we used a visual short-term memory (VSTM) paradigm first developed by Todd and Marois (2004). Two slightly different versions of this localizer were used across participants: all participants except for six in Experiment 4 completed one version, and the remaining six completed an alternative version (due to refinement of the localizer paradigm over time). Both versions used an event-related design (as in Xu and Jeong, 2015) where participants viewed a sample display of several objects, and after a delay, judged whether a new probe object matched one of the sample objects. In the first version, the sample display included 1–4 everyday objects, and the probe object appeared in the same location as one of the sample objects. A match occurred in half of the trials. Objects were gray-scaled images from four categories (shoes, bikes, guitars, and couches; alternative version included couches, lamps,

guitars, scissors, shoes, teapots, or umbrellas). In the sample display, objects could be placed above, below, to the left, or to the right of the central fixation ~4.0° away from the fixation (center to center). Four dark-gray rectangular placeholders, subtending 4.5° × 3.6°, marked all the possible object positions and were always present during the trial (rectangles not present in alternative version). The entire display subtended 12° × 12°. Each trial lasted 6 s and consisted of a fixation period of 1000 ms, a sample display period of 200 ms, a delay of 1000 ms, a test display period of 2500 ms in which participants provided their responses, and a feedback period of 1300 ms. Each run contained 15 trials for each set size and 15 fixation trials in which only the fixation dot appeared for 6 s. The trial order was predetermined using a counterbalanced trial history design (Todd and Marois, 2004; Xu and Chun, 2006). Two filler trials appeared at the beginning and one at the end of each run for practice and trial history balancing purposes. Each participant completed two runs of this localizer run, each lasting 8 min.

The alternative localizer was largely similar, with a few small changes. The set size varied from 1-6 instead of 1-4 (categories included couches, lamps, guitars, scissors, shoes, teapots, or umbrellas), and sample objects could be present in one of eight locations (top, bottom, left, right, or one of the four intermediate diagonals, forming an octagon of possible placements), this time without rectangular placeholders. Sample stimuli were now presented sequentially (rather than simultaneously) over the course of 1200ms, each for a 100ms slot randomly inserted into this 1200ms window (50ms between slots). The retention period lasted 950ms, The test display lasted 2000ms (instead of 2500ms), the feedback period lasted 1600ms (instead of 1300ms), and the fixation period lasted 250ms; thus, the total trial length was still 6s. The probe object now appeared in the center of the display, rather than in one of the positions previously

occupied by one of the objects. Participants completed three runs of this localizer, each lasting 8 minutes as in the other version.

MRI Methods

MRI data were collected using a Siemens MAGNETOM Trio, A Tim System 3T scanner, with a 32-channel receiver array head-coil. Participants lied on their back inside the MRI scanner and viewed the back-projected display through an angled mirror mounted inside the head coil. The display was projected using an LCD projector at a refresh rate of 60 Hz and a spatial resolution of 1024 × 768. An Apple Macbook Pro laptop was used to generate the stimuli and collect the motor responses. All stimuli were created using MATLAB and Psychtoolbox (Brainard, 1997), except for the topographic mapping stimuli which were created using VisionEgg (Straw, 2008).

A high-resolution T1-weighted structural image (1.0 × 1.0 × 1.3 mm) was obtained from each participant for surface reconstruction. For the first version of the superior IPS localizer scans, 24 axial slices parallel to the AC-PC line (5 mm thick, 3 × 3 mm in-plane resolution with no skip) were collected covering most of the brain, except the anterior temporal and frontal lobes (TR = 1.5 s, TE = 29 ms, flip angle = 90°, matrix = 72 × 72). For the second version of the superior IPS localizer scans, 64 interleaved axial-oblique slices taken 25 degrees towards coronal from ACPC alignment (2.3 mm isotropic voxels) were collected covering the whole brain (TR = 650ms, TE = 34.8ms, flip angle = 52°, matrix = 90 × 90). For topographic mapping, 42 slices (3 mm thick, 3.125 × 3.125 mm in-plane resolution with no skip) just off parallel to the AC-PC line were collected covering the whole brain (TR = 2.6 s, TE = 30 ms, flip angle = 90°, matrix = 64 × 64). Different slice prescriptions were used here for the different localizers to be consistent with

the parameters we used in previous studies. Because the localizer data were projected into the volume view and then onto individual participants' flattened cortical surface, the exact slice prescriptions used had minimal impact on the final results. For all functional scans, T2*-weighted gradient-echo, echo-planar sequences were used. For Experiments 1-3, 33 axial slices parallel to the AC-PC line (3 mm thick, 3 × 3 mm in-plane resolution with 20% skip) were collected covering the whole brain (TR = 2 s, TE = 29 ms, flip angle = 90°, matrix = 64 × 64). For Experiment 4, 84 axial slices parallel to the AC-PC line (1.5mm isotropic) were collected covering the whole brain (TR = 2 s, TE = 30 ms, flip angle = 80°, matrix = 136 × 136).

Data analysis

The same analysis pipeline was applied in all four experiments. fMRI data were analyzed using FreeSurfer (https://surfer.nmr.mgh.harvard.edu), fsfast (Dale et al., 1999), and in-house Python scripts. fMRI data preprocessing included 3D motion correction, slice timing correction, and linear and quadratic trend removal. No spatial smoothing was applied.

ROI Definitions

Retinotopic maps in early visual cortex were defined using the procedure outlined by Swisher et al. (2007). We identified V1, V2, V3, and V4 for each participant. The union of these ROIs was taken to form a composite ROI, V1-V4. Following Todd and Marois (2004), superior IPS was identified in each participant using that participant's behavioral VSTM capacity K score (Cowan, 2001) (Fig. 2B). The statistical threshold for selecting superior IPS voxels was set to $p < 0.001$ (uncorrected). This yielded an ROI with between 27 and 859 voxels (mean 267 voxels)

for Experiments 1-3, and an ROI with between 604 and 4138 voxels (mean 2238 voxels) for Experiment 4.

Generalized Linear Model

A generalized linear model (GLM) was used to estimate the responses of each voxel on every trial of the experiment. The GLM analysis was run in native voxel space, using the three motion parameters as nuisance regressors. Beta weights were estimated for each trial, where each regressor consisted of a single 8s boxcar function convolved with the default SPM hemodynamic response function. For Experiments 1-3, this resulted in a total of 256 beta values per voxel, and in Experiment 4, this resulted in a total of 160 beta values per voxel.

Pattern Difference Decoding

We devised a novel analysis technique to examine the extent to which various ROIs exhibit an interaction effect between combinations of task and stimulus variables, which we call *Pattern Difference Decoding*. The method examines whether voxel pattern differences between two tasks (e.g., the oddball and oneback task) are discriminable across different categories, and whether differences between two categories are discriminable across the two tasks. Importantly, this method allows for the detection of interaction effects that may be small and heterogeneous across voxels.

To implement the analysis, the patterns of beta values associated with the two different tasks in an experiment were extracted for two different categories, and z-normalized to equate their mean and standard deviation. Next, the *pattern difference vectors* between the two tasks were computed by subtracting one vector from the other within each category, resulting in, e.g., a

set of difference vectors for the difference between the oddball and oneback task for elephants, and a set of difference for the difference between the oddball and oneback task for cats. These difference vectors were then used to train an SVM to decode the task difference vectors for one category from the task difference vectors for the other category, using leave-two-runs-out cross-validation. The reverse analysis was also run: the pattern difference vectors between the two *categories* were computed within each *task*, and then these difference vectors were used to train an SVM to decode the category difference for one task from the category difference vectors for the other task, also using leave-two-runs-out cross-validation. The results of these two analyses were averaged, since they both measure the presence of category-by-task interaction effect within voxels. This analysis was performed across every possible pair of categories, and these results were further averaged.

