Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

I'll Know It When I See It: Bridging Perception and Conceptual Learning in Minds and Machines

Cliona O'Doherty

School of Psychology and Trinity College Institute of Neuroscience
Trinity College Dublin

A thesis submitted to
Trinity College Dublin, University of Dublin
For the degree of
Doctor of Philosophy

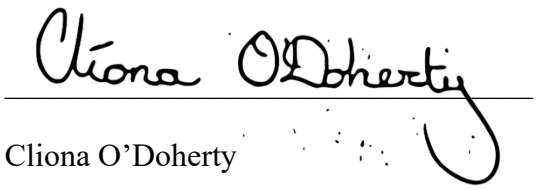
2025

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).



Cliona O'Doherty

Summary

To operate successfully within the world, we must be able to make sense of the things we see. This ability, rooted in our capacity to form conceptual understanding from visual inputs, is crucial from infancy to adulthood. Recent advances in neuroimaging and computational models of the ventral visual stream have provided new insights into visual recognition in both biological and computer vision, but much remains unknown regarding the developmental mechanisms that underpin this process. Further understanding of how conceptual knowledge emerges from perceptual inputs, especially from a young age with minimal labelled guidance, is crucial not only for advancing studies of human cognitive development but also for informing the design of artificial intelligence systems. This thesis explores these questions using emerging techniques in developmental cognitive computational neuroscience.

In **Chapter 1**, I review the relevant literature from cognitive neuroscience, psychology, philosophy and machine learning to provide an integrative account of our current understanding of concept representations and visual object recognition. I highlight that high-level vision provides a link between perception and semantic cognition, and that a statistical view of concept formation is a useful way forward for understanding minds and for engineering machines that understand. I discuss limitations in our current models of high-level vision, and propose that these can be overcome by looking for inspiration in the infant brain.

Chapter 2 describes computational work to investigate the promise of infant-inspired learning algorithms for improved conceptual representation learning. Taking inspiration

from development, I explore what can be learned from the extended temporal statistics of the environment by implementing a time-based learning objective into a deep neural network (DNN) model and examining its representation learning. I find that the time that links two input images provides a signal for either object or context level information, constraining the knowledge that can be embedded into the DNN during with a contrastive learning objective. Furthermore, I find that high object classification accuracy is not indicative of whether relational or contextual information has been learned. These findings highlight a disparity in how computer and biological vision systems represent concepts, and the importance of extended temporal experience for learning the wider context of the things we see.

In **Chapter 3**, I explore the role of fast and slow timescales for perceptual and categorical visual processing. Based on the findings from Chapter 2, I hypothesise that the anterior VVS – which encodes high-level category information – will be tuned to more slowly changing, recognisable inputs when placed against a backdrop of naturalistic perceptual noise. Additionally, I explore whether this preference is reflected in the intrinsic neural timescale (INT) of these high-level visual regions using a functional MRI (fMRI) paradigm. Adult participants viewed recognisable objects at either fast or slow intervals, appearing against an unrecognisable but perceptually intact video. INTs along VOTC were estimated from an independent resting state scan and the association between this underlying neural activity and task-based responses was explored. I find that the INT of a region and not the timescale of the input stimuli determines the direction of repetition responses, whether suppression or enhancement.

Without a thorough understanding of the starting point for the VVS, it is difficult to truly understand how object understanding develops through learning. **Chapter 4** presents results from the FOUNDCOG large-scale infant fMRI study which characterises the early beginnings of visual representations in VVS in response to a rich set of visual categories. The scale of our data collection enabled DNN models of vision to be applied to the developing ventral stream, marking the first application of this technique to infant functional activity in response to a diverse set of stimuli. I use our novel dataset to test opposing theories of visual category development: is this a bottom-up sensory-driven process or is it top-down, with innate beginnings of category knowledge? I discuss the theoretical implications of a top-down developmental mechanism, placing it within the context of existing research on visual development.

Finally, in **Chapter 5**, I provide a general discussion the wider implications of the results presented in this thesis and how developmental cognitive computational neuroscience can push forward the research landscape within this domain, with mutual benefit for studies of human and machine intelligence.

Acknowledgements

Of course I'll begin with where it all began. Thank you to my supervisor Prof. Rhodri Cusack. You asked me, when I first walked into your office for my final year project "so, what are we going to do?". I don't think that the expansive reach of that first opportunity quite hit me until now, five years later, when I can look back and see that actually – we did quite a lot. I'm not only grateful for the support, trust and belief you have shown me throughout, but for providing me with a genuinely exciting and stimulating beginning to my career. Despite my multiple efforts to veer another way, thank you for creating a research environment that has kept me where I clearly need to be.

Which brings me to Anna. I've told you a few times that I wouldn't be here if it wasn't for you – and we joke that we're not sure if that's a good thing – but I am so thankful that you've been by my side since day one. Thinking about how much we've achieved over these past few years really is mind boggling. It wasn't easy, and we definitely weren't always so sure we'd get here, but we did it. I have so much gratitude that I got to do it with you as a friend and mentor.

Áine, who knew at those NeuroAI seminars that it would lead to us down this crazy path together. Thank you for being such a special part of all of this. Thank you to everyone from the Cusack Lab past and present who have made contributions small and large to this research. From being there for 200 baby scans, achieving something incredible together with FOUNDCOG, to helping me solve minute technical matters or indulging me in big discussions about the work. Graham, Chiara, Anna K., Anna B., Lorijn, Tarek, Enna, Jessica, Keelin, Adrienne, Tammy, Marie. Sojo, for the hours spent with me at the scanner, and to every family that participated in FOUNDCOG.

Thank you to the Irish Research Council and European Research Council for funding the research and to my thesis committee Prof. Redmond O'Connell and Prof. Aljosa Smolic for the annual encouragement. Also to the wider community in TCIN, those of you behind that wall in the O'Connell lab (Jade, for that Sprite and Harvey for providing the vocal entertainment every evening). To every person who says hello in the corridor, there are too many to list but thank you for the little moments in the Lloyd that remind us all we're in this together. Tina and Chris for a happy greeting each morning, Elaine, Barbara and Aisling for

administrative help throughout the years. To Trinity in general, for being my home for almost ten years.

People who know me, know how important my work is to me. But people who really know me, know that my friends are what's most important. Aisling, thinking of us coming home from school compared to where we are today always makes me swell with pride. I'm so lucky to have you as a friend. Aideen, I really don't think these four years would have been the same without you there to understand what strange things I'm dealing with. Aoife, thank you for each caring check in, in the form of a weird video or something cat related. Ruth, Ruth, Ellen, Elaine – I mean it when I'm say I'm so proud of us all as women and as friends. I look forward to having the guaranteed weekends put aside for birthdays, keeping me grounded for many years to come. Jane, I don't even need to say it because our psychic connection will cover it. Mary, you've been such a caring friend, cheerleader and support system throughout, always there to listen when I need it. Thank you for being such an inspiring woman. Liam, we've done so much together from surviving undergrad, travelling the world, staying out all night and going through many seasons of change. We did it all with each other to rely on. These couple of years were no different, and I can't wait to see what our next chapter will be.

Without really being aware of it, Greg Walsh and the community at Temple Bar Yoga have been a part of my PhD journey from start to finish. Especially thanks to the Wednesday group, who have been an unknowing backbone of my life for these past few years. My weekly pilgrimage to that little courtyard sanctuary amidst the busy Dublin streets – or to the sanctuary of the zoom grid during COVID - has provided me with such a consistent outlet of joy and wellness that has been indispensable throughout my PhD.

I'm so blessed to have the family that I do. Thank you to Phil and Niamh for inspiring me since I was small. Edel, Tash and of course little Seóna and Sienna for providing me with the stability, support and confidence to be able to work as hard as I do. I really got lucky to be the baby of the family, because it means that I get to look up to you all.

Mam and Dad. Finding the words is almost impossible. Thank you. Thank you for being there, for providing me with everything I could ever need, unconditionally, for all things big and small. I would not be who I am without you both. I am so lucky to be your daughter and to be proud knowing that no matter what crazy thing I do next, I can do it because of you.

List of Publications and Presentations

Papers

O'Doherty, C., Dineen, Á.T., Truzzi, A., King, G, Zaadnoordijk, L., Harrison, K., D'Arcy, E.L., White, J., Caldinelli, C., Holloway, T., Kravchenko, A., Tarrant, A., Byrne, A.T., Foran, A., Molloy, E., Cusack, R. (*under consideration*), Infants Have Rich Visual Categories in Ventrotemporal Cortex at 2 Months Old.

Conference proceedings

O'Doherty, C., Dineen, Á.T., Truzzi, A., King, G, Zaadnoordijk, L., Harrison, K., D'Arcy, E.L., White, J., Caldinelli, C., Holloway, T., Kravchenko, A., Tarrant, A., Byrne, A.T., Foran, A., Molloy, E., Cusack, R. (2024), Deep Neural Networks Models of Infant Visual Cortex, Cognitive Computational Neuroscience (CCN), Boston, USA.

O'Doherty, C., Cusack, R., (2022), Objects or Context? Learning From Temporal Regularities in Continuous Visual Experience With an Infant-inspired DNN. Cognitive Computational Neuroscience, San Francisco, California, USA.

O'Doherty, C., Cusack, R., (2021), SemanticCMC: Contrastive Learning of Meaningful Object Associations from Temporal Co-occurrence Patterns in Naturalistic Movies. Computer Vision and Pattern Recognition (CVPR) Women in Computer Vision Workshop, Virtual Conference.

O'Doherty, C., Cusack, R., (2020), SemanticCMC – improved semantic self-supervised learning with naturalistic temporal co-occurrences. NeurIPS workshop on Self Supervised Learning: Theory and Practice, Virtual Conference.

Oral presentations

O'Doherty, C. (2024), *Development of object recognition in the visual cortex from 2 to 9 months using awake fMRI and deep learning.* International Congress of Infant Studies, Glasgow, Scotland.

O'Doherty, C. (2024), *Deep Neural Networks Model the Development of Visual Recognition in Infants.* Trinity College Dublin School of Psychology Research Symposium, Dublin, Ireland.

O'Doherty, C. (2024), *Hierarchical Correspondence Between Deep Neural Networks and Emerging Representations in the Infant Ventral Visual Stream.* Trinity College Institute of Neuroscience Research Day, Dublin, Ireland.

Infants and Machines in Interdisciplinary Dialogue, Budapest CEU Conference of Cognitive Development. Co-chairs: Rhodri Cusack (Trinity College Dublin), Moira Dillon (New York University). Discussant: Matt Botvinick (DeepMind, University College London). Speakers: Shannon Yasuda (New York University), Koleen McCrink (Barnard College), **Cliona O'Doherty** (Trinity College Dublin); January 5th 2023

Posters

O'Doherty, C., Dineen, Á., Truzzi, A., King, G, Harrison, K., Zaadnoordijk, L., D'Arcy, E., White, J., Holloway, T., Caldinelli, C., Kravchenko, A., Joseph, S., Molloy, E., Byrne, A., Foran, A., Tarrant, A., Cusack, R. (2024), Deep Neural Networks Model the Development of Visual Recognition in Infants. Trinity College Dublin School of Psychology Research Symposium, Dublin, Ireland. - ***Winner of People's Choice Poster Prize***

O'Doherty, C., Dineen, Á., Truzzi, A., King, G, Harrison, K., Zaadnoordijk, L., D'Arcy, E., White, J., Holloway, T., Caldinelli, C., Kravchenko, A., Joseph, S., Molloy, E., Byrne, A., Foran, A., Tarrant, A., Cusack, R. (2024), The development of object recognition in infancy: findings from neuroimaging and deep learning. Organization for Human Brain Mapping; Seoul, Korea.

O'Doherty, C., Dineen, A., King, G., Truzzi, A., Caldinelli, C., Zaadnoordijk, L., D'Arcy, E., White, J., Kravchenko, A., Ambre, C., Wadhwa, A., Healy, M., Burke, A., Joseph, S., Foran, A., Tarrant, A., Byrne, A., Molloy, E. Cusack, R. (2023). Characterising development in the ventral visual stream in young infants with neuroimaging and deep neural networks. Neuro AI Talks (NEAT), Osnabrück, Germany.

Dineen, Á.T., King, G, **O'Doherty, C**., Truzzi, A., Caldinelli, C., D'Arcy, E., White, J., Kravchenko, A., Ambre., C., Wadhwa, A., Healy, M., Burke, A., Joseph, S., Molloy, E., Foran, A., Tarrant, A., Byrne, A.T., Cusack, R. (2023). *Evaluating awake fMRI in one hundred 2- month-olds*. Fetal, Infant and Toddler Neuroimaging Group Conference, Santa Rosa, California, USA - **Oral presentation by Á.D**

O'Doherty, C., Cusack, R., (2023), Is the anterior ventral visual stream tuned to more slowly changing inputs? Organization for Human Brain Mapping, Montréal, Canada.

Dineen, Á., King. G, **O'Doherty, C.**, Truzzi, A., Caldinelli, C., D'Arcy, E., White, J., Kravchenko, A., Zaadnoordijk, L., Burke, A., Healy, M., Joseph, S., Molloy, E., Foran, A., Tarrant, A., Byrne, A., Cusack, R., (2023). 100 Babies: Lessons from Awake Infant fMRI at 2 Months. Organization for Human Brain Mapping, Montréal, Canada.

King, G., Dineen, A., **O'Doherty, C.**, Truzzi, A., Caldinelli, C., D'Arcy, E., White, J., Kravchenko A., Zaadnoordijk, L., Burke, A., Healy, M., Joseph, S., Molloy, E., Foran, A., Tarrant, A., Byrne, A., Cusack, R., (2023), 100 Babies: Lessons from Awake Infant fMRI at 2 months. Trinity College Dublin School of Psychology Research Symposium, Dublin, Ireland.

O'Doherty, C., Cusack, R., (2023), Objects or Context? Learning From Temporal Regularities in Continuous Visual Experience With an Infant-inspired DNN. Trinity College Dublin School of Psychology Research Symposium, Dublin, Ireland.

Dineen, Á.T., King, G, **O'Doherty, C.,** Truzzi, A., Caldinelli, C., D'Arcy, E., White, J., Kravchenko, A., Ambre., C., Wadhwa, A., Healy, M., Burke, A., Joseph, S., Molloy, E., Foran, A., Tarrant, A., Byrne, A.T., Cusack, R. (2023) 100 Babies: an evaluation of awake infant fMRI at 2-months corrected age. Computational Psychiatry Conference, Trinity College Dublin, Ireland.

O'Doherty, C., Truzzi, A., Cusack, R., (2020), Semantic relationships emerge from visual temporal co-occurrences; a statistical analysis of a learning mechanism in early infancy. International Conference of Infant Studies, Online.

List of Abbreviations

AI: Artificial Intelligence
ANN: Artificial neural network
AUC: Area under curve
BOLD: Blood-oxygen-level-dependent
CI: Confidence interval
CNN: Convolutional neural network
CMC: Contrastive Multiview Coding
DNN: Deep neural network
DOF: Degrees of freedom
EEG: Electroencephalography
EVC: Early visual cortex
FFC: Fusiform face complex
fMRI: Functional magnetic resonance imaging
GLM: Generalised linear model
IT: Inferotemporal Cortex
ISI: Inter stimulus interval
INT: Intrinsic neural timescale
LCH: Leacock Chodorow
LLM: Large language model
LO(C): Lateral occipital (complex)
LOTC: Lateral occipitotemporal cortex
MDS: Multidimensional scaling
MEG: Magnetoencephalography
NICU: Neonatal Intensive Care Unit
NCE: Noise contrastive estimation
PHA: Parahippocampal area
PIT: Posterior inferotemporal cortex
RDM: Representational Dissimilarity Matrix
ROI: Region of interest
RSA: Representational Similarity Analysis
RT: Response time
SD: Semantic dementia
TE: Echo time
TR: Repetition time

V1-8: Visual area 1-8

vcov: Variance/covariance

ViT: Vision transformer

VMV: Ventromedial visual area (1-3)

VOTC: Ventrooccipitotemporal Cortex

VTC: Ventrotemporal Cortex

VVC: Ventral visual complex

VVS: Ventral Visual Stream

Chapter 1 A review of conceptual representation learning in biological and computational vision systems.

Much of what we know is learned from what we see. Human experience is defined by our marked ability to learn about the world around us, drawing deep understanding from rich streams of sensory input. We do this before we can even speak, easily developing the ability to link recognition with conceptual understanding. We know that a dog is a friendly animal that is similar yet distinct from a cat. That a tiger – although a cat – is dangerous and to be feared; that it is expected to appear amongst trees in the jungle. Questions on the nature of this conceptual understanding go back as far as history can track, in a conversation that spans multiple research fields from philosophy and psychology to cognitive science, neuroscience and computer intelligence. Focusing on how conceptual knowledge may emerge from visual recognition, I will review where the fields of cognitive neuroscience, psychology and computer vision stand in terms of accounting for visual understanding. Deep learning networks continue to excel as models for human vision, but do they fully account for the wealth of semantic information present in our neural representations? The human vision literature is vast, but are we missing key insights from statistical signals that occur in naturalistic, contextualised settings? Importantly, what is the starting point for this system, and with what tools are we equipped in our early years to aid in concept formation? I argue that progress can be made in both computer and biological vision research by looking for inspiration in the most efficient learning systems that we know of, the human infant, by thinking about how our early environment and neural processes guide and constrains emerging knowledge.

1.1 What is a concept, and how might this emerge from visual perception?

The physical information we perceive must be transduced from sensory input into abstract or long-term knowledge. As such, conceptual understanding of the things we see should have a storage unit within the brain, or, within an artificially intelligent system. This idea of storing knowledge is formalised in the notion of mental representation. Representational theory of mind, which dates back to Aristotle, describes how cognitive states and processes are constituted by the occurrence, transformation and storage in the mind and brain of information-bearing symbols or structures called representations (Pitt, 2022). In this way, representational states can be activated and used when we encounter certain stimuli or environmental cues. For example, the perception of a dog involves the observation of many types of input from the animal's motion and the textures of its fur to the sound of its bark. All of this sensory input must be interpreted and transduced by the brain, characterised in a cognitive or neural mechanism. Indeed, these representations house myriad semantic information such as what a dog is, how it relates to other animals (for example, that it is similar to a cat in that it is expected to be a pet) and perhaps even emotionally valent memories of one's own dog. A mental representation is, therefore, not simply the fact that the perceived object is attached to the verbal label. Instead, mental representations house rich semantic knowledge for the concept that extends beyond the basic functional output of attaching the label to the referent.

The notion of a representation, its definition and usefulness for different fields, is highly debated (Baker et al., 2021; Rowlands, 2017). One such point of contention is the role of sensory experience. Traditional approaches assume that representations of knowledge are housed in the form of an amodal, internal symbol that lies independent from the brain's modal sensory regions, although this extreme perspective has always been met with

opposing arguments (Markie & Folescu, 2021). While the purely rationalist views of Plato and more modern philosophers such as Descartes, Leibnitz and Kant argue against concepts that are grounded in sensory experience, a complete lack of consideration for sensory components seems unfounded; the entry point for most information is through a sensory modality be it viewing a novel object, feeling a new texture or listening to an interesting fact. The importance of modal inputs for the mind's representations of knowledge goes back again to ancient philosophers such as Aristotle and more recent empiricists including Locke and Hume argue that all concepts should be derived from sensory experience. The importance of perception is stated in David Hume's principle of association, as he believed that all knowledge is derived from experience and must be analysable in terms of perceptual content (Hume, 2000), although his views went so far as to refute the existence of any innate ideas or theories. More recently, Lawrence Barsalou presents influential scientific arguments for knowledge being grounded in experience. The grounded cognition framework wholly rejects the need for amodal symbols and instead argues for an embodied view of concepts in which all knowledge comes from some form of sensory input (Barsalou, 2008). Considering each of these perspectives, I take the view in this thesis that representations originate in sensory experience – specifically, vision – but exist as a more abstract entity in the biological or computational system, taking the form of unique patterns of functional activity. These representational patterns can be activated in a reliable manner to access conceptual knowledge, thereby facilitating recognition and understanding.

This view ties naturally with the definition of knowledge as it pertains to memory of useful information, something that can be accessed and employed for a specific purpose. Semantic memory regards knowledge of general facts and information about the world and semantic cognition refers to our ability to use, manipulate and generalise knowledge as we experience

our life and environment. Studies of memory formation in the brain focus on the hippocampus, which is beyond the scope of this thesis, but representations in high-level visual regions are similarly concerned with semantic knowledge of visual categories (Haxby et al., 2001; Huth et al., 2012). In this way, high-level vision offers a chance to study the boundary of perceptual and conceptual representation in the brain. This cortical structure culminates in the anterior temporal lobe, which is causally linked to semantic cognition as evident from clinical lesions in semantic dementia (SD) (Czarnecki et al., 2008; Mummery et al., 2000). SD patients exhibit a behavioural deficit in the ability to access conceptual knowledge despite otherwise preserved cognitive function. For example, the patient may fail to retrieve the word 'dog' when shown an image of a dog, and they may not even understand what it is they are being shown. This finding is highly reliable and pathology is very predictable, with patients showing deficits in accessing meaning across all conceptual domains rather than specific categorical deficits (Hodges & Patterson, 2007). Together, these findings point to a clear role of ventrotemporal cortex (VTC) in semantic processing of concepts, especially as it pertains to recognition of visual categories. But what is the structure of a concept within the mind, and how can we empirically test it? Experiments in psychology have shed light on this, especially for the cognitive function of categorisation.

1.2 Psychological accounts of categories and concept formation

Many psychologists have tried to define what a category is, and find useful theoretical heuristics for how humans form such conceptual groupings. The classical theory of concepts seeks concrete definitional structure, for example that a triangle is something with three sides (Margolis & Laurence, 2023), but most attempts at finding such definitions fail. Moreover, a definition-based approach cannot account for empirical findings from psychological experiments such as the highly-reproduced typicality effect. Participants

consistently rate certain members of a category to be more representative of that category than others; for example, reaction times are faster for responding ‘True’ to the sentence ‘A robin is a bird’ than to the sentence ‘A chicken is a bird’, suggesting that a robin is seen as more typical for the concept of bird (McCloskey & Glucksberg, 1978). Although the classical theory is not inconsistent with such findings, it does nothing for explaining them.

Wittgenstein suggested instead that family resemblance may be a means of clustering concepts together so that we can be pointed towards relevant features. This idea of family resemblance is further developed in Eleanor Rosch’s Prototype theory (Rosch & Mervis, 1975). An experiment in which participants were asked to list from memory attributes of categories across a range of typical examples revealed there are very few shared properties across the instances of a category, as would be predicted by the definitional approach. For example, both a chair and a telephone could be considered furniture but lack any clear feature similarities. Instead, instances that would be considered more typical of a furniture, the chair in this example, share many attributes with more group members. This can be quantified with a family resemblance score, which increases alongside typicality, meaning that a prototypical example of the concept can be found which is the statistically average member of that category. Thus, Rosch’s prototype theory describes how a mental prototype is formed that represents the average picture of a concept even if such a thing has never been experienced. This probabilistic feature-based approach to concept formation is further formalised in the Conceptual Structure Account, which describes a potential neural instantiation semantic object processing (Taylor et al., 2007). This cognitive model focuses on the internal structure of a concept by relational structure, proposing that distinct features are stored in a distributed manner across the brain. Concepts with highly-correlated features

will be represented similarly as their high degree of feature correlation will activate more similar neural patterns.

A probabilistic views of concepts make sense when we consider how information is extracted from the environment, linking to evidence that infants engage with statistical learning (Saffran et al., 1996). Moreover, the statistical focus is parsimonious with the recent success in Artificial Intelligence (AI) using machine learning; evidently, learning from statistical regularities in the sensory environment can aid in concept formation. This may operate at many levels of complexity and can manifest in multiple systems. For example, at the pixel level for a computer vision model, at the feature level for infants identifying novel categories (Younger & Fearing, 1998) or across categories to rapidly process relational structure (Hafri & Firestone, 2021). The work presented in this thesis leans heavily on this approach, examining how visual concepts emerge from statistical signals in our environment and often using Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008) to examine the relational structure of human and computer vision.

1.3 From images to visual categories in brains and machines

Recognition is the first step in concept formation, and visual recognition plays a key role for the kinds of categories and semantic knowledge which I will address in this thesis. The site of object recognition in human visual cortex is on the ventral occipitotemporal pathway known as the Ventral Visual Stream (VVS). This ‘what’ pathway of vision is tasked with recognising objects in space and is separate to the ‘where’ dorsal stream which focuses on tasks related to motion or vision for action (Goodale & Milner, 1992; Mishkin et al., 1983). As discussed above, this cortical pathway has specific functional activity that encodes diffuse semantic structure (Haxby et al., 2001; Huth et al., 2012), with specificity in

particular regions for categories such as faces, body parts and scenes (Downing et al., 2006; Epstein et al., 1999; Kanwisher et al., 1997). In this way, the VVS represents visual categories across diffuse patterns of activation and also with reliable clusters of categorical regions, culminating in the locus of semantic cognition at the anterior temporal pole (Mummery et al., 2000). The VVS shows a degree of evolutionary conservation, with comparable structures in macaques (Orban et al., 2004). Structured along a posterior to anterior gradient, earlier regions in this hierarchical ventral pathway such as V1 and V2 respond to simpler critical features than more anterior regions (V4, IT) (Kobatake & Tanaka, 1994) and invariance to transformations in the input image increases along the pathway to enable robust recognition performance (Rust & DiCarlo, 2010). The complexity of the information also increases from posterior to anterior. Early visual cortex (EVC) processes features like luminance or orientation whereas abstract features are processed later in the pathway within anterior regions such as IT in macaques and lateral or ventral occipitotemporal cortex (LOTC/VOTC) in humans (Orban et al., 2004; Tanaka, 1996).

Fascinatingly, this hierarchy has also emerged in artificial neural network (ANN) models capable of object recognition. Early connectionist approaches to computer vision took inspiration from the human visual system (Fukushima, 1988), but a revolution in AI did not come until 2012 with the introduction of AlexNet (Krizhevsky et al., 2012) and large-scale datasets such as ImageNet. This deep convolutional neural network (CNN) that was capable of state-of-the-art performance on a common computer vision benchmark: the ImageNet Large Scale Visual Recognition Challenge (Russakovsky et al., 2015). The convolutional structure is loosely inspired by biological vision's pooling mechanism from complex to simple receptive field sizes (Hubel & Wiesel, 1962), but really the innovation was in the depth of the network and the size of the datasets on which the model was trained. The deep

learning revolution led to huge leaps in machine object recognition but, serendipitously, their improved behaviour from training on larger and larger datasets revealed their candidacy as models for human and macaque visual systems.

Computational models for neural data were now being optimised on the task of object recognition and not directly constrained by physiological recordings. This shift towards modelling goal-driven behaviour led Yamins and colleagues to describe the first quantitatively accurate image computable model of spiking responses in IT cortex (Yamins et al., 2014), following up with a study showing that the then emerging deep neural networks rivalled the representational performance of IT in a visual recognition task (Cadieu et al., 2014). Further elegant studies went on to affirm the power of these CNNs in modelling neural responses (Cichy et al., 2016; Eickenberg et al., 2017; Khaligh-Razavi & Kriegeskorte, 2014), finding striking correspondences between the hierarchical models and the VVS (Güçlü & Gerven, 2015). Indeed, deep neural networks are now well-established as leading models for the brain in cognitive computational neuroscience (Richards et al., 2019; Storrs & Kriegeskorte, 2019) and can be used to generate specific neuroscientific hypotheses to provide novel insight into brain function (Doerig et al., 2023). Here, I explore one such direction by examining how statistical signals in perceptual input are transformed into visual representations, and what features are extracted to facilitate object recognition and understanding.

1.4 Learning from the statistics of the environment in high-level vision

There are numerous statistical regularities in visual input that can guide formation of semantically meaningful representations from purely perceptual input, often discussed in terms of statistical learning. This is a domain-general learning mechanism that spans

modalities, to which we are sensitive from early infancy (Fiser & Aslin, 2002; Kirkham et al., 2002; Saffran et al., 1996). In the brain, signatures of statistical learning are present in numerous regions including medial temporal lobe and frontal cortex; there is evidence for learning from concurrent sequences in as early as V1 (Rosenthal et al., 2018) and suppressed activation to predictable object pairs are present throughout the ventral stream (Richter et al., 2018). Recent work in macaques used a behavioural training paradigm followed by fMRI recordings to show that more regularly recurring pairs of objects resulted in different processing patterns to random pairs of objects in occipitotemporal cortex and EVC (Vergnieux & Vogels, 2020). This probabilistic structure could take the form of consistent feature associations within an object, how an object relates to the wider context of its physical or social environment, or its co-occurrence with other objects (Hafri & Firestone, 2021; Sadeghi et al., 2015). These signals act as clues regarding the object's conceptual significance, which can in turn guide our understanding alongside the ability to recognise from perceptual similarity.

Considering the case of contextual associations, early psychological experiments reveal that successful recognition of an object is facilitated having first seen its congruent context (E. Palmer, 1975) and semantic relations between entities in a scene are rapidly accessed to facilitate identification (Biederman et al., 1982). Similarly, the brain is sensitive to these associations across space and time. EEG analyses reveal that expected spatial configurations of objects facilitate the extraction of contextual associations between objects that tend to co-occur, with a larger signal arising in occipitotemporal cortex when associated objects are typically positioned (Quek & Peelen, 2020). The general spatial configuration of a scene enables rapid "gist" perception; humans can quickly understand a scene's general meaning even with limited information, provided that the global spatial structure is consistent (Oliva

& Torralba, 2006). On the time dimension, items that are temporally associated give rise to similar internal representations causing them to be perceived as more similar (Schapiro et al., 2013).

To summarise, evidence suggests that the relevant machinery is in place for us to be sensitive to patterns in the things we see which informs our understanding of the objects we encounter in the world. The rapidly emerging success in machine learning supports this approach for developing intelligent behaviour in an artificial system. However, the presence of these mechanisms in the fully developed brain tells us little about the foundational processes that shape our ability to perceive and interpret the world. While the mature brain excels at extracting meaningful patterns from the environment, this proficiency does not reveal how these capabilities develop from the ground up. For more holistic understanding, we must turn to the early stages of these systems; namely, infancy. By studying development, perhaps we can uncover the initial signals and mechanisms that guide concept formation and pattern recognition, offering a window into the mechanisms of representation learning.

1.5 Infant development provides insight into visual representation learning

The foundations of visual and conceptual representations in infancy is discussed extensively in developmental science. Researchers including Spelke and Carey argue for the role of innately specified, core cognitive functions (Carey, 2009; Spelke & Kinzler, 2007), in opposition to sensory-driven bottom-up views of development from classical empiricists such as Locke or Piaget (Locke, 1948; Piaget, 1952). Others believe that cognitive abilities are rooted in innate perceptual mechanisms, but much of conceptual development is shaped by environmental interaction (Mandler, 2004). Nonetheless, empirical evidence shows that infants can distinguish categories in some circumstances; for example, 3-4 month olds show

different looking patterns for dogs and cats (Quinn & Eimas, 1998), and more recent studies in developmental neuroscience shed light on the emergence of visual categories throughout the first year of life (Kosakowski et al., 2022; Xie et al., 2022; Yan et al., 2023). This suggests that the infant brain is engaging in complex function, actively interacting with the environment *via* perceptual processes. By studying the structure of this environment in early life, we begin to understand the kinds of representations that may emerge through development and what this can tell us about learning of concepts in general.

There is compelling evidence that infants' visual input is tailored in towards facilitating better representation learning. Using headcams attached to infants in a naturalistic play setting, Linda Smith and colleagues have revealed the interesting ways that young children create their own optimal "training data" (Pereira et al., 2014). The authors show that when playing with toys and having parents assign labels or names to the objects, those words that were successfully learned post-play had a distinct visual signature. Infants manipulated the object so that it was centred, taking up a large portion of the visual field and sustained in view before and after word naming. These types of manipulations are increasingly common with development and are prevalent in results from headcam studies in older toddlers (Yu & Smith, 2012). When examining headcam data from mealtimes, it was shown that despite a typically cluttered visual environment a very small set of objects are present much more often; few things in the environment are very common and many things are rare with the most frequently encountered objects going on to become earlier learned words (Clerkin et al., 2017).

This differs starkly to the input provided to machine learning models, which typically involves massive datasets with diverse image examples. However, there is growing interest

in incorporating developmental insights into CNNs with the hypothesis that doing so could enhance performance in machine learning (Smith & Slone, 2017; Zaadnoordijk et al., 2022). For instance, data from infant headcams has been shown to facilitate more robust learning and generalisation in CNNs compared to models trained solely on adult-derived images (Bambach et al., 2018). More recent efforts have leveraged recordings from a single child to train a multimodal vision and language model, achieving impressive results despite having access to significantly less data (Vong et al., 2024). This egocentric video-based approach is now being expanded to include recordings from many more infants (Long et al., 2024), marking an exciting advancement in our understanding of infants' early visual environments and the potential learning signals they harbour, but a significant data gap still exists between infant and machine learning (Frank, 2023).

Current intuition in the field is that the subset of data selected by infants through their self-generated views of the world provide more suitable training inputs for learning about objects, but equally, this efficiency could be ascribed to optimised learning mechanisms or better inductive biases that are present in the innate wiring of the brain. Regardless, by examining the early environments of infants we highlight how the data we are currently using to train CNNs are impoverished; not only do we input unnaturalistic content with an unrealistic distribution of categories, but these images have been subject to adult judgements of what is a “good image”. Perhaps, the infant generated views are inherently more instructive to visual learning and expanding to these naturalistic data will introduce the temporal and spatial statistics that are informative for concept learning.

1.6 Limits of CNN models for high-level vision

Do current CNN models for object recognition fully capture the semantic structure of concepts? This seems to be an underappreciated oversight in the field. The engineering approach to AI focuses on successful performance, for example high accuracy in an object labelling task. But this does not tell us about the underlying structure that is driving this behaviour. A notable example of this dissonance is the shape vs. texture bias in CNNs (Geirhos et al., 2022). Models that can both recognise objects and predict neural data are strongly biased towards texture-based recognition, in contrast to human recognition which is driven by shape. The models are susceptible to adversarial attack, making incorrect predictions after tiny perturbations to input images (Brendel et al., 2018) whereas humans show excellent generalisation across many visual tasks (Geirhos et al., 2018). This is indicative of the specificity in model representation learning, and their lack of focus on configural based structure at the global level which can facilitate wider conceptual understanding.

Similarly, CNNs can fall short as models of the brain when applied to high-level visual cortex. Using fMRI data with improved signal to noise, 14 popular CNNs were tested which had been pretrained on the standard ImageNet dataset (Xu & Vaziri-Pashkam, 2021). While this replicated the mapping of lower neural network layers to earlier regions in the ventral stream, an improved estimation of the noise ceilings found higher layers of the networks were predictive of brain responses in regions such as LOTC and VOTC but the quality of the prediction was poor. The highest amount of explainable variance was 60% in high-level regions whereas many of the CNNs tested could fully capture the variance of lower visual areas (e.g. V1-V4). This finding raises one interesting limitation of the deep learning models of the brain, of which there are many (Bowers et al., 2023). In this instance, it appears that

their learned representations are missing something in the high-level information that would be present in semantic, anterior visual cortex.

When we consider the manner in which these neural network models are trained, it's not surprising that they fail to fully capture neural representations. Most of the networks found to be predictive of brain data in the above studies were trained with a supervised machine learning approach using millions of images and their associated labels. This aligns neither with how humans learn to recognise, nor the environment of the infant as measured in headcam studies. We spend our early days observing and interacting with the world around us, with very little in the way of supervision except for when our caregivers explicitly name things for us, which, when considering experience as a whole, is not very often. Human learning is more akin to self-supervised or semi-supervised machine learning methods, and our training data is vastly different. It is feasible that the probabilistic spatial and temporal signals in the environment are learned through an unsupervised process, thereby facilitating the representation of meaningful relational structure that has been discussed as a basis of conceptual understanding.

The importance of relations for concepts is highlighted in modelling work that examines the alignment of systems across modalities (Roads & Love, 2020). The authors show that the distinct signature of a concept in one system (say, a visual CNN) is recapitulated in another (a text-based model), such that the idiosyncratic signals between the concepts within each can be used to align the two systems. This results in a more holistic and meaningful representation. What's really key here is that the relational structure guides understanding; if everything was perceived as equidistant then alignment across systems would not be possible. Furthermore, a system comprising few concepts is better aligned across modalities

by restricting efforts to those words learned earlier in life. Similar to the benefits seen with training on infant-generated data (Bambach et al., 2018; Vong et al., 2024) this finding again points towards some optimal system in infant conceptual learning.

1.7 Aims and objectives

In this review, I have discussed the structure and form of conceptual representations in the mind, brain and machines. I highlighted how high-level vision provides a link between perception and semantic cognition, but that much remains unknown regarding the mechanisms of this process. I argue that the ability to understand global structure at the level of relations between objects guides understanding, and that it is learned through a self-supervised process from spatiotemporal statistics in naturalistic experience. Finally, I have discussed limitations of current CNN models for high-level vision, and I propose that these can be overcome by looking for inspiration in the infant brain. In this thesis, I will explore this topic using an interdisciplinary approach.

In Chapter 2, I propose an infant-inspired deep learning model for semantic representation learning. Noting a considerable gap in the temporal scale on which CNNs are trained, I test the hypothesis that access to long-range temporal co-occurrences from in naturalistic experience will result in learning a more semantically meaningful embedding. I approach this by creating a novel dataset from movies, and implementing a contrastive temporal learning objective into a model trained for object recognition.

Following from this, I hypothesise in Chapter 3 that semantic information is processed by anterior VVS when categories are presented at a long timescale. I explore the neural capacity of the brain to respond to these temporal signals. I test the tuning of visual cortex to stimuli

occurring at different temporal scales, and if this is linked with the intrinsic neural activity of the regions. This was executed using task-based and resting state fMRI in a cohort of healthy adults.

Chapter 4 presents novel findings regarding the origins of visual representations in the 2-month-old brain. Using the FOUNDCOG dataset, I apply advanced multivariate techniques to characterise VVS representations and, for the first time, I successfully apply DNN models of the ventral stream to an infant cohort. In this study, two prominent theories of visual development are explored to find how much bottom-up sensory experience is required for rich visual responses, and how much is innately specified.

It's important to note that throughout this thesis, I discuss concepts primarily in relation to visual perception. However, this focus does not imply that vision is essential for conceptual understanding, as many individuals who are born blind can process concepts independently of visual input. Although this topic is touched upon in various chapters, it falls beyond the scope of the work presented here. Nonetheless, it is crucial to acknowledge that vision is not a prerequisite for developing conceptual understanding.

The overarching aim of this work is to bridge the fields of machine learning, functional neuroimaging, and infant developmental neuroscience. This interdisciplinary approach paves the way for a new research program in developmental cognitive computational neuroscience. I strongly believe that continued efforts in this domain will yield mutual benefits: advancing our understanding of the inner workings of the mind and enhancing the capabilities of the next generation of AI.

Chapter 2 Infant-inspired contrastive learning of semantic structure from temporal regularities in visual experience.

Deep convolutional neural networks (DNNs) capable of object recognition have been instrumental in driving the past decade of computer and biological vision research, but the methods by which they are trained differ starkly from human learning. Many of these models ignore the extended dynamics of naturalistic experience that we learn from as young infants, and are trained instead on highly curated datasets of static images and their corresponding labels. Furthermore, much of the focus in machine learning is on chasing benchmarks for object classification accuracy which ignores the features that underly accurate performance. For example, we do not know if a DNN uses perceptual or semantic features to recognise or if the wider context of the category is understood. Contrastive self-supervised models are capable of learning useful representations from the hidden distributional structure of input data, without a reliance on labelled datasets. This offers a means to test what kinds of representations can be extracted from naturalistic input data using an objective that is more aligned with human learning. Here, we implement an infant-inspired contrastive learning mechanism into a self-supervised DNN that leverages commonalities in naturalistic video over various timescales. We hypothesise that the distance in time between two training images will determine the kind of representation learned by the network. Using representational similarity analysis, we find that commonalities across longer timescales reflect relational structure and scene context which change relatively slowly in the dataset. This was despite attenuation in object classification performance, revealing that this metric does not always indicate that the wider meaning of a concept has been learned.