To increase power and better equate the number of voxels across ROIs, for each training iteration of the classifier (i.e., each left-out pair of runs), voxel selection was performed to identify the most sensitive voxels (prior to z-normalizing). To do this, a univariate ANOVA was performed on each voxel across the runs used to train the classifier, and the 300 voxels showing the highest interaction effect were selected to train the classifier. This analysis does not constitute "double dipping", since it was applied only to the training but not the test runs.

After computing the mean classification accuracy in this manner for every experiment, participant, and ROI, various analyses were performed to examine how pattern difference decoding varies across experiments and ROIs. First, results from V1, V2, V3, V4 were pooled into a single V1-V4 macro-ROI, after a two-way (experiment X ROI) ANOVA revealed no significant interaction between experiment and ROI ($F(3, 36) = .65$, $p = .69$), suggesting a similar decoding profile in each region; the same was true of LO and pFs ($F(2, 12) = .52$, $p = .61$

for experiment-by-ROI interaction), so they were similarly averaged together into an LOC macro-ROI. After this pooling, there were thus three ROIs: V1-V4, LOC, and supIPS. One-tailed t-tests were used to compare decoding in each region to chance. Additionally, partially-overlapping t-tests (Derrick et al., 2017) were used to compare decoding between experiments in each ROI; this version of the t-test accounts for cases when only some subjects are shared between conditions, which was appropriate since only some of the participants in the Oddball/Oneback experiment participated in the other three experiments.

In order to better interpret some of the null results we observed, specifically the fact that LOC did not exhibit above-chance pattern difference decoding in the Oddball/Oneback experiment, we conducted a power analysis to assess the adequacy of the sample size we used in this experiment. Specifically, we estimated an effect size by computing the effect size of the most similar experiment to this one, the Colored Object experiment (since both experiments had participants attending to the object in both of their tasks), and used this effect size, in combination with the sample size and alpha level, to compute the power.

Chapter 5: Conclusion

Throughout this thesis, I have investigated the question: how might an intelligent visual system, whether biological or artificial, encode combinations of visual features (using color and form coding as a case study), and bring these features to bear in task-relevant processing? In Study 1, I undertook a comprehensive examination of how color and form are encoded across the ventral visual pathway. In Study 2, I examined the joint coding format of color and form in convolutional neural networks. In Study 3, I extended one of the analysis frameworks I developed for studying color/form coding in the human brain (*pattern difference decoding*) to study the joint representation of stimulus and task information across the human visual system, providing a proof-of-principle that *nonlinear mixed selectivity coding*--a neural coding scheme that confers distinct computational benefits--can be studied noninvasively in the human brain. In this final chapter, I summarize how these studies advance our understanding of multifeature coding, identify what remains to be learned, and chart some future directions.

Color and Form Processing Across the Ventral Visual Pathway: An Updated View

In Study 1, I examined how color and form information are encoded both in early visual cortex, and in higher-level ventral stream regions defined based on their sensitivity to the presence of either color or form information. I took the approach of using stimuli that were carefully controlled to equate their luminance and retinotopic footprint, and manipulated either a simple form feature (orientation) and a more midlevel form feature (curvature). In addition to examining whether information about each feature was present in each region, I also performed several analyses to document the coding relationship between color and form: that is, are these features encoded in an independent, additive manner, or in a more interactive fashion?

In summary, I found that nearly every region I examined exhibited above-chance decoding of both color and form information, with the sole exception of color-sensitive regions in central ventral temporal cortex. Additionally, this color decoding in retinotopic and form sensitive regions was not solely due to any overlap with color-sensitive regions: above-chance color decoding remained even when their overlap with color-sensitive regions was removed (albeit more weakly), and moreover, area LO exhibited above-chance color decoding despite no overlap with the color-sensitive regions whatsoever. That said, despite the anatomical commingling of color and form information, in most cases we examined it appeared to be encoded in an additive, independent manner (i.e., with mostly "main effects"); the exception was found for the conjunction of color and orientation in early visual regions, where we evidence for interactive color/orientation coding with two independent datasets and several different methods.

These data in some cases corroborate and illuminate past findings in the literature, but in some cases interestingly contrast with them. One potentially provocative finding is the null result for form decoding in the central ventral color region; this region stands alone, in that it exhibited significant decoding for one feature but not the other. This finding is an interesting contrast with results from a study that examined how the (plausibly) homologous regions in macaques respond to different object categories presented in different colors, which found that neurons in this central region encode not just color, but also appear to encode the shape of stimuli, as well as exhibiting a color-shape interaction (Chang et al., 2017). How can we explain this difference? The differences in species, stimuli, and recording method make several deflationary explanations possible: perhaps this region functions differently in macaques and humans; perhaps this region encodes shape for complex, naturalistic stimuli but not the simple, artificial stimuli used in this study; or perhaps shape information is present in the region in humans, but the tuning

distribution of neurons across voxels is not sufficiently heterogeneous for fMRI to detect (work in other domains has shown this to be a genuine concern: one study found that face identity information is detectable in face-sensitive regions when single neuron recordings are used, but not when fMRI is used; (Dubois et al., 2015). On the other hand, the use of naturalistic object stimuli by Chang et al. (2017) does introduce a potential confound: since these stimuli were not precisely controlled in the retinotopic distribution of color information (whereas the stimuli in my study precisely controlled this factor using the phase-alternating spiral and tessellation stimuli), it remains theoretically possible that the "shape" tuning they observe could arise from a spatial representation of the colors in the image rather than shape per se. For instance, imagine a neuron whose tuning reflects the color present in its receptive field (i.e., it prefers red to blue), and the area subtended by that color (i.e., it prefers a large blob of red to a tiny speck of red). If the different values of "shape" queried involve the colored regions of the stimulus tending to fall in different regions of the visual field, it could lead to the appearance of "shape" decoding where there is none. Under this interpretation, perhaps this region truly contains no shape information at all *per se*. Consistent with this possibility, Lafer-Sousa et al. (2016) found that this region responds equally to coherent objects and scrambled stimuli. In a currently ongoing study, I am presenting different object categories in different colors, and also in two different sizes, which may help to clarify the nature of the (either) shape or spatial information present in this region; for example, if size decoding succeeds in this region, but shape decoding within a size does not, it would be consistent with a "spatial color representation but no shape representation" view of this region.

While various ambiguities remain to be worked out for this region, it is tempting to speculate about the role it might play. The evidence from achromatopsia and from visual search

seem to imply some level of representation at which color is independently accessible from shape: might this central color region encode the separate feature map for color implied by Feature Integration Theory (Treisman & Gelade, 1980)? Several observations are consistent with this possibility. First, this region spatially coincides with retinotopic regions VO1 and VO2 (Brewer et al., 2005), suggesting it may contain the spatial information required to implement a retinotopic representation of color information. Second, this region appears to bear an especially tight connection with conscious color perception: this region but not V4 responds to color afterimages (Hadjikhani et al., 1998), and an intracranial study in this region found that the color that optimally drove neurons in the recording site roughly matched the subject's illusory color percept when this region was stimulated (Murphey et al., 2008). Third, although it is difficult to match regions across studies, damaging cortex in this vicinity can lead to achromatopsia with relatively spared form processing (Bouvier & Engel, 2006). Altogether, then, if there really does exist a specific brain region encoding a retinotopic, shape-free, conscious color representation—a region that can make a red target among blue distractors visible at a glance, a region whose disruption turns the world monochrome—then it is worth considering that it might be exactly this one (though the notion that our color qualia inhere in a single region may turn out to be overly simplistic).