This work has been presented at:
Neural Information Processing Systems (2020); workshop on Self Supervised Learning: Theory and Practice, Virtual Conference.
Computer Vision and Pattern Recognition (2021); Women in Computer Vision Workshop, Virtual Conference.
Cognitive Computational Neuroscience (2022), San Francisco, California, USA.

2.1 Introduction

Object recognition is an important function for both biological and computer vision. In recent years, deep convolutional neural networks (DNNs) have advanced considerably and achieved performance levels comparable to adult humans (He et al., 2015). This technology is now part of our daily lives – for example, the photo app on your phone can seamlessly recognise your dog from your cat and group them into a smart photo album – and DNNs are used increasingly to understand the foundations of human visual recognition (Yamins & DiCarlo, 2016). However, the focus in machine learning has historically been on chasing benchmarks and reliance on classification accuracy as a defining metric of success. This has led to success in computer vision with methods such as supervised learning from large labelled datasets (Krizhevsky et al., 2012) or training without labels but on a vast amounts of data (Liu et al., 2021). But for a system to be considered intelligent, does it merely need to recognise something and label it or must it also understand the meaning of things in context? Human object recognition is defined not only by having distinct responses for particular categories (Epstein et al., 1999; Haxby et al., 2001; Kanwisher et al., 1997), but is organised by similarity structure (Contier et al., 2024) and contextual cues (Bar et al., 2006). Moreover, human categorisation occurs at multiple levels from superordinate to basic (Rosch & Mervis, 1975); a car can be classified as a vehicle, and also the specific brand and model of one's own car. By optimising DNNs on the singular goal of object classification, are they missing this broader relational structure for the categories they can recognise?

Models for object recognition are typically trained large datasets of static images comprising many classes across a broad distribution. In contrast, naturalistic experience contains more slowly changing inputs in fewer contexts. This is especially true during the period where we must rapidly learn about the world around us: early infancy (Smith & Slone, 2017). As young

infants, we are exposed to a much narrower distribution of visual inputs compared to a DNN during training. Instead, we interact with the environment to learn a lot about only a few things, building categorical distinctions for categories such as cats vs. dogs from as young as 3-months-old (Quinn & Eimas, 1998). This is achieved from a continuous stream of visual input, without labels. It's the regularities in input that guide conceptual learning, extracted from the environment through a process of statistical learning (Kirkham et al., 2002; Saffran et al., 1996). Early learning from the latent structure of the environment – for example consistent feature associations (Fiser & Aslin, 2002), object co-occurrences (Sadeghi et al., 2015) or relations (Hafri & Firestone, 2021) – enables us to learn not only the name of an object we see, but a wider understanding of the concept. Computer vision still struggles with flexible and generalisable representation learning that facilitates more holistic conceptual understanding. There is a known local feature bias in DNNs compared to recognition at the global level in humans (Brendel & Bethge, 2019). Furthermore, the representational similarity structure present in performance-optimised DNNs is indeed explanatory for human behavioural and neural responses (Yamins et al., 2014), but falls short in capturing all of the variance of high-level visual regions that drive human recognition (Xu & Vaziri-Pashkam, 2021). This suggests that DNNs are missing a something in way of the relations that are important for capturing the wider contextual and semantic structure of visual concepts.

This is not surprising when we consider what is missing from the training data provided to these networks. Naturalistic experience contains many spatial and temporal regularities, the latter of which cannot be captured from static images. We propose that naturalistic temporal structure, especially over longer periods, will help a DNN to learn how an objects relate and facilitate semantic and contextual representation learning. For example, when scanning our

eyes in the living room the exact perceptual input to the retina changes rapidly by the moment but the legs and arms of the chair in the corner remain consistent, allowing us to perceive the chair as a whole. Beyond a singular snapshot of the living room scene, we may observe that the chair and coffee table remain beside each other and are consistently associated with the context of the room itself. Visually distinct objects are therefore bound by co-occurrence through time which may provide a useful cue for learning how objects relate in a broader context. By training DNNs on only static images, we preclude representation learning of this holistic scene understanding.

Temporal signals have previously been used to drain DNNs for object recognition, but with a focus instead on perceptual similarity at short timescales across successive video frames. The slowness principle of feature learning for object recognition used the coherence from frame to frame to learn invariances for features such as translation, rotation and contrast (Wiskott & Sejnowski, 2002) and the idea of temporal coherence was combined with deep learning for pose and face recognition (Mobahi et al., 2009). Further interesting work used insights from how infants would naturally interact with objects to generate views across time as augmentations in a contrastive learning task (Aubret et al., 2022), again across a short timescale that is representative of manipulating an object in one's hands. The intuition behind these models is that the perceptual similarities across frames will be similar enough to learn useful object representations, while having enough difference to do so in an invariant manner. We propose that by extending this intuition to dozens of seconds or even minutes will enable a DNN to learn the associative structure that is informative for semantic and contextual relations.

Inspired by human infant learning, we constructed a unique movie image dataset in which the temporal structure between images is preserved across short and long timescales. We explored whether perceptual and semantic features prevail at different timescales in the movie dataset and use these temporal signals to train a self-supervised contrastive learning network, SemanticCMC. By constructing positive pairs for contrastive learning based on the associations between images at various time lags we test whether semantic and contextual information are maximally informative at longer timescales, and how this influences the representations learned by SemanticCMC compared to models trained on conventional image datasets that lack temporal depth.

2.2 Methods

2.2.1 Generation of a movie image dataset

Feature length movies were used to obtain an image dataset with preserved temporal co-occurrence patterns. This allowed us to manipulate the time between images over a long timescale and across different scenes. 100 movies totalling 158.4 hr (mean movie length = 95.04 min) were collated and included in the dataset. To approximate infant visual experience, the selection criteria for movies were that they should be live-action, set in a naturalistic scenario and without fantastical visual content; for example, no superheroes or fantasy movies with heavy special effects.

An image database was generated from the 158.4 hr of videos by taking a still image for every movie at a sampling rate of 1 image per second using FFmpeg. This resulted in 572,949 images, a large enough dataset size to explore the timeseries of object occurrences and for later training of a deep learning network. The temporal structure of the videos was preserved in their sequence such that two images separated by a specified time lag were

related. For example, two images 1 s apart would have minor perceptual differences whereas two images separated by 5 min may be from the same context but have very different low-level visual statistics. The movies were concatenated into a single dataset before sampling. At the boundaries between movies two images would have very different statistics, but this case was rare (0.017%) and can be neglected. The same applies to scene changes which can be abrupt, contrasting with the smooth transitions we experience in the real world. However, an infant may experience something similar, where context changes occur while sleeping or traveling in a car or carrier, which obstructs their view.

To test our first hypothesis that semantic similarity will persist at longer timescales than perceptual similarity, the objects in an image were found using automatically generated labels. Images were tagged using Amazon Rekognition and a set of labels was returned for every 200 ms of video, resulting in 2,851,272 label sets comprising 41,330,953 labels of which 2,466 were unique. These were then used to test the hypothesis that objects persisted for long periods of time, over and above perceptual similarity between images. Some tags occurred infrequently and so only the 150 most frequently occurring labels with a noun part-of-speech tag in the WordNet database were included. This ensured that objects with a representative sample were included in further statistical analyses and limited computational expense.

2.2.2 Estimating the timeseries of low-level and high-level visual information

The automatically generated labels were used to estimate the persistence of object associations across time. Labels were transformed into a one-hot encoded binary array. Using an autoregressive ridge regression model ($\alpha = 1.0$), the probability of appearance of the 150 objects at one timepoint (t_0) was predicted from the set of objects present at a lagged

interval earlier (t_{lag}) ranging from 1 lag of 200 ms to an increasing lag distance ($\Delta t = t_{lag} - t_0$) of up to 1 hr. The coefficient of determination (R^2 score) of the autoregressive model was used as an estimate of how well t_0 could be predicted from t_{lag}, assuming that higher R^2 means stronger associations between objects at that lag distance. To measure the analogous timeseries of perceptual change, the root mean squared (RMS) difference between two images was calculated over an increasing lag distance. For every movie, RMS was calculated between the first image and another at an increasing lag up to 60 min. Increasing RMS difference indicates a larger difference in perceptual similarity; the same image compared to itself would have RMS = 0. For visualisation purposes and comparison to the decreasing R^2 scores of the object regression, the mean normalised negative RMS score ($-RMS_{norm}$) was plot to view the decrease in similarity over time. To estimate of the 95% confidence intervals of the R^2 metric, separate regression models were fit on 16 subsamples of the movie dataset. The perceptual $-RMS_{norm}$ metric was estimated across all movies. The rate of decay for the perceptual ($-RMS_{norm}$) and semantic (R^2 score) across increasing lag distance were estimated using non-linear least squares fit of the following exponential decay function in Equation 2.1:

$$f(t_{lag}) = Ae^{-t_{lag}/b} + C \qquad (2.1)$$

Where A is a scaling factor, C is the offset, t_{lag} is the lag distance in minutes and b is the time scale of the rate of decay expressed in minutes.

To measure if the objects that were more associated across time were grouped into meaningful categorical clusters, Hierarchical clustering with Ward linkage was performed on the pairwise matrix of regression coefficients for the 150 objects. The resulting clusters

were interpreted by the researcher to investigate if meaningful superordinate category clusters could be assigned.

2.2.3 Training a deep neural network on long temporal statistics

A new temporal contrastive learning objective, SemanticCMC, was implemented to test our second hypothesis that a DNN could learn semantically relevant representations from the extended temporal signals in the movie dataset, but at short timescales the perceptual signals would dominate. SemanticCMC was implemented as an adaptation of Contrastive Multiview Coding (CMC) (Tian et al., 2020). This self-supervised model learns mutual information across modalities or viewpoints by leveraging co-occurrence patterns across multiple augmentations of the training data. Noise contrastive estimation (NCE) loss (Gutmann & Hyvärinen, 2010) is used to estimate the input distribution against a defined noise distribution, by comparing two data points in the latent space. CMC offers flexibility in the choice of modality, encoding network and choice of auxiliary task. We introduce SemanticCMC, which implements a new temporal learning objective. Two images separated by a specified lag distance Δt, which is input as a hyperparameter, are passed to the encoder head of the model. The same full-size AlexNet architecture (Krizhevsky et al., 2012) was used here, but this could in theory be any visual network. NCE loss is calculated in the embedding space by identifying the positive pair of images, t_0 and t_{lag} from a large pool of random negative samples in the training dataset. Positive pairs are defined as the two matching input images provided within a batch, separated by the Δt hyperparameter. Negative pairs are obtained by random sampling from a uniform distribution of 16,384 images. Critically, these could be from any distance in time across the entire dataset and will not have associated temporal structure. It is assumed that the network will learn the information that is common across the images at Δt. Training SemanticCMC requires that

images have meaningful temporal co-occurrence patterns, necessitating the movie dataset as described above.

An important factor in contrastive learning is the augmentations applied to the data to increase variation the amount of training data and embed invariances in the learned representation. In agreement with previous temporal contrastive learning work, it was found that colour-based augmentations are important for training (Aubret et al., 2022). Random resize cropping, random horizontal flipping and the colour jitter method described in (Chen et al., 2020) were applied to all input images, which were then normalised to an appropriate range using a specified mean and standard deviation. During evaluation, the same augmentations as training were applied but with centre cropping instead of random resize crops.

2.2.4 Baseline networks

Each instantiation of SemanticCMC was compared to a set of baseline networks. This was to compare differences in the learned representational structure of the models with access to longer temporal co-occurrences and those trained on static images. Firstly, CMC as published by Tian et al. (2020) trained on the split-channel *L*a*b* colour augmentation task which had been shown to achieve very high ImageNet classification accuracy. CMC was then trained from scratch, as described in Tian et al. (2020), but using the movie image dataset. This acted as a control for the quality of the new dataset for object representation learning on a proven learning objective. A third baseline was supervised AlexNet, loaded from torchvision v0.15.2. Finally, an untrained AlexNet model was included, as the architectural structure of a deep learning model can provide sufficient inductive bias to extract object-relevant information at a low level of abstractedness (Gaier & Ha, 2019).

2.2.5 Implementation details

All analyses were coded in Python 3.7 using PyTorch 1.4.0 with CUDA v10.2, and run on RTX 6000 GPUs each with 24 GB in a Lambda quad workstation with 28 CPU cores and 128 GB of RAM. The movie-CMC baseline on the *L*a*b* task ran to convergence at 200 epochs with a batch size of 128 and a learning rate of 0.03 with decay by 0.1 at epochs 120 and 160 using a SGD optimiser with 0.9 momentum. As originally described by Tian et al. (2020) the encoder was a SplitBrain AlexNet architecture. Batch normalisation was used and images were transformed into the *L*a*b* space with a corresponding mean and standard deviation used for input normalisation. Linear decoding was performed on top of AlexNet convolutional layer 5 for 60 epochs using an SGD optimiser with an initial learning rate of 0.1 and decay by 0.2 at epochs 30, 40 and 50.

Temporal training was performed using one full-sized AlexNet as the encoder network. AlexNet was initialised with the published weights for CMC, as trained by the *L*a*b*-ImageNet task and then finetuned on our temporal objective. The network was trained from scratch on the SemanticCMC objective, but it emerged that the computational expense did not confer any benefit. Moreover, semantic analyses revealed that when training from scratch, coding of semantic relations was not as strong as with the fine tuning procedure (although the trend in the results reported below persisted). This led us to conclude that prior knowledge of object features is a useful basis for learning semantic structure, and motivated our choice of finetuning procedure. Temporal finetuning was run for 80 epochs, with a batch size of 128 and a learning rate of 0.03 with decay by 0.1 at epochs 30, 50 and 70. Batch normalisation was used, and a SGD optimiser with momentum of 0.9. Note that Δt was a hyperparameter and unique to each training instance. We hypothesised that at short values

for Δt the perceptual similarity between the two training images would dominate, but that as Δt increases semantic features would be learned. Hence, models were trained with Δt = 1 s, Δt = 10 s, Δt = 60 s and Δt = 5 min. This range of values for Δt considerably extends the timescale to which DNNs typically have access during training, while ensuring that two images drawn from the movie dataset during training are likely to have associated contexts. Four instances of each Δt model were run to generalise across initial randomisations (Mehrer et al., 2020).

2.2.6 Representational similarity analysis

It was hypothesised that the networks trained on positive pairs of images that are associated by extended temporal lags would enable more semantic representation learning. Object-classification accuracy, the evaluation for which these models were designed, would not suffice to test this and so we turned to Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008). This method characterises a system's representation by the pairwise distance or similarity matrix of its response patterns to a given set of stimuli. Using the 150 most frequent labels from the regression analysis that overlapped with ImageNet, network activations in response to 25 classes (50 images per class) were recorded from each pretrained network, while keeping the weights frozen. The pairwise correlation distance between class activations in the embedding space were calculated and structured as a representational dissimilarity matrix (RDM). The clusters found from the hierarchical clustering of the movie label regression analysis (Section 2.2.2) were then used to define a hypothesis based binary RDM which tested for the presence of the category cluster structure in the network representations. This was implemented as a binary RDM such that two objects from the same category were given a value of 1, and 0 otherwise. The superordinate RDM models a scenario in which the learned representation is entirely explained by the

superordinate clusters found from the association of objects through time, as measured with the regression analysis. To test the relationship between the superordinate clusters and network activations a Mantel test was conducted using the “vegan” package in R, with Pearson's product moment correlation (number of permutations = 999).

The above analysis was limited in the number of categories tested and relied on the data-driven measure of the clusters from the regression analysis. Therefore, the evaluation was extended to 256 randomly selected ImageNet classes (n = 150 images per class). The mean network activations were obtained to each of the 256 classes, and the pairwise distances between each class’s activations in the frozen weights networks were calculated giving a 256 x 256 RDM. A separate model RDM to test for semantic content was constructed using the WordNet Leacock Chodorow (LCH) similarity scores for every pair of the ImageNet classes. LCH quantifies the shortest distance between two classes in the WordNet hierarchy taking into account the depth of taxonomy, making it a suitable measure for quantifying semantic similarity. A 256 x 256 pairwise semantic LCH model was constructed from the WordNet similarity scores. Using a Mantel test with Pearson's product moment correlation (number of permutations = 999) the relationship between LCH and network activations were found. To account for the contribution of low-level perceptual features that are extracted by the networks’ convolutional architecture, and not its training, we measured activations from a randomly initiated, untrained AlexNet. A partial Mantel test was then used to test the correlation between network representations and the LCH semantic model, controlling for the untrained activations.

2.2.7 Object classification performance

The above analyses probed the semantic structure of the representations in each of the pretrained networks but did not test if these were useful for the task of object recognition. For this, the transfer learning performance on a 1000-wat ImageNet recognition task was performed for the baseline model (CMC trained on the *L*a*b* objective using the movie image dataset), and SemanticCMC (trained with Δt = 60 s). A linear decoder for the 1000-way ImageNet recognition task was trained on top of convolutional layer 5, for 60 epochs using an SGD optimiser with an initial learning rate of 0.1 and decay by 0.2 at epochs 30, 40 and 50.This replicated the Linear Probing analysis reported in Tian et al. 2020, which received high classification accuracy for CMC trained on ImageNet.

2.2.8 Testing the content of the learned representations

The results of the transfer learning task (Section 2.3.3) led us to hypothesise that the models trained with SemanticCMC were learning crude background features rather than object-relevant features. Thus, RSA was repeated using blurred images to measure if the semantic structure of the learned representations persisted when test images were degraded to low spatial frequency features using a Gaussian filter (σ = 10, kernel size = (31,31)) applied to the ImageNet exemplars.

As a further test of whether the networks had learned object or background features, RSA was extended to a pre-existing test set of stimuli designed to probe background encoding in deep learning models (Xiao et al., 2020). This dataset comprised images with congruent object/background pairs, images with the foreground object removed, and images with random backgrounds from an incongruent class. As per the ImageNet-9 test set described in Xiao et al. (2020) 9 object classes were tested with network activations calculated for 450

per class. The mean activation in response to each class was calculated, and a 9 x 9 RDM constructed for each network. This was performed three times per network: once for the original images (intact-RDM), secondly for those with the object removed (remove-object-RDM), and finally using those with a random background (random-background-RDM). For each model, the Pearson correlation (r) was calculated between the intact-RDM and the remove-object-RDM, and the intact-RDM with the random-background-RDM. Representational change after object removal or background change was then measured as $1 - r$.

2.3 Results

2.3.1 Object associations persist for much longer than perceptual features

An autoregressive model was fit to the movie labels to estimate the similarity of object as a function of lag. The association between objects declined steadily over the course of 40 min (Fig 2.2). In contrast, the perceptual similarity between movie images (572,949 images) dropped off much more rapidly within 40 s. As hypothesised, the semantic content of naturalistic visual experience decays more slowly than low-level perceptual features. This was quantified by the rate of decay for an exponential decay function fit to the perceptual and semantic timeseries, estimated to be 0.64 min and 20.69 min respectively. This difference in temporal signal presents as a potential training signal for a DNN.

A hierarchical clustering analysis of the label regression coefficients revealed that objects which were more correlated across time could be grouped into interpretable, semantically meaningful clusters at the superordinate level. For example, one cluster contained the labels lamp, couch, chair, closet, piano, pillow, desk, banister and window which appear to be home furniture. It was hypothesised that training a self-supervised network on the appropriate

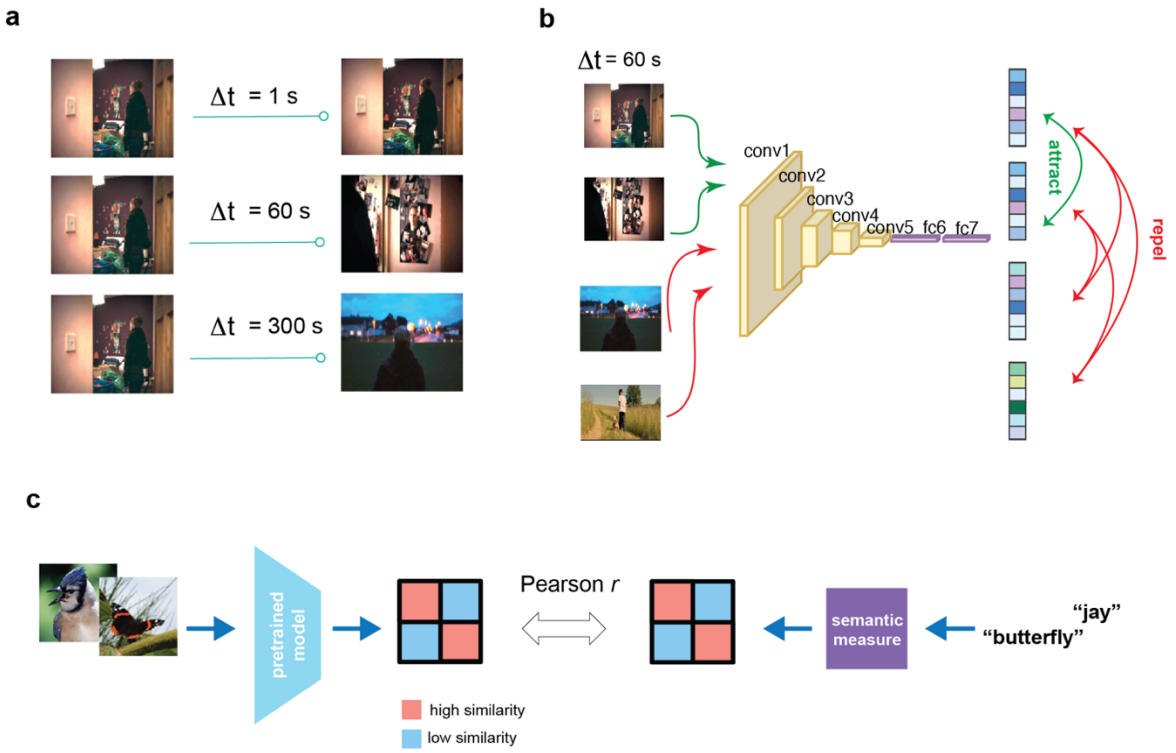


Figure 2.1. Schematic of methodology. (a) A movie dataset was generated that contained meaningful temporal structure over long timescales. The *Δt* between two images defined the semantic and contextual similarity of the scene. **(b)** SemanticCMC was implemented as a new contrastive training objective, building from the CMC network (Tian et al., 2020). The hyperparameter *Δt* defined the positive pair of images that must be identified from all other negative pairs which were randomly sampled images from the dataset. The encoder network was one full-sized AlexNet architecture. **(c)** Representational similarity analysis was used to test the relational structure after training. Pretrained models were evaluated with various sets of image stimuli, and their similarity structure were compared to various hypothesis-based similarity matrices.

temporal scale would lead to representations at a more global level, driven by superordinate relations and not entirely by object-based local structure.

2.3.2 Longer temporal self-supervised signals result in semantic representation learning

Multiple instantiations of CMC were trained, with *Δt* = 1 s, *Δt* = 10 s, *Δt* = 60 s, *Δt* = 5 min, Interestingly, as *Δt* increased the value to which loss converged increased (1 s loss =

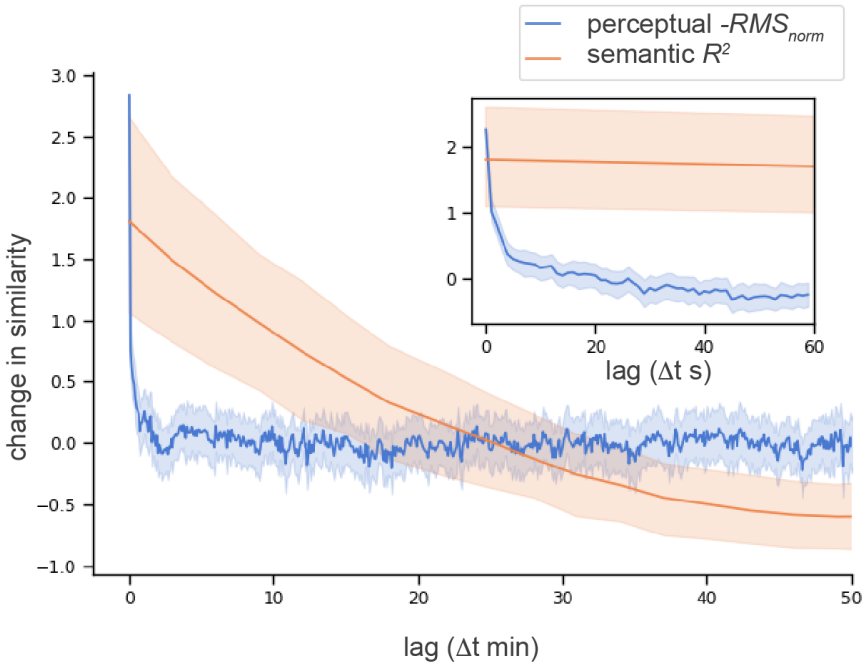


Figure 2.2. Time course of perceptual and semantic change in the movie dataset. Labels from the movie dataset were used to measure the semantic content of a scene, and the associations between these labels were measured through an autoregression. The normalised R^2 score of the fit is plot versus increasing lag distance (Δt), showing that object associations persist for up to 40 min. Error bands indicate the error across 16 subsets of the movies in the dataset. The perceptual difference across increasing lag distance was measured using the normalised -RMS_{norm} between two images. This attenuated much more quickly, reaching a stable asymptote within the first 40 s of Δt. The inset shows the first 60 s, to highlight this difference in perceptual and semantic timescales.

6.20, 10 s loss = 9.29, 60 s loss = 10.97, 5 min loss = 11.29) giving a first indication that different values of Δt provide unique information in their training signals. RDMs were constructed for the 4 instantiations of SemanticCMC and each of the baseline networks by calculating the layer wise activations in response to selected ImageNet classes from the validation dataset.

Using the classes that were in both ImageNet and the superordinate clusters found from the regression analysis, a hypothesis-based similarity matrix was defined. This modelled a

scenario where a given representation is explained entirely by these temporally extracted superordinate clusters. The extent to which each network's activations captured these superordinate semantics was quantified by correlating the pretrained network RDMs to the superordinate model using a Mantel test with Pearson's product moment correlation. Significant results were those with $p < 0.05$ from $n = 999$ permutations, Bonferroni corrected for multiple comparisons across 7 AlexNet layers. It was found that, across all layers, the randomly initiated networks and those trained on the colour augmentation task did not reach significant correlation with the superordinate model (Fig. 2.3a). This was also the case for SemanticCMC trained on Δt = 1 s, in which the information shared by the positive pair of images in the contrastive objective would likely be dominated by perceptual similarity. However, representations in the deeper layers of SemanticCMC trained with Δt of 10 s, 60 s and 5 min were all significantly correlated with the superordinate model. The 60 s network best captured this global semantic structure, with a maximum correlation on convolutional layer 5 ($r = 0.27$, $p < 0.05$). However, this did not test if the semantic structure was due entirely to the superordinate grouping or if object-based semantics at this basic level were also learned. Thus, each model's activations were measured in response to 256 ImageNet classes and compared to a measure of semantic similarity based on distance in the WordNet hierarchy.

Superordinate structure peaked in convolutional layer 5, and so this layer was chosen for further evaluation. Correlations between the semantic model and network activations were calculated using a partial Mantel test with Pearson's product moment correlation, controlling for the activations from a randomly initiated network with the same architecture. This measured the representational structure that was due to the training process, exclusive of the perceptual information that can be extracted from the convolutional structure of the model

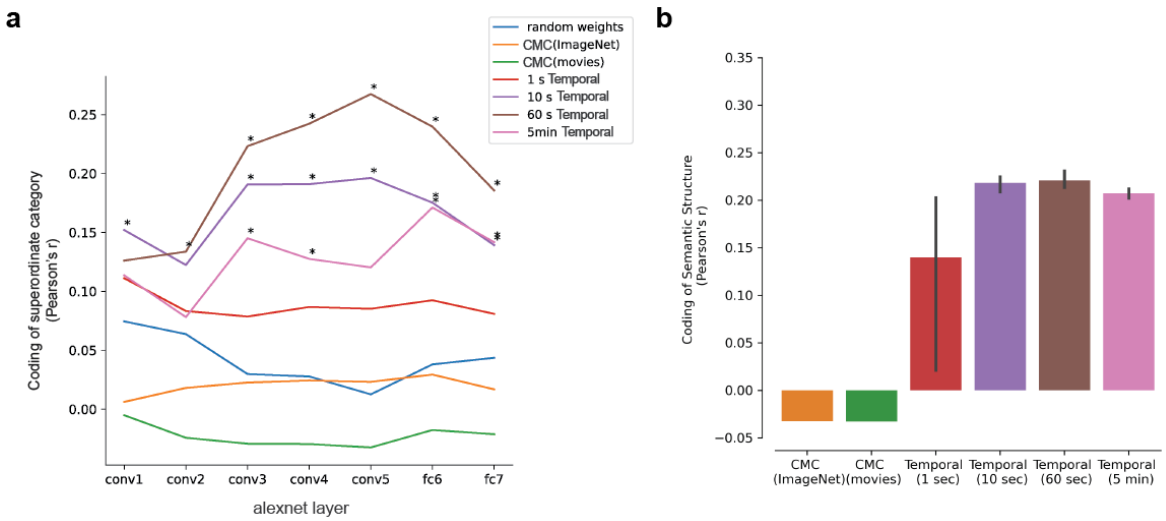


Figure 2.3. Representational coding of semantic information in pretrained models. (a) Each pretrained model's representation was correlated with a superordinate model, which defined class membership through hierarchical clustering from the regression analysis in Fig. 2.2. The self-supervised networks trained with intermediate Δt (10 s and 60 s) were best at capturing superordinate clusters, especially in mid to deep layers. * indicates significant result from a Mantel test, Bonferroni corrected for multiple comparisons across network layers. **(b)** Pretrained models' representational structures were compared to a semantic model which measured the taxonomic distance between 256 ImageNet classes. Partial correlations are plot, controlling for the representational structure of a randomly initiated network. Self-supervised networks trained on a colour-augmentation task did not learn this semantic structure when controlling for low-level perceptual features. The networks trained on our new temporal objective learned to embed semantic structure into their relational geometry, particularly at intermediate values for Δt. Error bars indicate the spread across 4 training instances of each SemanticCMC model.

(Gaier & Ha, 2019). Again, we find that the networks with access to temporal information at intermediate Δt were those that were significantly correlated to the semantic measure (Fig. 2.3b). At short time intervals, the perceptual similarity between images is confounded with the object co-occurrences, meaning that global conceptual or semantic structure is not learned. This was overcome when learning from temporally extended windows, as the persistent statistical signal is that of object co-occurrences and contextual correspondence (Fig 2.2).

Table 2.1. Classification accuracy is not indicative semantic relational geometry.

Network (AlexNet)	Task and dataset	Top-1 accuracy	Semantic correlation	
			Mantel test	Partial mantel
CMC	L vs. ab ImageNet	42.6%	0.1196 (***)	-0.02185 (*n.s.*)
CMC	L vs. ab Movies	32.38%	0.1403 (***)	0.0001595 (*n.s.*)
SemanticCMC – 60 s	t_0 vs. t_{lag} movies	1.89%	0.2668 (***)	0.2195 (***)

2.3.3 Object classification accuracy does not imply semantic representation learning

The trained models were evaluated on a transfer learning task, to test if the learned representations were also informative for object classification. A linear decoder for the 1000-way ImageNet recognition task was trained on top of convolutional layer 5 for three models. CMC published by Tian et al. (2020) achieved good accuracy at 42.60% top-1 performance on the 1000-way ImageNet task. To test if the movie dataset was sufficient for a comparable accuracy, the transfer learning performance of CMC trained with the movie image dataset on the same colour augmentation was tested. This achieved 32.38% top-1 accuracy, a slight dip likely owing to the smaller size of the dataset but a good demonstration of the efficacy of the movie images for learning useful object representations. Finally, we evaluated the performance of the SemanticCMC model with the strongest semantic representations, convolutional layer 5 of the 60 s model. Classification accuracy was 1.89%, showing very poor transfer learning capacity for this representation despite strong semantic coding. This was in contrast to the high-accuracy models, in which semantic coding was not significant

when controlling for randomly initiated network activations (Table 2.1). This result highlights a key dissociation when interpreting classification-based evaluations of computer vision models; the capacity to label visual inputs does not imply that broad semantic structure has been learned.

2.3.4 Temporal distance dictates whether objects or context are learned

We found that training on images with meaningful long-range temporal co-occurrences led to semantic representation learning, measured by hypothesis-based RDMs at the basic and superordinate level. However, this does not directly test the sensitivity of the networks to pixel-based measures of the feature selectivity. Additionally, it cannot reveal if the objects are driving the representational structure, or if its driven by global statistics and backgrounds. Scene context changes much more slowly than objects (Fig 2.1a), leading to the hypothesis that training on longer timescales would lead to a representation with these commonalities.

Firstly, to test if the global visual features of an image were learned by the longer temporally trained networks, the models were evaluated using the above 256 class RSA but with Gaussian blur applied. This tested the extent to which global statistics were contributing to semantic encoding akin to the human ability to understand the "gist" of a scene from very little visual information (Oliva & Torralba, 2006). The temporally trained models persisted in their semantic coding, even when images were blurred during evaluation (Fig. 2.4a). Next, to determine if the models were sensitive to foreground objects or background contexts, representational change was measured in response to object removal and background swap using the images from the ImageNet-9 background challenge (Xiao et al., 2020).

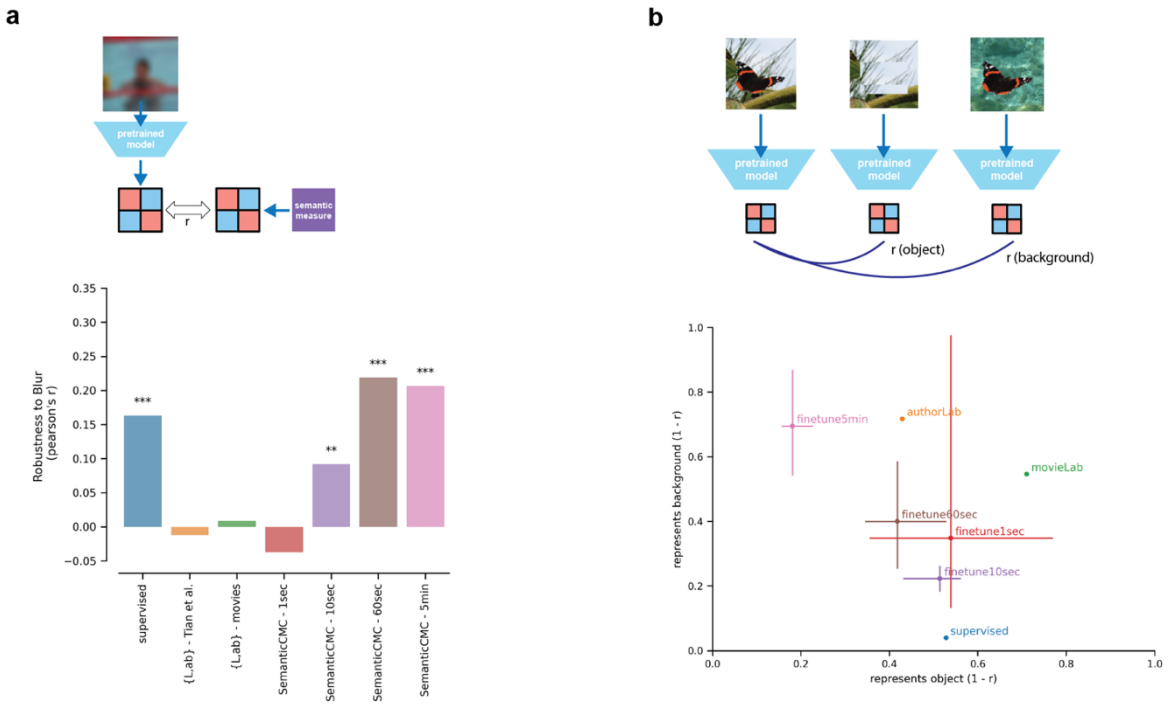


Figure 2.4. Representational coding of global features in pretrained models. (a) The models were evaluated with blurred images to test if the semantic structure found in Fig 2.3 was due to global or local features. After controlling for the representational similarity of a randomly initiated network, only the supervised and SemanticCMC network trained on sufficient Δt preserved their coding of semantic structure. **(b)** To test if the networks learned object or context based embeddings, they were evaluated with a set of ImageNet classes with either the object removed, or the background changed to that of another class. Similarity of was calculated between the intact-RDM and the remove-object-RDM, and the intact-RDM with the random-background-RDM. Representational change after object removal or background change was then measured as $1 - r$. In general, movie images provided good balance between object and background based embeddings. The longer Δt of 5 min resulted in backgrounds being learned whereas the supervised network relied on objects. Error bars indicate the range of representational change across 4 model training instances for the SemanticCMC networks.

Representational change was quantified by calculating $1 - r$ for intact image vs. test image RDMs using 1000 bootstrap samples across the 450 images in each class.

We found that all networks trained with movie images learned a better balance between object and background encoding when compared to a supervised baseline network that is

mainly object-centric (Fig. 2.4b). At the longest temporal lag of 5 min, where the main commonality between the positive pair of training images would be the scene context, backgrounds were mostly driving the representational structure. Again, the representation learned by the model trained with Δt = 60 s found good balance between the objects and contexts. Interestingly, the network trained on the movie images with the colour augmentation task was most sensitive to both object removal and background change. This model showed moderate transfer learning performance (Table 2.1) and sensitivity to both objects and contexts, but did not encode the relational geometry that corresponded to semantics, highlighting a dissociation across these two choices of evaluation for DNNs. Nonetheless, we find that extended temporal training from a naturalistic movie dataset captures global statistics that are relevant for both objects and backgrounds in an image. Depending on the learning objective and the timescale of information presented to a system, different concept-relevant features can be formed.

2.4 Discussion

Inspired by human infants, we hypothesised that semantic representations could be learned from long-range temporal structure in the visual environment. Using a novel dataset of images from movies, we found that semantic and perceptual visual features exist at long and short timescales respectively, presenting as a potential learning signal. Our findings suggest that the persistence of objects compared to perceptual similarity can be leveraged to learn more semantic and contextually relevant representations in a self-supervised contrastive learning network. Notably, this approach highlights a dissociation between semantically meaningful representation learning and traditional metrics such as object classification accuracy. This indicates that object recognition may not always be the best measure of a model's real-world performance or its ability to capture meaningful concepts from vision.

The analyses presented in this chapter place particular emphasis on relations between object classes, assessed through RSA. Prior work has discussed the importance of capturing this relational structure for conceptual learning (Roads & Love, 2020) describing how the distances between classes in a system's representational space form a distinct signature, which can be recapitulated in another system or modality facilitating alignment to form one holistic concept. Our results highlight that learning from static images without meaningful temporal structure constrains how much of this structure can be learned.

The use of temporal contrastive learning, inspired by an infant, has been proposed in the Contrastive Learning Through Time framework (Aubret et al., 2022). The authors use natural interactions of an agent with various objects, informed by findings that infants' manipulations of objects form the optimal viewpoints from which to learn invariant object representations (Bambach et al., 2018; Pereira et al., 2014; Yu & Smith, 2012). This allows for successful object shape generalisation, as well as improved invariance to background clutter, highlighting the utility of naturalistic temporal signatures for robust visual representation learning. Our work complements this, extending the window from which the network learns with the specific hypothesis that longer temporal co-occurrences will be informative for global semantic structure. Future work should examine the interplay of the slowness principle (Wiskott & Sejnowski, 2002) alongside this semantic temporal signal using egocentric video from infants (Vong et al., 2024) to form a robust understanding of the statistical signals that may be guiding infant learning.