Notably, the converse situation is different: both of the form-sensitive regions analyzed in the study, LO and pFs, contain information about color as well, despite the fact that LO exhibits zero overlap with the color-sensitive cortex we examined, and despite the fact that color decoding in pFs persists when color-sensitive voxels are removed. What should be made of this, given that the various agnosias arising from damage to these regions appear to leave color perception intact, at least in the experimental paradigms generally used (Benson & Greenberg,

1969; Goodale & Milner, 2004)? The most deflationary hypothesis is that the color information decodable in these regions is merely epiphenomenal: perhaps it is inherited from the upstream regions (such as V2 and V4) feeding into them, and this "spillover" information does not interfere enough with the shape-processing functions of these regions to make it worth discarding. A way of probing this question would be to document the color coding space of these regions more thoroughly using a wider variety of colors; if the color similarity space encoded by these regions bears little resemblance to behavioral color performance, it would support the epiphenomenal view. Another piece of evidence for the epiphenomenon view would be if these regions fail to exhibit hue decoding that generalizes across luminance (which is one technique that has been used to characterize the color information present in a region; e.g., Conway et al., 2007). Alternatively, however, the color information present in these might be put to use in shape processing, without itself being accessible to conscious perception. As summarized in the introduction, there are various demonstrated cases where color information seems to be used in extracting shape information, such as in extracting color-defined contours (Barbur et al., 1994; Heywood et al., 1991, 1998; Mandelli & Kiper, 2005); an interesting further study, then, could be to compare how these regions encode luminance versus color-defined shapes; for instance, does encoding of shape in this regions exhibit cross-decoding between cases where these shapes are defined with luminance information and cases where they are defined with isoluminant color information? That said, this could be achieved without color information per se being present in these regions: for instance, the color-luminance cells in V1 (Johnson et al., 2001), which respond to both color- and luminance-defined edges, could pass along this edge information, stripped of information about its specific origin (whether chromatic or luminance-based), allowing higher-level regions to use color-derived information in their shape computations without themselves

encoding color. Whatever the case, the results from this study unequivocally confirm the presence of distributed color information in these shape-sensitive regions, setting the stage for future work to probe the nature of these color representations in more detail.

All that said, despite the *presence* of color information in the shape regions, and of shape information in one of the color regions, the relative *levels* of decoding in these regions varied in a way that reflected their univariate preference: specifically, LO and pFs showed higher decoding of curvature than of color (though this was only true for the latter when its overlap with the color-sensitive regions was removed), and the color-sensitive regions showed higher decoding of color than of either form feature. In this sense, then, the data are indeed suggestive of specialization for either feature within these regions.

The analyses in this study that decode color information in V4 and pFs when their overlap with the color-sensitive regions is removed partially corroborate past work: Conway et al. (2007) found that neurons within "globs" in macaque V4 and IT defined based on their sensitivity to color information had much stronger color tuning and weaker form tuning than "interglob" neurons, with the latter showing no evidence of color tuning at all. By contrast, our analyses demonstrated that color information is present in these regions even when the overlap with color-sensitive regions is removed, although it does significantly drop. All the same, however, our results are consistent with an organization where there is some clustering of color-tuned neurons in these regions, although the presence of color tuning does not appear to be "all or none".

As mentioned, in addition to decoding single features, we performed various analyses where we probed the *format* of the joint color/shape representations in the regions that encoded both: that is, are these features encoded in an independent, additive, linear manner, or in a more

interactive manner? Using three different methods and two different datasets, we found evidence that the joint representation of color and form seems to be predominantly additive, with one exception: color and orientation appear to exhibit a degree of interactive coding in early visual cortex. This finding is consistent with psychophysical results suggesting automatic coding for color/orientation conjunctions (Holcombe & Cavanagh, 2001; McCollough, 1965), as well as several neural studies suggestive of this (Engel, 2005; Seymour et al., 2010). However, the present work suggests that these interactive color/form representations may be limited in two ways: first, they are present in early visual cortex but not in the higher-level ventral stream regions that we probed; second, they are present for the conjunction of color with a simple form feature, orientation, but not with a salient mid-level form feature, curvature.

These findings have several important implications. First, they suggest that a "conjunction detector" strategy—in which there exist neural populations tuned nonlinearly to particular feature combinations—is at least not the primary method used by the brain to encode color/form conjunctions. This is sensible on logical grounds: clearly the brain cannot have a dedicated cell for every possible color/form combination that could potentially occur, given how many possible values there are for each feature alone. Thus, this "conjunction detector" strategy could be a format that only applies in special cases. Local orientation is a more tractable feature than many other form features, in that it can be parametrized by a single value (orientation), and so there is less of a "combinatorial explosion" involved in positing neural populations that are tuned to a reasonable tiling of the range of possible orientation/color conjunctions. There is therefore a feasibility argument that the brain *can* wire up color/form conjunction detectors for certain form features but not others. However, this leaves open the question of what functional role this seemingly rather limited coding format might play; that is, why *should* this coding

format exist? The aforementioned psychophysical demonstrations (Holcombe & Cavanagh, 2001; McCollough, 1965) suggest that this automatic coding of color/orientation conjunctions is evident in visual experience under carefully contrived experimental circumstances, but plainly the brain did not wire up these neurons purely so they could produce the McCollough effect in the laboratory. One possibility is that these color/orientation neurons might serve as useful elements in a low-level visual "alphabet" for the low-level encoding of an image; one way to explore this possibility would be to examine whether applying an efficient coding principle to natural image statistics (as in Olshausen & Field, 1996) yields kernels with this sort of tuning.

While many questions remain, the results of the present study usefully constrain theorizing about the relationship between color and form processing throughout the ventral visual pathway, making it an apt time to attempt to synthesize the vast range of evidence available on this topic—psychophysical, neuropsychological, and neurophysiological—into an overarching model consistent with all available evidence. This attempt is complicated by the fact that these literatures often proceed independently of one another, where, for instance, neurophysiology papers often document their results without tying them to findings in psychophysics and neuropsychology. Even several attempts to synthesize findings into a fuller picture often neglect one line of evidence or another; for instance, two reviews (Di Lollo, 2012; Rentzeperis et al., 2014) arguing for a more intermingled, rather than separate, encoding of color and form completely omit seemingly germane neuropsychological evidence, such as the existence of achromatopsia. Below, I present my attempt at such a model.

I hypothesize that color and form information evolve through three different relationships over the course of processing: *entangled*, then *separated*, then *bound*. The *entangled* stage begins at the level of the retina: here, there is no explicit information about the shapes or colors present

in the image; while there are spatial patterns of cone activations, there is no readily accessible information about object shape (as in the framework of DiCarlo & Cox, 2007), and while there are differential activations of the three cone types, providing wavelength information about incident light, none of the color constancy operations that define our color experience have yet taken place. In a sense, color and form information is present in the same *units* (e.g., a photoreceptor whose activity could be exploited to be informative about color could also be informative about shape), but the format of this information renders it unuseful.

The visual system then confronts the following challenge: color and form information are both useful properties of an image, but different computations are involved in the extraction of either, so the next stage is to extract color and form into *separated* representations; this occurs over the course of processing throughout V1, V2, V4, and onto regions in IT that are relatively more specialized for the processing of either feature, whose impairment leads to the characteristic focal deficits involved in visual form agnosia or cerebral achromatopsia. However, although separate computational processes are involved in extracting the final representation of form and color, either feature can play a role in the computation of the other; for example, color can be used in the extraction of form information, and 3D surface information can play a role in computing color constancy. Neurally, this could occur via neural populations that exhibit tuning to both features (e.g. the color information in LO or the form information in the posterior color regions), or via some form of coupling between the brain regions encoding either feature (though fMRI is not ideally suited to examine such a coupling process). The result of the untangling process are separate, explicit representations of color and form. One point to emphasize is that this separate, explicit representation of different features is a genuine *achievement* of visual

processing: it does not come for free. Thus, the "initially separate feature maps" posited by Feature Integration Theory may not actually come into play until this later stage of processing.

After generating explicit, separate representations of either feature, the brain faces one final challenge, corresponding to the canonical version of the binding problem: correctly linking the features associated with the same objects. The possibility of illusory conjunctions—either in lesion patients, or in healthy subjects under impaired attention—suggests that this process does not "come for free" either. Much evidence suggests that the parietal cortex plays an important role in this process (Robertson, 2003), since lesioning parietal cortex can result in illusory conjunctions. The results from Study 1 do not directly bear on the role parietal cortex might play, although they do illuminate the nature of the pre-existing ventral stream feature representations over which parietal cortex might operate.

Just what is the role of parietal cortex in feature binding, then? Earlier accounts posited that it might play a role in spatially linking the feature representations encoded in ventral regions (Robertson, 2003), without itself encoding the underlying features, but more recently, it has become clear that various parietal regions encode not just "spatial pointers", but high-level object form information as well, suggesting a more "content-rich" view of parietal function (Konen & Kastner, 2008; Vaziri-Pashkam & Xu, 2017, 2019). One group has reported representation of color and spatial frequency conjunctions in human parietal cortex (Baumgartner et al., 2013; Pollmann et al., 2014), and another group has reported integration of task-relevant color and motion information in macaque LIP (Ibos & Freedman, 2014), raising the interesting possibility that parietal cortex itself directly encodes conjunction information (perhaps via "uploading" feature information from the ventral pathway), rather than simply spatially linking feature representations encoded in the ventral pathway. That said, literature on this exact topic is sparse,

but as our understanding of parietal cortex evolves, it is important for our understanding of its role in feature binding to evolve along with it.

Evidence from neuropsychology does offer some provocative clues as to the role of parietal cortex, however. Notably, the units over which it operates often appear to be discrete objects: note that illusory conjunctions are not merely "free-floating features" that can end up anywhere, but are bound to the *wrong* objects; evidently, the operations of parietal cortex are required to ensure that features are correctly bound to the right objects. The notion of discrete object representations with associated feature information has a rich history (Kahneman et al., 1992; Xu & Chun, 2009), and this level of representation is one that I did not specifically examine in the studies I have presented. However, one theory of the role of parietal cortex is that it somehow takes the high-level, fully-extracted feature information delivered by the ventral pathway and correctly packages this information into discrete object representations as required. The exact nature of this dorsal-ventral interaction is a crucial topic for further study.

The Joint Representation of Color and Form Information in Convolutional Neural Networks

In Study 2, I studied the format of the joint color/form representations in convolutional neural networks, and arrived at one consistent finding: as processing in each network proceeds, color representation becomes increasingly different across different objects, suggesting an increasingly interactive color/form representation. This was much less true for networks with randomly initialized weights than for the networks trained to recognize objects, suggesting that it is not a mere byproduct of network architecture.

These results contrast interestingly with those of Study 1: whereas in the neural data of Study 1, the only evidence of interactive color/form coding was found early in processing, and the only evidence of completely separate coding was later in processing (for the central color regions), here we evidently find the opposite pattern, where color and form are initially encoded in an orthogonal manner, but become increasingly interactive over the course of processing. This finding has several important implications.

First, independent representations of color and form do not inevitably arise in a feedforward network optimized for object recognition. Instead, color and form increasingly interact over the course of processing. Why might this be the case? One possibility is that the sole goal of the network is to recognize objects, not to recognize colors as such; thus, color information will be utilized in a way that assists in the object recognition task (e.g., distinguishing yellow and green might be especially important for lemon/lime shapes), but there is no pressure for the network to extract shape-invariant color information. One interesting way to test this possibility would be to train a network for different tasks, such as extracting color independent of shape, or extracting *both* color and shape, and to see whether this changes the representations it encodes. Additionally, given the seemingly automatic interactive coding of color and form that develops, it is striking that this mode of representation appears to be largely absent in higher-level ventral stream regions, emphasizing the point that an independent, factorized representation of different features is a genuine achievement of visual processing rather than something that comes for free.

Second, interactive color/form coding can emerge even in the absence of feedback or recurrent mechanisms; thus, when such effects are observed in the brain, the involvement of such mechanisms cannot be assumed.

Third, an additional finding of the study was that the color similarity space also transforms *within* a given object over the course of processing, and that this transformation occurs in a similar manner across different networks. This suggests that it may be adaptive for the networks to transform the color representation from its initial RGB representation into a coding format that may be more diagnostic for discriminating among different objects.

Finally, these results illuminate novel principles of information processing in CNNs: their "default" scheme for encoding multiple features appears to be a conjunctive, rather than independent coding scheme.

That said, there are important methodological differences between this study and Study 1, namely that Study 1 used artificial stimuli, whereas this study used complex, real-world object stimuli. Thus, in order to compare more directly between Study 1 and Study 2, in an ongoing fMRI experiment we are examining responses across the ventral visual pathway to different real-world objects presented in different colors. In this way, the responses of ventral visual regions and CNN layers can be compared using the same stimuli and the same analytical approaches.

Although in this case we specifically examined the coding of color and form as a case study, the method we use—examining how the coding space for one feature varies across other features—is perfectly general and can be applied across any pair of features, such as texture, size, position, and orientation. By applying the same analytical approach to different pairs of features, we can find whether the conjunctive coding hallmark we observe across multiple networks for color and form also applies to other features.

Mixed Selectivity for Task and Stimulus Information in the Human Visual System

In Study 3, we applied an fMRI analysis technique developed in Study 1, *pattern difference decoding,* to examine the joint coding format not of color and form, but of stimulus and task information. This study had several important takeaways, both methodological and substantive.

First, the method developed in this study makes it possible to probe for nonlinear mixed selectivity coding in the human brain, even when interaction effects may be heterogeneous and subtle across individual voxels. As this coding motif has been shown to apply in various neural domains in other species (e.g., Diomedi et al., 2020; Grunfeld & Likhtik, 2018; Ledergerber et al., 2020; Rigotti et al., 2013), it is important to assess how widely applicable it is in the human brain, which this analysis technique enables.

Second, on a substantive level, the study documents the prevalence and heterogeneity of stimulus/task mixed selectivity across the human visual system. Namely, such mixed selectivity appears to exist in early visual cortex and LOC only when the two tasks involve differences in spatial, feature, or object-based attention, with such interactions disappearing when the two tasks are equated in these demands; by contrast, in the superior IPS it exists even when these attentional demands are carefully equated. A possible explanation for this pattern of results is that V1-V4 and LOC might selectively encode certain stimulus details when they are attended to, resulting in a stimulus-by-task interaction effect, but once the attentional "gating" of different features is equated, they encode stimulus details in an identical way across tasks. By contrast, superior IPS is even more flexible in its handling of visual information: even after equating spatial, object, and feature-based attention, it can process the encoded stimulus details in different ways depending on the task. This highlights an important point: nonlinear mixed

selectivity coding for stimulus/task combinations can emerge for different reasons in different contexts, and establishing the role it plays can requires carefully manipulating the nature of the tasks being compared.