The relevance of this finding lies in a known limitation of DNNs trained with traditional learning objectives. These models are biased towards local image features, unable to capture

the global relational structure that is important for conceptual learning (Geirhos et al., 2020) or lacking robustness to large transformations of images that would otherwise greatly impair human recognition (Brendel & Bethge, 2019). The disparities found in this work did not preclude the network from reaching high ImageNet classification performance, operating at an accuracy comparable to state-of-the-art models at the time. In contrast, human scene recognition can proceed rapidly by focusing on the global image features that provide a summary of the spatial layout of the visual input in a low dimensional code, allowing a rapid "gist" of the scene to be understood and thereby constrain the subsequent local feature analysis for more specific object recognition (Oliva & Torralba, 2006). In this way, the scene as a whole is first perceived as a single entity and then the local details are processed at a finer scale. These findings reiterate the aforementioned dissociation between classification performance and global image understanding, placing a limit on the concepts formed by artificial systems. Our temporal training method overcame some of these limitations as these models coded superordinate category information and were robust to blurring of input images.

The work presented here offers many opportunities for future directions. We found that models with semantic relational structure performed poorly at an object-recognition transfer learning task. While this result is interesting in highlighting the disparity of these evaluation metrics, a system that is capable of both levels of understanding would be preferable. However, the models trained on the temporal task learned global structure and a balance of object and background features. This suggests that that these representations may be better suited for scene classification and more in depth transfer learning tests should be explored. It also remains to be seen if this type of training leads to better correspondences between computer and biological vision and future work could examine this link to reveal if time-

based training is indeed more relevant for human visual concepts. Additionally, newer transformer-based architectures would act as interesting baselines to which the temporal CNN can be compared and large language models may be improved models for the semantic measures presented here. In sum, we have shown that the long-range temporal structure of naturalistic experience can be leveraged to learn semantic representational structure, despite poor object classification accuracy. This shows the promise of incorporating naturalistic statistics into DNNs to help bridge limitations in their representation learning.

Chapter 3 Exploring the temporal tuning of anterior ventral visual stream to perceptual and categorical inputs.

Our visual environment unfolds through extended, dynamic experiences, with different types of information dominating at varying timescales. While low-level perceptual input to the retina changes rapidly from moment to moment, the context in which we find ourselves is more persistent. In the brain, higher-order regions exhibit slower preferred intrinsic timescales of neural activity compared to faster sensory regions. This hierarchical organisation may extend to the posterior-to-anterior gradient of feature complexity along the "what" pathway of vision, known as the ventral visual stream (VVS). However, it remains unclear whether this temporal hierarchy influences the VVS's tuning to specific perceptual and categorical inputs presented at particular timescales, especially in the presence of noisy perceptual input. In this fMRI study, we tested the hypothesis that category information would be more strongly encoded by anterior ventral regions when inputs were presented at slower timescales. Healthy adults viewed a sequence of recognisable objects presented at fast and slow intervals against an unrecognisable, but perceptually intact, video background. Additionally, participants completed a resting-state scan, allowing us to characterise the intrinsic neural timescales of these visual regions. We found that the magnitude and direction of repetition responses were associated with a region's intrinsic neural timescale. Regions with long intrinsic timescales exhibited repetition enhancement, while regions with short intrinsic timescales showed repetition suppression. These findings suggest that the underlying temporal dynamics of regions in the VVS play a crucial role in fMRI adaptation patterns and how these dynamics interact with temporal tuning along the visual processing hierarchy.

3.1 Introduction

Recognising an object requires the brain to distinguish its identity from a complex array of perceptual and semantic information entering the retina from the environment. This process unfolds along a hierarchy that spans from the initial perceptual input to the eventual semantic understanding of the object. The ventral visual stream (VVS) plays a crucial role in this, serving as a pathway that is fundamental for object recognition. Organised hierarchically, early visual areas, such as V1 and V2, process basic features like edges and textures, while later stages, integrate these features into more complex representations, eventually leading to the recognition of objects (Felleman & Van Essen, 1991; Mishkin et al., 1983). However, the levels of visual information in the environment also unfold over different timescales. While perceptual inputs to the retina can change rapidly—often within milliseconds—the features that define a semantic concept and its contextual associations persist for much longer. This raises a critical question: does the hierarchical organisation of the VVS reflect this disparity in timescales? Specifically, does this organisation act as a mechanism to extract relevant features for object recognition from the continuous and noisy stream of sensory input?

The results of the modelling experiment in Chapter 2 quantified a dissociation in the timescale of semantic and perceptual information in a movie dataset (Fig. 2.2). Semantic representations were learned by a DNN trained on images separated time lags at which associations between objects were strong, but perceptual similarity was low. Conversely, at a short interval of 1 s the perceptual similarity between two images was dominant and no semantic representational structure was learned by the DNN (Fig. 2.3). This led to the hypothesis that semantic information can be distinguished, over and above perceptual noise, only when the temporal distance between objects is sufficient. We predicted that regions

further along the VVS processing hierarchy would respond preferentially to recognisable objects compared to an unrecognisable background at long, but not short, lag intervals. This can be thought of as a kind of signal to noise problem in which the meaningful signal of objects is confounded by perceptual similarity at short time intervals, only becoming distinct when the temporal window is increased (Fig. 2.2).

Previous research has demonstrated that the temporal tuning of neural responses indeed varies along the visual hierarchy. Early visual areas such as V1 and V2 respond optimally to rapidly presented stimuli, while higher-level areas are more attuned to slower presentation rates. For example, areas V1-V3 exhibit peak responses to face and house stimuli presented at rates of 18-25 items per second, while area V4 peaks at 9 items per second, and FFA and PPA peak at much slower rates of 4-5 items per second (Gauthier et al., 2012; McKeeff et al., 2007). These differences suggest that lower-level visual areas are optimised for processing quick, transient stimuli, whereas higher-level regions are more suited to processing information that unfolds over longer periods, reflecting the more complex semantic content that they encode. This may be a result of learned tuning to the temporal dynamics of these features in the environment, but alternatively could be attributed to the neural dynamics of the regions themselves.

Temporal tuning within visual cortex can be studied by varying the rate of stimulus presentation, but fMRI studies have shown that neural responses in the visual cortex can adapt to repeated stimuli in a phenomenon known as repetition suppression (Grill-Spector et al., 2006; Henson, 2003). This adaptation typically manifests as a suppressed BOLD response upon repeated exposure to the same stimulus, the exact mechanisms of which are not fully understood. It is clear, however, that the extent of suppression is influenced by

several factors including the recognisability of the stimulus (Grill-Spector et al., 2000), the visual quality of the image (Turk-Browne et al., 2007), the level of attention directed towards it (Eger et al., 2004; S. O. Murray & Wojciulik, 2004), and the duration for which the stimulus is presented (Zago et al., 2005). In the controlled environment of a laboratory, these factors can be tightly regulated. In the more chaotic and noisy real world, where perceptual inputs are less predictable and changing more rapidly, the influence of these factors—and consequently the pattern of repetition suppression—might differ significantly. These considerations, in conjunction with the findings from the modelling experiment, informed the design of the below fMRI experiment. Using a warping method applied to naturalistic video we constructed a stimulus in which recognition is impaired but perceptual statistics are preserved (Stojanoski & Cusack, 2014). Fully recognisable objects could then be overlaid on this noisy perceptual background at various time intervals, allowing us to test if the rapidly changing perceptual information dominated object responses in VVS at short time intervals, as was hypothesised from the autoregressive (Fig. 2.2) and computational models (Fig. 2.3).

Furthemore, we wished to consider the mechanisms underlying repetition effects, for which several theories have been proposed. The accumulation hypothesis suggests that neural responses to a stimulus become more efficient upon repetition because the neural representation of that stimulus is already primed (Henson et al., 2002; James et al., 2000; Noguchi et al., 2004). This priming leads to a more rapid and brief trajectory of the neural response curve, which, when integrated across a population of neurons, results in decreased area under the curve of neural activity. The accumulation hypothesis also accounts for cases where repeated exposure leads to enhanced neural responses, called repetition enhancement. In this scenario, say where visual inputs are degraded (Turk-Browne et al., 2007), the

priming response evolves more slowly with a longer peak in the neural response curve. This appears as an increase in activity when responses are integrated population level. However, this theory does not fully account for the brain's intrinsic responsiveness to varying timescales, which varies by region.

A largely separate literature has studied Intrinsic neural timescales (INTs), the natural rhythm or timing at which different brain regions process information independently of external stimuli (Ito et al., 2020; J. D. Murray et al., 2014). These intrinsic timescales are thought to reflect the temporal windows over which different regions of the brain integrate information. For instance, regions involved in early sensory processing, such as the primary auditory cortex, tend to operate on very short timescales, responding to momentary input. In contrast, higher-level regions, such as those in the parietal and frontal cortices, integrate information over much longer periods, which is necessary for understanding complex stimuli like narratives (Hasson et al., 2008; Lerner et al., 2011). The current models of fMRI adaptation do not typically account for these intrinsic timescales, particularly in the context of noisy perceptual environments where distinguishing relevant features for recognition becomes more challenging. This omission is significant because the intrinsic timescale of a given brain region likely influences how that region responds to repeated stimuli. For example, a region with a short intrinsic timescale might quickly adapt to repeated presentations of a stimulus, resulting in rapid repetition suppression. Conversely, a region with a longer intrinsic timescale might continue to process the stimulus over an extended period, leading to a more gradual or even enhanced response upon repetition.

In the present study, we aimed to fill this gap by examining how the INTs of regions within the VVS influenced their response to semantically recognisable objects amidst a background

of unrecognisable perceptual noise. We employ both task-based and resting-state fMRI to characterise the INTs across different regions of the VVS. Additionally, we measure repetition effects by comparing neural responses to recognisable objects against a background of unrecognisable video. We hypothesise that regions in the anterior VVS, which are involved in higher-level object recognition, will only exhibit object-selective responses when the stimuli are presented at slower timescales. At faster timescales, the rapidly changing perceptual information is expected to overshadow the semantic content, preventing the brain from effectively distinguishing objects from the background noise. We anticipate that this pattern of results will correlate with increasing INTs along the ventral stream, with anterior regions displaying longer intrinsic timescales compared to posterior regions.

3.2 Methods

3.2.1 Participants

A cohort of healthy young adults (n = 25, 19 female and 6 male) were recruited from Trinity College Dublin, through advertisement on campus and *via* the School of Psychology's research management system. Informed consent was obtained from all participants, who received either monetary reimbursement or research credits in exchange for their time. One participant completed only 3 task-based runs due to discomfort in the scanner, and one participant did not complete the resting state scan due to time constraints. The study was approved by Trinity College Dublin's School of Psychology Research Ethics Committee (Approval ID: SPREC102021-20) and all scans were conducted at Trinity College Institute of Neuroscience.

3.2.2 MRI Acquisition

MRI data were collected on a Siemens MAGNETOM Prisma 3T scanner using the Siemens 64 channel head and neck coil. Anatomical scans were a T1 weighted MPRAGE sequence, collected in the sagittal plane with TE 2.98 ms, TR 2300 ms, flip angle 9 degrees, FOV 248 x 256, voxel size of 1 x1 x 1 mm, slice thickness 1 mm of which there were 176 slices totalling a 6.75 min scan. fMRI sequences were multiband echo planar images, with a multiband acceleration factor of 4, collected in an anterior to posterior phase encoding direction, TE 30 ms, echo spacing 0.54 ms, TR 656 ms, flip angle 40 degrees, FOV 192 x 192 with a voxel size of 3 x 3 x 3 and 40 slices with 3 mm thickness. For the task based fMRI, 1201 volumes were collected totalling 13.13 min of acquisition, and for the resting state scan 910 volumes were collected totalling 9.95 min. Participants were instructed to close their eyes during the resting state session while remaining awake. Two single band spin-echo planar images were collected in opposite phase encoding directions (anterior to posterior and posterior to anterior) with TE 30 ms, echo spacing 0.5 ms, TR 2510 ms, flip angle 40 degrees, FOV 192 x 192 with a 3 x 3 x 3 mm voxel size and 40 slices and 3 mm thickness. 10 volumes were collected totalling 25 s of acquisition.

3.2.3 Stimulus presentation

An unrecognisable but perceptually intact background was created using diffeomorphic warping (Stojanoski & Cusack, 2014) applied to a video from first-person perspective of walking through a museum. The warping method ablated recognisability while preserving the low-level perceptual features of the naturalistic setting. The video played across the entire screen throughout the stimuli presentation blocks and each block was separated by 5 s of a blank screen. Images of intact, recognisable objects appeared overlaid on the perceptual background at various time intervals. These were drawn from the basic level

categories plane, pizza, cloud, sheep, lighthouse, hand, fish, wheel, brain, tree, pencil and leaf. The categories were chosen to be spread across the animate/inanimate organising principle of VOTC (Konkle & Caramazza, 2013), and contained further subsets of perceptually similar but semantically different pairs; for example, plane and fish or pizza and wheel. These subsets were not evaluated in the present analysis, but were included as part of the design for future work. Eight objects appeared on top of the perceptual background either chosen at random, or in a block comprising one basic level category; for example, eight exemplars of "plane" were presented in succession. This enabled tests for category specific repetition effects. Each block was either assigned to the short or slow condition, defined by the inter stimulus interval (ISI). ISI ranged from 0.5 s – 1.6 s (mean ISI = 1.1 s) and 5.0 s – 16.0 s (mean ISI = 11 s) in the fast and slow blocks respectively. All objects were rescaled such that the longest axis was 640 pixels. The first object always appeared 5 s after the perceptual background began playing, and the background played for 5 s at the end of each block. To ensure participants remained engaged with the task, they were instructed to press a button when an object appeared upside down. This occurred with 20% chance, excluding the first and last objects which were never rotated. Participant engagement was assessed using accuracy on this task, defined as the sum of true positive and true negative button presses divided by the total number of trials. Response time (RT) was recorded for each trial. Condition-dependent differences in the average RT was tested using a t-test across subjects, testing for the alternative that the slow condition has faster RT, assuming repetition suppression in this condition.

3.2.4 Preprocessing

Results included in this manuscript come from preprocessing performed using fMRIPrep v0.0.1 (Esteban et al., 2019) (RRID:SCR_016216), which is based on Nipype 1.6.1 (Gorgolewski et al., 2011) (RRID:SCR_002502).

A total of 1 T1-weighted (T1w) images were found within the input BIDS dataset. The T1-weighted (T1w) image was corrected for intensity non-uniformity (INU) with N4BiasFieldCorrection (Tustison et al., 2010), distributed with ANTs 2.3.3 (Avants et al., 2011) (RRID:SCR_004757), and used as T1w-reference throughout the workflow. The T1w-reference was then skull-stripped with a Nipype implementation of the antsBrainExtraction.sh workflow (from ANTs), using OASIS30ANTs as target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and grey-matter (GM) was performed on the brain-extracted T1w using fast (Zhang et al., 2001) (FSL 5.0.11, RRID:SCR_002823). Brain surfaces were reconstructed using recon-all (Dale et al., 1999) (FreeSurfer 6.0.1, RRID:SCR_001847), and the brain mask estimated previously was refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of the cortical grey-matter of Mindboggle (Klein et al., 2017) (RRID:SCR_002438). Volume-based spatial normalization to one standard space (MNI152NLin2009cAsym) was performed through nonlinear registration with antsRegistration (ANTs 2.3.3), using brain-extracted versions of both T1w reference and the T1w template. The following template was selected for spatial normalization: ICBM 152 Nonlinear Asymmetrical template version 2009c (Fonov et al., 2009) (RRID:SCR_008796; TemplateFlow ID: MNI152NLin2009cAsym).

For each of the 5 BOLD runs found per subject (across all tasks and sessions), the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated using a custom methodology of fMRIPrep. Susceptibility distortion correction (SDC) was omitted. The BOLD reference was then co-registered to the T1w reference using bbregister (FreeSurfer) which implements boundary-based registration (Greve & Fischl, 2009). Co-registration was configured with six degrees of freedom. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using mcflirt (Jenkinson et al., 2002) (FSL 5.0.11). BOLD runs were slice-time corrected using 3dTshift from AFNI 20210112 (Cox, 1996) (RRID:SCR_005927). The BOLD time-series were resampled onto the following surfaces (FreeSurfer reconstruction nomenclature): fsaverage5. The BOLD time-series (including slice-timing correction when applied) were resampled onto their original, native space by applying the transforms to correct for head-motion. These resampled BOLD time-series will be referred to as pre-processed BOLD in original space, or just pre-processed BOLD. The BOLD time-series were resampled into standard space, generating a pre-processed BOLD run in MNI152NLin2009cAsym space. First, a reference volume and its skull-stripped version were generated using a custom methodology of fMRIPrep. Several confounding time-series were calculated based on the pre-processed BOLD: framewise displacement (FD), DVARS and three region-wise global signals. FD was computed using two formulations following Power (absolute sum of relative motions, (Power et al., 2014) and Jenkinson (relative root mean square displacement between affines (Jenkinson et al., 2002) FD and DVARS are calculated for each functional run, both using their implementations in Nipype (following the definitions by Power et al. 2014). The three global signals are extracted within the CSF, the WM, and the whole-brain masks. Additionally, a set of physiological regressors were extracted to allow for

component-based noise correction (CompCor,(Behzadi et al., 2007). Principal components are estimated after high-pass filtering the pre-processed BOLD time-series (using a discrete cosine filter with 128s cut-off) for the two CompCor variants: temporal (tCompCor) and anatomical (aCompCor). tCompCor components are then calculated from the top 2% variable voxels within the brain mask. For aCompCor, three probabilistic masks (CSF, WM and combined CSF+WM) are generated in anatomical space. The implementation differs from that of Behzadi et al. in that instead of eroding the masks by 2 pixels on BOLD space, the aCompCor masks are subtracted a mask of pixels that likely contain a volume fraction of GM. This mask is obtained by dilating a GM mask extracted from the FreeSurfer's aseg segmentation, and it ensures components are not extracted from voxels containing a minimal fraction of GM. Finally, these masks are resampled into BOLD space and binarized by thresholding at 0.99 (as in the original implementation). Components are also calculated separately within the WM and CSF masks. For each CompCor decomposition, the k components with the largest singular values are retained, such that the retained components' time series are sufficient to explain 50 percent of variance across the nuisance mask (CSF, WM, combined, or temporal). The remaining components are dropped from consideration. The head-motion estimates calculated in the correction step were also placed within the corresponding confounds file. The confound time series derived from head motion estimates and global signals were expanded with the inclusion of temporal derivatives and quadratic terms for each (Satterthwaite et al., 2013). Frames that exceeded a threshold of 0.5 mm FD or 1.5 standardised DVARS were annotated as motion outliers. All resamplings can be performed with a single interpolation step by composing all the pertinent transformations (i.e. head-motion transform matrices, susceptibility distortion correction when available, and co-registrations to anatomical and output spaces). Gridded (volumetric) resamplings were performed using antsApplyTransforms (ANTs), configured with Lanczos interpolation

to minimize the smoothing effects of other kernels (Lanczos, 1964). Non-gridded (surface) resamplings were performed using mri_vol2surf (FreeSurfer).

Many internal operations of fMRIPrep use Nilearn 0.8.0 (Abraham et al., 2014) (RRID:SCR_001362), mostly within the functional processing workflow. For more details of the pipeline, see the section corresponding to workflows in fMRIPrep's documentation.

3.2.5 Intrinsic Neural Timescales

Following the method in (Ito et al., 2020), the resting state functional scans were used to estimate the INTs in each region in the Glasser parcellation (Glasser et al., 2016). Within each subject, the parcellated and pre-processed functional data were cleaned using nuisance regression, as described in (Ciric et al., 2017). The first five frames were removed from each run, the mean was subtracted and motion parameters and physiological noise were included as covariates. This included six motion regressors as well as their derivatives and quadratics and aCompCor time series extracted from the white matter and ventricle voxels as well as their derivatives and quadratics, totalling 14 motion and 40 physiological noise regressors. A given region's INT was quantified using the decay coefficient of the autocorrelation of the BOLD timeseries across a window of 100 lags. An exponential decay $R(k\Delta)$ function was fit to the timeseries of the autocorrelation using non-linear least squares estimation with Equation 3.1:

$$R(k\Delta) = A[exp(-\frac{k\Delta}{\tau}) + B] \qquad (3.1)$$

where A is a scaling factor, B reflects the offset for contribution of timescales longer than the observation window, and τ corresponds to the intrinsic timescale (i.e., rate of decay)

expressed in seconds using Δ, the TR of the resting state scan (0.656 s). To avoid the biological implausibility of finding negative scaling factor and intrinsic timescale, parameters were constrained to the bounds $A \in [0,\infty)$, $B \in [-\infty,\infty)$ and $\tau \in [0, \infty)$. Outliers over the 95th percentile were excluded when calculating the average τ value per ROI across subjects.

3.2.6 Generalised Linear Model

A custom analysis pipeline was written in Nilearn v0.10.1 using Python 3.9.0 to estimate various BOLD response time courses using a generalised linear model including the confounds of 6 motion parameters, CSF, white matter and global signal returned from fMRIPrep. All functional data were on the fsaverage5 cortical surface. Button presses and the perceptual background were modelled as separate regressors. The regressors of interest were included at various scales of coarseness, depending on the hypothesis. To test if responses to the recognisable objects were greater than the unrecognisable background, and whether this varied by timescale, each of the exemplars in the block were collapsed, excluding the rotated objects giving four conditions of interest. These were "basic fast", "basic slow", "mixed fast" and "mixed slow". Following contrasts at this level, a separate GLM was run in which each of the 8 objects in a block were modelled as separate regressors. This allowed us to test the first and last objects, as well as the 6 interim objects. Category identity was excluded, but the block type – basic or mixed and fast or slow – was preserved. Models were fit to each run of individual subjects' recorded BOLD timeseries, and first-level contrasts were calculated.

3.2.7 Contrasts

Statistical t-maps for each condition in the design matrix were obtained at the individual level for each voxel in the brain, and then averaged across runs. Contrasts were calculated by linearly combining the conditions of interest, according to various tests described below. Second level analyses tested group-level effects using a one-sided t-test across subjects, with the alternative that $t > 0$ in a given voxel.

Initial contrasts tested for differences at the block level, defined by category type and time. These were either basic (b) or mixed (m) categories and either fast (F) or slow (S) times. The unrecognisable perceptual background (bg) was included as a single regressor and was played during the entirety of a trial block (25.83 s for fast blocks, 95.12 sec for slow blocks). To test if object > background responses were attributed to category specific differences at the block level, a whole-brain contrast for ((0.5 * bF + 0.5 * bS) - bg) - ((0.5 * mF + 0.5 * mS) - bg) was calculated. This resulted in no significant voxels, and henceforth basic and mixed blocks were grouped for subsequent contrasts.

Voxels which responded selectively to objects vs background were tested using a contrast of (0.25 * mF +0.25 * bF + 0.25 * mS + 0.25 * bS) - bg. In a separate model, the response to individual objects in a block were estimated. There were 8 objects per block, labelled as first (f), interim (i) 1-6 and last (l). A t-contrast tested object-sensitive repetition suppression throughout the block. It was hypothesised that the last object presented in a block would elicit a suppressed response relative to the first object, and that this would be greater than the background response. In the fast condition this was implemented as ((0.5 * mFf + 0.5 * bFf) - (0.5 * mFl + 0.5 * bFl) - bg), and similarly for the slow condition ((0.5 * mSf + 0.5 * bSf) - (0.5 * mSl + 0.5 * bSl) - bg). Finally, the interaction between object and time

conditions, relative to the background, were calculated using the contrast (((0.5 * mFf + 0.5 * bFf) – (0.5 * mFl + 0.5 * bFl)) – ((0.5 * mSf + 0.5 * bSf) – (0.5 * mSl + 0.5 * bSl)) – bg). Individual subject z-maps for each contrast were then used in a second level design, implemented as a one sample t-test across participants. Significant voxels were found by correcting for multiple comparisons using Bonferroni correction at $\alpha = 0.01$.

3.2.8 Measuring changes in BOLD response with object repetition

The contrasts testing for repetition suppression (first object > last object) revealed significantly negative results in some voxels, suggestive of repetition enhancement, especially in the slow condition. The curve of beta estimates for two large ROIs, posterior early visual cortex (EVC) and anterior ventral visual (VVC) regions, were used to characterise the time course of beta estimates throughout the 8 stimulus presentations. Objects in the fast condition had an ISI of 0.5 – 1.6 s, which is less than the time taken for the hemodynamic response, meaning that individual beta estimates for the fast objects were not interpretable due to multicollinearity of the GLM. Thus, analyses on the interim betas were restricted to the slow trials and the difference between the first and last condition were only used in the fast trials. Differences between first and last objects' beta estimates were calculated using a t-test across subjects. Individual object beta estimates were averaged across the mixed and basic conditions, as no category-specific differences were found between these blocks, as well as being averaged across runs within each subject. The beta value for each condition was plot as a function of the stimulus number, revealing an initial sharp rise in response followed by a slow decline. This asymmetric response was representative of a gamma function that may typically be associated with task evoked BOLD activity (Friston et al., 1994) and was modelled using non-linear least squares optimisation with Equation 3.2:

$$f(x) = a \cdot e^{-t_0 x} \cdot x^c \qquad (3.2)$$

Where a is *a* scaling factor, t_0 is the rate parameter controlling the decay of the response, *x* is the condition number and *c* is the shape parameter that controls the power of the response curve. Observed differences in t_0 between the two large ROIs (EVC: V1, V2, V3, V4 and VVC: V8, FFC, PIT, VMV1, VMV2, VMV3,VVC) prompted investigation at the individual ROI level, defined using the Glasser atlas surface parcellation. The sampling distribution of t_0 within each ROI was estimated using bootstrap resampling across subjects, to observe if the rate of repetition response decay varied as a function of distance along the visual hierarchy.

3.2.9 Testing the association between INTs and repetition response

The relationship between INT (τ estimates) and repetition response (difference between first and last beta estimates in each condition, β_{first} - β_{last}) across different brain regions was investigated using linear methods. Firstly, the association between beta differences per ROI and the region's estimated τ was tested using Pearson correlation. The interaction between τ per ROI and the fast/slow condition was tested using a separate multiple regression. A binary variable labelled the distinction between the fast and slow conditions, and an interaction term was computed by multiplying τ and Condition. This interaction term was included in the regression model to assess whether the relationship between τ and repetition response differs according to the stimulus lag interval. The significance of the model coefficients was evaluated using p-values, with a threshold of $p < 0.05$. indicating statistical significance.

Results

Participants were engaged with the stimuli, with a mean accuracy of 93% on a distractor task in which they were instructed to press a button when an object appeared upside down. Despite the long length of the scanning session, this accuracy rate did not vary by run (run-1: 0.931, run-2: 0.921, run-3: 0.929, run-4: 0.926) and did not differ according to the time condition (fast: 0.930, slow: 0.924). RTs also did not vary according to ISI (fast: 1.0089 s, slow: 1.0085 s, p = 0.053), indicating that recognisability was at ceiling in both fast and slow blocks. A whole-brain group-level contrast for object > background response showed the expected activation in anterior ventral visual regions (Fig. 3.1b), revealing successful sensitivity to objects in the expected brain regions, despite the continuous background noise. Objects appeared in one of two block category types: basic blocks had 8 exemplar stimuli from 1 of 12 basic-level categories, and mixed blocks had 8 exemplar stimuli drawn randomly from the 12 basic categories. This allowed us to test for category-specific adaptation or enhancement, finding that no voxels were selective for objects in the basic condition compared to the mixed condition, relative to the background (Fig. 3.1c). Thus, all subsequent analyses grouped basic and mixed blocks and were conducted at the level of objects vs. background.

3.2.10 Intrinsic neural timescales increase from medial to lateral on the ventral surface

INTs were estimated for each region in the surface parcellation by estimating the decay of the autocorrelation function (τ) for each voxel at rest and finding the mean timescale within a region (Fig 3.2). Contrary to our previous literature, INTs did not increase from anterior to posterior along the ventral surface; ventromedial regions of the temporal lobe show faster

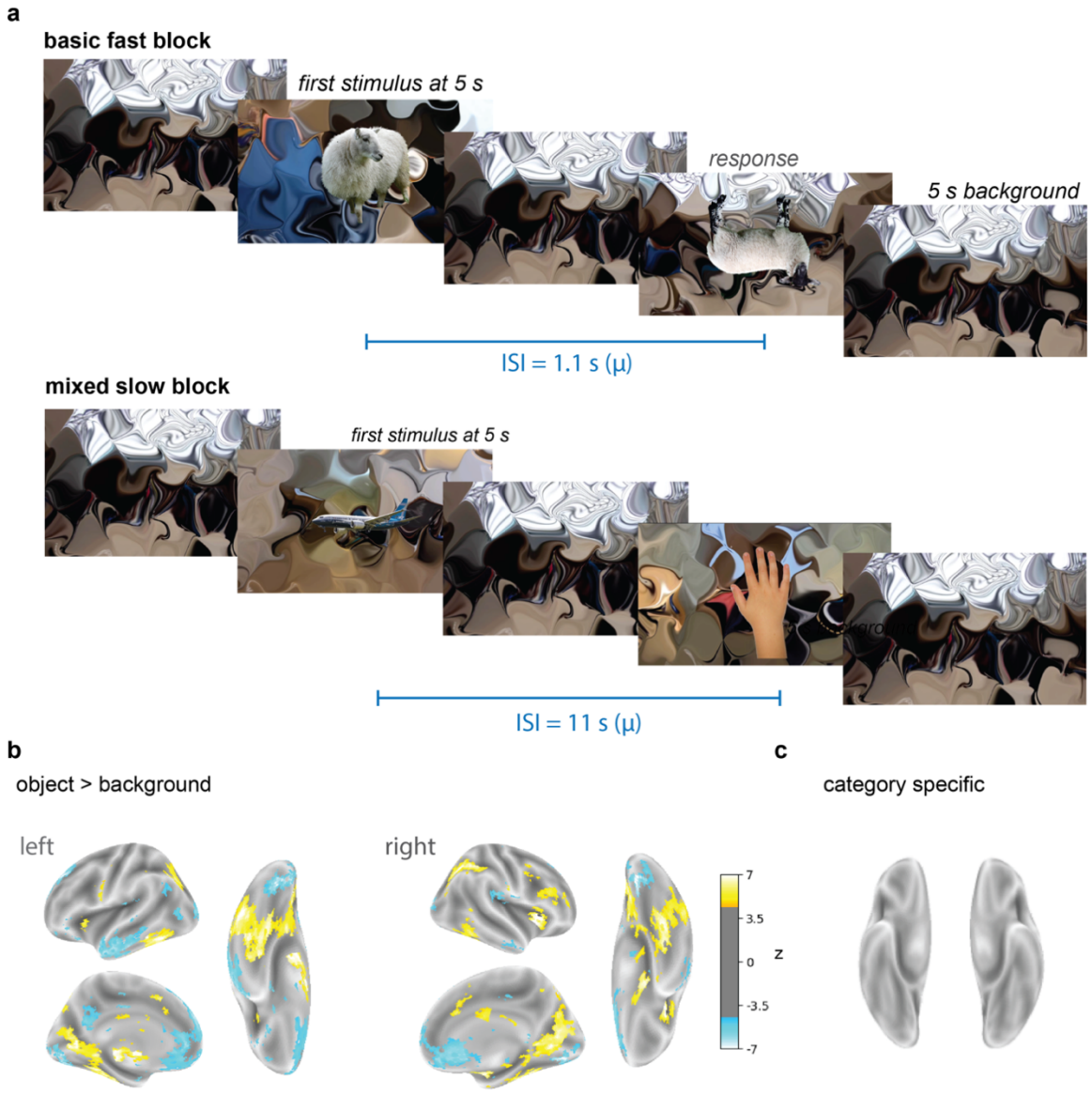


Figure 3.1 Task design and basic contrasts. (a) An unrecognisable video background (bg), made with diffeomorphic warping, played during an experimental block which was defined by category level (basic or mixed) and time (Fast: mean ISI of 1.1 s, Slow: mean ISI of 11 s). Objects were presented in four possible block types: basic Fast (bF), basic Slow (bS), mixed fast (mF) and mixed slow (mS). During the block, 8 recognisable objects appeared overlaid on the background with pseudorandom ISI jittering. Objects appeared rotated with 20% chance, and participants were prompted to press a button when this occurred, ensuring maintained engagement with the stimuli without drawing attention to the condition of interest. **(b)** Whole-brain group level z-maps, Bonferroni corrected at $\alpha = 0.01$. The expected object response was observed in anterior ventral visual cortex, tested using the contrast (0.25 * bF + 0.25 * mF + 0.25 * bS + 0.25 * mS) > bg. **(c)** No voxels showed significant activation for objects in the basic blocks compared to the mixed blocks, relative to the background [((0.5 * bF + 0.5 * bS) > bg) > ((0.5 * mF + 0.5 * mS) > bg)].

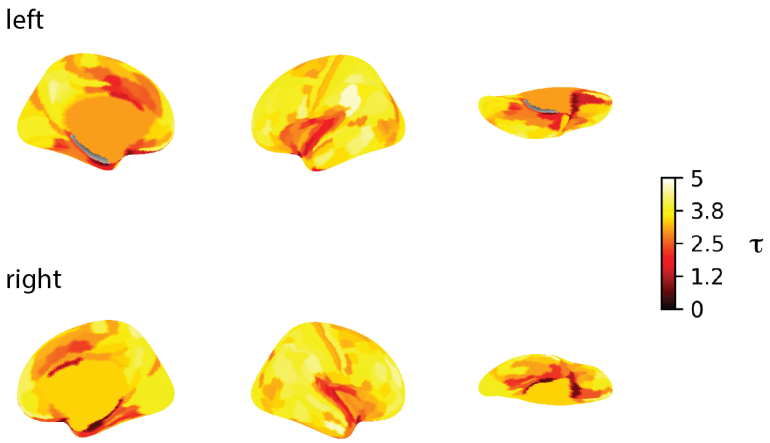


Figure 3.2 Intrinsic neural timescales. The timescale (τ) of neural activity within each ROI of the Glasser parcellation, estimated using temporal autocorrelation of the BOLD response during

intrinsic activity with increasing τ in the lateral direction. Typically, regions which process higher order information have slower intrinsic neural activity (Ito et al., 2020; J. D. Murray et al., 2014) and this was expected within the ventral visual processing hierarchy. We find the inverse relationship here, with ventromedial visual areas having faster INTs than EVC. However, this pattern has been previously observed in similar studies of INTs which include the ventral surface (Truzzi & Cusack, 2023). Ventromedial areas are associated with the processing of object identity and other high-level visual features (e.g., colour and form) and faster INTs in these regions could facilitate the quick extraction of relevant information from the visual input. Conversely, the slower INTs in lateral areas may reflect a greater reliance on temporal integration over longer periods, potentially supporting more complex or context-dependent processing, such as integrating information about object location or motion (Mishkin et al., 1983). These differences in INT are relevant when considering repetition effects, as the intrinsic activity will influence the rate of adaptation to repeated presentations of the stimuli.

3.2.11 Response inhibition for objects vs. scenes varies by the temporal lag of inputs

The time course of BOLD responses to repeated recognisable objects, relative to the background stimulus, was tested using a repetition suppression analysis. Reduced activation for the last versus first object was tested using a group level whole-brain contrast of (first – last) > background in each time condition (Fig 3.3a). Based on the findings in Chapter 2, we predicted that semantically recognisable objects would only be distinguishable in the VVS from the quickly changing perceptual background when presented at sufficiently long intervals. However, as in the INT analysis, we observed the opposite effect. Anterior ventral regions showed response inhibition in the fast condition only. In the slow condition, there was a general pattern of negative beta estimates for the (first – last) > background contrast. This was evident in early and mid-level ventral visual regions, lateral occipital complex (LOC) and visual motion areas in the dorsal regions (Fig 3.3a, slow), indicating a stronger effect of the background scene in the slow condition. A contrast that tested for the interaction between object > background responses the time (fast > slow) again showed significant preference in VVC for objects displayed at the fast timescale, even after strict thresholding with Bonferroni correction at $\alpha = 0.01$. We observed a consistent left lateralised effect in visual association area TE2a, which has been implicated in recognition memory (Ho et al., 2011; Málková et al., 1995), indicating that responses were not due to impaired recognisability in either condition as was evident from the behavioural results.

The clusters of significant negative voxels in the slow blocks led us to investigate the role of repetition enhancement in the object response. The direction of repetition effects depend on visual quality (Turk-Browne et al., 2007), which may have been influenced by the stronger response of the background noise in the slow condition. Beta estimates for each of the 8 objects in the block were used to plot a response curve over the course of object

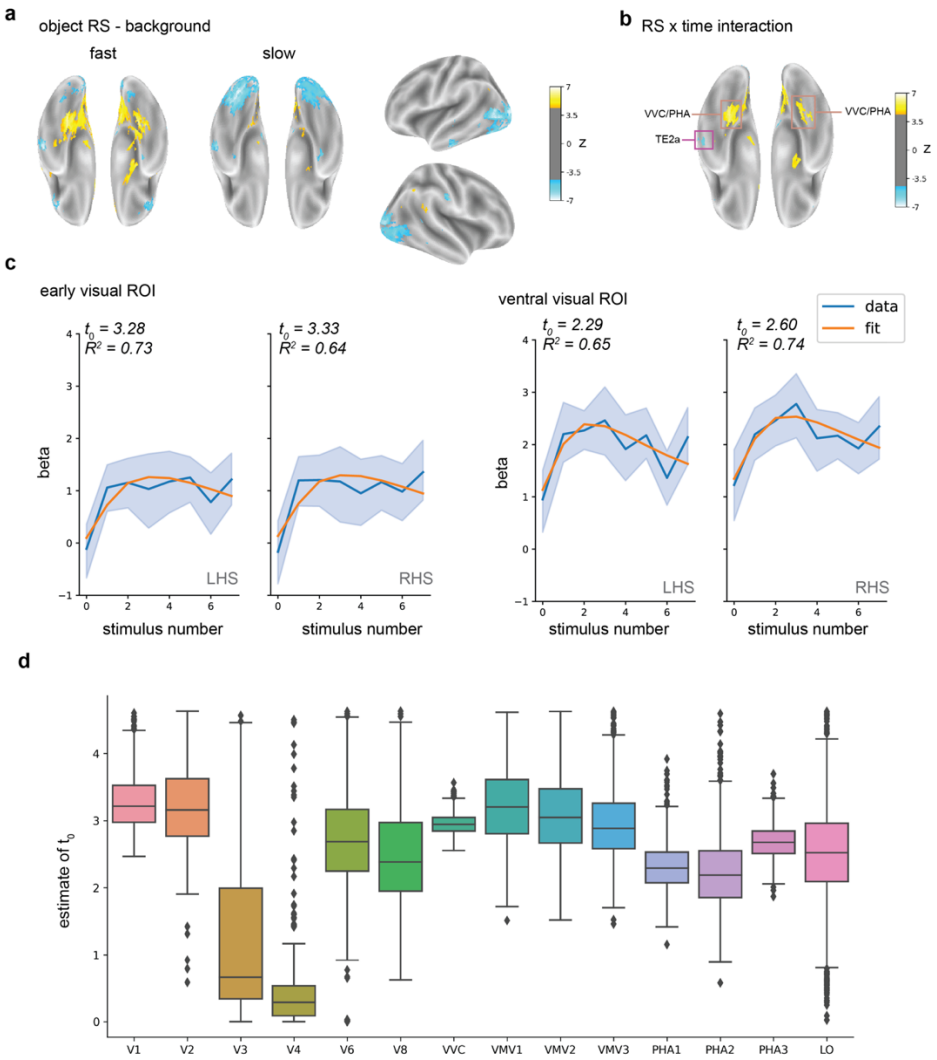


Figure 3.3 Repetition effects for slow condition. (a) Group level z-maps for voxels that show significant activation (Bonferroni corrected at $\alpha = 0.05$) for the contrast (first object – last object) > background in each of the time conditions. **(b)** Interaction contrast at the group level (Bonferroni corrected at $\alpha = 0.01$) for the repetition effect of object > background and the time (fast > slow) condition. Bilateral ventral and parahippocampal visual cortex, and left anterior temporal association cortex survive rigorous thresholding. **(c)** The beta estimates for individual objects in the slow condition, in two large ROIs. Error bands indicated 95% confidence interval estimated using 999 permutations across subjects. A gamma function was fit to each response time course, and the rate of decay (t_0) observed for each ROI. **(d)** Estimated sampling distributions ($n = 1000$ bootstraps across $n = 25$ subjects, averaged across two hemispheres) of t_0 for individual ventral-occipitotemporal ROIs from the Glasser parcellation. Note that error in (d) is estimated using non-parametric bootstrapping, and not across individual subjects as in other plots. Boxes show quartiles of the sampling distribution and whiskers extend to the remainder of the distribution. Individual points are outliers. Larger values for t_0 indicate a slower rate of decay in the response curve to repeated objects.

repetition, revealing increases in the BOLD response with repeated stimuli presentations. A gamma function was fit to the response curves, and the rate of decay (t_0) was used as a metric for how quickly the object > background response was responding to repeated trials. The equivalent analysis for the fast condition was not possible because the short design of these trials led to unstable individual parameter fits for each of the 8 objects.

A slight difference in slow block t_0 estimates was found for each of the curves when analysed across two large ROIs (Fig. 3.3c), with further differences evident when analysed at the individual ROI level (Fig. 3.3d). The gamma function only described a lesser portion of response curves in mid-level visual regions, especially V3 (R^2 = 0.47), V4 (R^2 = 0.40), PHA3 (R^2 = 0.38) and LO (R^2 = 0.43) but explained a large proportion of the variance in in V1 (R^2 = 0.89), V2 (R^2 = 0.70) and PHA2 (R^2 = 0.73). The gamma curve was most explanatory for VVC responses (R^2 = 0.81). Here, we observed a tight distribution of t_0 estimates when resampling with replacement across subjects (μ = 2.95, σ = 0.15) compared to earlier regions V1 (μ = 3.28, σ = 0.40), indicating that VVC responses to stimuli at a long temporal lag reliably follow the gamma decay function. The observed value of t_0 was again faster in VVC than in earlier regions, but was not significantly different from the mean of V1 estimates for t_0 [95% confidence interval for $t_{0,V1} - t_{0,VVC}$ = (-0.37, 1.36)].

Although the gamma model explained 73% of the variance in PHA2, the direction of the curve was indicative of repetition suppression (Supplementary Fig 3.1, 3.2). To observe if the direction of adaptation effects differed by region, the distribution across subjects of first and last beta estimates, relative to the background, were visualised in VTC ROIs for both the fast and slow conditions (Fig 3.4). Scene selective PHA exhibited suppressed response in both conditions, revealing a contribution of the background scene in all cases. However,

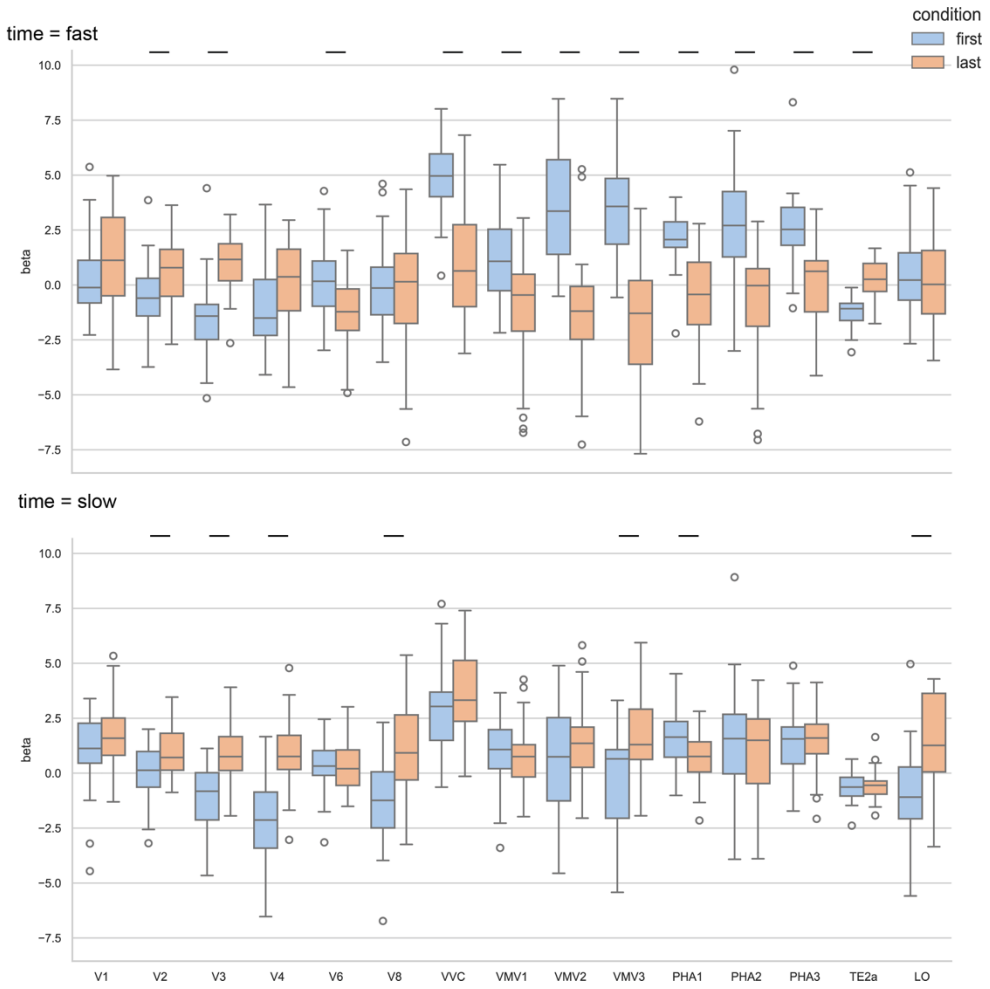


Figure 3.4 ROI-based repetition effects. Beta estimates for first and last presentation of a stimulus in a block in the fast and slow conditions. Mid and high-level visual ROIs show either repetition suppression or enhancement depending on the ISI of the block. Boxes show quartiles of the distribution and whiskers extend to the remainder of the distribution. Individual points are outliers. Horizontal lines indicate a significant difference between the mean beta for first and last objects within an ROI, tested using a t-test across $n = 25$ subjects.

LO showed significant repetition enhancement to objects presented at a slow timescale. As demonstrated in the above whole-brain contrast (Fig 3.3a), anterior ventral visual regions had attenuated BOLD responses upon object repetition at fast temporal lags (VVC, VMV1, VMV2, VMV3). When stimuli were presented at a longer lag interval, ventromedial visual areas trended instead towards repetition enhancement, with the last object's beta estimate being significantly larger than the first object in VMV3. This reveals that object responses

were encoded by ventral visual regions, relative to the noisy perceptual background, when images appeared at fast lag intervals.

3.2.12 The direction of repetition effects is determined by intrinsic neural timescale

Contrary to our hypothesis that semantic information would be better distinguished from perceptual noise in the slow condition, we find that recognisable objects are processed over and above the background in the fast trials. However, we also found the opposite results to our expectation in the INT analysis, where regions lower in the visual processing hierarchy had slower intrinsic activity. A linear regression analysis revealed a negative relationship between INT and repetition response (Fig 3.5): regions with shorter INTs (τ) show repetition suppression ($\beta_{first} > \beta_{last}$) in response to repeated objects and vice versa for regions with long INTs and repetition enhancement. Pearson correlation with permutation testing revealed this relationship was significant most hemispheres and time conditions (Fast: Left Hemisphere r = -0.81, p = 0.0025, Right Hemisphere r = -0.78, p = 0.00015; Slow: Left Hemisphere r = -0.49, p = 0.051, Right Hemisphere r = -0.57, p = 0.015). A multiple linear regression revealed no significant interaction between the time condition and relationship between a given region's INT and repetition response, although the relationship is weaker for the objects in the slow trial. This is likely due to minimised response for the objects in this condition relative to the background in general. We therefore conclude that the repetition response profile along the ventral visual hierarchy is largely driven by the intrinsic neural activity of the region.

3.3 Discussion

The present study investigated the role of temporal processing in the VVS for recognising objects amidst perceptual noise, and how this relates to the intrinsic neural activity of the

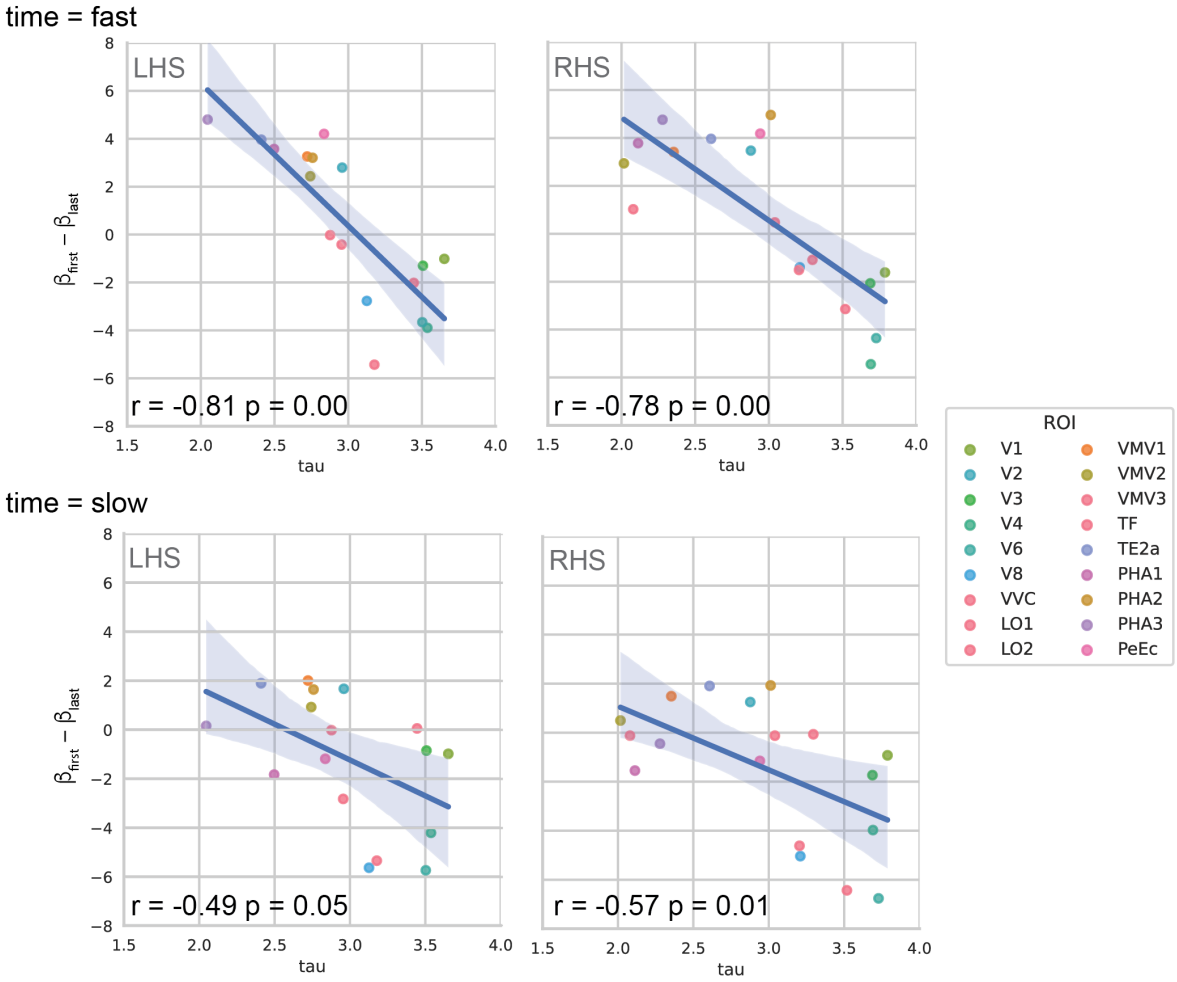


Figure 3.5 Association between tau and repetition response. The difference between first and last objects' beta response (β_{first} - β_{last}) on the y axis, and each region's estimated INTs (τ) on the x axis. Data are split by time condition and hemisphere, and averaged across voxels within each ROI from the Glasser parcellation, and then across subjects (n = 23). Positive values for β_{first} - β_{last} indicate repetition suppression in a given region, and negative values repetition enhancement. The line shows results of a linear regression, and error bands were estimated using bootstrap resampling across ROIs to obtain the 95% confidence interval. r and p values on each panel indicate the results of a Pearson correlation between β_{first} - β_{last} and τ. A linear multiple regression showed no interaction effect between β_{first} - β_{last}, τ and time condition.

visual cortex at rest. The models trained with a temporal learning objective in Chapter 2, revealed that semantic representations were learnable from long, but not short, lag intervals. We therefore hypothesised that object responses in anterior regions of the VVS would be distinguishable from persistent perceptual noise only at long lag intervals, and that regions in the posterior to anterior hierarchy would exhibit increasingly long intrinsic neural activity.

Contrary to our hypothesis, we observed a gradient of INTs from medial to lateral regions with object-selective areas in anterior VVS not necessarily displaying longer INTs than earlier visual regions. Furthermore, we found that in ventral regions the object responses were suppressed when repeated at a short ISI of 1.1 s (on average), but not at a longer lag of 11 s (on average). We find a negative linear correlation between a region's INT and its repetition response magnitude, underscoring the significance of intrinsic neural activity in interpreting fMRI adaptation effects.

The unexpected repetition response observed in the fast condition challenges the assumption that semantic features operate exclusively on slower temporal scales than perceptual features. It is known that humans can recognise objects rapidly (Grill-Spector & Kanwisher, 2005) and that the neural response to semantic features can begin within 120 ms (Clarke et al., 2013), meaning that our ISI of 1.1 s might still be slow enough to engage category-selective regions effectively. For instance, the semantic meaning and context of words can be processed by the parafovea within 100 ms (Pan et al., 2024). Prior research suggests that category-selective regions are sensitive to stimulus frequencies of around 4-5 item / s (McKeeff et al., 2007), and our presentation rate likely exceeded this threshold, irrespective of the perceptual noise from the background video. Future studies should address this by testing faster presentation rates, though the limited temporal resolution of fMRI remains a challenge.

Interestingly, objects presented at long lag intervals did not elicit the expected attenuated responses in the VVS but instead showed repetition enhancement. This enhancement was not due to reduced object recognisability, as task performance and activation in temporal association areas, involved in recognition memory, were consistent across conditions (Ho et

al., 2011; Málková et al., 1995). The response curve in most visual regions was well-modelled by a gamma function, with region-specific variations in the decay rate. In the VVC, a faster rate of decay was observed relative to early visual regions, indicating a consistent but non-significant response at the group level. According to the accumulation hypothesis of repetition suppression, this pattern suggests a slower response curve, leading to a larger area under the curve (AUC) from a broader peak amplitude, which could be interpreted as less effective priming from repeated objects.

Instead of the expected object-selective repetition suppression in VVC, slow trials triggered a linear curve of repetition suppression in PHA and significant activation in visual motion areas. This result suggests that at longer intervals, the background video dominated and overshadowed object recognition. Previous studies have identified long-lasting suppression effects after several minutes (Henson et al., 2000) or even days (van Turennout et al., 2000). However, our findings question whether such effects occur when repeated objects are encoded against a naturalistic background of perceptual interference, or simply indicate a lack of power in the design to detect these long-lag effects. An alternative explanation for the slow ISI responses could involve the significant activation observed in lateral regions like the LOC, known to construct scene meaning by integrating component objects (MacEvoy & Epstein, 2011). At slower presentation rates, the brain might be tuning into potential associations between recognisable objects and the background, leading to the suppression response in LOC. Although this hypothesis is not directly testable with our current data, it presents an intriguing avenue for future research.

The most compelling explanation for the observed repetition effects lies in the intrinsic neural activity of the regions themselves, rather than the stimulus conditions. We identified

a significant negative linear relationship within the ventral occipitotemporal cortex between INTs and repetition suppression. Specifically, regions with longer INTs exhibited repetition enhancement, while those with shorter INTs showed suppression. These INTs were estimated independently using a resting-state scan, confirming that the observed effects are not driven by task stimuli but are inherent to the regions' temporal dynamics. This suggests that the underlying temporal characteristics of a region are critical for understanding stimulus-driven repetition effects. Our findings align with the accumulation theory of fMRI adaptation (James & Gauthier, 2006), which posits that repetition enhancement occurs when a neural response is integrated over a longer duration. Traditionally, this theory has focused on task-based responses, but our results indicate that a region's intrinsic temporal activity plays a key role. Regions with longer INTs, where neural responses unfold more slowly, could allow repeated stimuli to build upon the initial response and leading to enhancement. Conversely, regions with shorter INTs, which process information more rapidly, may quickly habituate to repeated stimuli and exhibit suppression. This highlights offers new insight on the role of INTs for fMRI adaptation and its role in cognitive processes.

In conclusion, our study as designed does not support emergent hypotheses from Chapter 2 but does provide novel insights into the relationship between INTs and repetition effects within VVS. The findings challenge traditional models of fMRI adaptation by highlighting the critical role of a region's intrinsic temporal dynamics in shaping the response to repeated stimuli. This work not only advances our understanding of neural adaptation but also opens new avenues for investigating how intrinsic timescales influence cognitive processes in complex, real-world environments. Future research should further explore these dynamics, particularly in conditions with varied temporal structures and perceptual complexities.

Chapter 4 Deep neural networks reveal top-down development of category representations in the infant ventral stream.

What are the foundations of visual categories in the human brain? One theory proposes a bottom-up emergence in which they are extracted from statistical regularities in the environment, constrained by a proto-organisation. In contrast, the early presence of some categories suggests a role for an innate template. While infant looking behaviour has been used to characterise the development of categorisation it cannot measure covert neural representation or distinguish the underlying mechanism. For this, we need rich neuroimaging data from young infants and the capacity to apply advanced computational models of vision. Here, we present the largest cohort of infants in an awake fMRI study collected to date and find that categorical structure is present in high-level visual cortex from 2-months-old. This precedes its emergence in earlier visual cortex, supporting a top-down development of category representations. A deep neural network model aligned with infants' representational geometry, indicating that the features comprising this early category template span a range of complexities, and are of the kind that can be learned from the statistics of visual input. Our results reveal the existence of complex function in ventral visual cortex at 2-months-old, raising interesting questions regarding the role of innateness and learning in the formation of concepts.

This work is currently under review with the title "Infants Have Rich Visual Categories in Ventrotemporal Cortex at 2 Months Old"

Cliona O'Doherty[1,2], Áine T. Dineen[1,2], Anna Truzzi[1,2], Graham King[1,2], Lorijn Zaadnoordijk[1,2], Keelin Harrison[1,2], Enna-Louise D'Arcy[1,2], Jessica White[1,2], Chiara Caldinelli[1,2], Tamrin Holloway[1,2], Anna Kravchenko[1,2], Joern Diedrichsen[3,4], Ailbhe Tarrant[7,8], Angela T. Byrne[6,9], Adrienne Foran[7,8], Eleanor J. Molloy[5,6,9], Rhodri Cusack[1,2]

1 Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin 2, Ireland
2 School of Psychology, Trinity College Dublin, Dublin 2, Ireland

3 Departments of Statistical and Actuarial Sciences and Computer Science, Western University, London, Ontario N6A 5B7, Canada.
4 Western Institute of Neuroscience, Western University, London, Ontario N6A 3K7, Canada

5 Paediatrics and Child Health, Trinity College Dublin, Dublin 2
6 The Coombe Hospital, Cork Street, Dublin 8
7 The Rotunda Hospital, Parnell Square, Dublin 1
8 Children's Health Ireland at Temple Street, Dublin 1
9 Children's Health Ireland at Crumlin, Dublin 12

Contributions

Conceptualisation: COD, ÁTD, A Truzzi, GK, LZ, RC.

Methodology: COD, ÁTD, A Truzzi, GK, ELD, JD, RC. Recruitment and Resources: ÁTD, A Truzzi, GK, ELD, JW, CC, AK, A Tarrant, ATB, AF, EJM, RC.

Investigation: COD, ÁTD, A Truzzi, GK, ELD, JW, CC, TH, AK, A Tarrant, ATB, AF, EJM. Ethics and Project administration: A Truzzi, GK, LZ, ELD, JW, TH, A Tarrant, ATB, AF, EJM, RC.

Data, Curation and Preprocessing: COD, ÁT
D, A Truzzi, GK, RC.

Formal analysis and Validation: COD, RC . **Writing – original draft: COD**, KH, RC.

Writing – review & editing: COD, ÁTD, A Truzzi, KH, LZ, JD, RC.

Visualisation: COD, ÁTD, RC. Supervision: EJM, RC. Funding acquisition: RC

4.1 Introduction

Infants develop many concepts in their first year, laying the foundations for cognition. Contrasting theories on how they achieve this differ in emphasis on sensory-driven bottom-up statistical learning (Locke, 1948; Piaget, 1952), innate knowledge (Carey, 2009; Spelke & Kinzler, 2007) or specific innate learning algorithms (Mandler, 2004). The sensory-driven bottom-up emergence of visual categories may be guided by a "proto-organisation" that selectively connects regions tuned to specific low-level visual features to a particular category selective region (Arcaro & Livingstone, 2017; Hasson et al., 2002). For instance, faces are typically viewed at the fovea and comprise predominantly low-spatial frequencies, while scenes are generally seen in the periphery and include high-spatial frequencies (Kamps et al., 2020; Levy et al., 2001; Livingstone et al., 2017). A proto-organisation based on eccentricity and spatial frequency could, therefore, highlight category-relevant features and facilitate statistical learning in face and scene selective regions. However, in contrast to this bottom-up mechanism, there is evidence for innate templates for faces (Goren et al., 1975; Johnson et al., 1991; Reid et al., 2017), suggesting a more top-down learning process. Other category templates could potentially exist, though some argue against this based on the principle of parsimony (Mandler, 2004).

The emergence of visual categories can be evaluated by measuring the habituation of looking behaviour in response to sequences of images. Young infants have been found to be sensitive to global structure and perceptual features (Turati et al., 2003; Younger & Fearing, 2000), while 10-month-olds are sensitive to basic-level categories and conceptual features (Younger & Cohen, 1986). However, looking time cannot reveal the covert learning of conceptual knowledge before it manifests in behaviour. Furthermore, behavioural measures can be degenerate, as multiple mechanisms could lead to the same pattern of behaviour.

Neural measures address this by providing an alternative window into the mechanisms at play. In humans, electro- and magneto- encephalography (EEG and MEG), have found a staggered development of categories (Xie et al., 2022; Yan et al., 2023) that mirrors the transition in feature complexity found in behavioural studies. However, it is unclear which visual regions drive these results due to limits in the spatial resolution of EEG and MEG. Methodological advances in awake infant fMRI have the potential to overcome this limitation (Deen et al., 2017; Ellis et al., 2020). While low-level retinotopic maps predicted from the proto-organisation theory are present in young infants (Ellis et al., 2021), no evidence has been found for a bottom-up progression of category selective regions (Kosakowski et al., 2022). These results suggest instead that category responses in high-level visual cortex may not depend on the maturation of earlier stages in the ventral processing hierarchy, leaving an open question of when category representations emerge in the ventral stream. A further critical question is how the bottom-up and top-down mechanisms proposed for canonical categories, such as faces and houses, can be extended to the broader range of categories that human infants rapidly learn to recognise.

To provide robust insight into how these functions emerge, we believed it was necessary to acquire better neural data and to apply computational models of vision. Stimuli used in infant neuroimaging paradigms to-date have not been rich enough to separate low-level visual features from high-level category information, focusing instead on category-selective localised responses to strong stimuli such as faces and scenes in small sample sizes (Deen et al., 2017; Ellis et al., 2020; Kosakowski et al., 2022; Yan et al., 2023). This has precluded the measurement of distributed response patterns to a broad range of objects (Haxby et al., 2001) and restricted which computational models can be differentiated, specifically, comparison to deep neural networks (DNNs). These artificial neural networks (ANNs) are

now cemented as effective models of the adult brain in several domains (Doerig et al., 2023; Richards et al., 2019), acting as complex feature detectors that learn the statistical regularities in visual input to encode an increasingly complex range of multidimensional features in their hierarchical layers - comparable to adult visual cortex (Güçlü & Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014). Particular progress has been made using representational similarity analysis (RSA) (Kriegeskorte et al., 2008), where the distributed codes for diverse stimuli are compared between the brain and the DNN to characterise the features encoded in a visual region. Until now, this has not been possible in infants due to the methodological challenges of collecting sufficient quality neural data.

We acquired a large-scale longitudinal MRI dataset of functional brain activity from awake infants viewing a rich set of stimuli. This bridges the gap between large studies of sleeping or sedated infants (Howell et al., 2019; Makropoulos et al., 2018; Morris et al., 2020) and the handful of pioneering studies that test awake infants with small sample sizes (Deen et al., 2017; Ellis et al., 2020; Kosakowski et al., 2022). In early visual regions, we observe a transition from simple to complex feature representations. This contrasts to the ventral visual stream (VVS) where we find that the neural foundations of category knowledge are present from the first few weeks of life. A DNN model of vision reveals that the features in infant representations can be learned from the statistics of visual input, raising interesting questions regarding the role of innateness and learning in the formation of concepts.

4.2 Methods

4.2.1 Participants

Infants were recruited from participating maternity hospitals and attended two scanning sessions at Trinity College Institute of Neuroscience, the first of which was at 2-months of age. At 9-months-old, 28 participants' caregivers chose not to continue in the study and 3 infants were not invited back due to poor tolerance of the scanning at 2-months. Awake functional MRI was successful in 97% of 2-month-olds and 64% of 9-month-olds, a success rate that exceeded our projections based on previous awake infant scanning efforts (Deen et al., 2017; Ellis et al., 2020; Wild et al., 2017). The final dataset consisted of 130 2-month-olds (52 female, 78 male, 1.5 – 4.7 months corrected gestational age [CGA], mean = 2.4 months), 65 9-month-olds (30 female, 35 male, 7.5 – 10.9 months CGA, mean = 9.4 months) and was composed of infants born healthy at full-term (n = 96 2-months, n = 54 9-months) as well as a smaller proportion who were born pre-term and spent time in the neonatal intensive care unit (NICU) (n = 29 2-months, n = 11 9-months). For this initial study, both groups were combined but future work will test for differences between infants born full and preterm. Ethical approval was obtained from the Trinity College Dublin School of Psychology Research Ethics Committee, The Rotunda Ethics Committee and The Coombe Ethics committee.

4.2.2 Infant fMRI data collection

We iterated upon scanning procedures from pioneering methods (Ellis et al., 2020), using projection onto the bore ceiling and the bottom half of the head coil to allow the infant an uninterrupted view. Additionally, for participants at the 2-month timepoint, a 4-channel surface coil was placed above the infant's forehead. Similarly to Ellis and colleagues, we

built our own stimulus display program using PsychoPy (Peirce et al., 2019) v2022.1.4 in Python v3.7.12 which was essential to successful scanning, enabling constant stimulation as well as fast and flexible switching between tasks depending on the temperament of the infant. We deployed an in-bore MRI compatible camera (MRC Systems HighResolution), which streamed a video of the infant's face to the scanning team and the parents, allowing us to monitor the infant's comfort levels closely, put caregivers at ease, and record gaze.

MRI data were collected on a Siemens MAGNETOM Prisma 3T scanner using the posterior array of the 64-channel head/neck adult coil and four channel flex-coil in the 2-month old infants. The task-based fMRI sequences used multiband accelerator factor of 4, voxel size 3 x 3 x 3 mm, FOV 192 x 192 mm, phase encoding direction anterior to posterior, repetition time (TR) 610 ms, echo time (TE) 32 ms, flip angle 40 degrees, echo spacing 0.54 ms, 36 slices per volume. 510 volumes were collected for the pictures task. Two short reference sequences of single-band spin-echo volumes with opposite (anterior-posterior and posterior-anterior) phase encoding directions were also collected, with a TR of 2260 ms, a TE of 32 ms, and identical geometry to the fMRI, with a scanning time of 29 s. These two reference scans were used to apply susceptibility distortion correction using FSL's TOPUP. T2 weighted images were a noise-reduced scan consisting of the following parameters: 100 contiguous near-axial slices, GRAPPA acceleration factor of 3, voxel size 1 mm × 1 mm x 1 mm, coverage of the whole head and a field of view 192 mm, TR 6090 ms, TE 84 ms, flip angle 150 degrees, echo spacing 12 ms, and duration 4 m and 5 s. Later in the study, a T1 MPRAGE sequence was added with 144 sagittal slices, voxel size 1 mm × 1 mm x 1 mm, coverage of the whole head and a field of view 256 mm, TR 2300 ms, TE 2.98 ms, TI 900 ms, flip angle 9 degrees, echo spacing 7.1 ms, and duration 6 m and 28 s.

The set of data collected in each session varied depending on infant temperament, technical capacity and time constraints. In an ideal scan session, we collected multiple repetitions of the awake pictures task (228 runs from *n* =114 2-month-olds, 82 runs from *n* = 51 9-month-olds) in which infants viewed looming objects from a variety of categories spanning the animate/inanimate distinction and an awake naturalistic video fMRI task which is not reported upon here (final numbers collected: 346 runs from *n* = 131 2-month-olds, 122 runs from *n* = 64 9-month-olds). Runs that were deemed to be of poor quality by attending researchers were manually excluded from later analyses (23 runs from *n* = 18 2-month-olds and 8 runs from *n* = 6 9-month-olds). In cases where infants fell asleep during scanning T2 (*n* = 104 2-month-olds, *n* = 32 9-month-olds) and T1 (*n* = 44 2-month-olds, *n* = 23 9-month-olds, added later in the study after we observed high tolerance) anatomical scans were collected as well as 10 mins of asleep resting state fMRI (*n* = 105 2-month-olds, *n* = 27 9-month-olds). Extensive manual data curation and quality control were crucial in standardising the infant fMRI data. The final dataset consisted of 130 2-month-olds and 65 9-month-olds. The sample included in the MVPA analyses was further refined after rigorous motion censoring and modelling restraints, giving a sample of 103 and 38 2- and 9-month-olds respectively.

4.2.3 Preprocessing

MRI data were pre-processed using an in-house pipeline written in Python 3.8.10 with the NiPype 1.8.5 processing framework (Gorgolewski et al., 2011) to avail of the neuroimaging software packages FSL 6.0 (Jenkinson et al., 2012) and ANTs 2.4.4 (Avants et al., 2011). Two fieldmap reference scans with opposite phase encoding (A/P and P/A) were input to FSL TOPUP (Andersson et al., 2003) to estimate the susceptibility-induced off-resonance field. EPI images were motion corrected to a middle reference volume using FSL MCFLIRT

(Jenkinson et al., 2002) and converted to SPM format from which framewise displacement and DVARS were calculated. The field distortion estimation from topup was then applied to the EPI data to correct for susceptibility distortions. Brain images were co-registered to the mean of a chosen reference scan using an affine transform with FSL FLIRT (Jenkinson & Smith, 2001). Due to limited movement in the infants when they fell asleep, resting state scans were the preferred choice for the functional reference. In the absence of a resting state scan, awake fMRI runs were used instead. The degrees of freedom (DOF) used for the affine transform were chosen by manual inspection for each infant, with the vast majority of participants having a successful registration with DOF = 12. Scans in native space were then normalised to a common space with FSL FLIRT, using an age-appropriate NIHPD template (Fonov et al., 2011). The Schaefer 400 parcel 7 network atlas (Schaefer et al., 2018) was transformed into the infant space to define two ROIs: posterior early visual cortex (EVC) and anterior ventrotemporal cortex (VTC). This was achieved by first identifying the subset of visual network regions in the age-adjusted atlas. Those regions covering caudal occipital cortex up to the parieto-occipital sulcus were allocated to the EVC, and ventral visual regions up to anterior temporal lobe were allocated to VTC.

4.2.4 Stimuli selection and adult piloting

Picture stimuli were three exemplars from 12 categories, chosen to be something an infant may encounter in the first year of life, in person or in books, to correspond to words with an early age of acquisition, and to be balanced across the animate, small inanimate and large inanimate categories. These were cat, crab, bird, squirrel, rubber duck, dishware, fence, food, supermarket shelves, shopping cart, towel and tree. Piloting in adults helped inform the final 12 categories tested. Adult participants (n = 18) were recruited through advertisements and word of mouth in Trinity College Dublin under the approval of the

School of Psychology Ethics Committee. One participant was excluded due to technical difficulties during the session, resulting in a final sample of n = 17. Scanning consisted of 3 separate fMRI runs of the pictures task, with twice as many stimuli as in the infant paradigm and only one repetition of a stimulus per run (6 contexts x 4 object types x 3 instances x 1 repetition). The final infant stimuli were chosen from this larger set of adult pilot stimuli through an iterative modelling process to find the stimuli that gave the strongest and most separable brain responses in adults, while adhering to known organising principles of the ventral stream.

Adult MRI data were collected using the full 64-channel Siemens Head Neck coil. The acquisition parameters were as follows: a multiband acceleration factor of 4 was used with a voxel size of 3 x 3 x 3 mm and a FOV of 224 x 245; the phase encoding direction was left to right; repetition time was 656 ms and echo time was 30 ms; flip angle 50 degrees and echo spacing of 0.54 ms. 40 slices were collected per volume and 561 volumes were taken due to the longer design for this piloting study. As per the infant protocol, DICOM data were converted into BIDS format (Gorgolewski et al., 2016) using HeuDiConv v0.10.0 (Halchenko et al., 2021). The adult dataset was pre-processed using fMRIPrep (Esteban et al., 2019) with FSL.

4.2.5 Task fMRI

Pictures appeared in pseudo-random order against a grey background for 3 s followed by a fixation cross, with the jittered ISI ranging between 3.5 – 4.5 s. Images started at half of their final size and loomed towards participants, increasing logarithmically in time, doubling in size within the presentation window. There were 3 exemplars for each of the 12 categories and each was repeated twice per fMRI run. This gave a final design of 12 categories x 3

exemplars x 2 repetitions, totalling 5 min of scanning. During the task, a series of short nursery rhymes played through the infant headphones to maintain engagement. Images were obtained online, selecting exemplars from a variety of viewpoints for each category. Backgrounds were removed and the images were rescaled so that the largest dimension, either width or height, touched the image border. Then, images were resized and padded to a standard size of 640 x 360 pixels and pre-distorted to ensure they appeared normal when displayed on the scanner's curved surface, given the projection geometry.

4.2.6 First level model and Multivariate Pattern Analysis

A volumetric whole-brain generalised linear model (GLM) was fit for each functional run with custom Python code using the Nilearn package (v0.9.2). A design matrix was constructed with 36 regressors of interest, one for each object, convolved with a canonical Glover HRF function. To account for motion, the framewise displacement values calculated during pre-processing were used to add spike nuisance regressors, giving censoring in frames that exceeded a threshold of 1.5 mm, thereby removing these frames from parameter estimation. This method has previously shown to lead to statistical improvements in high motion cohorts (Siegel et al., 2014). If greater than 50% of the frames in a run were above this motion threshold, the entire run was discarded from further analyses. This was the case for 35 runs in the 2-month-olds and 1 run in the 9-month-olds. Three translation and three rotation motion regressors calculated during pre-processing were included as covariates, as well as a linear trend regressor and a cosine drift model with a high pass filter of 0.01 Hz.

The run-wise parameter estimates (betas) for each condition within each ROI were calculated as well as the variance/covariance (vcov) matrix across timepoints of the design, multiplied by the residual mean square image to estimate dispersion. To control for

differences in baseline signal within each fMRI run, we performed run-level mean centring of estimates by subtracting each voxel's run-level mean across all trials from all estimates within a run, as has been shown to improve MVPA results (Lee & Kable, 2018). Due to the high motion in some infant runs and the use of censoring, a small percentage of model fits resulted in unstable parameter estimates and very noisy betas (5% of 2-month data points and 3% of 9-month data points). To overcome this, the vcov estimates were used to threshold the voxel wise betas. If a particular voxel had a vcov dispersion >10 it was excluded from the MVPA, this value was chosen by cross-validation of a range of thresholds and their effect on the resulting beta distribution.

The pairwise correlations between the 36 objects' voxel wise response patterns were calculated across all unique pairs of subjects (2-months: 13 443 pairs, 9-months: 1 308 pairs, adults: 136 pairs). These across subject-pair RDMs were then averaged to calculate a group visual representation. The sampling distribution was estimated using bootstrap resampling (1000 bootstraps) across pairs of subjects' RDMs and used this to calculate test statistics. To measure the bound on performance for the computational models from the fMRI data, the split-half noise ceiling was calculated using the Spearman-Brown prophecy formula (Lage-Castellanos et al., 2019). Pairwise distances were projected into a 2-dimensional space using multidimensional scaling (MDS). Embedding spaces were aligned hierarchically while minimising stress using the SMACOF algorithm. Embedding models were fit across ages and regions, then separately refined for each region collapsed across ages and finally for age.

4.2.7 Variance partitioning

The raw correlation ranges varied in scale with age [2-months (-0.02, 0.03); 9-months (-0.06, 0.08); adults (-0.09, 0.12)]. This could reflect a strengthening of the representational resources required to form concepts (Carey, 2004) but it may also be attributed to noisier data from the infant cohort. By averaging across subjects for all possible pairs of activation correlations, our approach ensures that the resulting RDMs are mostly driven by signal with minimal effects of subject- and run-specific noise. However, these other sources of variance could be contributing to differences across the groups, which would not be detectable here. We conducted a variance partitioning analysis to measure changes in session-based, group-level and subject-specific idiosyncratic signals. To differentiate potential sources of variance in the visual representations, RDMs were partitioned by group, individual across session and individual within session components using covariance as the distance metric, instead of correlation, to leverage its additive property. Within-run repetitions of a single stimulus were required and when fitting a model that had 3 exemplars and 2 repetitions as separate regressors, the estimates became unstable. Therefore, we decided to use a GLM at the 12 category level, which had two repetitions per category in a run. All subjects and runs were then aggregated into a complete stimulus x voxel matrix, and covariance RDM was calculated. This was composed of three types of comparisons: (i) across pairs of subjects (ii) within subjects, but across stimulus repetitions and (iii) within subjects, but within stimulus repetitions. RDMs for group, individual and session variance could then be partitioned from these covariance matrices, and the trace of each used as an estimate of each source of noise. Error estimates on the variance partitions were calculated with bootstrap resampling across subjects/runs, which gave an estimate for the developmental changes in group, idiosyncratic and session noise. We find that session-based noise did decrease with age, but this was accompanied by an increase in both the individual and group components (Supplementary

Fig. 4.2). It is possible that the increasing idiosyncratic representational alignment could reflect individual experience (Charest et al., 2014; Kovack-Lesh et al., 2014) and increased group variance could reflect common visual experience in the world. However, the origins of the signal strength differences – whether neural or hemodynamic (Arichi et al., 2012) – remained unclear. We therefore focused on representational content by z-scoring across each RDM. This approach allowed us to focus the presence or absence of distinct responses to a broad set of categories in the ventral stream, tested using second order correlations in RSA.

4.2.8 Representational similarity analysis

The correlation between visual representations and a set of theoretical feature RDMs were used to assess the presence of perceptual features and hypothesis-based semantic distinctions. The perceptual feature RDMs - to which infants in the 4-months to 19-months range are sensitive (Spriet et al., 2022) - were calculated from the pairwise distances between images' size (number of foreground pixels – number of background pixels), elongation (max of height-to-width or width-to-height ratios), colour (the average of the 3 colour channels of the image) and compactness (the ratio between the area of the shape and the area of the disk with the same perimeter). Semantic RDMs were built using categorical distinctions. Namely, within vs. between exemplar (implemented in practice as 1 - an identity matrix) and generalisation across exemplars but within category (within exemplar contrast weights are 0, within category but across exemplar weights are -1, and across category weights are 1) . Lastly, broad semantic structure was tested using the tripartite distinction, as our stimuli incorporated animate objects, large inanimate objects and small inanimate objects.