This study also advances the state of our understanding regarding the coding format used by the superior IPS: multiple studies have now shown that this region encodes both stimulus and task information, but it is has not revealed the exact format of this joint representation, such as whether it is additive or interactive (Vaziri-Pashkam & Xu, 2017, 2019; Xu & Vaziri-Pashkam, 2019), a gap that this study fills. That said, demonstrating an interaction effect is only the beginning: while this finding carves an important "joint" regarding the coding format used by this region, further work is needed to more finely characterize the neural code used in this region.

Summing Up

I began this dissertation by raising the question of how an intelligent visual system, whether biological or synthetic, might represent multiple features, both in conjunction with each other and in conjunction with task information. I have approached this venerable question by developing new methods to apply the framework of nonlinear mixed selectivity coding to data from fMRI and CNNs, focusing on the case study of color/form coding as well as stimulus/task coding. However, I note that the methods I develop, including pattern difference decoding and the correlation-of-correlations approach I use to examine how coding of color varies across shape in CNNs, are perfectly general, and can be applied to any other pair of features in any recording modality. I hope, then, that this work will serve as a valuable step towards understanding the cognitively ubiquitous challenge of encoding multiple features with multiple neurons.

References

Adams, D. L., & Zeki, S. (2001). Functional Organization of Macaque V3 for Stereoscopic Depth. *Journal of Neurophysiology*, *86*(5), 2195–2203. https://doi.org/10.1152/jn.2001.86.5.2195

Allefeld, C., & Haynes, J.-D. (2014). Searchlight-based multi-voxel pattern analysis of fMRI by cross-validated MANOVA. *NeuroImage*, *89*(Supplement C), 345–357. https://doi.org/10.1016/j.neuroimage.2013.11.043

Asaad, W. F., Rainer, G., & Miller, E. K. (1998). Neural Activity in the Primate Prefrontal Cortex during Associative Learning. *Neuron*, *21*(6), 1399–1407. https://doi.org/10.1016/S0896-6273(00)80658-3

Bannert, M. M., & Bartels, A. (2013). Decoding the Yellow of a Gray Banana. *Current Biology*, *23*(22), 2268–2272. https://doi.org/10.1016/j.cub.2013.09.016

Bannert, M. M., & Bartels, A. (2018). Human V4 Activity Patterns Predict Behavioral Performance in Imagery of Object Color. *Journal of Neuroscience*, *38*(15), 3657–3668. https://doi.org/10.1523/JNEUROSCI.2307-17.2018

Bao, P., She, L., McGill, M., & Tsao, D. Y. (2020). A map of object space in primate inferotemporal cortex. *Nature*, *583*(7814), 103–108. https://doi.org/10.1038/s41586-020-2350-5

Barbur, J. L., Harlow, J., & Plant, G. T. (1994). Insights into the different exploits of colour in the visual cortex. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *258*(1353), 327–334. https://doi.org/10.1098/rspb.1994.0181

Bartels, A., & Zeki, S. (2000). The architecture of the colour centre in the human visual brain: New results and a review*. *Eur J Neurosci*, *12*(1), 172–193.

Baumgartner, F., Hanke, M., Geringswald, F., Zinke, W., Speck, O., & Pollmann, S. (2013). Evidence for feature binding in the superior parietal lobule. *Neuroimage*, *68*, 173–180.

Beauchamp, M. S., Haxby, J. V., Jennings, J. E., & DeYoe, E. A. (1999). An fMRI Version of the Farnsworth-Munsell 100-Hue Test Reveals Multiple Color-selective Areas in Human Ventral Occipitotemporal Cortex. *Cereb. Cortex*, *9*(3), 257–263. https://doi.org/10.1093/cercor/9.3.257

Benson, D. F., & Greenberg, J. P. (1969). Visual Form Agnosia: A Specific Defect in Visual Discrimination. *Archives of Neurology*, *20*(1), 82–89. https://doi.org/10.1001/archneur.1969.00480070092010

Bhandari, A., Gagne, C., & Badre, D. (2018). Just above Chance: Is It Harder to Decode Information from Prefrontal Cortex Hemodynamic Activity Patterns? *Journal of Cognitive Neuroscience*, *30*(10), 1473–1498. https://doi.org/10.1162/jocn_a_01291

Blackman, R. K., Crowe, D. A., DeNicola, A. L., Sakellaridi, S., MacDonald, A. W., & Chafee, M. V. (2016). Monkey Prefrontal Neurons Reflect Logical Operations for Cognitive

Control in a Variant of the AX Continuous Performance Task (AX-CPT). *Journal of Neuroscience*, *36*(14), 4067–4079. https://doi.org/10.1523/JNEUROSCI.3578-15.2016

Bloj, M. G., Kersten, D., & Hurlbert, A. C. (1999). Perception of three-dimensional shape influences colour perception through mutual illumination. *Nature*, *402*(6764), 877–879. https://doi.org/10.1038/47245

Bouvier, S. E., & Engel, S. A. (2006). Behavioral Deficits and Cortical Damage Loci in Cerebral Achromatopsia. *Cerebral Cortex*, *16*(2), 183–191. https://doi.org/10.1093/cercor/bhi096

Brewer, A. A., Liu, J., Wade, A. R., & Wandell, B. A. (2005). Visual field maps and stimulus selectivity in human ventral occipital cortex. *Nat Neurosci*, *8*(8), 1102–1109. https://doi.org/10.1038/nn1507

Brouwer, G. J., & Heeger, D. J. (2009). Decoding and Reconstructing Color from Responses in Human Visual Cortex. *J. Neurosci.*, *29*(44), 13992–14003. https://doi.org/10.1523/JNEUROSCI.3577-09.2009

Brouwer, G. J., & Heeger, D. J. (2013). Categorical Clustering of the Neural Representation of Color. *Journal of Neuroscience*, *33*(39), 15454–15465. https://doi.org/10.1523/JNEUROSCI.2472-13.2013

Bushnell, B. N., & Pasupathy, A. (2012). Shape encoding consistency across colors in primate V4. *Journal of Neurophysiology*, *108*(5), 1299–1308. https://doi.org/10.1152/jn.01063.2011

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., & DiCarlo, J. J. (2014). Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLOS Computational Biology*, *10*(12), e1003963. https://doi.org/10.1371/journal.pcbi.1003963

Cavanagh, P., & Anstis, S. (1991). The contribution of color to motion in normal and color-deficient observers. *Vision Research*, *31*(12), 2109–2148. https://doi.org/10.1016/0042-6989(91)90169-6

Cavanagh, P., Hénaff, M.-A., Michel, F., Landis, T., Troscianko, T., & Intriligator, J. (1998). Complete sparing of high-contrast color input to motion perception in cortical color blindness. *Nature Neuroscience*, *1*(3), 242–247. https://doi.org/10.1038/688

Chang, L., Bao, P., & Tsao, D. Y. (2017). The representation of colored objects in macaque color patches. *Nature Communications*, *8*(1), 2064. https://doi.org/10.1038/s41467-017-01912-7

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*(1), 27755. https://doi.org/10.1038/srep27755

Cohen, A., & Rafal, R. D. (1991). Attention and feature integration: Illusory conjunctions in a patient with a parietal lobe lesion. *Psychological Science*, *2*(2), 106–110.