The Spearman correlation was calculated between the vectorised upper triangles of brain and feature RDMs, and the sampling distribution was estimated using bootstrap resampling

(1000 bootstraps) across pairs of subjects that entered the group-average RDM. Group differences were calculated by testing significant difference from 0, using the 95% confidence intervals of the difference between the sampling distributions for pairs of age groups. Partial correlations of the semantic models, accounting for all perceptual models as covariates, determined the extent of semantic information encoded in the visual representations beyond basic perceptual features. Final correlation values were expressed as a proportion of the noise ceiling for a given ROI and age-group. The leading diagonal was included in this RSA, as the individual exemplars were important for the category models, and the group-level RDMs calculated across subjects resulted in a meaningful diagonal. When excluding the main diagonal for RSA, the trend in the results reported in Fig. 4.2a were consistent, with slightly diminished correlation values. The singular exception to this was the colour model, which had no correlation to the EVC and VTC representations when excluding the main diagonal.

4.2.9 Word embedding model

Word embeddings were found for our 12 stimulus categories using the keyed vectors from *word2vec* (Mikolov et al., 2013), and the multimodal large language model *CLIP* (Radford et al., 2021). The *word2vec* model was trained on the *wordnet* corpus which does not contain our exact categories, therefore we chose the most semantically similar words. These were cat, crabs, bird, squirrel, duck, dishes, fence, food, shelves, carts, towel and tree with the most notable deviation from our visual stimuli being duck in place of rubber duck. However, this was less of a concern as we found that infant and adult fMRI responses for rubber duck grouped this object with other animate categories. The *word2vec* model was mostly dominated by the animate category, so we also tested a model which did not rely on the closest matching wordnet categories. *CLIP* embeddings were calculated for the words best

describing our visual categories; cat, crab, seagull, squirrel, rubber duck, dishware, plate of food, bath towel, garden fence, grocery shelves, shopping cart and tree. This multimodal model had the added advantage of incorporating visual and linguistic data into its representations, leading to increased correspondence to the brain in EVC as well as VTC (Supplementary Fig. 4.4). As above, RDMs for both models were calculated using the correlation distance between word embeddings. Because our stimuli corresponded to 12 distinct words, the word embedding model was correlated with fMRI RDMs calculated from a separate generalised linear model with the 12 categories collapsed across exemplars.

4.2.10 Deep neural network modelling

Deep learning analysis were written in Python v3.8 with PyTorch v2.0.1 and CUDA 11.7 and all models used the CNN AlexNet (Krizhevsky et al., 2012) as the architectural backbone. The untrained model weights were randomly initialised using Glorot initialisation (Glorot & Bengio, 2010). The supervised model was initiated using the default pretrained weights from torchvision v0.15.2, and finally the self-supervised model tested here used the pretrained weights provided by the authors of the self-supervised model instance-prototype contrastive learning (https://github.com/harvard-visionlab/open_ipcl). Both pre-trained models have previously been validated as models for the adult VVS (Khaligh-Razavi & Kriegeskorte, 2014; Konkle & Alvarez, 2022; Yamins et al., 2014). The 36 experimental stimuli, projected on a grey background, were input to each model and the activations from each layer were recorded while keeping model weights frozen. RDMs were then calculated using the pairwise correlation distances between the images' activation vectors. Finally, we performed RSA using the upper triangle of the model RDMs and the fMRI RDMs were used to measure the degree of similarity between the DNN and brain using Spearman correlation with bootstrapping. Statistics were as described in Section 4.2.8.

4.3 Results

4.3.1 Infant visual representations measured with awake fMRI at scale

Functional MRI (fMRI) was acquired longitudinally from 2-month-old (n = 130) and 9-month-old (n = 65) infants as they viewed 12 categories of common objects. These were chosen across animate, small inanimate and large inanimate classes that would typically be seen in the first year of life (in person or through books), corresponding to words with a young age-of- acquisition (Frank et al., 2017). For each category, we included three representative exemplars from diverse viewpoints. The success of awake fMRI depends upon infants staying content and relatively still in the scanner, so the pictures loomed to captivate attention. Most infants (63%) participated for at least four repetitions of each stimulus, totalling 144 pictures across 10 minutes of scanning. The distribution of head motion was acceptable [runs with a median framewise displacement of < 1.5mm at 2-months, 85%; 9-months 97%] (Fig. 4.1b,c) and allowed for rigorous scrubbing during fMRI analysis (see Methods). A cohort of adults (n = 17) was acquired for comparison. The BOLD response to each object exemplar was estimated with a generalised linear model and RSA was used to measure the representational geometry of the EVC and VTC. To do this, the similarities of voxel wise response patterns to each pair of conditions were calculated across subjects, giving a representational dissimilarity matrix (RDM) for each ROI and age group (Fig. 4.1d,e). The group mean RDMs were reliable, as estimated by a split-half noise ceiling that calculates the correlation that would be expected with a repeat of the entire experiment [2-months-old: $\rho(66)$=0.86, 0.90 for EVC and VTC, respectively; 9-months-old: 0.83, 0.90; adult: 0.89, 0.92]. The raw correlation ranges varied in scale with age [2-months (-0.02, 0.03); 9-months (-0.06, 0.08); adults (-0.09, 0.12)] and were z-scored to focus on representational content, in terms of relational geometry, rather than strength (see Methods).

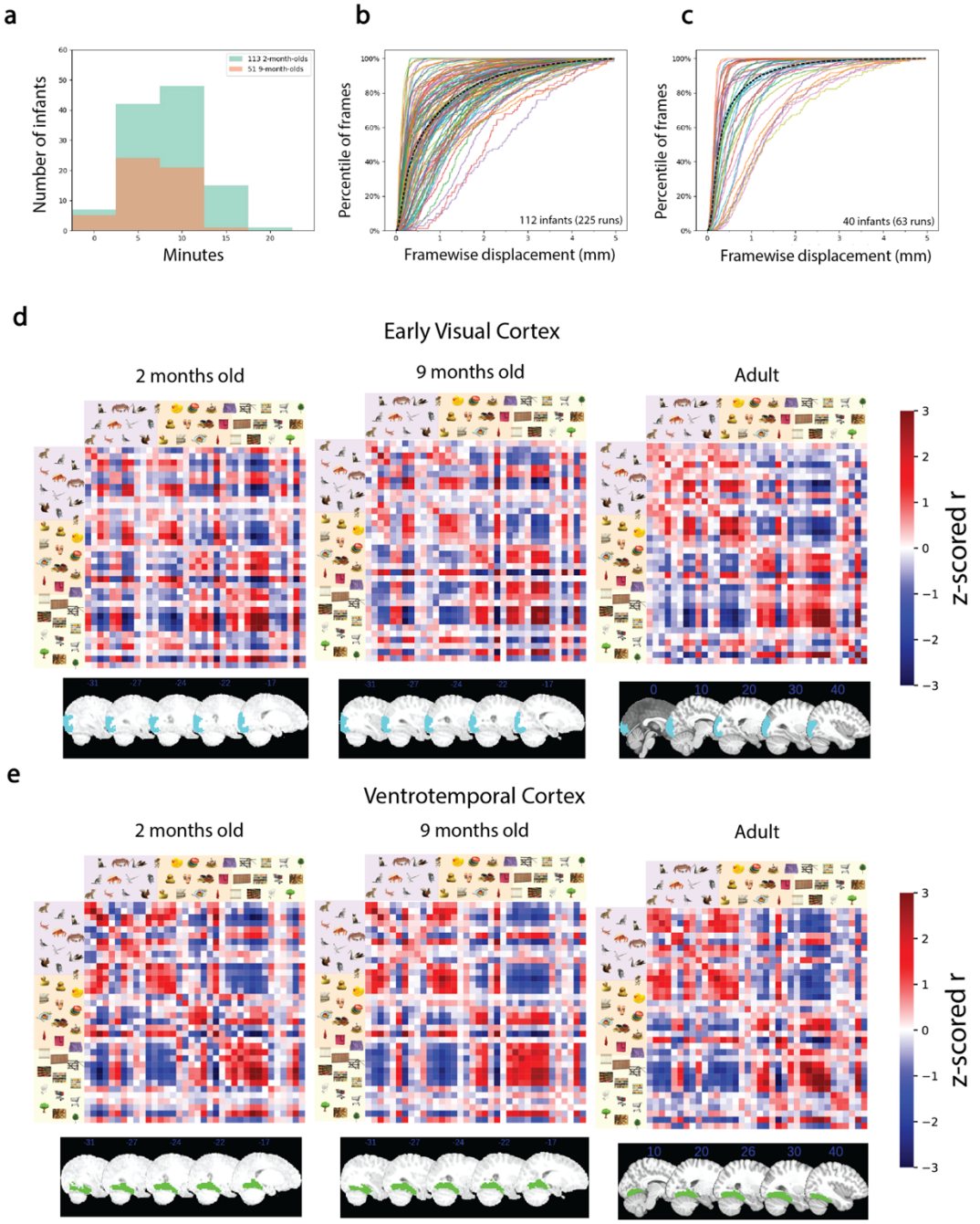


Figure 4.1. Visual representations from infancy to adulthood. **(a)** Frequency of time spent in the awake fMRI task for 2-month-olds and 9-month-olds. **(b,c)** Cumulative frequency plots showing the motion in a given run for 2-month-olds and 9-month-olds respectively. The black dotted line indicates median framewise displacement. **(d)** The group average, across subject pairwise correlation distance between voxel wise response patterns from EVC to each object. Conditions within the RDMs are nested by animacy class (animate, inanimate-small to inanimate-large), then category (4), and finally exemplar (3). To focus on representational content rather than strength, we plot the z-score of the correlation distance. Infant ROIs are displayed at x=-31,-27,-24,-22,-17 and adult ROIs at x=0,10,20,30,40. **(e)** Group visual representations for VTC. Infant ROIs at x=-31,-27,-24,-22,-17 and adult ROIs at x=10,20,26,30,40.

4.3.2 Basic and global categories are present in the ventral stream from 2-months-old

Infant representations in both EVC and VTC were surprisingly mature. Different exemplars evoked clearly distinct BOLD patterns in both regions, assessed by a contrast of the within-exemplar versus between-exemplar similarities (exemplar RDM). All statistics, including the across group differences and their significance, were calculated using bootstrap resampling across subjects to estimate the sampling distribution (see Methods). In EVC, representations showed a strengthening in their correlation to the exemplar RDM from 2-months-old [spearman correlation of brain RDM and exemplar RDM, r = 0.279, 95% bootstrap CI = (0.265, 0.285)] to 9-months-old [r = 0.297, CI = (0.282, 0.303)] and again to adulthood [r = 0.330, CI = (0.319, 0.334)]. In contrast, objects in VTC were as distinct at 2-months-old [r = 0.310, CI = (0.300, 0.315)] as they were at 9-months-old [r = 0.305, CI = (0.294, 0.310)] with a smaller difference from adult-like object distinctions [r = 0.326, CI = (0.319, 0.330)]. Our novel dataset revealed a clear stimulus-specific representational geometry in the infant brain which shows considerable continuity with adults. This maturity of individual object representations was evident in VTC prior to EVC, at an age when overt visual categorisation is restricted or undetectable (Spriet et al., 2022), visual acuity is still developing (Dobson & Teller, 1978) and experience with the world is limited (Clerkin et al., 2017).

Having found distinct representations for individual objects, we proceeded to test the features comprising this relational geometry through comparison with perceptual and categorical model RDMs (Supplementary Fig. 4.3). Informed by existing behavioural and EEG evidence (Spriet et al., 2022; Xie et al., 2022) we hypothesised there would be a transition from low-level perceptual organisation in the infants to high-level categorical organisation in adults. According to the bottom-up theory of visual development, this

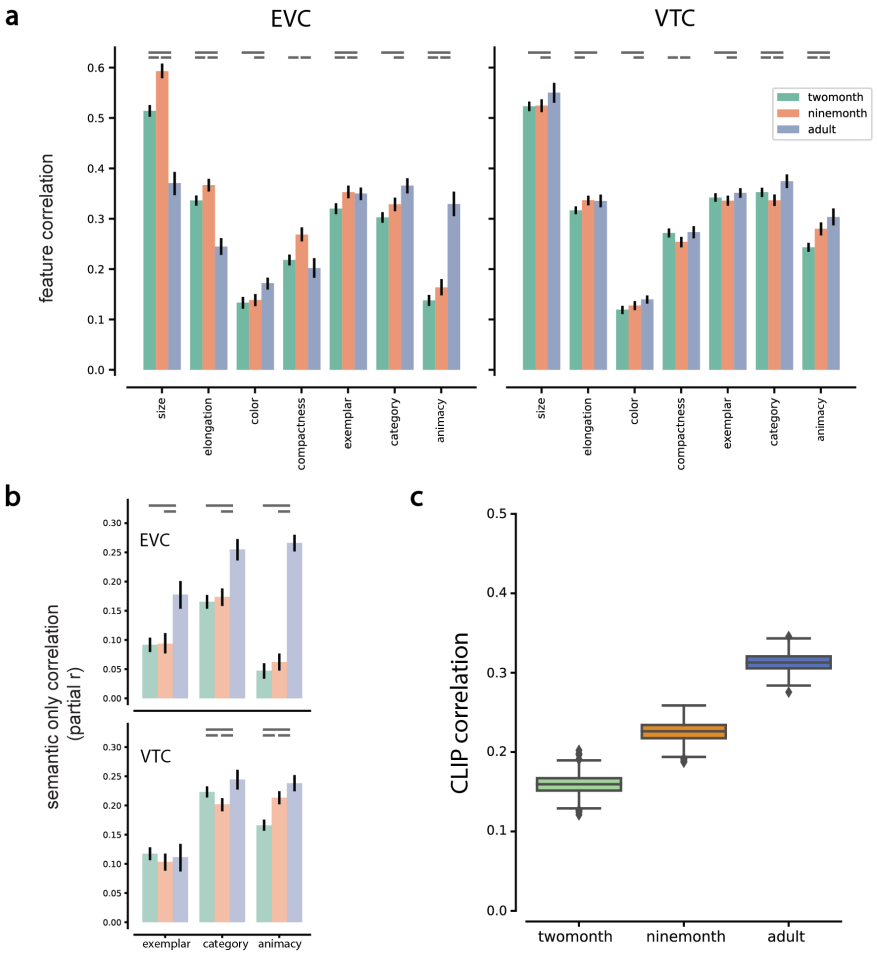


Figure 4.2. The development of perceptual and semantic feature representations. (a) Spearman correlation, adjusted for the noise ceiling, between visual cortex and hypothesis-based model RDMs. Features are ordered from perceptual to categorical. Error bars indicate the bootstrapped 95% confidence interval, calculated across subjects, and horizontal lines above the bars indicate significant differences between the age groups calculated using the differences between bootstrap distributions to test significant difference from 0. **(b)** Partial correlations for the categorical features tested in (a), controlling for all four perceptual features. Error bars and significance are as in (a). **(c)** Bootstrap distributions, calculated across subjects, for the correlation between visual response patterns in VTC and word embeddings for our 12 categories calculated

perceptual to categorical transition should occur in EVC prior to VTC. The perceptual RDMs chosen were size, elongation, colour and compactness – features which have previously been used to model the trajectory of infant looking time behaviour from 4- to 19-months (Spriet et al., 2022). Category based RDMs were the exemplar RDM described above, generalisation across different objects within the same basic-level category (category

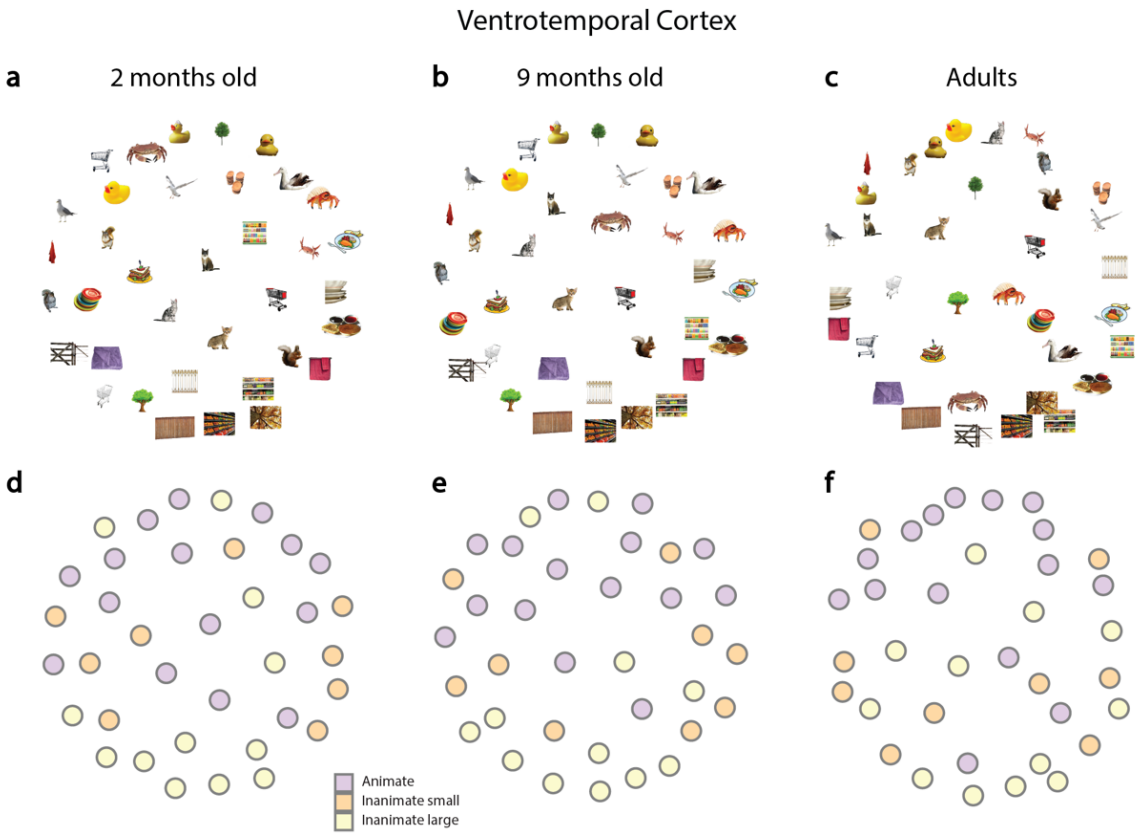


Figure 4.3. Global organisation by animacy is present in the ventral stream from 2-months-old. Multidimensional scaling plots of the VTC representations in Fig. 4.1e. **(a-c)** Embedding space of the pairwise distances for each age group in VTC. Embeddings have been aligned using the SMACOF algorithm. **(d-f)** The same embedding projections, displayed according to animate, inanimate large and inanimate small categories. Rubber ducks were grouped into the animate category in all age groups' visual representations and have been defined as such here.

RDM), and similarity within the global categories of animate, small inanimate and large inanimate objects (animacy RDM). This tripartite, animacy distinction is a known organising principle of adult visual cortex to which infants are not behaviourally sensitive until 10-months-old (Konkle & Caramazza, 2013; Spriet et al., 2022).

All features were present in each age group (Figure 4.2a), including distinct categorical representations across visual cortex at 2-months-old. Infant EVC representations were driven by perceptual features, transitioning to more categorical representations by adulthood. However, this was not the case in VTC where we found similar trends with

perceptual and categorical representations across the age-groups; rather than a transition from the representation being driven by low-level visual features to high-level information, VTC begins high-level and is refined with age. A partial correlation revealed that after controlling for the perceptual similarity between images, the correspondence between infant and adult category representations persists in VTC but not EVC (Figure 4.2b). This suggests that the developmental transition in feature complexity previously reported from behavioural and M/EEG studies is being driven by posterior visual regions, a finding that is possible due to the higher spatial resolution in our fMRI data. Moreover, the developmental transition of feature complexity in EVC – a lower region in the visual processing hierarchy – was not necessary for the category representations to be present in VTC, contrary to the bottom-up theory of visual development.

We extended this test to global categories using the tripartite distinction for animate and inanimate objects (Konkle & Caramazza, 2013) and find evidence for this organisation much earlier in the brain than has been previously reported from looking time. The tripartite organisation is mostly present in VTC from 2-months-old and becomes refined with age and experience (Figure 4.3) [partial spearman correlation, controlling for the four perceptual features 2-months-old r = 0.150 CI = (0.140, 0.159); 9-months-old r = 0.194 CI = (0.178, 0.205); adult r = 0.221 CI = (0.197, 0.240), significant change with each age, (Figure 4.2b)]. We do not see the same organisation and early development in infant EVC, which has quite disparate representation of the tripartite distinction between infants and adults [partial spearman correlation, controlling for the four perceptual features 2-months-old r = 0.041 CI = (0.030, 0.051); 9-months-old r = 0.052 CI = (0.038, 0.066); adult r = 0.242 CI = (0.216, 0.258), significant change from 9-months to adults only].

The presence of basic and global visual categories at 2-months-old prompts the question: when do more detailed distinctions develop? To investigate this, we used word embeddings of the 12 categories from a neural network trained on natural language and another trained on vision and language (Supplementary Fig. 4.4). These models are known to estimate the semantic distances between concepts that are represented in adult visual cortex, with overt evidence of semantic knowledge perhaps emerging with the first words at around 10-14 months. At 2-months-old, we found the weakest representation of fine-grained semantics in VTC, but by 9-months-old these distinctions had strengthened (Figure 4.2c). This is a time when infants recognise a few nouns (Bergelson & Swingley, 2012) and begin to associate words with an increasing number of visual categories (Pomiechowska & Gliga, 2019), just prior to the onset of speech.

4.3.3 Deep neural networks model infant visual cortex

We found the presence of high-level features in infant visual representations, as defined by within vs. between category distinctions. However, category representations could be shaped by many visual features with potentially complex multidimensional tuning in feature space. DNNs have shown great success in capturing these complex feature manifolds from the statistics of visual input, learning representations of objects that are simultaneously successful for recognition and relevant for adult visual cortex (Yamins et al., 2014). To date, parallels to these models have only been drawn to the "fully-trained" adult brain. With our novel dataset characterising infant visual representations, we can now compare the *learning* human brain to the *learning* model. Here, we tested if deep learning-based models of visual recognition can capture infant neural responses. Using the 36 images from the fMRI experiment, we calculated activations from a set of models with an AlexNet architecture (Supplementary Fig. 4.5-4.7) by testing the extremes of the training process – untrained and

fully-trained. The architecture of untrained neural networks can provide sufficient inductive biases for capturing object relevant information (Gaier & Ha, 2019), but training the network weights through visual input to the model confers significant advantage for representation learning. By testing correspondences to the models at these two stages, we can delineate the contribution of features that are learnable from the statistics of visual experience for explaining brain representations.

Infant representations correlated more with an untrained DNN than adult representations at both 2-months and 9-months (Figure 4.4a). However, the fully-trained network outperformed the untrained model across all age groups demonstrating that, from as early as 2-months-old, the distributional structure learned from experience is important for explaining visual representations. Even with less visual experience (Cusack et al., 2024), infants' representations are sophisticated enough to correspond well with those from a fully trained neural network. This correspondence increased with development, alongside a decrease in correlation with the untrained network, demonstrating the continued influence of visual input as we age after a strong initial start.

The DNN model above underwent supervised training, where each image had a corresponding label, something preverbal infants would not have had access to. Instead, infants are sensitive to comparisons and patterns within the stream of sensory input (Fiser & Aslin, 2002; Oakes et al., 2009), aligning better with self-supervised training. To test if the observed effect generalised to DNNs trained without labels, we compared to a DNN trained with a self-supervised algorithm (Konkle & Alvarez, 2022). Infants showed similar patterns of correlation to the model regardless of learning algorithm (Fig. 4.4d,e). This was not the case for adults. Adult VTC was more similar than infant VTC to deep layers of both

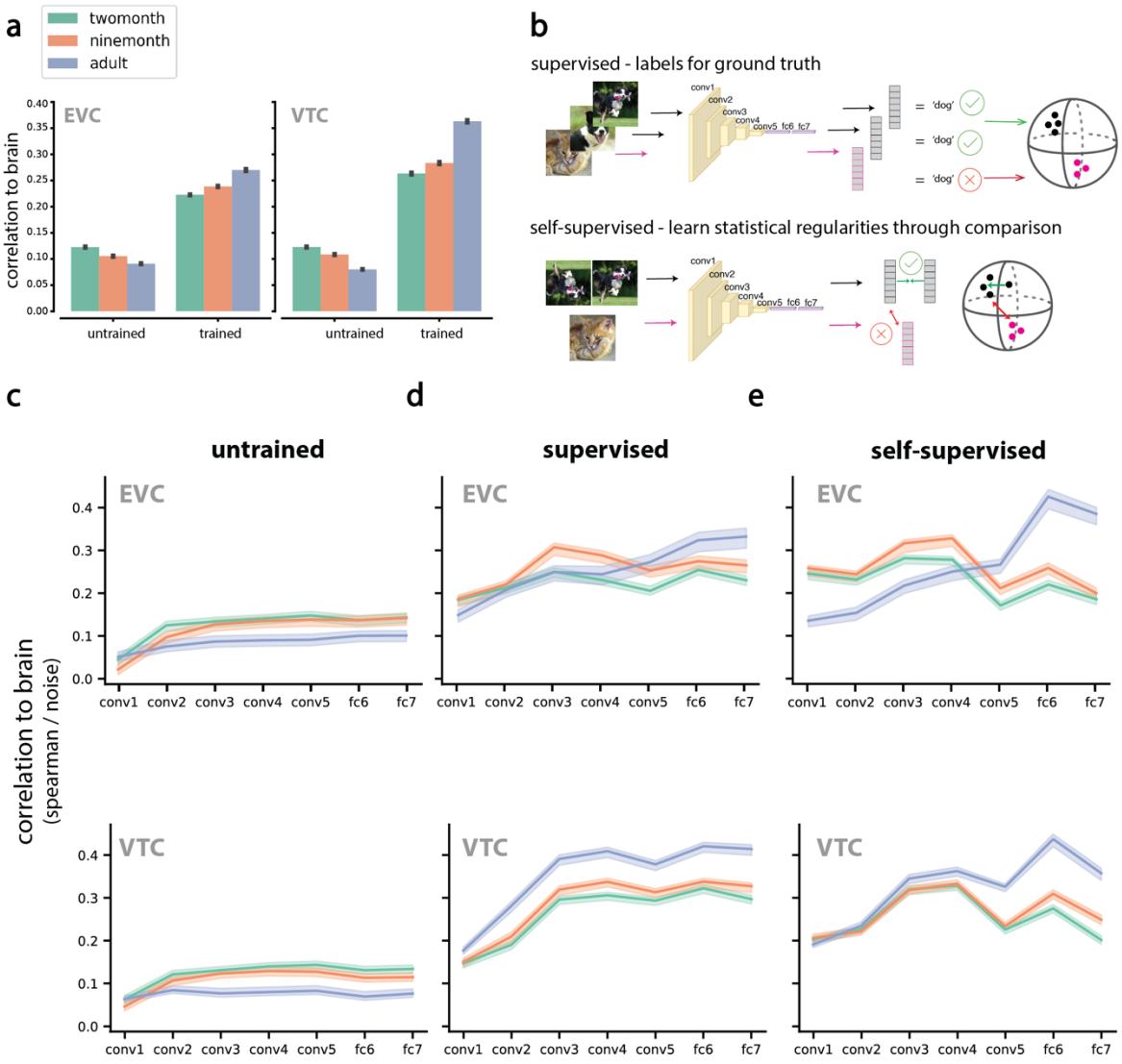


Figure 4.4. Deep neural networks model visual representations from infancy to adulthood. **(a)** Spearman correlations, adjusted for the MRI noise ceiling, for untrained and trained networks in each age group. Error bars indicate bootstrapped confidence intervals, calculated across subjects. **(b)** Schematic of supervised vs. label-free self-supervised learning. **(c-e)** Layer-wise Spearman correlations between the DNN AlexNet and visual cortex, adjusted for the noise ceiling within each age group and ROI. Activations were calculated from three DNNs – (c) untrained, initialised with random weights; (d) supervised, trained on a fully labelled object recognition task and (e) self-supervised, trained using instance protocol contrastive learning. Error bands depict the 95% confidence interval, calculated using bootstrap resampling across subjects.

fully-trained models, and to every layer of the supervised model, revealing an increased contribution of high-level semantic information through development. The absence of an expected hierarchical correspondence between early layers and adult EVC (Güçlü &

Gerven, 2015) is likely attributed to the large cortical area covered by our ROIs. The correlations of infants and adults to the DNN were generally more comparable in VTC compared to EVC regions, except for in deep layers. This regional dissociation supports the idea of a perceptual-to-conceptual transition in the early visual regions, whereas higher ventral areas begin with a more mature representation that becomes increasingly semantic with age and experience. The high-level feature representations in areas lower in the processing hierarchy appear later, alongside increased experience and development. Notably, the model trained with self-supervised learning explained infants' EVC significantly better than adults', demonstrating for the first time that different learning algorithms are better models for different developmental stages. This opens the future possibility of identifying the learning model that best explains infant brain data.

4.4 Discussion

Using a broad set of stimuli, we found that infant VTC contained category and animacy information at 2-months-old, which was not explained by low-level visual features. This category structure appears in later regions of the visual hierarchy before emerging in early regions, suggesting a top-down propagation of category representations. Previous studies of face and place categories have found that focal regions of selectivity emerge earlier in anterior regions compared to posterior regions during development (Kosakowski et al., 2022). We have extended this finding to a broad range of animate and inanimate categories, with important implications for the developmental mechanism.

Infant visual representations corresponded with a fully-trained DNN, indicating that complex features learnable from visual experience are present as early as 2-months. A critical question for the future will be to establish when this representation is set up. Is it

rapidly learned in 2-months of visual experience, despite the paucity of visual acuity and colour vision during this period (Dobson & Teller, 1978; Skelton et al., 2022), or has it been learned by past generations and then transcoded by evolutionary adaptation into an innate form? As the visual representations were calculated across subjects, they show what is stereotypical across the group. This suggests they are shaped by a common genetic code and not merely learned from experience. There are known examples of neural adaptations to the statistics of the visual world. For example, the centre-surround structure of retinal ganglion cells provides efficient coding for spatially extended patches in the visual environment by performing edge detection (Kuffler, 1953). We propose that neural adaptation has been extended to the identification of visual categories and their invariances.

Why might category representations develop in a top-down manner? An intriguing possibility arises from the idea that VTC categories could be innate. The genetic code, along with developmental constraints, limits the biological structure that can be passed through evolution. The brain's complex wiring exceeds this capacity (Koulakov et al., 2022), so simple rules must be encoded to guide its development. In category selective cortex, such a rule may benefit from a top-down mechanism in which VTC starts with the necessary structure to distinguish categories in general, approximately identifying what is seen. This response could then train connections to low-level regions, learning the detailed visual features necessary for understanding. Computational experiments simulating the genomic bottleneck have shown that simple rules built into an ANN enable basic recognition without visual experience, demonstrating this mechanism's potential (Koulakov et al., 2022). Beyond vision, cross-modal input could also contribute. Indeed, the long-range connectivity of future category regions has been found from early infancy (Cabral et al., 2022). A mechanism of innate responsiveness to categories combined with sufficient flexibility for

learning is also consistent with the emergence of distributed representations in the ventral stream of children before category-selective focal regions (Cohen et al., 2019), and is parsimonious with findings of highly plastic face-selective regions in ventral cortex of congenitally blind individuals (Ratan Murty et al., 2020; van den Hurk et al., 2017).

Our unique dataset allowed DNN models of ventral cortex to be applied to infants. Even more than for adult vision research, DNNs offer the potential for novel insight into visual development. The parallel between infant learning and machine learning is of increasing interest for researchers in both fields (Smith & Slone, 2017; Zaadnoordijk et al., 2022) with recent work showing that video data from an infant's perspective provides the necessary structure to learn word-visual referents (Vong et al., 2024). Our work brings this approach a step earlier in the developmental process, using DNNs to characterise the visual representations to which words are later attached. By incorporating rich awake infant neuroimaging paradigms into our toolkit, we unlock the substantial potential of computational modelling to understand the learning mechanisms implemented in the developing brain, allowing us to gain further insight of the first year of life as a time of intricate development and complex brain function.

Chapter 5 General Discussion

"I'll know it when I see it" – it's a common adage, to which we may not pay much attention. But how does this intrinsic link between vision and knowledge come to be? Many of our earliest milestones are defined by learning to recognise and name the things we see, and some of the most impressive feats of AI in recent years come from computer vision systems that convey recognition and understanding. In this thesis, I explored how visual perception bridges towards conceptual understanding in minds and machines. Focusing on development and the processes that guide visual representation learning in the brain and DNNs, I find that extended temporal statistics in the environment provide signals from which a system can learn relational structure in a self-supervised manner. Longer temporal signals are informative for contextual information, and both brains and neural network models have the capacity to learn global information from these temporal correspondences which are associated with the intrinsic neural activity of a given region. Looking towards the beginnings of visual representations in the brain, I find that rich category knowledge is present from a very young age which suggests a top-down developmental mechanism for category knowledge. In sum, the results presented here reveal a complex interplay between bottom-up, sensory driven inputs and innate activity within the brain for learning of visual concepts; more generally, the results presented in this thesis provide an important step forward for the emerging field of developmental cognitive computational neuroscience.

Although purely perceptual input is not sufficient for explaining visual representations (Chapter 3, 4), the statistical signals present in the environment do provide essential clues for learning of meaningful relational structure (Chapter 2). I propose that a foundational capacity for object recognition is crucial for intelligent systems to develop conceptual

understanding from experience, and that this understanding emerges from the rich statistics of object and contextual associations in the visual environment.

5.1 Thinking about early human environments to improve understanding in DNNs

The approach presented in Chapter 2 was developed by thinking about what current computer vision models might be lacking, given the huge disparity between how they are trained compared to early human experience. During the first year of life we are exposed to huge amounts of rich input through our experiences – our family homes, new toys and shows, our siblings or family pets, outings to the park or supermarket. It's only in the latter half of the first year, when caregivers adapt their labelling to match their infants' emerging communication skills (Poulin-Dubois et al., 1995) when we have access to explicit labels for the objects with which we are interacting. Prior to this, our learning is driven mostly through observation as the extent to which we can interact with and physically manipulate objects is limited. In contrast, a lot of success in computer vision has come from the advent of large labelled datasets such as ImageNet for object recognition (Deng et al., 2009), MSCOCO for scene captioning (Lin et al., 2014) or Kinetics for action recognition (Kay et al., 2017), although in recent years there has been a shift towards self-supervised methods such as transformers (Liu et al., 2021). DNNs can learn robust embeddings for object recognition by providing them with extensive training images across a broad distribution of categories and classes alongside their associated labels. Networks are trained from randomly initiated weights, and although the weight initiation can be important for the kinds of representation that are ultimately learned (Mehrer et al., 2020) this initial starting point is largely ignored.

This agnosticism towards the starting point for visual learning contrasts starkly with how humans learn during early infancy. We begin with only sensory inputs, and labels come later. Moreover, the distribution of classes that we observe are more limited in scope as results from infant headcam studies show that we tend to view a few objects very often, from many different viewpoints (Smith et al., 2018). As such, we must rely on hidden patterns in the environment which guide our learning through self-supervised statistical learning processes (Saffran et al., 1996). In Chapter 2, I highlighted one such signal – extended temporal co-occurrences between objects and contexts – and demonstrated the potential learning value from this for a DNN. As discussed, I argue that a disregard for these long-range temporal signals precludes traditional computer vision models from learning global relational structure and contextual grounding that is important for broad semantic understanding. If we care about building AI that has true understanding of the things it appears to know, then this kind of structure must be taken into account.

Categorisation and the ability to group individual objects into meaningful and useful concepts is a core cognitive function, but each group is meaningless without its relation to others. Humans can perceive relational information, even in the absence of direct perceptual input to the retina (Hafri & Firestone, 2021) and infants as young as 5-months-old have the capacity to distinguish object relations such as one object supporting another, or fitting into a container tightly or loosely (Hespos & Spelke, 2004). By learning the structure of how objects relate – at a feature, category and global level – both minds and machines can develop robust concepts, which can even be aligned across modalities (Roads & Love, 2020). I demonstrated in Chapter 2 that DNNs trained on data and learning objectives that are very different from naturalistic visual experience fail to capture this global relational structure that is important for semantic representation learning.

A handful of researchers have been using insights from infant behavioural research to explore if there is something unique to infant experience that could be useful for AI, with interest picking up in recent years. In parallel, there has been another revolution in the AI space with large language models (LLMs) becoming state-of-the-art for many tasks. Although an LLM such as ChatGPT appears to have advanced knowledge and understanding for many tasks, there is a huge data gap between its training and a child learning to speak (Frank, 2023). Humans are much more data efficient than an LLM, which receive four to five orders of magnitude more data than a child and consume thousands of time more energy than a baby in its first year (Cusack et al., 2024). One possible reason for this may be in the fact infants interact with and manipulate the things they see to generate the optimal views on their own "training data" to maximise their learning (Smith et al., 2018).

Research that takes this infant approach to developing machine learning is, in itself, still in its early days. There is promising evidence, however, that the visual and linguistic experience of a young child can be used to train large transformer models approaching the bound of performance with a model trained on the more typical hugely scaled datasets (Vong et al., 2024), even in the absence of further inductive biases in the models that may help representation learning (Orhan et al., 2020). Contrary to what might be expected from infants' strong capacity for generalisation these authors find that transfer learning performance of the network trained on infant data was not better than when trained with a standard computer vision dataset, as may have been expected from previous training with infant egocentric views (Bambach et al., 2018). Regardless, the benefit conferred by naturalistic early experience for vision transformers (ViT) can also be seen by comparing a ViT and newborn chicks in tightly controlled rearing environments (Pandey et al., 2023).

The simulated experience of a chick's view of a single object was sufficient for the model to learn object representations. Interestingly, most of the above work makes use of contrastive learning with some sort of temporal objective as was introduced in Chapter 2. However, these studies do not leverage the extended temporal dynamics that I propose are important for grounding object knowledge within wider contextual and semantic understanding. Perhaps a combination of training with infant ego-centric data at this temporally extended scale could shed further light on how to embed deeper conceptual understanding into the next generation of AI.

5.2 Using infant-inspired DNNs as a useful model for brain function

The finding that semantic structure was learned by a DNN when given access to visual information across an extended temporal windows informed our thinking about how this may manifest in the brain. Considering the hierarchy along the VVS from low-level to high-level information, I hypothesised that the perceptual similarity dominating the model's representation learning at short time intervals would overshadow object responses in anterior VVS. This was conceived specifically for when objects are presented against a background of perceptual noise. Moreover, I considered a potential underlying mechanism of this response by examining INTs measured at rest and how they relate to object processing in a repetition based paradigm. Contrary to the predictions that emerged from Chapter 2, I find that the sign of repetition responses – whether suppression or enhancement – is associated with how long a given region within VOTC remains similar to itself. More broadly, the approach taken across Chapters 2 and 3 highlight the complementary nature of using DNNs for neuroscience research. The developmentally inspired model in Chapter 2 resulted in a testable hypothesis that informed the design of the fMRI study in Chapter 3, in which we

asked if the hierarchy in feature complexity along the VVS reflects the different perceptual and semantic perceptual signals found in the computational work.

The role of temporal signals for neural processing has previously been researched, especially in the event segmentation or language domain. For example, temporal event structure at various timescales lead to different patterns of activity across the brain, demonstrating that there are temporal receptive fields similar to spatial ones with the cortex (Hasson et al., 2008), and a similar hierarchy existing for temporal receptive windows in a narrated story (Lerner et al., 2011). However, typically the role of temporal processing for discrete events is explored in terms of episodic memory in the medial temporal lobe (Schapiro et al., 2012), with a special role of the hippocampus for learning sequences of events and therefore encoding conceptually similar events (Ranganath & Hsieh, 2016).