Conway, B. R. (2001). Spatial Structure of Cone Inputs to Color Cells in Alert Macaque Primary Visual Cortex (V-1). *Journal of Neuroscience*, *21*(8), 2768–2783. https://doi.org/10.1523/JNEUROSCI.21-08-02768.2001

Conway, B. R. (2009). Color Vision, Cones, and Color-Coding in the Cortex. *The Neuroscientist*, *15*(3), 274–290. https://doi.org/10.1177/1073858408331369

Conway, B. R. (2018). The Organization and Operation of Inferior Temporal Cortex. *Annual Review of Vision Science*, *4*(1), 381–402. https://doi.org/10.1146/annurev-vision-091517-034202

Conway, B. R., Chatterjee, S., Field, G. D., Horwitz, G. D., Johnson, E. N., Koida, K., & Mancuso, K. (2010). Advances in Color Science: From Retina to Behavior. *Journal of Neuroscience*, *30*(45), 14955–14963. https://doi.org/10.1523/JNEUROSCI.4348-10.2010

Conway, B. R., Moeller, S., & Tsao, D. Y. (2007). Specialized Color Modules in Macaque Extrastriate Cortex. *Neuron*, *56*(3), 560–573. https://doi.org/10.1016/j.neuron.2007.10.008

Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, & Li Fei-Fei. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Denys, K., Vanduffel, W., Fize, D., Nelissen, K., Peuskens, H., Van Essen, D., & Orban, G. A. (2004). The Processing of Visual Shape in the Cerebral Cortex of Human and Nonhuman Primates: A Functional Magnetic Resonance Imaging Study. *J. Neurosci.*, *24*(10), 2551–2565. https://doi.org/10.1523/JNEUROSCI.3569-03.2004

Di Lollo, V. (2012). The feature-binding problem is an ill-posed problem. *Trends Cogn Sci*, *16*(6), 317–321. https://doi.org/10.1016/j.tics.2012.04.007

DiCarlo, J. J, & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, *11*(8), 333–341.

DiCarlo, James J., Zoccolan, D., & Rust, N. C. (2012). How Does the Brain Solve Visual Object Recognition? *Neuron*, *73*(3), 415–434. https://doi.org/10.1016/j.neuron.2012.01.010

Diomedi, S., Vaccari, F. E., Filippini, M., Fattori, P., & Galletti, C. (2020). Mixed Selectivity in Macaque Medial Parietal Cortex during Eye-Hand Reaching. *IScience*, *23*(10), 101616. https://doi.org/10.1016/j.isci.2020.101616

Dubois, J., Berker, A. O. de, & Tsao, D. Y. (2015). Single-Unit Recordings in the Macaque Face Patch System Reveal Limitations of fMRI MVPA. *The Journal of Neuroscience*, *35*(6), 2791–2802. https://doi.org/10.1523/JNEUROSCI.4037-14.2015

Duncan, J. (2001). An adaptive coding model of neural function in prefrontal cortex. *Nature Reviews Neuroscience*, *2*(11), 820–829. https://doi.org/10.1038/35097575

Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017). Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, *152*, 184–194.

https://doi.org/10.1016/j.neuroimage.2016.10.001

Engel, S. A. (2005). Adaptation of Oriented and Unoriented Color-Selective Neurons in Human Visual Areas. *Neuron*, *45*(4), 613–623. https://doi.org/10.1016/j.neuron.2005.01.014

Flachot, A., & Gegenfurtner, K. R. (2018). Processing of chromatic information in a deep convolutional neural network. *JOSA A*, *35*(4), B334–B346. https://doi.org/10.1364/JOSAA.35.00B334

Friedman, H. S., Zhou, H., & Heydt, R. von der. (2003). The coding of uniform colour figures in monkey visual cortex. *The Journal of Physiology*, *548*(2), 593–613. https://doi.org/10.1111/j.1469-7793.2003.00593.x

Fusi, S., Miller, E. K., & Rigotti, M. (2016). Why neurons mix: High dimensionality for higher cognition. *Current Opinion in Neurobiology*, *37*, 66–74. https://doi.org/10.1016/j.conb.2016.01.010

Gallant, J. L., Braun, J., & Essen, D. V. (1993). Selectivity for polar, hyperbolic, and Cartesian gratings in macaque visual cortex. *Science*, *259*(5091), 100–103. https://doi.org/10.1126/science.8418487

Gallant, Jack L, Shoup, R. E., & Mazer, J. A. (2000). A Human Extrastriate Area Functionally Homologous to Macaque V4. *Neuron*, *27*(2), 227–235. https://doi.org/10.1016/S0896-6273(00)00032-5

Garner, W. R. (1976). Interaction of stimulus dimensions in concept and choice processes. *Cognitive Psychology*, *8*(1), 98–123. https://doi.org/10.1016/0010-0285(76)90006-2

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2019). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *ArXiv:1811.12231 [Cs, q-Bio, Stat]*. http://arxiv.org/abs/1811.12231

Goodale, M. A., & Milner, A. D. (2004). *Sight unseen: An exploration of conscious and unconscious vision* (pp. ix, 135). Oxford University Press.

Grill-Spector, K., Kushnir, T., Edelman, S., Itzchak, Y., & Malach, R. (1998). Cue-invariant activation in object-related areas of the human occipital lobe. *NEURON-CAMBRIDGE MA-*, *21*, 191–202.

Grunfeld, I. S., & Likhtik, E. (2018). Mixed selectivity encoding and action selection in the prefrontal cortex during threat assessment. *Current Opinion in Neurobiology*, *49*, 108–115. https://doi.org/10.1016/j.conb.2018.01.008

Hadjikhani, N., Liu, A. K., Dale, A. M., Cavanagh, P., & Tootell, R. B. H. (1998). Retinotopy and color sensitivity in human visual cortical area V 8. *Nature Neuroscience*, *1*(3), 235–241.

Hardcastle, K., Maheswaranathan, N., Ganguli, S., & Giocomo, L. M. (2017). A multiplexed, heterogeneous, and adaptive code for navigation in medial entorhinal cortex. *Neuron*,

94(2), 375-387.e7. https://doi.org/10.1016/j.neuron.2017.03.025

Hasantash, M., Lafer-Sousa, R., Afraz, A., & Conway, B. R. (2019). Paradoxical impact of memory on color appearance of faces. *Nature Communications*, *10*(1), 3010. https://doi.org/10.1038/s41467-019-10073-8

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. *Science*, *293*(5539), 2425–2430. https://doi.org/10.1126/science.1063736

Heywood, C. A., Cowey, A., & Newcombe, F. (1991). Chromatic Discrimination in a Cortically Colour Blind Observer. *European Journal of Neuroscience*, *3*(8), 802–812. https://doi.org/10.1111/j.1460-9568.1991.tb01676.x

Heywood, C. A., Kentridge, R. W., & Cowey, A. (1998). Form and motion from colour in cerebral achromatopsia. *Experimental Brain Research*, *123*(1), 145–153. https://doi.org/10.1007/s002210050555

Holcombe, A. O., & Cavanagh, P. (2001). Early binding of feature pairs for visual perception. *Nature Neuroscience*, *4*(2), 127–128. https://doi.org/10.1038/83945

Hong, H., Yamins, D. L. K., Majaj, N. J., & DiCarlo, J. J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature Neuroscience*, *19*(4), 613–622. https://doi.org/10.1038/nn.4247

Houck, M. R., & Hoffman, J. E. (1986). Conjunction of color and form without attention: Evidence from an orientation-contingent color aftereffect. *Journal of Experimental Psychology: Human Perception and Performance*, *12*(2), 186–199. https://doi.org/10.1037/0096-1523.12.2.186

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, *160*(1), 106–154. https://doi.org/10.1113/jphysiol.1962.sp006837

Huk, A. C., Katz, L. N., & Yates, J. L. (2017). The Role of the Lateral Intraparietal Area in (the Study of) Decision Making. *Annual Review of Neuroscience*, *40*(1), 349–372. https://doi.org/10.1146/annurev-neuro-072116-031508

Ibos, G., & Freedman, D. J. (2014). Dynamic integration of task-relevant visual features in posterior parietal cortex. *Neuron*, *83*(6), 1468–1480.

Jeong, S. K., & Xu, Y. (2016). Behaviorally Relevant Abstract Object Identity Representation in the Human Parietal Cortex. *Journal of Neuroscience*, *36*(5), 1607–1619. https://doi.org/10.1523/JNEUROSCI.1016-15.2016

Johnson, E. N., Hawken, M. J., & Shapley, R. (2001). The spatial transformation of color in the primary visual cortex of the macaque monkey. *Nature Neuroscience*, *4*(4), 409–416. https://doi.org/10.1038/86061

Johnston, W. J., Palmer, S. E., & Freedman, D. J. (2019). Nonlinear mixed selectivity supports

reliable neural computation. *BioRxiv*, 577288. https://doi.org/10.1101/577288

Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology*, *24*(2), 175–219.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLOS Computational Biology*, *10*(11), e1003915. https://doi.org/10.1371/journal.pcbi.1003915

Komatsu, H., & Ideura, Y. (1993). Relationships between color, shape, and pattern selectivities of neurons in the inferior temporal cortex of the monkey. *J Neurophysiol*, *70*(2), 677–694.

Konen, C. S., & Kastner, S. (2008). Two hierarchically organized neural systems for object information in human visual cortex. *Nature Neuroscience*, *11*(2), 224–231.

Kourtzi, Z., & Kanwisher, N. (2001). Representation of Perceived Object Shape by the Human Lateral Occipital Complex. *Science*, *293*(5534), 1506–1509. https://doi.org/10.1126/science.1061133

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, *17*(8), 401–412.

Lafer-Sousa, R., & Conway, B. R. (2013). Parallel, multi-stage processing of colors, faces and shapes in macaque inferior temporal cortex. *Nature Neuroscience*, *16*(12), 1870–1878. https://doi.org/10.1038/nn.3555

Lafer-Sousa, R., Conway, B. R., & Kanwisher, N. G. (2016). Color-Biased Regions of the Ventral Visual Pathway Lie between Face- and Place-Selective Regions in Humans, as in Macaques. *Journal of Neuroscience*, *36*(5), 1682–1697. https://doi.org/10.1523/JNEUROSCI.3164-15.2016

Ledergerber, D., Battistin, C., Blackstad, J. S., Gardner, R. J., Witter, M. P., Moser, M.-B., Roudi, Y., & Moser, E. I. (2020). Task-dependent mixed selectivity in the subiculum. *BioRxiv*, 2020.06.06.129221. https://doi.org/10.1101/2020.06.06.129221

Lehky, S. R., & Tanaka, K. (2016). Neural representation for object recognition in inferotemporal cortex. *Current Opinion in Neurobiology*, *37*, 23–35. https://doi.org/10.1016/j.conb.2015.12.001

Lennie, P., Krauskopf, J., & Sclar, G. (1990). Chromatic mechanisms in striate cortex of macaque. *Journal of Neuroscience*, *10*(2), 649–669. https://doi.org/10.1523/JNEUROSCI.10-02-00649.1990

Leventhal, A. G., Thompson, K. G., Liu, D., Zhou, Y., & Ault, S. J. (1995). Concomitant sensitivity to orientation, direction, and color of cells in layers 2, 3, and 4 of monkey striate cortex. *Journal of Neuroscience*, *15*(3), 1808–1818. https://doi.org/10.1523/JNEUROSCI.15-03-01808.1995

Lindsay, G. W., Rigotti, M., Warden, M. R., Miller, E. K., & Fusi, S. (2017). Hebbian Learning

in a Random Network Captures Selectivity Properties of the Prefrontal Cortex. *Journal of Neuroscience*, *37*(45), 11021–11036. https://doi.org/10.1523/JNEUROSCI.1222-17.2017

Linsley, D., Eberhardt, S., Sharma, T., Gupta, P., & Serre, T. (2017). *What Are the Visual Features Underlying Human Versus Machine Vision?* 2706–2714. https://openaccess.thecvf.com/content_ICCV_2017_workshops/w40/html/Linsley_What_Are_the_ICCV_2017_paper.html

Livingstone, M., & Hubel, D. (1988). Segregation of form, color, movement, and depth-Anatomy, physiology, and perception. *Science*, *240*(4853), 740–749.

Malach, R., Reppas, J. B., Benson, R. R., Kwong, K. K., Jiang, H., Kennedy, W. A., Ledden, P. J., Brady, T. J., Rosen, B. R., & Tootell, R. B. (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *92*(18), 8135–8139. https://doi.org/VL - 92

Mandelli, M.-J. F., & Kiper, D. C. (2005). The local and global processing of chromatic Glass patterns. *Journal of Vision*, *5*(5), 2–2. https://doi.org/10.1167/5.5.2

McCollough, C. (1965). Color Adaptation of Edge-Detectors in the Human Visual System. *Science*, *149*(3688), 1115–1116. https://doi.org/10.1126/science.149.3688.1115

McMahon, D. B. T., & Olson, C. R. (2009). Linearly Additive Shape and Color Signals in Monkey Inferotemporal Cortex. *J Neurophysiol*, *101*(4), 1867–1875. https://doi.org/10.1152/jn.90650.2008

Meister, M. L. R., Hennig, J. A., & Huk, A. C. (2013). Signal Multiplexing and Single-Neuron Computations in Lateral Intraparietal Area During Decision-Making. *Journal of Neuroscience*, *33*(6), 2254–2267. https://doi.org/10.1523/JNEUROSCI.2984-12.2013

Murphey, D. K., Yoshor, D., & Beauchamp, M. (2008). Perception Matches Selectivity in the Human Anterior Color Center. *Current Biology*, *18*(3), 216–220. https://doi.org/10.1016/j.cub.2008.01.013

Nogueira, R., Rodgers, C. C., Bruno, R. M., & Fusi, S. (2021). The non-linear mixed representations in somatosensory cortex support simple and complex tasks. *BioRxiv*, 2021.02.11.430704. https://doi.org/10.1101/2021.02.11.430704

Olshausen, B. A., & Field, D. J. (1996). Natural image statistics and efficient coding. *Network: Computation in Neural Systems*, *7*(2), 333–339. https://doi.org/10.1088/0954-898X_7_2_014

Park, I. M., Meister, M. L. R., Huk, A. C., & Pillow, J. W. (2014). Encoding and decoding in parietal cortex during sensorimotor decision-making. *Nature Neuroscience*, *17*(10), 1395–1403. https://doi.org/10.1038/nn.3800

Pasupathy, A., & Connor, C. E. (2001). Shape Representation in Area V4: Position-Specific Tuning for Boundary Conformation. *J Neurophysiol*, *86*(5), 2505–2519.