DNNs have been credited as the current best candidate models that we have for biological vision, and testing the predictions of these artificial neural networks (ANN) for the brain has resulted in what is essentially the birth of a new field (Kubilius et al., 2019; Mehrer et al., 2021; Richards et al., 2019; Storrs & Kriegeskorte, 2019; Zhuang et al., 2021). Despite this success, the value of this prediction based modelling and the discrepancies between DNNs and biological vision has been called into question (Bowers et al., 2023). For example, Bowers and colleagues point out the limits of correlational methods, the sensitivity of these networks to adversarial attack or their insensitivity to global shape and configural processing. Indeed, it has been shown that huge differences in architectures and learning objectives lead to very similar brain predictivity scores in a large-scale survey of candidate vision models (Conwell et al., 2023) and that it is the data on which the model is trained, rather, that is important for capturing brain representations. By connecting the approach in

Chapters 2 and 3, I believe that I have demonstrated the value of DNNs as models of human vision outside of these limitations. Instead, I believe there is great value in using these computational models as a means to test new theories of learning, providing inspiration for hypotheses on how this may manifest in the brain and testing this with empirical work.

This approach has been proposed in the neuroconnectionist research programme (Doerig et al., 2023) which formalises ANNs as computational models that operate in a sweet spot for understanding the brain. They are complex enough to show a degree of biological plausibility and engage with cognitive tasks, while still being simple enough for algorithmic interpretability. In this way, the research programme is highly progressive for generating new theoretical insight from the computational work leading to novel predictions that can be tested with human experiments. Throughout the work in this thesis, I found this to be a natural and informative approach, especially when considering complex ideas such as conceptual understanding, meaning and the nature of early infancy.

5.3 What are the developmental foundations of visual understanding?

Using novel methods in awake infant fMRI and deep neural network modelling, the results in Chapter 4 demonstrate that visual representations to a diverse set of categories are present from 2-months-old. This provides important insight into the neural foundation of the recognition behaviour demonstrated in both the DNN (Chapter 2) and adult cohort (Chapter 3). Infant VTC is not a blank slate from which all visual representations develop through experience, but begins with an innate capacity for recognition of objects at the category level. This stands in contrast to the purely bottom-up learning of a DNN learning from naturalistic visual inputs, even if infant-inspired as in Chapter 2, and is something to

consider if we wish to gain mechanistic understanding of how INTs interact with object responses to facilitate recognition (Chapter 3).

This finding of an innate category template in VTC contributes to the age-old debate of nature vs. nurture. While devout empiricists such as Locke would argue that infants develop from tabula rasa, there is ample evidence that demonstrates the capacity of very young infants to engage with cognitive functions and categorisation (Quinn et al., 2001; Spelke & Kinzler, 2007) and neural signatures of category selectivity are found in infant VTC (Kosakowski et al., 2022). The general consensus among developmental scientists is that both innate and learned components play a role in cognitive development, but the exact components of each is still contested (Smith, 1999; Spelke, 1994). Our results add to this landscape, demonstrating a role for innate category structure which develops throughout visual cortex in a top-down manner. As discussed in Chapter 4, this correspondence is at a time when infants' prior visual acuity has been quite poor and the distributional statistics of their environment has been limited. One may argue that a more controlled test of the role for visual experience for early-emerging category representations would be to record awake functional activity in newborns. However, neonates limited visual capacity and the impoverished structure of the visual environment in the first two months of life is unlikely to facilitate category learning for the rich stimulus set tested in this study.

The presence of category representations in high-level visual cortex so early on in development raises an interesting question about the optimal starting point for a system trying to learn concepts from perceptual input. Infants are born with many inductive biases for things like biological motion (Simion et al., 2011) or towards feature configurations that are informative for faces (Johnson et al., 1991) and the inclusion of these inductive biases

into computational models has been proposed to confer a possible advantage for more intelligent machine learning (Zaadnoordijk et al., 2022). By studying the developmental foundations of VTC category responses we can better understand the starting point for the system, and how this may facilitate future conceptual understanding which may be useful if considered when building AI. The developmental theory described in Chapter 4 posits that initial category responsiveness in high-level vision is used to train the category knowledge across visual cortex in a top-down manner through experience, but only after an initial strong start. This aligns with a formulation of VTC as a continuous semantic space (Huth et al., 2012) that encodes behaviourally relevant feature dimensions across the cortex. In this way, category specific responses in regions such as FFA emerge as special cases of sparse tuning within a wider organising principle that is defined by broader semantic categories that define similarity ratings (Contier et al., 2024). Our results suggest that this organising principle begins with category representations in anterior VTC and is then propagated back down the visual hierarchy through experience.

How might the signals in perceptual input interact with neural activity to facilitate this learning mechanism? I find that training a DNN on temporally extended signals leads to a contextual representation being learned, and in Chapter 3 that longer timescales the regions responsible for encoding contexts from objects activate at long stimulus presentation intervals (MacEvoy & Epstein, 2011). Taken together, this suggests a unique role of temporally extended events for learning object representations alongside their associated context, forming a more holistic concept. This is interesting taken in tandem with findings that infants have longer INTs than adults (Truzzi & Cusack, 2023) and segment continuous experience into longer windows or events (Yates et al., 2022).

In general, infants seem to be existing at a slower timescale – both neurally and in the statistics of their visual environment. The findings from Chapters 2 and 3 suggest that extended temporal dynamics are important for learning the meaning of objects in a broader context. For example, the model trained on two images separated by 60 s learned a representation that balanced objects and contexts (Fig. 2.4b) and LOC (which can construct scenes from objects [MacEvoy & Epstein, 2011]) exhibited enhancement to slowly repeating objects. Crucially, this was accompanied by slower INTs in the regions that integrate objects and contexts (Fig. 3.5). Could the slowness of the underlying neural dynamics be a mechanism for learning the contextual meaning of the things we see? If infants are slower in general, what does that mean for development as we learn to navigate the world?

The infant brain already has a distinct capacity to respond to visual categories, but the slower dynamics of the system is suggestive of global associative learning. Perhaps it is this information that cannot be innately specified – that must be developed top-down - and is learned through an extended developmental period within the individual's environment. I propose that we begin with a capacity to respond to objects in general, but the fine-grained semantics of what these are and the social or cultural context in which we find them is something that is better learned from individual experience. It's possible that this is enacted computationally within the VVS through extended intrinsic timescales, or could equally be attributed to another computational mechanism. The process itself is akin to language beginning with phonetic building blocks (Werker, 2024), or no preferences for faces of a particular ethnic background in the early years (McKone et al., 2012). It is only through experience that the most contextually relevant preferences for our individual experience develop and we learn the specifics of which perceptual features are most important for the concepts that we need for our own lives.

5.4 Infant-inspired AI, what can developmental neuroscience give us?

The work presented in this thesis brings together developmental science, neuroimaging and machine learning research. In a complementary process of discovery, I used prior infant literature to inspire a new deep learning model which, in turn, led to a testable hypothesis for brain function. Awake functional MRI and DNNs were used to explore the earliest beginnings of visual concepts in the brain, with theoretical discussion of how the findings from the computational and adult work can inform our understanding of early brain function. To close out this generative and complementary process, I propose that further efforts in infant developmental neuroscience can inform the next generation of AI.

The current trend in AI is to focus on scaling, with most of those who are building these systems turning away from algorithmic innovation and instead focusing on datasets of larger and larger size (Hoffmann et al., 2022; Orhan, 2023). This is working, in the sense that we are seeing improvements in the end product. However, I believe there is a disregard for the sustainability of this approach. As discussed above, a human learner uses much less data and much less energy than a foundation model (Cusack et al., 2024) and does not require the entirety of the internet, including proprietary data, to learn complex concepts. The current AI hype is sure to lose ground when the legal, social, environmental and economic impacts of training at this scale become even more and more evident.

The slower pace of science in the field of NeuroAI and computational cognitive neuroscience offer an alternative path to sustainable innovation, and I believe that furthering our understanding of the initial foundations of brain development will provide inspiration for smarter, better, faster AI. This does not have to come in the form of direct biological plausibility, or fitting to neural data which has not been hugely successful or informative in

machine learning. Rather, I believe that a complementary research programme in which we think about the emergence of intelligence in general, using DNNs to study the brain and using insights about the brain to inspire model development, is a fruitful and exciting path forward. I have already referenced many such pioneering works where this approach has been explored and has shown success (Aubret et al., 2022; Cusack et al., 2024; Doerig et al., 2023; Smith & Slone, 2017; Vong et al., 2024; Zaadnoordijk et al., 2022).

To provide a more concrete example, we may consider how to improve the model presented in Chapter 2 given the findings in subsequent chapters. Evidently, infants and DNNs begin learning from very different foundations. While the models are trained from random weights, infants in fact have category structure in visual cortex from a young age. Interestingly, we decided to implement the models presented in Chapter 2 as a fine tuning procedure on top of ImageNet trained weights. This was a decision made for purely computational reasons, but hints further at the importance of beginning with a capacity for object recognition when trying to learn meaningful representations. Going forward, I would explore how to align the models' initialisation with the structure in infant VTC to find an optimal starting point. Further, these models don't have the temporal structure that I propose is important for learning objects in context. By introducing recurrent temporal dynamics into the model we may further improve its ability to learn semantic structure from the movie dataset, and may even boost object recognition performance and alignment with human visual cortex (Kietzmann et al., 2019).

5.5 Limitations

One may ask, what use is there in using one highly complex system – the brain – to understand another complex system – DNNs, neither of which we fully understand. This

criticism is valid, but depends on the level at which we are interested in exploring these questions. For example, none of the work presented here considers the models and the brain to be working on the same computational processes. Rather, I have used them both as a system to explore learning from visual experience and the kinds of representations that can be developed from this. This can tell us about the structure of knowledge and concepts within biological vision, but of course we cannot say that human vision came to these concepts in the exact same way as the artificial network. Nonetheless, I have found this methodological approach to be highly progressive in informing my understanding and exploration of perception, categories, concepts and knowledge more generally. Without the tools used in these studies, thinking about the developmental mechanism of visual representations in a top-down manner would not have been possible. Without thinking about the structure of infants' early experience, I would not have been able to find faults in CNN models of vision. Now that I see differences in how the visual system begins compared to the tabula rasa nature of DNNs, I see that more efficiency could possibly be found by finding a better starting point for the models. Perhaps, like infants, beginning with the necessary structure for category responses in general and then learning from temporal signals in the environment would provide the necessary information to learn more robust, contextualised concepts from perceptual input.

The work presented here of course has its limitations. As briefly addressed in Chapter 1, I do not consider cases where understanding emerges in the absence of perceptual input to the retina. Those who are born without sight of course go on to develop conceptual knowledge, although it may be driven by another modality. For example, face selectivity is still observed in the fusiform gyrus of congenitally blind individuals, instead being driven by auditory input (Mahon et al., 2009; van den Hurk et al., 2017). Such cases were not considered here,

but do align with a theory of concept development which begins with the capacity for categorical distinction and then is tailored towards individual experience based on sensory input. Similarly, the role of multimodal inputs is not discussed. This is a very important distinction between the brain and computational models of vision. The ventral visual stream (VVS) is highly interconnected to the rest of the brain, with the long-range connectivity of future category selective regions being present in neonates (Cabral et al., 2022). This cannot be compared to current DNNs for vision, despite recent advances in multimodal AI (Girdhar et al., 2023). Indeed, machine learning progresses rapidly and the models presented in Chapter 2 of this thesis are no longer state-of-the-art. Future development of the infant-inspired model would benefit from expansion to a more updated architecture, and it would be interesting to compare the learned representations to infant and adult visual cortex.

5.6 Conclusion

Using methods in machine learning, neuroimaging and developmental science I have explored the question of how visual perception emerges in conceptual understanding. Using an interdisciplinary approach, I asked how infant-inspired training can allow a DNN to learn broader understanding of its inputs. I found that, depending on the timescale of perceptual input, either contexts or objects can be learned. Having access to images which are temporally associated over an extended period of time allowed a DNN to learn a more semantically meaningful embedding. Following from this, I tested the temporal tuning of VVS in human adults to objects, relative to background noise, over varying input timescales. I find that, rather than being specifically tuned towards short or long lagged objects, the pattern of repetition suppression or enhancement in VVS is determined by a region's underlying intrinsic neural activity. I discuss the potential implications of this for neural theories of repetition effects. Finally, I presented novel findings that explore the earliest

beginnings of visual recognition, finding that infants begin life with rich category responses in VTC. These become refined through experience, and semantic details develop top-down along the visual hierarchy. Together, the results of this thesis provide novel methodological advances in the field of developmental cognitive computational neuroscience, and shed light on the potential mechanisms of concept development in minds and machines alike.

Bibliography

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., & Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, *8*. https://doi.org/10.3389/fninf.2014.00014

Andersson, J. L. R., Skare, S., & Ashburner, J. (2003). How to correct susceptibility distortions in spin-echo echo-planar images: Application to diffusion tensor imaging. *NeuroImage*, *20*(2), 870–888. https://doi.org/10.1016/S1053-8119(03)00336-7

Arcaro, M. J., & Livingstone, M. S. (2017). A hierarchical, retinotopic proto-organization of the primate visual system at birth. *eLife*, *6*, e26196. https://doi.org/10.7554/eLife.26196

Arichi, T., Fagiolo, G., Varela, M., Melendez-Calderon, A., Allievi, A., Merchant, N., Tusor, N., Counsell, S. J., Burdet, E., Beckmann, C. F., & Edwards, A. D. (2012). Development of BOLD signal hemodynamic responses in the human brain. *NeuroImage*, *63*(2), 663–673. https://doi.org/10.1016/j.neuroimage.2012.06.054

Aubret, A., Ernst, M., Teulière, C., & Triesch, J. (2022). *Time to augment self-supervised visual representation learning* (arXiv:2207.13492). arXiv. https://doi.org/10.48550/arXiv.2207.13492

Avants, B. B., Tustison, N. J., Wu, J., Cook, P. A., & Gee, J. C. (2011). An Open Source Multivariate Framework for n-Tissue Segmentation with Evaluation on Public Data. *Neuroinformatics*, *9*(4), 381–400. https://doi.org/10.1007/s12021-011-9109-y

Baker, B., Lange, R., Achille, A., Cao, R., Kriegeskorte, N., Schwartz, O., & Pitkow, X. (2021). *What makes representations "useful"?* https://openreview.net/forum?id=HHk5tZ-Pgzk

Bambach, S., Crandall, D., Smith, L. B., & Yu, C. (2018). Toddler-Inspired Visual Object Learning. *Advances in Neural Information Processing Systems*, *31*. https://proceedings.neurips.cc/paper_files/paper/2018/hash/48ab2f9b45957ab574cf005eb8a76760-Abstract.html

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmid, A. M., Dale, A. M., Hämäläinen, M. S., Marinkovic, K., Schacter, D. L., Rosen, B. R., & Halgren, E. (2006). Top-down facilitation of visual recognition. *Proceedings of the National Academy of Sciences*, *103*(2), 449–454. https://doi.org/10.1073/pnas.0507062103

Barsalou, L. W. (2008). Grounded Cognition. *Annual Review of Psychology*, *59*(1), 617–645. https://doi.org/10.1146/annurev.psych.59.103006.093639

Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage*, *37*(1), 90–101. https://doi.org/10.1016/j.neuroimage.2007.04.042

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, *109*(9), 3253–3258. https://doi.org/10.1073/pnas.1113380109

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, *14*(2), 143–177. https://doi.org/10.1016/0010-0285(82)90007-X

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., Puebla, G., Adolfi, F., Hummel, J. E., Heaton, R. F., Evans, B. D., Mitchell, J., & Blything, R. (2023). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, *46*, e385. https://doi.org/10.1017/S0140525X22002813

Brendel, W., & Bethge, M. (2019). *Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet* (arXiv:1904.00760). arXiv. https://doi.org/10.48550/arXiv.1904.00760

Brendel, W., Rauber, J., & Bethge, M. (2018). Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models. *International Conference on Learning Representations*. https://arxiv.org/abs/1712.04248

Cabral, L., Zubiaurre-Elorza, L., Wild, C. J., Linke, A., & Cusack, R. (2022). Anatomical correlates of category-selective visual regions have distinctive signatures of connectivity in neonates. *Developmental Cognitive Neuroscience*, *58*, 101179. https://doi.org/10.1016/j.dcn.2022.101179

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). https://doi.org/10.1371/journal.pcbi.1003963

Carey, S. (2004). Bootstrapping & the Origin of Concepts. *Daedalus*, *133*(1), 59–68.

Carey, S. (2009). *The origin of concepts* (pp. viii, 598). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195367638.001.0001

Charest, I., Kievit, R. A., Schmitz, T. W., Deca, D., & Kriegeskorte, N. (2014). Unique semantic space in the brain of each beholder predicts perceived similarity. *Proceedings of the National Academy of Sciences*, *111*(40), 14565–14570. https://doi.org/10.1073/pnas.1402594111

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A Simple Framework for Contrastive Learning of Visual Representations. *Proceedings of the 37th International Conference on Machine Learning*, 1597–1607. https://proceedings.mlr.press/v119/chen20j.html

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), Article 1. https://doi.org/10.1038/srep27755

Ciric, R., Wolf, D. H., Power, J. D., Roalf, D. R., Baum, G. L., Ruparel, K., Shinohara, R. T., Elliott, M. A., Eickhoff, S. B., Davatzikos, C., Gur, R. C., Gur, R. E., Bassett, D. S., & Satterthwaite, T. D. (2017). Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. *NeuroImage*, *154*, 174–187. https://doi.org/10.1016/j.neuroimage.2017.03.020

Clarke, A., Taylor, K. I., Devereux, B., Randall, B., & Tyler, L. K. (2013). From Perception to Conception: How Meaningful Objects Are Processed over Time. *Cerebral Cortex*, *23*(1), 187–197. https://doi.org/10.1093/cercor/bhs002

Clerkin, E. M., Hart, E., Rehg, J. M., Yu, C., & Smith, L. B. (2017). Real-world visual statistics and infants' first-learned object names. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), 20160055. https://doi.org/10.1098/rstb.2016.0055

Cohen, M. A., Dilks, D. D., Koldewyn, K., Weigelt, S., Feather, J., Kell, A. JE., Keil, B., Fischl, B., Zöllei, L., Wald, L., Saxe, R., & Kanwisher, N. (2019). Representational similarity precedes category selectivity in the developing ventral visual pathway. *NeuroImage*, *197*, 565–574. https://doi.org/10.1016/j.neuroimage.2019.05.010

Contier, O., Baker, C. I., & Hebart, M. N. (2024). Distributed representations of behaviour-derived object dimensions in the human visual system. *Nature Human Behaviour*, 1–15. https://doi.org/10.1038/s41562-024-01980-y

Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., & Konkle, T. (2023). *What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines?* (p. 2022.03.28.485868). bioRxiv. https://doi.org/10.1101/2022.03.28.485868

Cox, R. W. (1996). AFNI: Software for Analysis and Visualization of Functional Magnetic Resonance Neuroimages. *Computers and Biomedical Research*, *29*(3), 162–173. https://doi.org/10.1006/cbmr.1996.0014

Cusack, R., Ranzato, M., & Charvet, C. J. (2024). Helpless infants are learning a foundation model. *Trends in Cognitive Sciences*, *0*(0). https://doi.org/10.1016/j.tics.2024.05.001

Czarnecki, K., Duffy, J. R., Nehl, C. R., Cross, S. A., Molano, J. R., Jack, C. R., Jr, Shiung, M. M., Josephs, K. A., & Boeve, B. F. (2008). Very Early Semantic Dementia With Progressive Temporal Lobe Atrophy: An 8-Year Longitudinal Study. *Archives of Neurology*, *65*(12), 1659–1663. https://doi.org/10.1001/archneurol.2008.507

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical Surface-Based Analysis: I. Segmentation and Surface Reconstruction. *NeuroImage*, *9*(2), 179–194. https://doi.org/10.1006/nimg.1998.0395

Deen, B., Richardson, H., Dilks, D. D., Takahashi, A., Keil, B., Wald, L. L., Kanwisher, N., & Saxe, R. (2017). Organization of high-level visual cortex in human infants. *Nature Communications*, *8*(1), Article 1. https://doi.org/10.1038/ncomms13995

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (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

Dobson, V., & Teller, D. Y. (1978). Visual acuity in human infants: A review and comparison of behavioral and electrophysiological studies. *Vision Research*, *18*(11), 1469–1483. https://doi.org/10.1016/0042-6989(78)90001-9

Doerig, A., Sommers, R. P., Seeliger, K., Richards, B., Ismael, J., Lindsay, G. W., Kording, K. P., Konkle, T., van Gerven, M. A. J., Kriegeskorte, N., & Kietzmann, T. C. (2023). The neuroconnectionist research programme. *Nature Reviews Neuroscience*, *24*(7), 431–450. https://doi.org/10.1038/s41583-023-00705-w

Downing, P. E., Peelen, M. V., Wiggett, A. J., & Tew, B. D. (2006). The role of the extrastriate body area in action perception. *Social Neuroscience*, *1*(1), 52–62. https://doi.org/10.1080/17470910600668854

Eger, E., Henson, R., Driver, J., & Dolan, R. J. (2004). BOLD Repetition Decreases in Object-Responsive Ventral Visual Areas Depend on Spatial Attention. *Journal of Neurophysiology*, *92*(2), 1241–1247. https://doi.org/10.1152/jn.00206.2004

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

Ellis, C. T., Skalaban, L. J., Yates, T. S., Bejjanki, V. R., Córdova, N. I., & Turk-Browne, N. B. (2020). Re-imagining fMRI for awake behaving infants. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-18286-y

Ellis, C. T., Yates, T. S., Skalaban, L. J., Bejjanki, V. R., Arcaro, M. J., & Turk-Browne, N. B. (2021). Retinotopic organization of visual cortex in human infants. *Neuron*, *109*(16), 2616-2626.e6. https://doi.org/10.1016/j.neuron.2021.06.004

Epstein, R., Harris, A., Stanley, D., & Kanwisher, N. (1999). The parahippocampal place area: Recognition, navigation, or encoding? *Neuron*, *23*(1), 115–125. https://doi.org/10.1016/s0896-6273(00)80758-8

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., & Gorgolewski, K. J. (2019). fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1), Article 1. https://doi.org/10.1038/s41592-018-0235-4

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex (New York, N.Y.*, *1*(1), 1–47. https://doi.org/10.1093/cercor/1.1.1-a

Fiser, J., & Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences*, *99*(24), 15822–15826. https://doi.org/10.1073/pnas.232472899

Fonov, V., Evans, A. C., Botteron, K., Almli, C. R., McKinstry, R. C., & Collins, D. L. (2011). Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage*, *54*(1), 313–327. https://doi.org/10.1016/j.neuroimage.2010.07.033

Fonov, V., Evans, A., McKinstry, R., Almli, C., & Collins, D. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, *47*, S102. https://doi.org/10.1016/S1053-8119(09)70884-5

Frank, M. C. (2023). Bridging the data gap between children and large language models. *Trends in Cognitive Sciences*, *27*(11), 990–992. https://doi.org/10.1016/j.tics.2023.08.007

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*, *44*(3), 677–694. https://doi.org/10.1017/S0305000916000209

Friston, K. J., Jezzard, P., & Turner, R. (1994). Analysis of functional MRI time-series. *Human Brain Mapping*, *1*(2), 153–171. https://doi.org/10.1002/hbm.460010207

Fukushima, K. (1988). Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural Networks*, *1*(2), 119–130. https://doi.org/10.1016/0893-6080(88)90014-7

Gaier, A., & Ha, D. (2019). Weight Agnostic Neural Networks. *Advances in Neural Information Processing Systems*, *32*. https://proceedings.neurips.cc/paper/2019/hash/e98741479a7b998f88b8f8c9f0b6b6f1-Abstract.html

Gauthier, B., Eger, E., Hesselmann, G., Giraud, A.-L., & Kleinschmidt, A. (2012). Temporal Tuning Properties along the Human Ventral Visual Stream. *Journal of Neuroscience*, *32*(41), 14433–14441. https://doi.org/10.1523/JNEUROSCI.2467-12.2012

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, *2*(11), 665–673. https://doi.org/10.1038/s42256-020-00257-z

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2022). *ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness* (arXiv:1811.12231). arXiv. https://doi.org/10.48550/arXiv.1811.12231

Geirhos, R., Temme, C. R. M., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. *Advances in Neural Information Processing Systems*, *31*. https://proceedings.neurips.cc/paper_files/paper/2018/hash/0937fb5864ed06ffb59ae5f9b5ed67a9-Abstract.html

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., & Misra, I. (2023). *ImageBind: One Embedding Space To Bind Them All*. 15180–15190. https://openaccess.thecvf.com/content/CVPR2023/html/Girdhar_ImageBind_One_Embedding_Space_To_Bind_Them_All_CVPR_2023_paper.html

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C. F., Jenkinson, M., Smith, S. M., & Van Essen, D. C. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, *536*(7615), 171–178. https://doi.org/10.1038/nature18933

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 249–256. https://proceedings.mlr.press/v9/glorot10a.html

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, *15*(1), 20–25. https://doi.org/10.1016/0166-2236(92)90344-8

Goren, C. C., Sarty, M., & Wu, P. Y. K. (1975). Visual Following and Pattern Discrimination of Face-like Stimuli by Newborn Infants. *Pediatrics*, *56*(4), 544–549. https://doi.org/10.1542/peds.56.4.544

Gorgolewski, K., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., Flandin, G., Ghosh, S. S., Glatard, T., Halchenko, Y. O., Handwerker, D. A., Hanke, M., Keator, D., Li, X., Michael, Z., Maumet, C., Nichols, B. N., Nichols, T. E., Pellman, J., … Poldrack, R. A. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*, *3*(1), Article 1. https://doi.org/10.1038/sdata.2016.44

Gorgolewski, K., Burns, C., Madison, C., Clark, D., Halchenko, Y., Waskom, M., & Ghosh, S. (2011). Nipype: A Flexible, Lightweight and Extensible Neuroimaging Data Processing Framework in Python. *Frontiers in Neuroinformatics*, *5*. https://www.frontiersin.org/articles/10.3389/fninf.2011.00013

Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, *48*(1), 63–72. https://doi.org/10.1016/j.neuroimage.2009.06.060

Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: Neural models of stimulus-specific effects. *Trends in Cognitive Sciences*, *10*(1), 14–23. https://doi.org/10.1016/j.tics.2005.11.006

Grill-Spector, K., & Kanwisher, N. (2005). Visual Recognition: As Soon as You Know It Is There, You Know What It Is. *Psychological Science*, *16*(2), 152–160. https://doi.org/10.1111/j.0956-7976.2005.00796.x

Grill-Spector, K., Kushnir, T., Hendler, T., & Malach, R. (2000). The dynamics of object-selective activation correlate with recognition performance in humans. *Nature Neuroscience*, *3*(8), 837–843. https://doi.org/10.1038/77754

Güçlü, U., & Gerven, M. A. J. van. (2015). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *Journal of Neuroscience*, *35*(27), 10005–10014. https://doi.org/10.1523/JNEUROSCI.5023-14.2015

Gutmann, M., & Hyvärinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 297–304. https://proceedings.mlr.press/v9/gutmann10a.html

Hafri, A., & Firestone, C. (2021). The Perception of Relations. *Trends in Cognitive Sciences*, *25*(6), 475–492. https://doi.org/10.1016/j.tics.2021.01.006

Halchenko, Y., Goncalves, M., Castello, M. V. di O., Ghosh, S., Salo, T., Hanke, M., Velasco, P., Dae, Kent, J., Brett, M., Amlien, I., Gorgolewski, K., Lukas, D. C., Markiewicz, C., Tilley, S., Kaczmarzyk, J., Stadler, J., Kim, S., Kahn, A., … Meyer, K. (2021). *Nipy/heudiconv:* (Version v0.10.0) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.5557588

Hasson, U., Levy, I., Behrmann, M., Hendler, T., & Malach, R. (2002). Eccentricity Bias as an Organizing Principle for Human High-Order Object Areas. *Neuron*, *34*(3), 479–490. https://doi.org/10.1016/S0896-6273(02)00662-1

Hasson, U., Yang, E., Vallines, I., Heeger, D. J., & Rubin, N. (2008). A Hierarchy of Temporal Receptive Windows in Human Cortex. *Journal of Neuroscience*, *28*(10), 2539–2550. https://doi.org/10.1523/JNEUROSCI.5487-07.2008

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 (New York, N.Y.)*, *293*(5539), 2425–2430. https://doi.org/10.1126/science.1063736

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification*. 1026–1034. https://openaccess.thecvf.com/content_iccv_2015/html/He_Delving_Deep_into_ICCV_2015_paper.html

Henson, R. (2003). Neuroimaging studies of priming. *Progress in Neurobiology*, *70*(1), 53–81. https://doi.org/10.1016/S0301-0082(03)00086-8

Henson, R., Price, C. J., Rugg, M. D., Turner, R., & Friston, K. J. (2002). Detecting Latency Differences in Event-Related BOLD Responses: Application to Words versus Nonwords and Initial versus Repeated Face Presentations. *NeuroImage*, *15*(1), 83–97. https://doi.org/10.1006/nimg.2001.0940

Henson, R., Shallice, T., & Dolan, R. (2000). Neuroimaging Evidence for Dissociable Forms of Repetition Priming. *Science*, *287*(5456), 1269–1272. https://doi.org/10.1126/science.287.5456.1269

Hespos, S. J., & Spelke, E. S. (2004). Conceptual precursors to language. *Nature*, *430*(6998), 453–456. https://doi.org/10.1038/nature02634

Ho, J. W.-T., Narduzzo, K. E., Outram, A., Tinsley, C. J., Henley, J. M., Warburton, E. C., & Brown, M. W. (2011). Contributions of area Te2 to rat recognition memory. *Learning & Memory*, *18*(7), 493–501. https://doi.org/10.1101/lm.2167511

Hodges, J. R., & Patterson, K. (2007). Semantic dementia: A unique clinicopathological syndrome. In *Lancet Neurology* (Vol. 6, Issue 11, pp. 1004–1014). Elsevier. https://doi.org/10.1016/S1474-4422(07)70266-1

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. de L., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., Driessche, G. van den, Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., … Sifre, L. (2022). *Training Compute-Optimal Large Language Models* (arXiv:2203.15556). arXiv. https://doi.org/10.48550/arXiv.2203.15556

Howell, B. R., Styner, M. A., Gao, W., Yap, P.-T., Wang, L., Baluyot, K., Yacoub, E., Chen, G., Potts, T., Salzwedel, A., Li, G., Gilmore, J. H., Piven, J., Smith, J. K., Shen, D., Ugurbil, K., Zhu, H., Lin, W., & Elison, J. T. (2019). The UNC/UMN Baby Connectome Project (BCP): An overview of the study design and protocol development. *NeuroImage*, *185*, 891–905. https://doi.org/10.1016/j.neuroimage.2018.03.049

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

Hume, D. (2000). *A Treatise of Human Nature*. Oxford University Press.

Huth, A. G., Nishimoto, S., Vu, A. T., & Gallant, J. L. (2012). A Continuous Semantic Space Describes the Representation of Thousands of Object and Action Categories across the Human Brain. *Neuron*, *76*(6), 1210–1224. https://doi.org/10.1016/j.neuron.2012.10.014

Ito, T., Hearne, L. J., & Cole, M. W. (2020). A cortical hierarchy of localized and distributed processes revealed via dissociation of task activations, connectivity changes, and intrinsic timescales. *NeuroImage*, *221*, 117141. https://doi.org/10.1016/j.neuroimage.2020.117141

James, T. W., & Gauthier, I. (2006). Repetition-induced changes in BOLD response reflect accumulation of neural activity. *Human Brain Mapping*, *27*(1), 37–46. https://doi.org/10.1002/hbm.20165

James, T. W., Humphrey, G. K., Gati, J. S., Menon, R. S., & Goodale, M. A. (2000). The effects of visual object priming on brain activation before and after recognition. *Current Biology*, *10*(17), 1017–1024. https://doi.org/10.1016/S0960-9822(00)00655-2

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. *NeuroImage*, *17*(2), 825–841. https://doi.org/10.1006/nimg.2002.1132

Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., & Smith, S. M. (2012). FSL. *NeuroImage*, *62*(2), 782–790. https://doi.org/10.1016/j.neuroimage.2011.09.015

Jenkinson, M., & Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, *5*(2), 143–156. https://doi.org/10.1016/S1361-8415(01)00036-6

Johnson, M. H., Dziurawiec, S., Ellis, H., & Morton, J. (1991). Newborns' preferential tracking of face-like stimuli and its subsequent decline. *Cognition*, *40*(1), 1–19. https://doi.org/10.1016/0010-0277(91)90045-6

Kamps, F. S., Hendrix, C. L., Brennan, P. A., & Dilks, D. D. (2020). Connectivity at the origins of domain specificity in the cortical face and place networks. *Proceedings of the National Academy of Sciences*, *117*(11), 6163–6169. https://doi.org/10.1073/pnas.1911359117

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. *The Journal of Neuroscience*, *17*(11), 4302–4311. https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., & Zisserman, A. (2017). *The Kinetics Human Action Video Dataset* (arXiv:1705.06950). arXiv. https://doi.org/10.48550/arXiv.1705.06950

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

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K. A., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. Proceedings of the National Academy of Sciences, 116(43), 21854–21863. https://doi.org/10.1073/pnas.1905544116

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition*, *83*(2), B35–B42. https://doi.org/10.1016/S0010-0277(02)00004-5

Klein, A., Ghosh, S. S., Bao, F. S., Giard, J., Häme, Y., Stavsky, E., Lee, N., Rossa, B., Reuter, M., Neto, E. C., & Keshavan, A. (2017). Mindboggling morphometry of human brains. *PLOS Computational Biology*, *13*(2), e1005350. https://doi.org/10.1371/journal.pcbi.1005350

Kobatake, E., & Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of Neurophysiology*, *71*(3), 856–867. https://doi.org/10.1152/jn.1994.71.3.856

Konkle, T., & Alvarez, G. A. (2022). A self-supervised domain-general learning framework for human ventral stream representation. *Nature Communications*, *13*(1), 491. https://doi.org/10.1038/s41467-022-28091-4

Konkle, T., & Caramazza, A. (2013). Tripartite Organization of the Ventral Stream by Animacy and Object Size. *Journal of Neuroscience*, *33*(25), 10235–10242. https://doi.org/10.1523/JNEUROSCI.0983-13.2013

Kosakowski, H. L., Cohen, M. A., Takahashi, A., Keil, B., Kanwisher, N., & Saxe, R. (2022). Selective responses to faces, scenes, and bodies in the ventral visual pathway of infants. *Current Biology*, *32*(2), 265-274.e5. https://doi.org/10.1016/j.cub.2021.10.064

Koulakov, A., Shuvaev, S., Lachi, D., & Zador, A. (2022). *Encoding innate ability through a genomic bottleneck* (p. 2021.03.16.435261). bioRxiv. https://doi.org/10.1101/2021.03.16.435261

Kovack-Lesh, K. A., McMurray, B., & Oakes, L. M. (2014). Four-month-old infants' visual investigation of cats and dogs: Relations with pet experience and attentional strategy. *Developmental Psychology*, *50*(2), 402–413. https://doi.org/10.1037/a0033195

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*. https://doi.org/10.3389/neuro.06.004.2008

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, *25*. https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

Kubilius, J., Schrimpf, M., Kar, K., Rajalingham, R., Hong, H., Majaj, N., Issa, E., Bashivan, P., Prescott-Roy, J., Schmidt, K., Nayebi, A., Bear, D., Yamins, D. L., & DiCarlo, J. J. (2019). Brain-Like Object Recognition with High-Performing Shallow Recurrent ANNs. *Advances in Neural Information Processing Systems*, *32*. https://proceedings.neurips.cc/paper/2019/hash/7813d1590d28a7dd372ad54b5d29d033-Abstract.html

Kuffler, S. W. (1953). Discharge patterns and functional organization of mammalian retina. *Journal of Neurophysiology*, *16*(1), 37–68. https://doi.org/10.1152/jn.1953.16.1.37

Lage-Castellanos, A., Valente, G., Formisano, E., & De Martino, F. (2019). Methods for computing the maximum performance of computational models of fMRI responses. *PLoS Computational Biology*, *15*(3), e1006397. https://doi.org/10.1371/journal.pcbi.1006397

Lanczos, C. (1964). Evaluation of Noisy Data. *Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis*, *1*(1), 76–85. https://doi.org/10.1137/0701007

Lee, S., & Kable, J. W. (2018). Simple but robust improvement in multivoxel pattern classification. *PLoS ONE*, *13*(11), e0207083. https://doi.org/10.1371/journal.pone.0207083

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic Mapping of a Hierarchy of Temporal Receptive Windows Using a Narrated Story. *Journal of Neuroscience*, *31*(8), 2906–2915. https://doi.org/10.1523/JNEUROSCI.3684-10.2011

Levy, I., Hasson, U., Avidan, G., Hendler, T., & Malach, R. (2001). Center–periphery organization of human object areas. *Nature Neuroscience*, *4*(5), 533–539. https://doi.org/10.1038/87490

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer Vision – ECCV 2014* (pp. 740–755). Springer International Publishing. https://doi.org/10.1007/978-3-319-10602-1_48

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). *Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows*. 10012–10022. https://openaccess.thecvf.com/content/ICCV2021/html/Liu_Swin_Transformer_Hierarchical_Vision_Transformer_Using_Shifted_Windows_ICCV_2021_paper

Livingstone, M. S., Vincent, J. L., Arcaro, M. J., Srihasam, K., Schade, P. F., & Savage, T. (2017). Development of the macaque face-patch system. *Nature Communications*, *8*(1), 14897. https://doi.org/10.1038/ncomms14897

Locke, J. (1948). An essay concerning human understanding, 1690. In *Readings in the history of psychology* (pp. 55–68). Appleton-Century-Crofts. https://doi.org/10.1037/11304-008

Long, B., Xiang, V., Stojanov, S., Sparks, R. Z., Yin, Z., Keene, G. E., Tan, A. W. M., Feng, S. Y., Zhuang, C., Marchman, V. A., Yamins, D. L. K., & Frank, M. C. (2024). *The BabyView dataset: High-resolution egocentric videos of infants' and young children's everyday experiences* (arXiv:2406.10447). arXiv. https://doi.org/10.48550/arXiv.2406.10447

MacEvoy, S. P., & Epstein, R. A. (2011). Constructing scenes from objects in human occipitotemporal cortex. *Nature Neuroscience*, *14*(10), 1323–1329. https://doi.org/10.1038/nn.2903

Mahon, B. Z., Anzellotti, S., Schwarzbach, J., Zampini, M., & Caramazza, A. (2009). Category-Specific Organization in the Human Brain Does Not Require Visual Experience. *Neuron*, *63*(3), 397–405. https://doi.org/10.1016/j.neuron.2009.07.012

Makropoulos, A., Robinson, E. C., Schuh, A., Wright, R., Fitzgibbon, S., Bozek, J., Counsell, S. J., Steinweg, J., Vecchiato, K., Passerat-Palmbach, J., Lenz, G., Mortari, F., Tenev, T., Duff, E. P., Bastiani, M., Cordero-Grande, L., Hughes, E., Tusor, N., Tournier, J.-D., … Rueckert, D. (2018). The developing human connectome project: A minimal processing pipeline for neonatal cortical surface reconstruction. *NeuroImage*, *173*, 88–112. https://doi.org/10.1016/j.neuroimage.2018.01.054

Málková, L., Mishkin, M., & Bachevalier, J. (1995). Long-term effects of selective neonatal temporal lobe lesions on learning and memory in monkeys. *Behavioral Neuroscience*, *109*(2), 212–226. https://doi.org/10.1037/0735-7044.109.2.212

Mandler, J. M. (2004). *The Foundations of Mind: Origins of Conceptual Thought*. Oxford University Press.