Pasupathy, A., Kim, T., & Popovkina, D. V. (2019). Object shape and surface properties are jointly encoded in mid-level ventral visual cortex. *Current Opinion in Neurobiology*, *58*, 199–208. https://doi.org/10.1016/j.conb.2019.09.009

Pollmann, S., Zinke, W., Baumgartner, F., Geringswald, F., & Hanke, M. (2014). The right temporo-parietal junction contributes to visual feature binding. *NeuroImage*, *101*, 289–297.

Rafegas, I., & Vanrell, M. (2018). Color encoding in biologically-inspired convolutional neural networks. *Vision Research*, *151*, 7–17. https://doi.org/10.1016/j.visres.2018.03.010

Rappaport, S. J., Humphreys, G. W., & Riddoch, M. J. (2013). The attraction of yellow corn: Reduced attentional constraints on coding learned conjunctive relations. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(4), 1016–1031. https://doi.org/10.1037/a0032506

Rappaport, S. J., Riddoch, M. J., Chechlacz, M., & Humphreys, G. W. (2015). Unconscious Familiarity-based Color–Form Binding: Evidence from Visual Extinction. *Journal of Cognitive Neuroscience*, *28*(3), 501–516. https://doi.org/10.1162/jocn_a_00904

Rentzeperis, I., Nikolaev, A. R., Kiper, D. C., & van Leeuwen, C. (2014). Distributed processing of color and form in the visual cortex. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00932

Rigotti, M., Barak, O., Warden, M. R., Wang, X.-J., Daw, N. D., Miller, E. K., & Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature*, *advance online publication*. https://doi.org/10.1038/nature12160

Robertson, L. C. (2003). Binding, spatial attention and perceptual awareness. *Nature Reviews Neuroscience*, *4*(2), 93–102.

Roe, A. W., Chelazzi, L., Connor, C. E., Conway, B. R., Fujita, I., Gallant, J. L., Lu, H., & Vanduffel, W. (2012). Toward a Unified Theory of Visual Area V4. *Neuron*, *74*(1), 12–29. https://doi.org/10.1016/j.neuron.2012.03.011

Saxena, S., & Cunningham, J. P. (2019). Towards the neural population doctrine. *Current Opinion in Neurobiology*, *55*, 103–111. https://doi.org/10.1016/j.conb.2019.02.002

Serre, T. (2019). Deep Learning: The Good, the Bad, and the Ugly. *Annual Review of Vision Science*, *5*(1), 399–426. https://doi.org/10.1146/annurev-vision-091718-014951

Seymour, K., Clifford, C. W., Logothetis, N. K., & Bartels, A. (2010). Coding and binding of color and form in visual cortex. *Cerebral Cortex*, *20*(8), 1946–1954.

Simmons, W. K., Ramjee, V., Beauchamp, M. S., McRae, K., Martin, A., & Barsalou, L. W. (2007). A common neural substrate for perceiving and knowing about color. *Neuropsychologia*, *45*(12), 2802–2810. https://doi.org/10.1016/j.neuropsychologia.2007.05.002

Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale

Image Recognition. *ArXiv:1409.1556 [Cs]*. http://arxiv.org/abs/1409.1556

Singh, A., Bay, A., & Mirabile, A. (2020). Assessing The Importance Of Colours For CNNs In Object Recognition. *ArXiv:2012.06917 [Cs]*. http://arxiv.org/abs/2012.06917

Storrs, K. R., & Kriegeskorte, N. (2019). Deep Learning for Cognitive Neuroscience. *ArXiv:1903.01458 [Cs, q-Bio]*. http://arxiv.org/abs/1903.01458

Stromeyer, C. F. (1969). Further studies of the McCollough effect. *Perception & Psychophysics*, *6*(2), 105–110. https://doi.org/10.3758/BF03210691

Treisman, A. (1999). Solutions to the binding problem: Progress through controversy and convergence. *Neuron*, *24*(1), 105–125.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.

Treisman, A., & Schmidt, H. (1982). Illusory conjunctions in the perception of objects. *Cognitive Psychology*, *14*(1), 107–141. https://doi.org/10.1016/0010-0285(82)90006-8

Ullman, S., Assif, L., Fetaya, E., & Harari, D. (2016). Atoms of recognition in human and computer vision. *Proceedings of the National Academy of Sciences*, *113*(10), 2744–2749. https://doi.org/10.1073/pnas.1513198113

Valenti, J. J., & Firestone, C. (2019). Finding the "odd one out": Memory color effects and the logic of appearance. *Cognition*, *191*, 103934. https://doi.org/10.1016/j.cognition.2019.04.003

Vaziri-Pashkam, M., & Xu, Y. (2017). Goal-directed visual processing differentially impacts human ventral and dorsal visual representations. *Journal of Neuroscience*, 3392–16. https://doi.org/10.1523/JNEUROSCI.3392-16.2017

Vaziri-Pashkam, M., & Xu, Y. (2019). An Information-Driven 2-Pathway Characterization of Occipitotemporal and Posterior Parietal Visual Object Representations. *Cerebral Cortex*, *29*(5), 2034–2050. https://doi.org/10.1093/cercor/bhy080

Walsh, V., Ashbridge, E., & Cowey, A. (1998). Cortical plasticity in perceptual learning demonstrated by transcranial magnetic stimulation. *Neuropsychologia*, *36*(4), 363–367. https://doi.org/10.1016/S0028-3932(97)00113-9

Wang, G., Tanaka, K., & Tanifuji, M. (1996). Optical Imaging of Functional Organization in the Monkey Inferotemporal Cortex. *Science*, *272*(5268), 1665–1668. https://doi.org/10.1126/science.272.5268.1665

Xu, Y. (2018a). A Tale of Two Visual Systems: Invariant and Adaptive Visual Information Representations in the Primate Brain. *Annual Review of Vision Science*, *4*(1), 311–336. https://doi.org/10.1146/annurev-vision-091517-033954

Xu, Y. (2018b). The Posterior Parietal Cortex in Adaptive Visual Processing. *Trends in Neurosciences*, *41*(11), 806–822. https://doi.org/10.1016/j.tins.2018.07.012

Xu, Y., & Chun, M. M. (2009). Selecting and perceiving multiple visual objects. *Trends Cogn Sci*, *13*(4), 167–174. https://doi.org/10.1016/j.tics.2009.01.008

Xu, Y., & Vaziri-Pashkam, M. (2019). Task modulation of the 2-pathway characterization of occipitotemporal and posterior parietal visual object representations. *Neuropsychologia*, *132*, 107140. https://doi.org/10.1016/j.neuropsychologia.2019.107140

Xu, Y., & Vaziri-Pashkam, M. (2020). Limited correspondence in visual representation between the human brain and convolutional neural networks. *BioRxiv*, 2020.03.12.989376. https://doi.org/10.1101/2020.03.12.989376

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365. https://doi.org/10.1038/nn.4244

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111

Zhang, C. Y., Aflalo, T., Revechkis, B., Rosario, E., Ouellette, D., Pouratian, N., & Andersen, R. A. (2020). Preservation of Partially Mixed Selectivity in Human Posterior Parietal Cortex across Changes in Task Context. *ENeuro*, *7*(2). https://doi.org/10.1523/ENEURO.0222-19.2019

Zhang, C. Y., Aflalo, T., Revechkis, B., Rosario, E. R., Ouellette, D., Pouratian, N., & Andersen, R. A. (2017). Partially Mixed Selectivity in Human Posterior Parietal Association Cortex. *Neuron*, *95*(3), 697-708.e4. https://doi.org/10.1016/j.neuron.2017.06.040