Margolis, E., & Laurence, S. (2023). Concepts. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2023). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2023/entries/concepts/

Markie, P., & Folescu, M. (2021). *Rationalism vs. Empiricism*. https://plato.stanford.edu/entries/rationalism-empiricism/?ref=hackernoon.com

McCloskey, M. E., & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets? *Memory & Cognition*, *6*(4), 462–472. https://doi.org/10.3758/BF03197480

McKeeff, T. J., Remus, D. A., & Tong, F. (2007). Temporal Limitations in Object Processing Across the Human Ventral Visual Pathway. *Journal of Neurophysiology*, *98*(1), 382–393. https://doi.org/10.1152/jn.00568.2006

McKone, E., Crookes, K., Jeffery, L., & Dilks, D. D. (2012). A critical review of the development of face recognition: Experience is less important than previously believed. *Cognitive Neuropsychology*, *29*(1–2), 174–212. https://doi.org/10.1080/02643294.2012.660138

Mehrer, J., Spoerer, C. J., Jones, E. C., Kriegeskorte, N., & Kietzmann, T. C. (2021). An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, *118*(8). https://doi.org/10.1073/pnas.2011417118

Mehrer, J., Spoerer, C. J., Kriegeskorte, N., & Kietzmann, T. C. (2020). Individual differences among deep neural network models. *Nature Communications*, *11*(1), 5725. https://doi.org/10.1038/s41467-020-19632-w

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *arXiv:1310.4546 [Cs, Stat]*. http://arxiv.org/abs/1310.4546

Mishkin, M., Ungerleider, L. G., & Macko, K. A. (1983). Object vision and spatial vision: Two cortical pathways. *Trends in Neurosciences*, *6*, 414–417. https://doi.org/10.1016/0166-2236(83)90190-X

Mobahi, H., Collobert, R., & Weston, J. (2009). Deep learning from temporal coherence in video. *Proceedings of the 26th Annual International Conference on Machine Learning*, 737–744. https://doi.org/10.1145/1553374.1553469

Morris, A. S., Wakschlag, L., Krogh-Jespersen, S., Fox, N., Planalp, B., Perlman, S. B., Shuffrey, L. C., Smith, B., Lorenzo, N. E., Amso, D., Coles, C. D., & Johnson, S. P. (2020). Principles for Guiding the Selection of Early Childhood Neurodevelopmental Risk and Resilience Measures: HEALthy Brain and Child Development Study as an Exemplar. *Adversity and Resilience Science*, *1*(4), 247–267. https://doi.org/10.1007/s42844-020-00025-3

Mummery, C. J., Patterson, K., Price, C. J., Ashburner, J., Frackowiak, R. S. J., & Hodges, J. R. (2000). A voxel-based morphometry study of semantic dementia: Relationship between temporal lobe atrophy and semantic memory. *Annals of Neurology*, *47*(1), 36–45. https://doi.org/10.1002/1531-8249(200001)47:1<36::AID-ANA8>3.0.CO;2-L

Murray, J. D., Bernacchia, A., Freedman, D. J., Romo, R., Wallis, J. D., Cai, X., Padoa-Schioppa, C., Pasternak, T., Seo, H., Lee, D., & Wang, X.-J. (2014). A hierarchy of intrinsic timescales across primate cortex. *Nature Neuroscience*, *17*(12), Article 12. https://doi.org/10.1038/nn.3862

Murray, S. O., & Wojciulik, E. (2004). Attention increases neural selectivity in the human lateral occipital complex. *Nature Neuroscience*, *7*(1), 70–74. https://doi.org/10.1038/nn1161

Noguchi, Y., Inui, K., & Kakigi, R. (2004). Temporal Dynamics of Neural Adaptation Effect in the Human Visual Ventral Stream. *Journal of Neuroscience*, *24*(28), 6283–6290. https://doi.org/10.1523/JNEUROSCI.0655-04.2004

Oakes, L. M., Kovack-Lesh, K. A., & Horst, J. S. (2009). Two are better than one: Comparison influences infants' visual recognition memory. *Journal of Experimental Child Psychology*, *104*(1), 124–131. https://doi.org/10.1016/j.jecp.2008.09.001

Oliva, A., & Torralba, A. (2006). Chapter 2 Building the gist of a scene: The role of global image features in recognition. In S. Martinez-Conde, S. L. Macknik, L. M. Martinez, J.-M. Alonso, & P. U. Tse (Eds.), *Progress in Brain Research* (Vol. 155, pp. 23–36). Elsevier. https://doi.org/10.1016/S0079-6123(06)55002-2

Orban, G. A., Van Essen, D., & Vanduffel, W. (2004). Comparative mapping of higher visual areas in monkeys and humans. *Trends in Cognitive Sciences*, *8*(7), 315–324. https://doi.org/10.1016/j.tics.2004.05.009

Orhan, E. (2023). *Scaling may be all you need for achieving human-level object recognition capacity with human-like visual experience* (arXiv:2308.03712). arXiv. https://doi.org/10.48550/arXiv.2308.03712

Orhan, E., Gupta, V., & Lake, B. M. (2020). Self-supervised learning through the eyes of a child. *Advances in Neural Information Processing Systems*, *33*, 9960–9971. https://proceedings.neurips.cc/paper/2020/hash/7183145a2a3e0ce2b68cd3735186b1d5-Abstract.html

Palmer, E. (1975). The effects of contextual scenes on the identification of objects. *Memory & Cognition*, *3*(5), 519–526. https://doi.org/10.3758/BF03197524

Pan, Y., Frisson, S., Federmeier, K. D., & Jensen, O. (2024). Early parafoveal semantic integration in natural reading. *eLife*, *12*, RP91327. https://doi.org/10.7554/eLife.91327

Pandey, L., Wood, S. M. W., & Wood, J. N. (2023). *Are Vision Transformers More Data Hungry Than Newborn Visual Systems?* (arXiv:2312.02843). arXiv. https://doi.org/10.48550/arXiv.2312.02843

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Pereira, A. F., Smith, L. B., & Yu, C. (2014). A Bottom-up View of Toddler Word Learning. *Psychonomic Bulletin & Review*, *21*(1), 178–185. https://doi.org/10.3758/s13423-013-0466-4

Piaget, J. (1952). *The origins of intelligence in children* (p. 419). W W Norton & Co. https://doi.org/10.1037/11494-000

Pitt, D. (2022). Mental Representation. In *The Stanford Encyclopedia of Philosophy* (Fall 2022). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2022/entries/mental-representation/

Pomiechowska, B., & Gliga, T. (2019). Lexical Acquisition Through Category Matching: 12-Month-Old Infants Associate Words to Visual Categories. *Psychological Science*, *30*(2), 288–299. https://doi.org/10.1177/0956797618817506

Poulin-Dubois, D., Graham, S., & Sippola, L. (1995). Early lexical development: The contribution of parental labelling and infants' categorization abilities. *Journal of Child Language*, *22*(2), 325–343. https://doi.org/10.1017/S0305000900009818

Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage*, *84*, 320–341. https://doi.org/10.1016/j.neuroimage.2013.08.048

Quek, G. L., & Peelen, M. V. (2020). Contextual and Spatial Associations Between Objects Interactively Modulate Visual Processing. *Cerebral Cortex*, *30*(12), 6391–6404. https://doi.org/10.1093/cercor/bhaa197

Quinn, P. C., & Eimas, P. D. (1998). Evidence for a Global Categorical Representation of Humans by Young Infants. *Journal of Experimental Child Psychology*, *69*(3), 151–174. https://doi.org/10.1006/jecp.1998.2443

Quinn, P. C., Eimas, P. D., & Tarr, M. J. (2001). Perceptual Categorization of Cat and Dog Silhouettes by 3- to 4-Month-Old Infants. *Journal of Experimental Child Psychology*, *79*(1), 78–94. https://doi.org/10.1006/jecp.2000.2609

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *Proceedings of the 38th International Conference on Machine Learning*, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html

Ranganath, C., & Hsieh, L.-T. (2016). The hippocampus: A special place for time. *Annals of the New York Academy of Sciences*, *1369*(1), 93–110. https://doi.org/10.1111/nyas.13043

Ratan Murty, N. A., Teng, S., Beeler, D., Mynick, A., Oliva, A., & Kanwisher, N. (2020). Visual experience is not necessary for the development of face-selectivity in the lateral fusiform gyrus. *Proceedings of the National Academy of Sciences*, *117*(37), 23011–23020. https://doi.org/10.1073/pnas.2004607117

Reid, V. M., Dunn, K., Young, R. J., Amu, J., Donovan, T., & Reissland, N. (2017). The Human Fetus Preferentially Engages with Face-like Visual Stimuli. *Current Biology*, *27*(12), 1825-1828.e3. https://doi.org/10.1016/j.cub.2017.05.044

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., … Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, *22*(11), Article 11. https://doi.org/10.1038/s41593-019-0520-2

Richter, D., Ekman, M., & Lange, F. P. de. (2018). Suppressed Sensory Response to Predictable Object Stimuli throughout the Ventral Visual Stream. *Journal of Neuroscience*, *38*(34), 7452–7461. https://doi.org/10.1523/JNEUROSCI.3421-17.2018

Roads, B. D., & Love, B. C. (2020). Learning as the unsupervised alignment of conceptual systems. *Nature Machine Intelligence*, *2*(1), Article 1. https://doi.org/10.1038/s42256-019-0132-2

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605. https://doi.org/10.1016/0010-0285(75)90024-9

Rosenthal, C. R., Mallik, I., Caballero-Gaudes, C., Sereno, M. I., & Soto, D. (2018). Learning of goal-relevant and -irrelevant complex visual sequences in human V1. *NeuroImage*, *179*, 215–224. https://doi.org/10.1016/j.neuroimage.2018.06.023

Rowlands, M. (2017). Arguing about representation. *Synthese*, *194*(11), 4215–4232. https://doi.org/10.1007/s11229-014-0646-4

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Rust, N. C., & DiCarlo, J. J. (2010). Selectivity and Tolerance ("Invariance") Both Increase as Visual Information Propagates from Cortical Area V4 to IT. *Journal of Neuroscience*, *30*(39), 12978–12995. https://doi.org/10.1523/JNEUROSCI.0179-10.2010

Sadeghi, Z., McClelland, J. L., & Hoffman, P. (2015). You shall know an object by the company it keeps: An investigation of semantic representations derived from object co-occurrence in visual scenes. *Neuropsychologia*, *76*, 52–61. https://doi.org/10.1016/j.neuropsychologia.2014.08.031

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science*, *274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Satterthwaite, T. D., Elliott, M. A., Gerraty, R. T., Ruparel, K., Loughead, J., Calkins, M. E., Eickhoff, S. B., Hakonarson, H., Gur, R. C., Gur, R. E., & Wolf, D. H. (2013). An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. *NeuroImage*, *64*, 240–256. https://doi.org/10.1016/j.neuroimage.2012.08.052

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X.-N., Holmes, A. J., Eickhoff, S. B., & Yeo, B. T. T. (2018). Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cerebral Cortex*, *28*(9), 3095–3114. https://doi.org/10.1093/cercor/bhx179

Schapiro, A. C., Kustner, L. V., & Turk-Browne, N. B. (2012). Shaping of Object Representations in the Human Medial Temporal Lobe Based on Temporal Regularities. *Current Biology*, *22*(17), 1622–1627. https://doi.org/10.1016/j.cub.2012.06.056

Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B., & Botvinick, M. M. (2013). Neural representations of events arise from temporal community structure. *Nature Neuroscience*, *16*(4), 486–492. https://doi.org/10.1038/nn.3331

Siegel, J. S., Power, J. D., Dubis, J. W., Vogel, A. C., Church, J. A., Schlaggar, B. L., & Petersen, S. E. (2014). Statistical improvements in functional magnetic resonance imaging analyses produced by censoring high-motion data points. *Human Brain Mapping*, *35*(5), 1981–1996. https://doi.org/10.1002/hbm.22307

Simion, F., Di Giorgio, E., Leo, I., & Bardi, L. (2011). Chapter 10 - The processing of social stimuli in early infancy: From faces to biological motion perception. In O. Braddick, J. Atkinson, & G. M. Innocenti (Eds.), *Progress in Brain Research* (Vol. 189, pp. 173–193). Elsevier. https://doi.org/10.1016/B978-0-444-53884-0.00024-5

Skelton, A. E., Maule, J., & Franklin, A. (2022). Infant color perception: Insight into perceptual development. *Child Development Perspectives*, *16*(2), 90–95. https://doi.org/10.1111/cdep.12447

Smith, L. B. (1999). Do infants possess innate knowledge structures? The con side. *Developmental Science*, *2*(2), 133–144. https://doi.org/10.1111/1467-7687.00062

Smith, L. B., Jayaraman, S., Clerkin, E., & Yu, C. (2018). The Developing Infant Creates a Curriculum for Statistical Learning. *Trends in Cognitive Sciences*, *22*(4), 325–336. https://doi.org/10.1016/j.tics.2018.02.004

Smith, L. B., & Slone, L. K. (2017). A Developmental Approach to Machine Learning? *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.02124

Spelke, E. S. (1994). Initial knowledge: Six suggestions. *Cognition*, *50*(1), 431–445. https://doi.org/10.1016/0010-0277(94)90039-6

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, *10*(1), 89–96. https://doi.org/10.1111/j.1467-7687.2007.00569.x

Spriet, C., Abassi, E., Hochmann, J.-R., & Papeo, L. (2022). Visual object categorization in infancy. *Proceedings of the National Academy of Sciences*, *119*(8), e2105866119. https://doi.org/10.1073/pnas.2105866119

Stojanoski, B., & Cusack, R. (2014). Time to wave good-bye to phase scrambling: Creating controlled scrambled images using diffeomorphic transformations. *Journal of Vision*, *14*(12), 6–6. https://doi.org/10.1167/14.12.6

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

Tanaka, K. (1996). *Inferotemporal Cortex and Object Vision*. 31.

Taylor, K. I., Moss, H. E., & Tyler, L. K. (2007). The conceptual structure account: A cognitive model of semantic memory and its neural instantiation. In *Neural basis of semantic memory* (pp. 265–301). Cambridge University Press. https://doi.org/10.1017/CBO9780511544965.012

Tian, Y., Krishnan, D., & Isola, P. (2020). Contrastive Multiview Coding. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), *Computer Vision – ECCV 2020* (pp. 776–794). Springer International Publishing. https://doi.org/10.1007/978-3-030-58621-8_45

Truzzi, A., & Cusack, R. (2023). The development of intrinsic timescales: A comparison between the neonate and adult brain. *NeuroImage*, *275*, 120155. https://doi.org/10.1016/j.neuroimage.2023.120155

Turati, C., Simion, F., & Zanon, L. (2003). Newborns' Perceptual Categorization for Closed and Open Geometric Forms. *Infancy*, *4*(3), 309–325. https://doi.org/10.1207/S15327078IN0403_01

Turk-Browne, N., Yi, D.-J., Leber, A., & Chun, M. (2007). Visual Quality Determines the Direction of Neural Repetition Effects. *Cerebral Cortex*, *17*(2), 425–433. https://doi.org/10.1093/cercor/bhj159

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4ITK: Improved N3 Bias Correction. *IEEE Transactions on Medical Imaging*, *29*(6), 1310–1320. IEEE Transactions on Medical Imaging. https://doi.org/10.1109/TMI.2010.2046908

van den Hurk, J., Van Baelen, M., & Op de Beeck, H. P. (2017). Development of visual category selectivity in ventral visual cortex does not require visual experience. *Proceedings of the National Academy of Sciences*, *114*(22), E4501–E4510. https://doi.org/10.1073/pnas.1612862114

van Turennout, M., Ellmore, T., & Martin, A. (2000). Long-lasting cortical plasticity in the object naming system. *Nature Neuroscience*, *3*(12), 1329–1334. https://doi.org/10.1038/81873

Vergnieux, V., & Vogels, R. (2020). Statistical Learning Signals for Complex Visual Images in Macaque Early Visual Cortex. *Frontiers in Neuroscience*, *14*. https://doi.org/10.3389/fnins.2020.00789

Vong, W. K., Wang, W., Orhan, A. E., & Lake, B. M. (2024). Grounded language acquisition through the eyes and ears of a single child. *Science*, *383*(6682), 504–511. https://doi.org/10.1126/science.adi1374

Werker, J. F. (2024). Phonetic perceptual reorganization across the first year of life: Looking back. *Infant Behavior and Development*, *75*, 101935. https://doi.org/10.1016/j.infbeh.2024.101935

Wild, C. J., Linke, A. C., Zubiaurre-Elorza, L., Herzmann, C., Duffy, H., Han, V. K., Lee, D. S., & Cusack, R. (2017). Adult-like processing of naturalistic sounds in auditory cortex by 3-and 9-month old infants. *NeuroImage*, *157*, 623–634.

Wiskott, L., & Sejnowski, T. J. (2002). Slow Feature Analysis: Unsupervised Learning of Invariances. *Neural Computation*, *14*(4), 715–770. Neural Computation. https://doi.org/10.1162/089976602317318938

Xiao, K., Engstrom, L., Ilyas, A., & Madry, A. (2020). *Noise or Signal: The Role of Image Backgrounds in Object Recognition* (arXiv:2006.09994). arXiv. https://doi.org/10.48550/arXiv.2006.09994

Xie, S., Hoehl, S., Moeskops, M., Kayhan, E., Kliesch, C., Turtleton, B., Köster, M., & Cichy, R. M. (2022). Visual category representations in the infant brain. *Current Biology*, *32*(24), 5422-5432.e6. https://doi.org/10.1016/j.cub.2022.11.016

Xu, Y., & Vaziri-Pashkam, M. (2021). Limits to visual representational correspondence between convolutional neural networks and the human brain. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-021-22244-7

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

Yan, X., Tung, S., Fascendini, B., Chen, Y. D., Norcia, A. M., & Grill-Spector, K. (2023). *When Do Visual Category Representations Emerge in Infants' Brains?* (p. 2023.05.11.539934). bioRxiv. https://doi.org/10.1101/2023.05.11.539934

Yates, T. S., Skalaban, L. J., Ellis, C. T., Bracher, A. J., Baldassano, C., & Turk-Browne, N. B. (2022). Neural event segmentation of continuous experience in human infants. *Proceedings of the National Academy of Sciences*, *119*(43), e2200257119. https://doi.org/10.1073/pnas.2200257119

Younger, B. A., & Cohen, L. B. (1986). Developmental Change in Infants' Perception of Correlations among Attributes. *Child Development*, *57*(3), 803–815. https://doi.org/10.2307/1130356

Younger, B. A., & Fearing, D. D. (1998). Detecting correlations among form attributes: An object-examining test with infants. *Infant Behavior and Development*, *21*(2), 289–297. https://doi.org/10.1016/S0163-6383(98)90007-8

Younger, B. A., & Fearing, D. D. (2000). A Global-to-Basic Trend in Early Categorization: Evidence From a Dual-Category Habituation Task. *Infancy*, *1*(1), 47–58. https://doi.org/10.1207/S15327078IN0101_05

Yu, C., & Smith, L. B. (2012). Embodied attention and word learning by toddlers. *Cognition*, *125*(2). https://doi.org/10.1016/j.cognition.2012.06.016

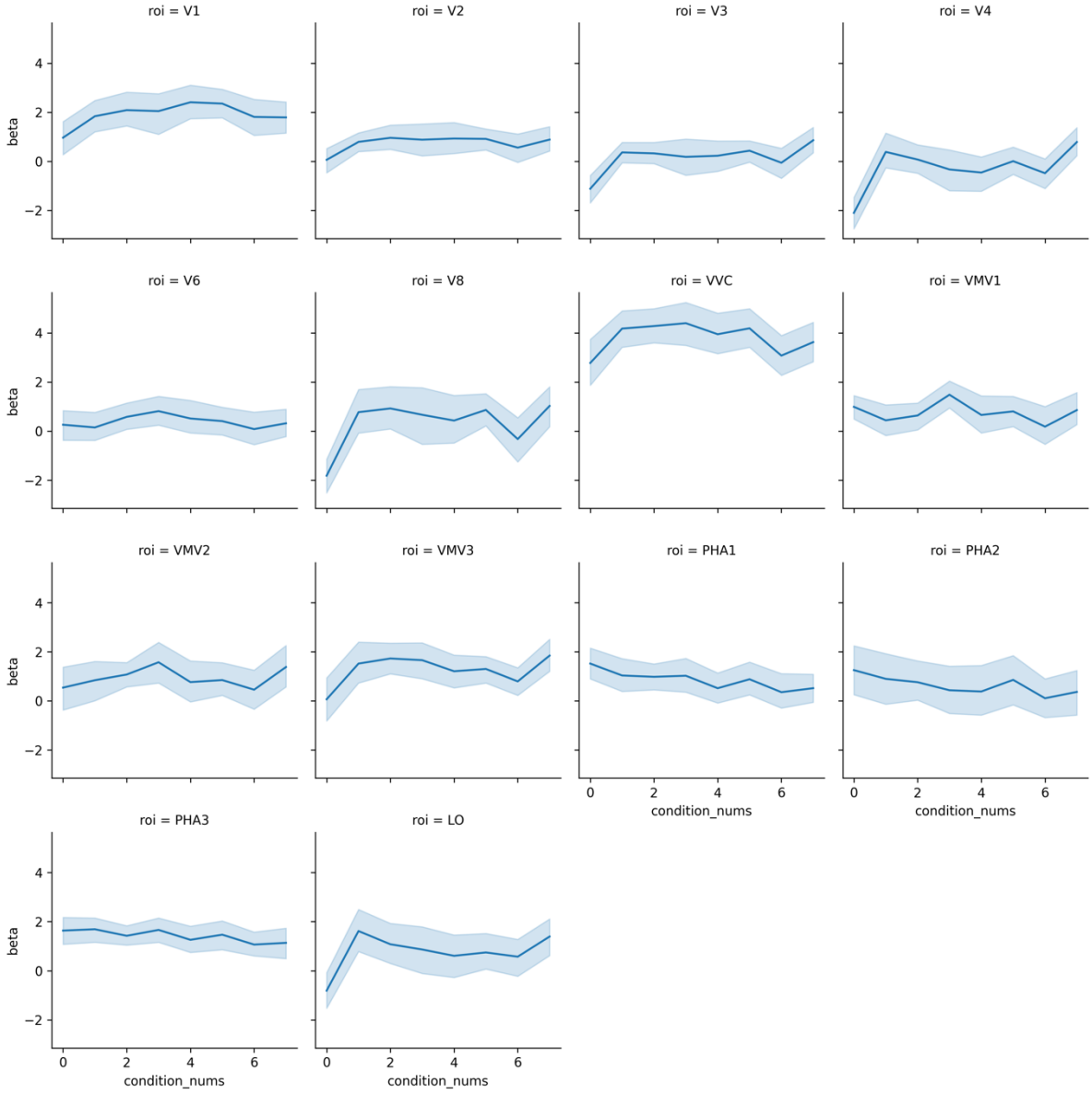
Zaadnoordijk, L., Besold, T. R., & Cusack, R. (2022). Lessons from infant learning for unsupervised machine learning. *Nature Machine Intelligence*, *4*(6), 510–520. https://doi.org/10.1038/s42256-022-00488-2

Zago, L., Fenske, M. J., Aminoff, E., & Bar, M. (2005). The Rise and Fall of Priming: How Visual Exposure Shapes Cortical Representations of Objects. *Cerebral Cortex*, *15*(11), 1655–1665. https://doi.org/10.1093/cercor/bhi060

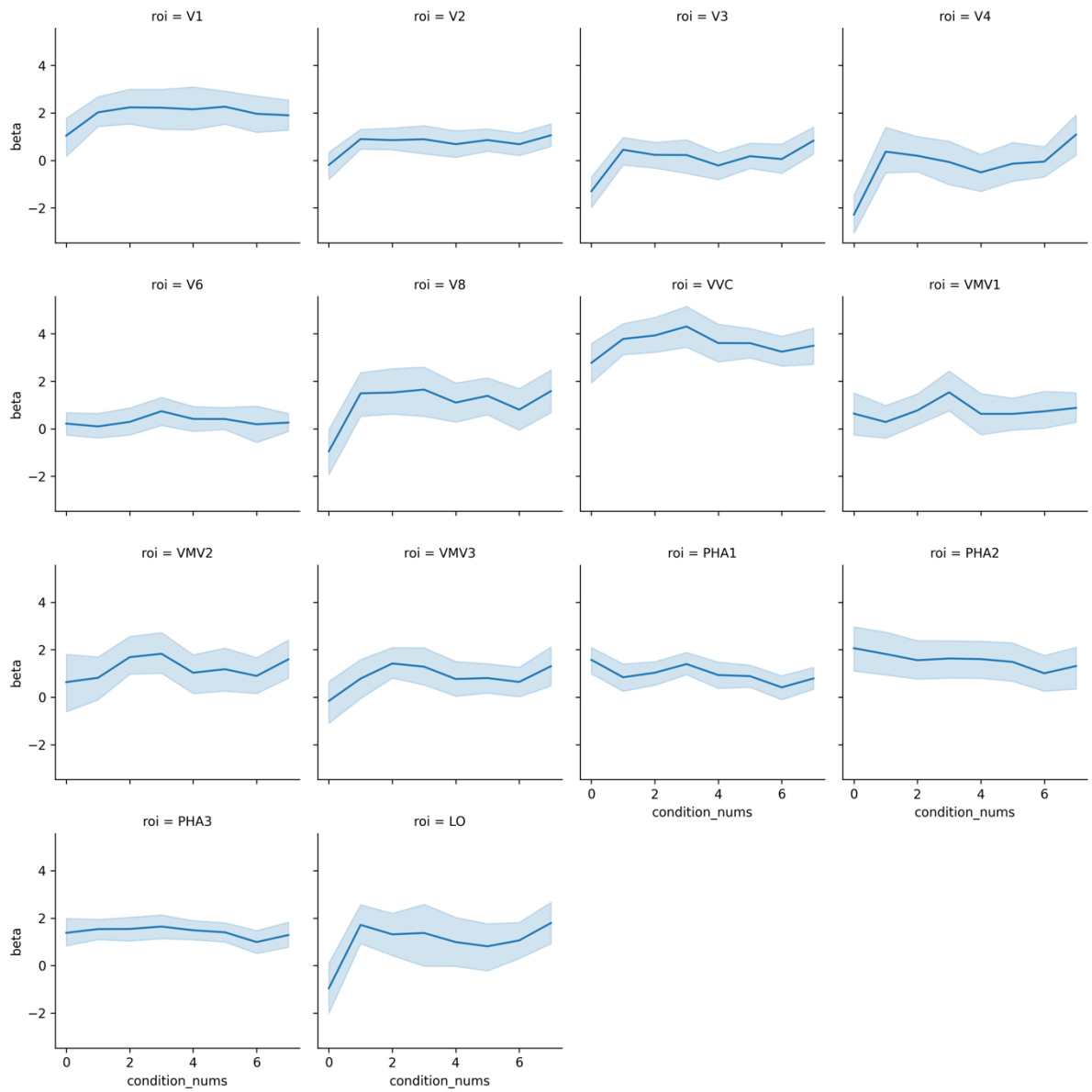
Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, *20*(1), 45–57. IEEE Transactions on Medical Imaging. https://doi.org/10.1109/42.906424

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., & Yamins, D. L. K. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, *118*(3), e2014196118. https://doi.org/10.1073/pnas.2014196118

Appendix



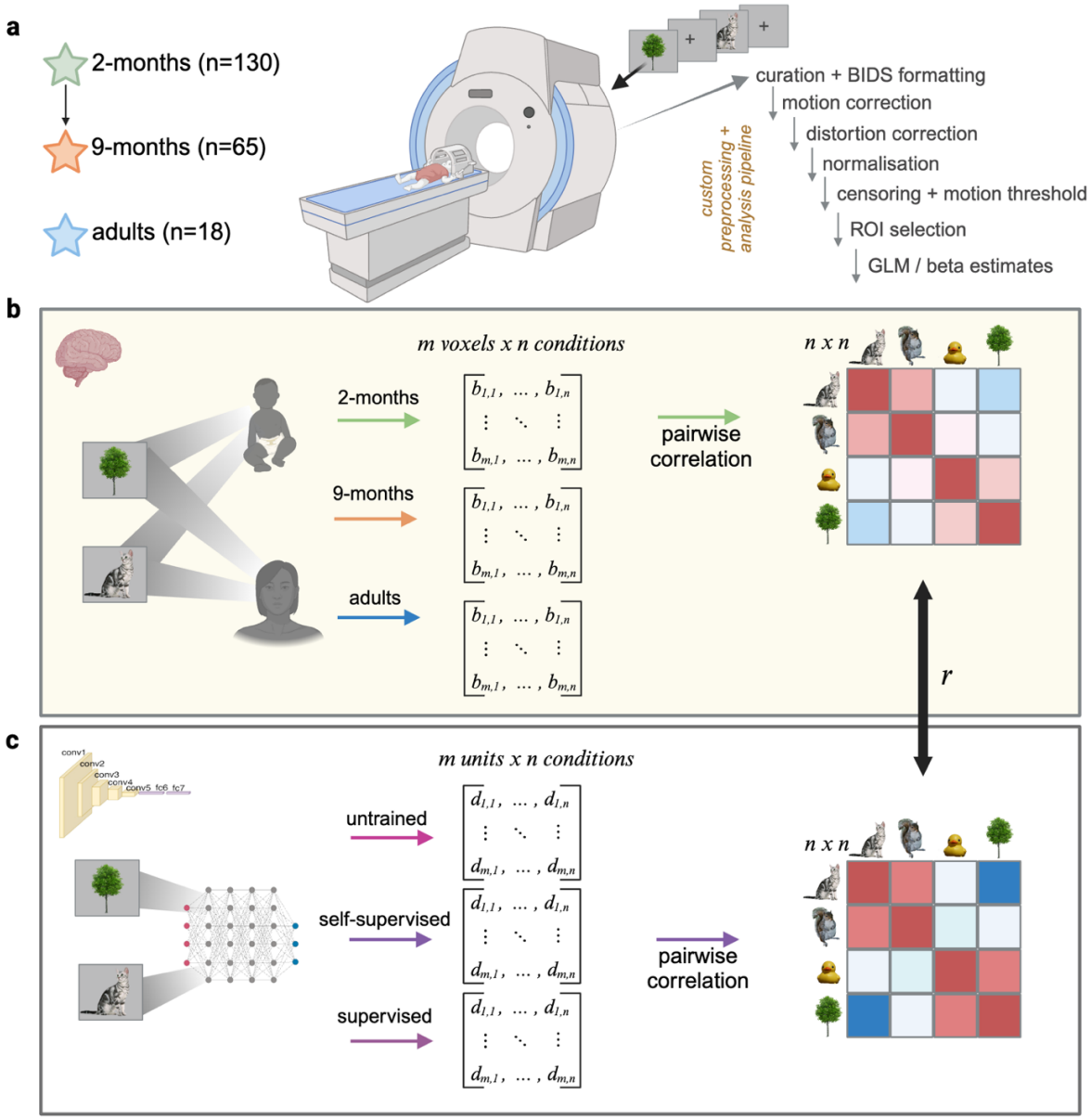
Supplementary Figure 3.1 Repetition response curves for left hemisphere in each ROI. The beta estimates for individual objects in the slow condition within each ROI. Regions were defined using the Glasser parcellation. Error bands indicated 95% confidence interval estimated using 999 permutations across $n = 25$ subjects.



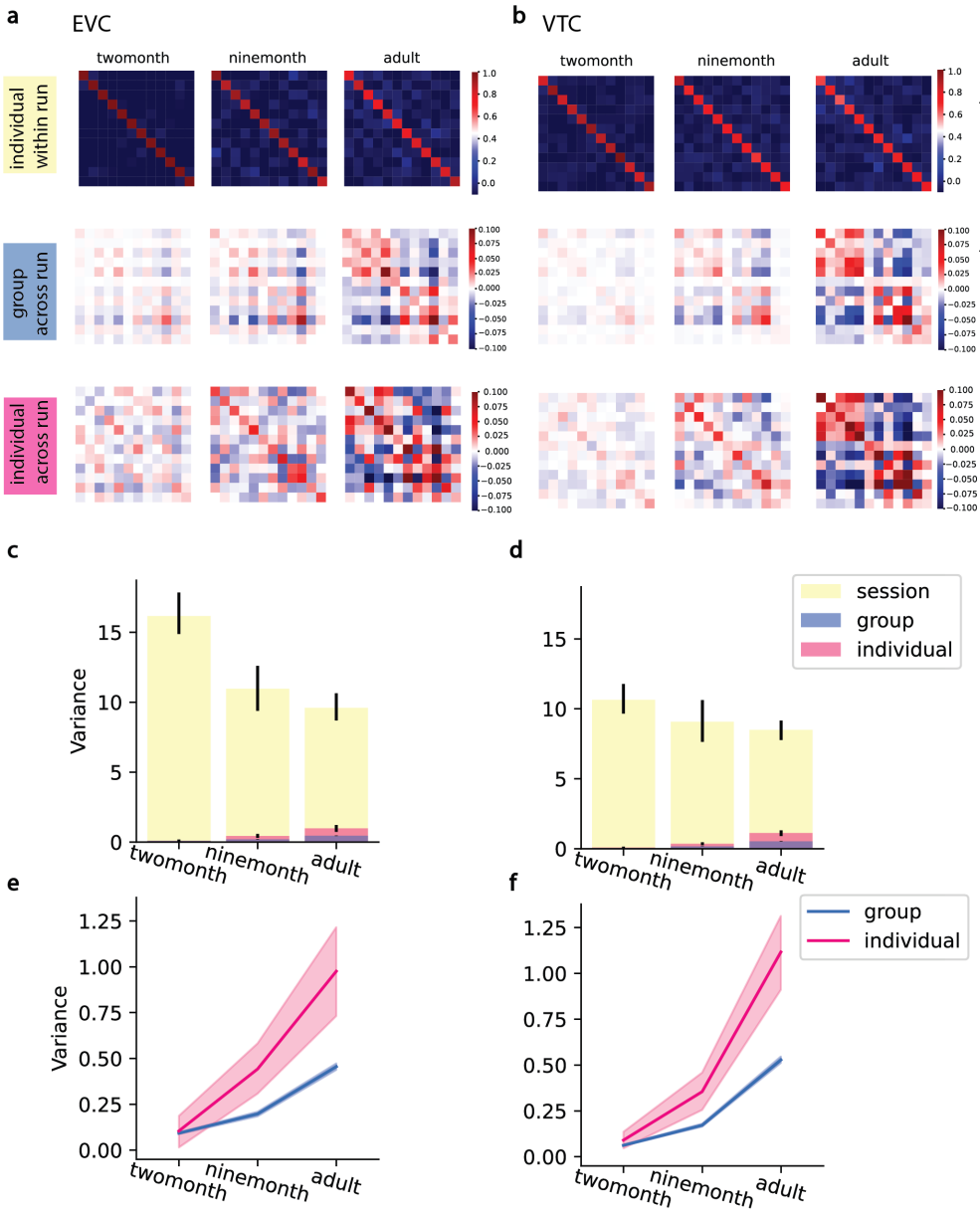
Supplementary Figure 3.2 Repetition response curves for right hemisphere in each ROI. The beta estimates for individual objects in the slow condition within each ROI. Regions were defined using the Glasser parcellation. Error bands indicated 95% confidence interval estimated using 999 permutations across n = 25 subjects.

Supplementary Table 3.1 Results of the multiple linear regression for the association between intrinsic neural timescale and repetition response.

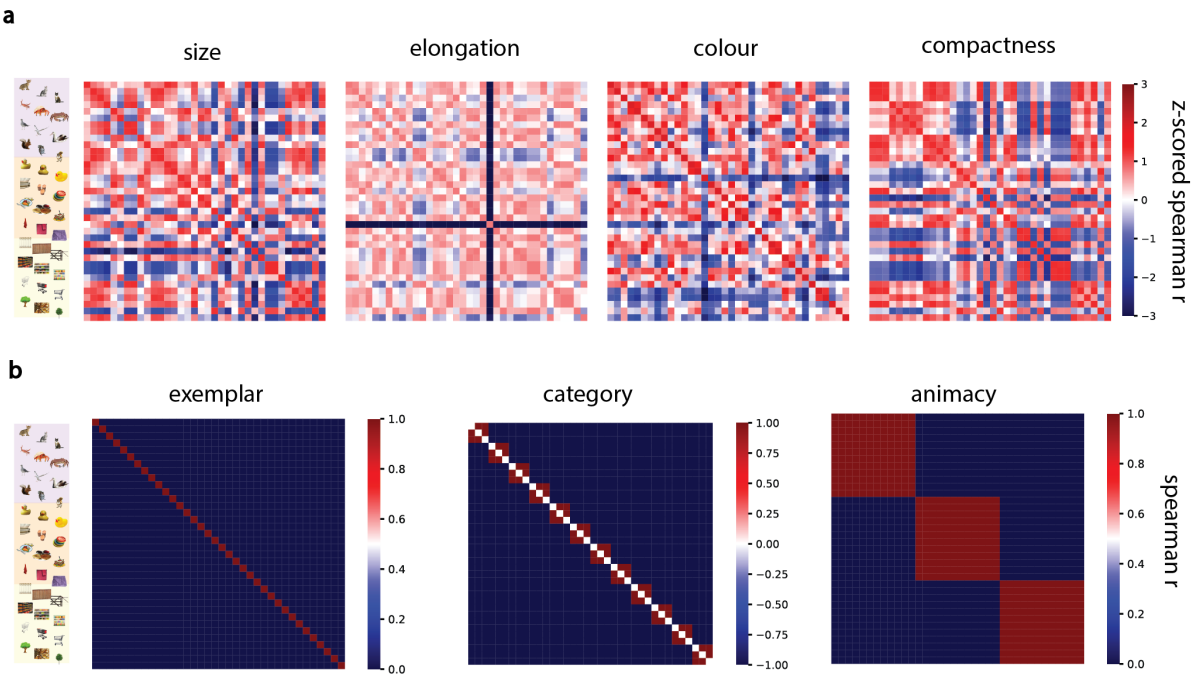
	Coefficient	Std. Error	t-value	p-value	CI Lower	CI Upper
Intercept	6.2646	2.7531	2.2755	0.0297	0.6569	11.8725
INT (τ)	-2.5945	0.9095	-2.8525	0.0075	-4.4472	-0.7418
Fast/slow	7.1533	3.8934	1.8373	0.0755	-0.7773	15.0839
Interaction	-1.6941	1.2863	-1.3170	0.1972	-4.3142	0.9260



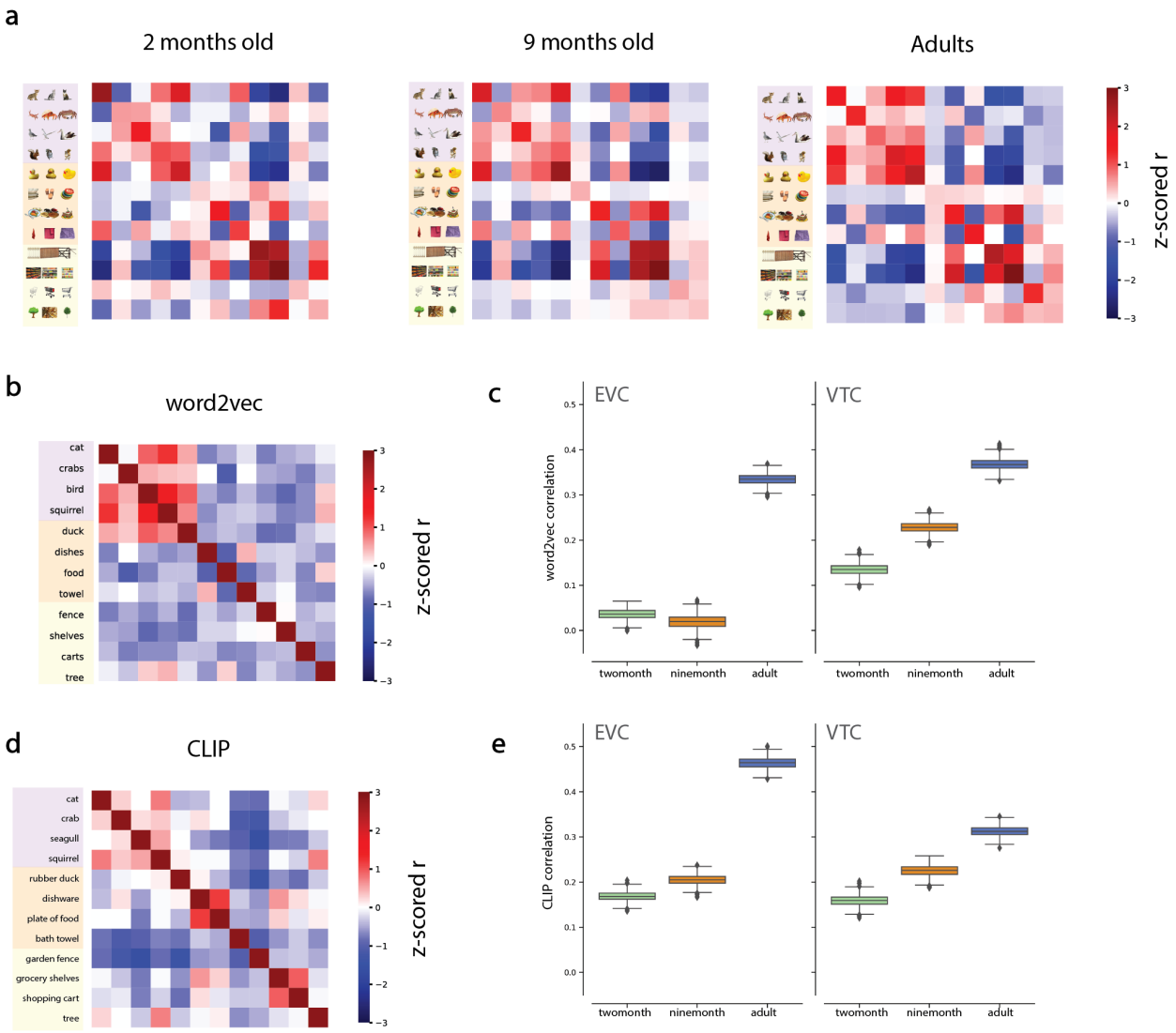
Supplementary Figure 4.1. Schematic overview of methods. (a) We conducted awake longitudinal fMRI in 2-month-olds and again at 9-months-old. A cohort of adults was collected for comparison. Infants and adults viewed 36 exemplars from 12 categories against a grey background, and their BOLD responses were recorded. After careful data curation and pre-processing with custom analysis pipelines, response patterns in two regions of interest were estimated using a generalised linear model. The final dataset, after motion thresholding, included n = 103 and n = 38 2 and 9-month-olds respectively **(b)** Visual representations were characterised per ROI and age-group using multivariate pattern analysis. The pairwise correlations between all *n-conditions* beta estimates in a *m-voxel* dimensional space gave a *n x n* similarity matrix. **(c)** The equivalent analysis was performed in a deep neural network. Activations were calculated from each layer, in response to the same images used during scanning. The resulting model similarity matrix could then be compared with the fMRI data used a Spearman correlation. *Created with* ***BioRender.com***



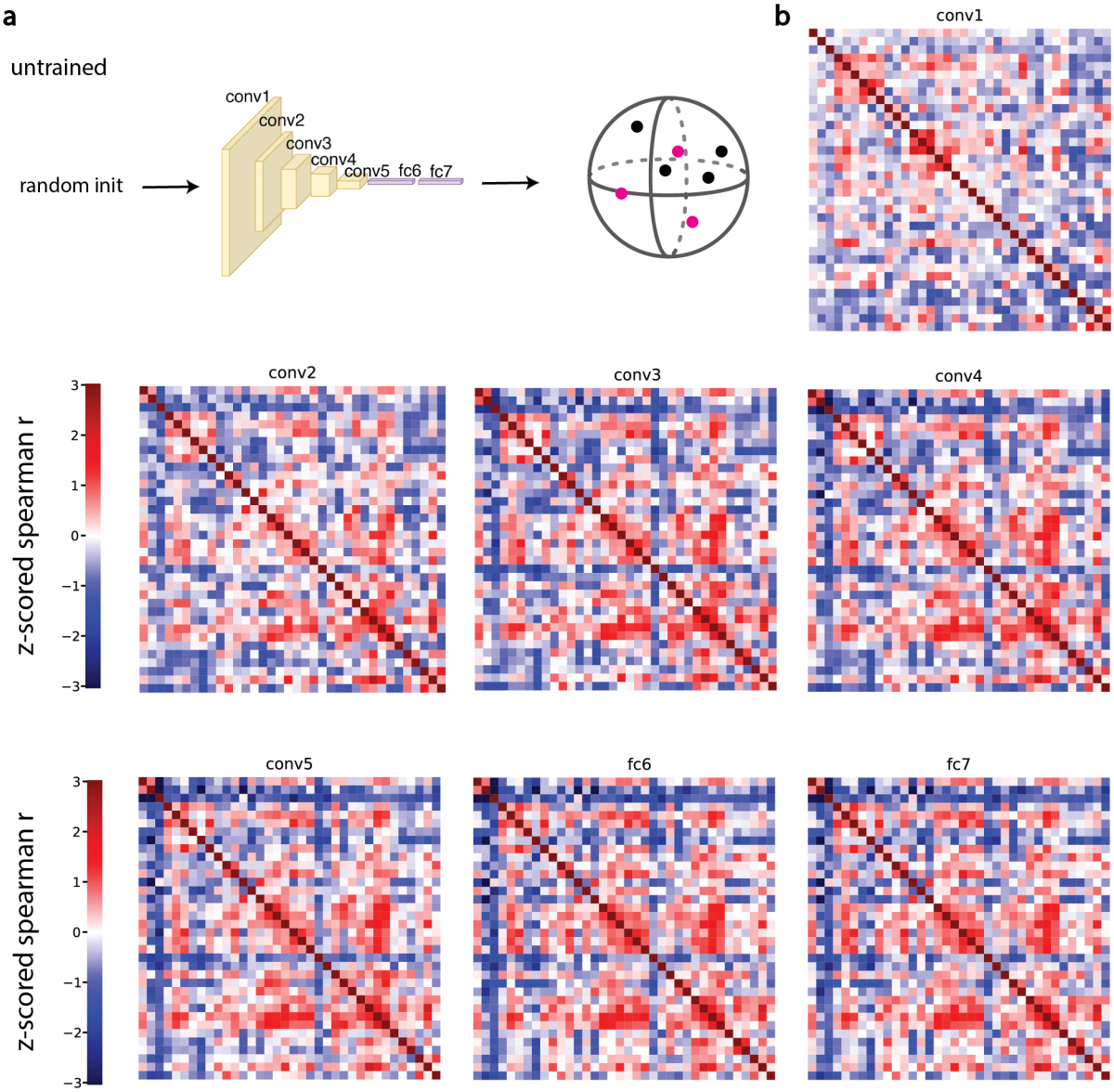
Supplementary Figure 4.2. Variance partitioning analysis for group, individual and session based variance. (a,b) Group average pairwise covariances across 12 categories' beta estimates in EVC and VTC respectively using a 12 category model that is collapsed across exemplars. Pairwise comparisons were calculated within subjects/within run, across subjects/runs and within subject/across runs to partition the session, group and subject variances respectively. Variance estimates are not z-scored, and X and Y axes of the matrices are identical. **(c,d)** Overall estimate of within run (session), across subjects (group) and across run (individual) noise calculated using the trace of the RDMs in (a) and (b) for EVC and VTC respectively. All sources of noise varied with age. Error bars are the 95% confidence intervals, calculated by bootstrap resampling across subject/run pairs in the group average RDM. **(e,f)** The same noise estimates in (c) and (d), with only the group and individual level variance plot to illustrate developmental changes in shared and idiosyncratic noise.



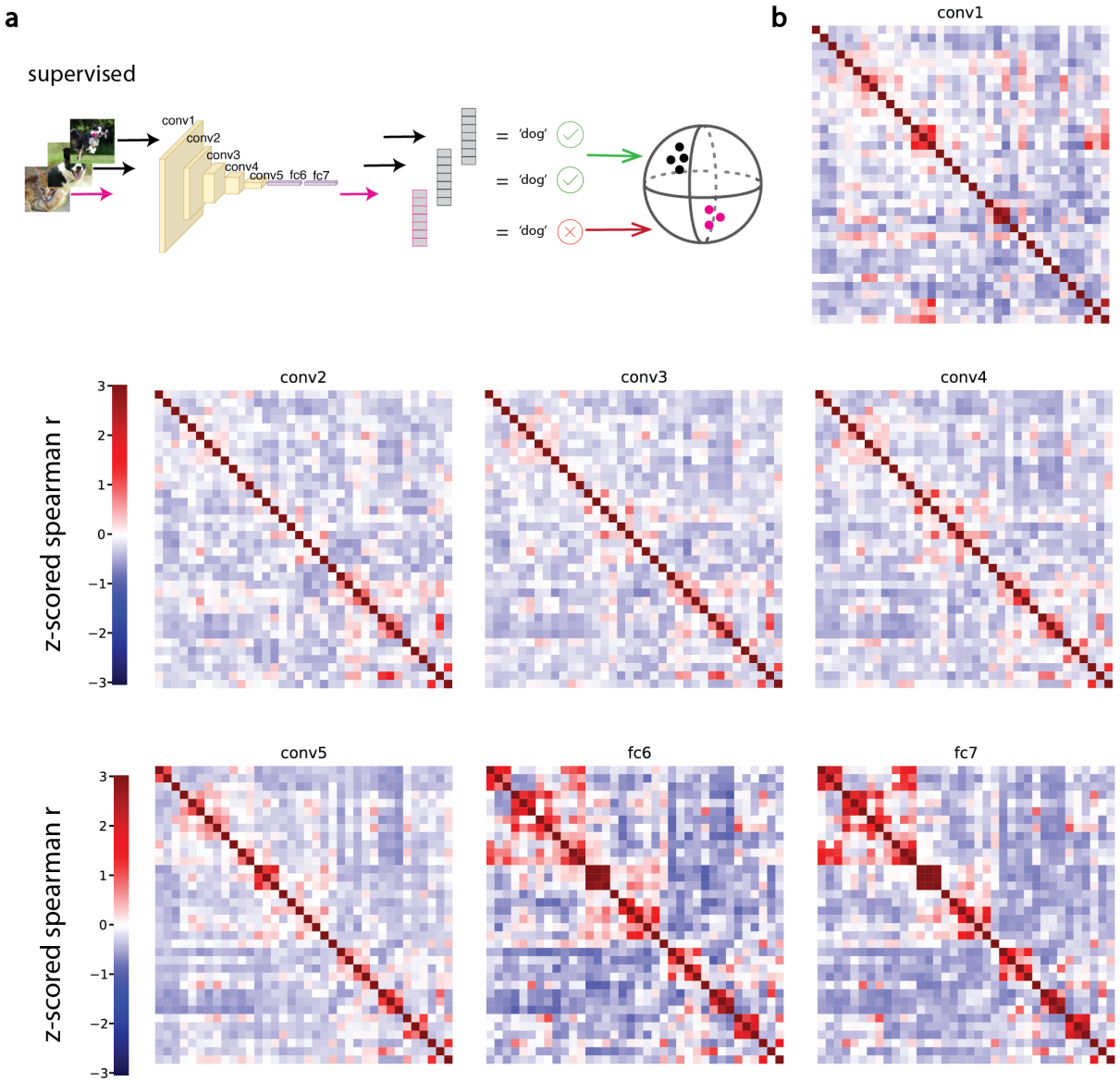
Supplementary Figure 4.3. Perceptual and semantic model RDMS. (a) Perceptual RDMs for the features of size, elongation, colour and compactness. The raw measurements of each feature were calculated for the 36 image stimuli, and pairwise distances entered as the distance value in the RDM. We plot the z-scored distances to aid comparison, as the raw values vary with feature. **(b)** Semantic feature RDMs. Raw values are plot to illustrate model definitions, but the appropriately weighted and normalised values were used when comparing to brain data. Exemplar RDM: defines within image similarities as 1 and between image similarities as 0. Category RDM: models generalisation across images within the same category by encoding within image similarities as 0, within category but across image similarities as 1 and across category similarities as 0. Animacy RDM: Encodes similarities within the groups of animate, inanimate small and inanimate big as 1 and across group similarities as 0. X and Y axes for all seven feature plots are identical, and are nested by tripartite class (animate, inanimate-small to inanimate-large), then category (4), and finally exemplar (3).



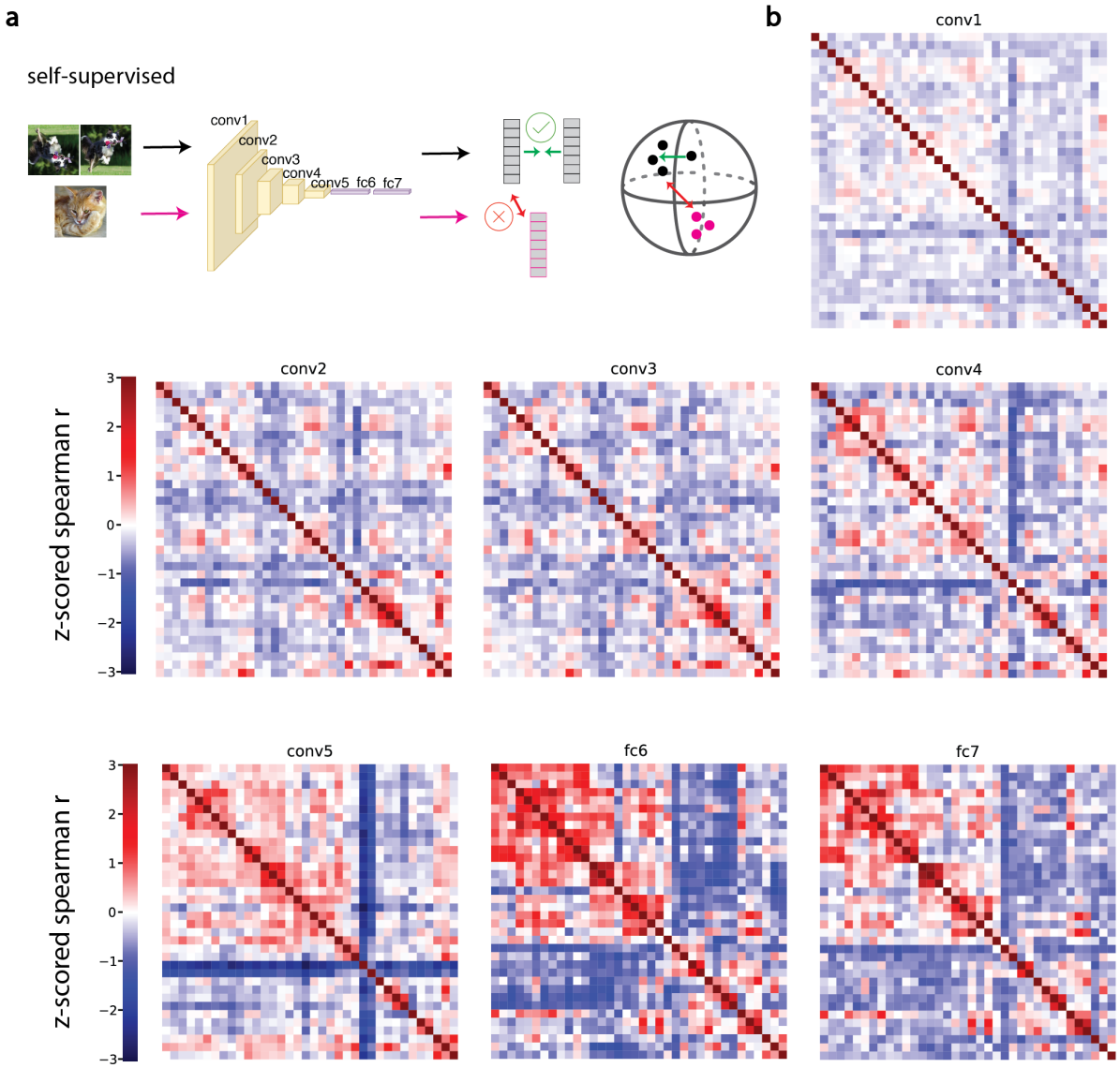
Supplementary Figure 4.4. Category level RDMs and word embedding model RDM. (a) VTC RDMs calculated using correlation as the distance metric between betas from a GLM at the 12-category level. Each row/column represents the average brain response across the 3 exemplars for each category. X and Y axes are identical and nested by tripartite class (3) and then category (4). We plot the z-scored spearman correlation values, as the raw correlation range varied in scale with age: 2 months: (-0.02,0.02), 9 months (-0.05,0.07), Adults (-0.08, 0.12). **(b)** The corresponding correlation based RDM for a word model trained on natural language, *word2vec*. We chose the 12 most similar words to our 12 visual categories that were in the model's corpus. For comparison, we plot the z-score of the raw spearman r values in the range of (0.002,0.66), exclusive of the diagonal which compares the same two words' activation vectors giving it a value of 1. **(c)** Correlation distributions, plot as a proportion of the noise ceiling, for each age-group/ROI to the *word2vec* RDM. **(d)** The corresponding plot for the multimodal visual-linguistic model *CLIP*. Raw correlations were in the range (0.62, 0.86). **(e)** The equivalent plot to (c), for the *CLIP* model.



Supplementary Figure 4.5. RDMs for untrained AlexNet. (a) AlexNet was initiated with random weights and had no exposure to visual stimuli. Layer wise activations were calculated in response to the 36 stimuli, while keeping weights frozen. **(b)** Untrained RDMs for each layer. X and Y axes are the 36 objects, nested by tripartite class (3), category (4) and exemplar (3). Pairwise correlations were calculated between activation functions in response to each stimulus. We plot the z-scored correlation to highlight representational content rather than strength, and to aid comparison to brain-derived RDMs. Raw correlation ranges: all diagonal values are 1, indicating within exemplar comparisons, off-diagonal ranges were conv1: (-0.1,0.5), conv2: (0.1,0.7), conv3: (0.3,0.9), conv4: (0.4,0.9), conv5: (0.4,0.9), fc6: (0.6,1) fc7: (0.7,1).



Supplementary Figure 4.6. RDMs for supervised AlexNet. (a) AlexNet was initialised with the pre-trained weights, learned during a supervised ImageNet recognition task. The model had access to all image labels during training. **(b)** Supervised RDMs for each layer, calculated using correlation between the activation vectors of the model in response to each stimulus image. X and Y axes are the 36 objects, nested by tripartite class, category and exemplar. Pairwise correlations were calculated between activation functions in response to each stimulus. We plot the z-scored correlation to highlight representational content rather than strength, and to aid comparison to brain-derived RDMs. Raw correlation ranges: all diagonal values are 1, off-diagonal values: conv1: (0,0.5), conv2: (0,0.3), conv3: (0,0.3), conv4: (0,0.3), conv5: (0,0.4), fc6: (-0.1, 0.7), fc7: (-0.1,0.8).



Supplementary Figure 4.7. RDMs for self-supervised AlexNet. (a) The self-supervised model was trained with Instance Protocol Contrastive Learning (https://github.com/harvard-visionlab/open_ipcl). The algorithm learned only from natural image structure by bringing together augmented versions of an image closer to an average "prototype" of the image in embedding space, while keeping the image's representation distinct from recently encountered images in memory. Activation vectors were calculated in response to each of the 36 images, while keeping training weights frozen. **(b)** Unsupervised RDMs for each layer, calculated using pairwise correlations between each image's activation vector. X and Y axes are the 36 objects, nested by tripartite class, category and exemplar. We plot the z-scored correlation to highlight representational content rather than strength, and to aid comparison to brain-derived RDMs. Raw correlation ranges: all diagonal values are 1, off-diagonal ranges: conv1: (0,0.3), conv2: (0,0.5), conv3: (0.1,0.5), conv4: (0.2,0.6), conv5: (0.8,0.9), fc6: (-0.1,0.8), fc7: (0,0.7).

Supplementary Table S4.1.

ROI	feature	age	diagonal included			diagonal excluded			
			Spearman *r*	95% CI (lower)	95% CI (upper)	Spearman *r*	95% CI (lower)	95% CI (upper)	
EVC	size	two month	0.514	0.503	0.525	0.446	0.434	0.457	
		nine month	0.593	0.577	0.609	0.523	0.506	0.541	
		adult	0.371	0.346	0.395	0.269	0.243	0.296	
	elongation	two month	0.336	0.326	0.347	0.238	0.227	0.249	
		nine month	0.367	0.353	0.38	0.261	0.247	0.275	
		adult	0.245	0.229	0.259	0.119	0.102	0.137	
	colour	two month	0.134	0.122	0.144	0.012	0.001	0.023	
		nine month	0.139	0.127	0.152	0.003	-0.01	0.017	*n.s.*
		adult	0.172	0.159	0.184	0.044	0.029	0.057	
	compactness	two month	0.218	0.207	0.229	0.11	0.099	0.122	
		nine month	0.269	0.255	0.283	0.152	0.136	0.167	
		adult	0.202	0.183	0.222	0.076	0.056	0.096	
	exemplar	two month	0.32	0.308	0.331	/	/	/	
		nine month	0.353	0.341	0.364	/	/	/	
		adult	0.35	0.337	0.361	/	/	/	
	category	two month	0.303	0.292	0.312	0.132	0.12	0.143	
		nine month	0.329	0.316	0.341	0.139	0.124	0.152	
		adult	0.366	0.349	0.382	0.19	0.172	0.209	
	animacy	two month	0.138	0.126	0.149	0.037	0.025	0.049	
		nine month	0.164	0.15	0.179	0.053	0.038	0.07	
		adult	0.329	0.306	0.352	0.241	0.216	0.265	
VTC	size	two month	0.523	0.514	0.532	0.449	0.439	0.458	
		nine month	0.525	0.511	0.537	0.452	0.438	0.467	
		adult	0.55	0.534	0.567	0.478	0.459	0.499	

elongation	two month	0.317	0.307	0.325	0.21	0.2	0.219	
	nine month	0.337	0.325	0.348	0.234	0.222	0.247	
	adult	0.335	0.322	0.349	0.226	0.21	0.243	
colour	two month	0.119	0.111	0.128	-0.019	-0.028	-0.01	
	nine month	0.128	0.118	0.137	*-0.007*	*-0.017*	*0.003*	*n.s.*
	adult	0.14	0.13	0.149	*0.002*	*-0.009*	*0.013*	*n.s.*
compactness	two month	0.272	0.263	0.28	0.159	0.15	0.169	
	nine month	0.254	0.243	0.264	0.141	0.129	0.152	
	adult	0.273	0.261	0.285	0.158	0.143	0.172	
exemplar	two month	0.342	0.333	0.349	/	/	/	
	nine month	0.336	0.327	0.345	/	/	/	
	adult	0.351	0.343	0.359	/	/	/	
category	two month	0.353	0.345	0.361	0.181	0.172	0.19	
	nine month	0.337	0.327	0.346	0.164	0.154	0.175	
	adult	0.374	0.362	0.387	0.203	0.189	0.217	
animacy	two month	0.243	0.234	0.253	0.145	0.135	0.155	
	nine month	0.28	0.267	0.293	0.19	0.175	0.203	
	adult	0.303	0.283	0.323	0.211	0.189	0.233	

Supplementary Table S4.2

ROI	feature	ages tested	diagonal included				diagonal excluded			
			Spearman *r*	95% CI (lower)	95% CI (upper)		Spearman *r*	95% CI (lower)	95% CI (upper)	
EVC	size	two vs nine	0.05	0.066	0.033		0.051	0.069	0.033	
		two vs adult	-0.112	-0.088	-0.135		-0.144	-0.117	-0.169	
		nine vs adult	-0.162	-0.138	-0.187		-0.195	-0.167	-0.224	
	elongation	two vs nine	0.015	0.029	0.002		0.012	0.026	-0.003	*n.s.*
		two vs adult	-0.072	-0.055	-0.088		-0.099	-0.081	-0.118	
		nine vs adult	-0.087	-0.07	-0.104		-0.11	-0.091	-0.129	
	colour	two vs nine	0	0.014	-0.013	*n.s.*	-0.008	0.006	-0.022	*n.s.*
		two vs adult	0.038	0.053	0.024		0.029	0.044	0.013	
		nine vs adult	0.038	0.054	0.024		0.037	0.053	0.021	
	compactness	two vs nine	0.035	0.05	0.02		0.031	0.047	0.013	
		two vs adult	-0.008	0.013	-0.027	*n.s.*	-0.027	-0.006	-0.048	
		nine vs adult	-0.043	-0.022	-0.064		-0.058	-0.035	-0.08	
	exemplar	two vs nine	0.017	0.032	0.003		/	/	/	
		two vs adult	0.036	0.05	0.021		/	/	/	
		nine vs adult	0.019	0.034	0.004		/	/	/	
	category	two vs nine	0.013	0.027	-0.001	*n.s.*	0.002	0.016	-0.013	*n.s.*
		two vs adult	0.065	0.082	0.047		0.056	0.075	0.037	
		nine vs adult	0.051	0.069	0.034		0.054	0.074	0.035	
	animacy	two vs nine	0.017	0.032	0.003		0.013	0.029	-0.002	*n.s.*
		two vs adult	0.174	0.197	0.151		0.183	0.207	0.158	
		nine vs adult	0.157	0.179	0.134		0.17	0.195	0.146	
VTC	size	two vs nine	0.001	0.016	-0.013	*n.s.*	0.003	0.019	-0.013	*n.s.*
		two vs adult	0.035	0.053	0.017		0.037	0.057	0.015	
		nine vs adult	0.034	0.053	0.015		0.033	0.056	0.01	

elongation	two vs nine	0.018	0.031	0.005		0.022	0.036	0.006	
	two vs adult	0.024	0.037	0.008		0.019	0.036	0.002	
	nine vs adult	0.006	0.021	-0.011	*n.s.*	-0.003	0.015	-0.022	*n.s.*
colour	two vs nine	0.007	0.019	-0.005	*n.s.*	0.011	0.023	-0.002	*n.s.*
	two vs adult	0.021	0.033	0.01		0.019	0.032	0.007	
	nine vs adult	0.014	0.025	0.002		0.008	0.022	-0.005	*n.s.*
compactness	two vs nine	-0.016	-0.003	-0.029		-0.016	-0.002	-0.03	
	two vs adult	0.007	0.02	-0.006	*n.s.*	0.002	0.016	-0.013	*n.s.*
	nine vs adult	0.023	0.038	0.009		0.018	0.035	0.001	
exemplar	two vs nine	-0.006	0.005	-0.016	*n.s.*	/	/	/	
	two vs adult	0.015	0.026	0.005		/	/	/	
	nine vs adult	0.021	0.032	0.01		/	/	/	
category	two vs nine	-0.014	-0.003	-0.025		-0.015	-0.002	-0.029	
	two vs adult	0.027	0.04	0.014		0.024	0.039	0.008	
	nine vs adult	0.041	0.055	0.028		0.039	0.055	0.024	
animacy	two vs nine	0.033	0.047	0.019		0.04	0.056	0.025	
	two vs adult	0.06	0.08	0.039		0.064	0.086	0.041	
	nine vs adult	0.027	0.049	0.006		0.024	0.047	0	*n.s.*

Supplementary Table S4.3

ROI	feature	age	diagonal included		
			Spearman *r*	95% CI (lower)	95% CI (upper)
EVC	exemplar	two month	0.08	0.067	0.09
		nine month	0.079	0.066	0.09
		adult	0.16	0.145	0.17
	category	two month	0.145	0.132	0.152
		nine month	0.147	0.131	0.156
		adult	0.229	0.21	0.243
	animacy	two month	0.041	0.03	0.051
		nine month	0.052	0.038	0.066
		adult	0.242	0.216	0.258
VTC	exemplar	two month	0.106	0.097	0.114
		nine month	0.095	0.083	0.103
		adult	0.103	0.09	0.115
	category	two month	0.203	0.192	0.209
		nine month	0.183	0.171	0.191
		adult	0.227	0.209	0.24
	animacy	two month	0.15	0.14	0.159
		nine month	0.194	0.178	0.205
		adult	0.221	0.197	0.24

Supplementary Table S4.4

ROI	feature	ages tested	diagonal included			
			Spearman *r*	95% CI (lower)	95% CI (upper)	
EVC	exemplar	two vs nine	-0.003	-0.019	0.012	*n.s.*
		two vs adult	-0.079	-0.097	-0.061	
		nine vs adult	-0.076	-0.093	-0.059	
	category	two vs nine	0.001	-0.013	0.016	*n.s.*
		two vs adult	-0.087	-0.106	-0.067	
		nine vs adult	-0.088	-0.103	-0.067	
	animacy	two vs nine	-0.009	-0.025	0.007	*n.s.*
		two vs adult	-0.197	-0.222	-0.174	
		nine vs adult	-0.188	-0.213	-0.164	
VTC	exemplar	two vs nine	0.104	-0.002	0.024	*n.s.*
		two vs adult	0.002	-0.014	0.019	*n.s.*
		nine vs adult	-0.023	-0.048	0	*n.s.*
	category	two vs nine	0.007	-0.005	0.02	*n.s.*
		two vs adult	-0.023	-0.041	-0.006	
		nine vs adult	-0.031	-0.048	-0.013	
	animacy	two vs nine	-0.048	-0.063	-0.032	
		two vs adult	-0.071	-0.094	-0.048	
		nine vs adult	-0.023	-0.048	0	

Supplementary Table S4.5

ROI	age	layer	diagonal included			diagonal excluded			
			Spearman *r*	95% CI (lower)	95% CI (upper)	Spearman *r*	95% CI (lower)	95% CI (upper)	
EVC	two month	conv1	0.044	0.032	0.054	-0.008	-0.018	0.002	*n.s.*
		conv2	0.124	0.111	0.134	-0.017	-0.028	-0.005	
		conv3	0.133	0.12	0.143	-0.009	-0.02	0.003	*n.s.*
		conv4	0.14	0.127	0.149	-0.003	-0.014	0.008	*n.s.*
		conv5	**0.148**	**0.135**	**0.157**	0.003	-0.007	0.014	*n.s.*
		fc6	0.136	0.124	0.147	-0.007	-0.017	0.004	*n.s.*
		fc7	0.143	0.13	0.153	-0.001	-0.011	0.011	*n.s.*
	nine month	conv1	0.022	0.009	0.036	-0.032	-0.043	-0.019	
		conv2	0.097	0.083	0.109	-0.054	-0.066	-0.04	
		conv3	0.127	0.111	0.137	-0.029	-0.042	-0.016	
		conv4	0.135	0.119	0.145	-0.022	-0.035	-0.01	
		conv5	0.138	0.123	0.148	-0.019	-0.032	-0.007	
		fc6	0.137	0.12	0.146	-0.022	-0.035	-0.011	
		fc7	**0.142**	**0.125**	**0.151**	**-0.017**	**-0.03**	**-0.006**	
	adult	conv1	0.051	0.041	0.063	0.002	-0.009	0.015	*n.s.*
		conv2	0.075	0.063	0.09	-0.056	-0.07	-0.041	
		conv3	0.087	0.074	0.1	-0.046	-0.061	-0.031	
		conv4	0.09	0.076	0.103	-0.047	-0.063	-0.031	
		conv5	0.091	0.077	0.104	-0.043	-0.059	-0.027	
		fc6	0.1	0.086	0.112	-0.038	-0.054	-0.023	
		fc7	**0.101**	**0.087**	**0.113**	-0.037	-0.052	-0.021	
VTC	two month	conv1	0.063	0.054	0.072	0.006	-0.003	0.014	*n.s.*

	conv2	0.121	0.111	0.131	-0.028	-0.036	-0.018	
	conv3	0.131	0.121	0.14	-0.02	-0.028	-0.011	
	conv4	0.14	0.13	0.149	-0.012	-0.02	-0.004	
	conv5	**0.144**	**0.134**	**0.152**	-0.009	-0.018	-0.001	
	fc6	0.131	0.121	0.14	-0.021	-0.029	-0.012	
	fc7	0.134	0.125	0.143	-0.018	-0.026	-0.01	
nine month	conv1	0.046	0.036	0.056	-0.008	-0.017	0.002	*n.s.*
	conv2	0.107	0.095	0.117	-0.038	-0.047	-0.028	
	conv3	0.123	0.112	0.133	-0.024	-0.032	-0.014	
	conv4	**0.129**	**0.118**	**0.139**	-0.019	-0.027	-0.01	
	conv5	0.128	0.117	0.137	-0.021	-0.029	-0.012	
	fc6	0.113	0.103	0.123	-0.034	-0.042	-0.024	
	fc7	0.115	0.105	0.125	-0.033	-0.04	-0.023	
adult	conv1	0.064	0.057	0.071	**0.026**	**0.015**	**0.035**	
	conv2	**0.085**	**0.077**	**0.093**	-0.049	-0.058	-0.037	
	conv3	0.078	0.069	0.088	-0.058	-0.067	-0.045	
	conv4	0.081	0.072	0.091	-0.056	-0.066	-0.043	
	conv5	0.084	0.075	0.094	-0.054	-0.063	-0.041	
	fc6	0.07	0.061	0.081	-0.065	-0.074	-0.051	
	fc7	0.077	0.068	0.087	-0.059	-0.068	-0.046	

Supplementary Table S4.6

ROI	age	layer	diagonal included			diagonal excluded		
			Spearman *r*	95% CI (lower)	95% CI (upper)	Spearman *r*	95% CI (lower)	95% CI (upper)
EVC	two month	conv1	0.184	0.172	0.193	0.115	0.104	0.125
		conv2	0.213	0.2	0.221	0.218	0.206	0.226
		conv3	0.25	0.237	0.259	0.207	0.194	0.216
		conv4	0.231	0.219	0.241	0.213	0.201	0.223
		conv5	0.206	0.196	0.216	**0.233**	**0.222**	**0.242**
		fc6	**0.255**	**0.242**	**0.265**	**0.233**	**0.222**	**0.242**
		fc7	0.231	0.218	0.24	0.251	0.24	0.26
	nine month	conv1	0.187	0.169	0.195	0.112	0.095	0.12
		conv2	0.221	0.203	0.229	0.227	0.208	0.234
		conv3	**0.308**	**0.289**	**0.318**	0.255	0.236	0.263
		conv4	0.289	0.271	0.3	0.265	0.249	0.275
		conv5	0.253	0.238	0.265	**0.284**	**0.267**	**0.294**
		fc6	0.275	0.258	0.287	0.252	0.235	0.263
		fc7	0.266	0.248	0.278	0.29	0.272	0.301
	adult	conv1	0.149	0.133	0.162	0.068	0.053	0.084
		conv2	0.208	0.191	0.221	0.194	0.176	0.207
		conv3	0.25	0.229	0.264	0.188	0.168	0.202
		conv4	0.245	0.22	0.263	0.202	0.181	0.22
		conv5	0.273	0.25	0.29	0.284	0.261	0.3
		fc6	**0.324**	**0.297**	**0.343**	0.284	0.261	0.304
		fc7	0.333	0.305	0.352	**0.325**	**0.301**	**0.344**
VTC	two month	conv1	0.147	0.136	0.155	0.079	0.07	0.087

	conv2	0.19	0.18	0.199	0.199	0.189	0.207
	conv3	0.296	0.285	0.304	0.243	0.234	0.252
	conv4	0.306	0.295	0.314	0.283	0.272	0.291
	conv5	0.293	0.283	0.301	**0.317**	**0.307**	**0.324**
	fc6	**0.322**	**0.311**	**0.331**	0.296	0.285	0.304
	fc7	0.296	0.286	0.306	0.316	0.306	0.325
nine month	conv1	0.151	0.139	0.159	0.084	0.074	0.093
	conv2	0.21	0.197	0.22	0.217	0.205	0.227
	conv3	0.319	0.306	0.328	0.266	0.255	0.275
	conv4	**0.337**	**0.326**	**0.346**	0.311	0.301	0.32
	conv5	0.313	0.301	0.321	0.335	0.324	0.343
	fc6	**0.337**	**0.324**	**0.345**	0.309	0.297	0.318
	fc7	0.327	0.313	0.335	**0.343**	**0.33**	**0.351**
adult	conv1	0.177	0.168	0.185	0.121	0.108	0.127
	conv2	0.281	0.267	0.291	0.272	0.255	0.281
	conv3	0.391	0.377	0.401	0.307	0.293	0.316
	conv4	0.409	0.395	0.419	0.347	0.331	0.357
	conv5	0.378	0.364	0.388	0.374	0.357	0.384
	fc6	**0.42**	**0.407**	**0.43**	0.378	0.361	0.389
	fc7	0.414	0.4	0.425	**0.404**	**0.386**	**0.416**

Supplementary Table S4.7

ROI	age	layer	diagonal included			diagonal excluded		
			Spearman *r*	95% CI (lower)	95% CI (upper)	Spearman *r*	95% CI (lower)	95% CI (upper)
EVC	two month	conv1	0.245	0.231	0.253	0.105	0.095	0.116
		conv2	0.231	0.218	0.239	0.077	0.067	0.087
		conv3	**0.281**	**0.268**	**0.289**	**0.119**	**0.108**	**0.129**
		conv4	0.277	0.265	0.285	0.116	0.105	0.127
		conv5	0.17	0.16	0.179	0.025	0.015	0.036
		fc6	0.219	0.208	0.23	0.113	0.103	0.124
		fc7	0.185	0.173	0.195	0.047	0.037	0.059
	nine month	conv1	0.258	0.24	0.264	0.105	0.089	0.114
		conv2	0.244	0.226	0.251	0.076	0.062	0.085
		conv3	0.316	0.294	0.324	0.135	0.117	0.143
		conv4	**0.328**	**0.309**	**0.336**	**0.145**	**0.129**	**0.155**
		conv5	0.211	0.196	0.221	0.044	0.031	0.056
		fc6	0.259	0.244	0.27	0.138	0.124	0.151
		fc7	0.2	0.187	0.212	0.047	0.035	0.06
	adult	conv1	0.135	0.121	0.146	0.015	0.001	0.027
		conv2	0.153	0.137	0.167	0.011	-0.006	0.024
		conv3	0.217	0.199	0.232	0.064	0.045	0.077
		conv4	0.25	0.229	0.264	0.087	0.07	0.1
		conv5	0.266	0.246	0.28	0.087	0.069	0.099
		fc6	**0.425**	**0.397**	**0.442**	**0.262**	**0.238**	**0.279**
		fc7	0.385	0.36	0.401	0.19	0.169	0.207
VTC	two month	conv1	0.203	0.193	0.21	0.064	0.055	0.072

	conv2	0.227	0.217	0.234	0.07	0.062	0.078
	conv3	0.319	0.308	0.327	0.145	0.136	0.154
	conv4	**0.329**	**0.318**	**0.336**	0.153	0.144	0.161
	conv5	0.225	0.216	0.233	0.065	0.056	0.072
	fc6	0.275	0.265	0.284	**0.155**	**0.147**	**0.165**
	fc7	0.201	0.192	0.211	0.053	0.046	0.063
nine month	conv1	0.206	0.196	0.215	0.068	0.059	0.077
	conv2	0.222	0.213	0.231	0.068	0.061	0.077
	conv3	0.318	0.309	0.329	0.147	0.139	0.157
	conv4	**0.333**	**0.324**	**0.342**	0.158	0.151	0.169
	conv5	0.234	0.224	0.243	0.072	0.065	0.082
	fc6	0.31	0.298	0.319	**0.189**	**0.179**	**0.199**
	fc7	0.25	0.238	0.26	0.097	0.088	0.108
adult	conv1	0.191	0.184	0.2	0.063	0.052	0.072
	conv2	0.235	0.225	0.243	0.056	0.047	0.068
	conv3	0.345	0.334	0.354	0.14	0.129	0.152
	conv4	0.363	0.351	0.371	0.155	0.144	0.166
	conv5	0.325	0.315	0.332	0.121	0.111	0.131
	fc6	**0.437**	**0.42**	**0.449**	**0.257**	**0.241**	**0.272**
	fc7	0.357	0.341	0.369	0.156	0.142	0.171