Visual semantic processing in primate frontotemporal cognitive network and machines

Morteza Mooziri[1,4], Meysam Zare[1,4], Milad Qolami[1], Farideh Shakerian[1], Mohammad Javan[1,2], Mohammad-Reza A. Dehaqani[3,5]*

1. Department of Brain and Cognitive Sciences, Cell Science Research Center, Royan Institute for Stem Cell Biology and Technology, ACECR, Tehran, Iran.
2. Department of Physiology, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran.
3. Cognitive Systems Laboratory, Control and Intelligent Processing Center of Excellence (CIPCE), School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran.
4. Equal contribution.
5. Lead contact.
* Correspondence: dehaqani@ut.ac.ir

Abstract

Visual processing includes extraction of physical features and abstract information. Unlike physical features, we have limited knowledge on how the primate brain extracts semantic information of the visual input. Here, we recorded the neuronal activities of inferior temporal (ITC) and ventrolateral prefrontal (PFC) cortices, while macaque monkeys viewed a series of natural and artificial visual stimuli. We found that mid-level semantic information, e.g., face vs. body distinction, is processed by ITC population, while high-level abstractions, e.g., animate vs. inanimate, are solved in the PFC. Additionally, bi-directional information flow between the two neuronal populations suggests that these distinct object features are being transferred between the two regions. Also, we show that the encoding axes corresponding to the two information types are orthogonally aligned in the PFC neural space in the early phase of exposure to the object, suggesting that separate neural subspaces of the PFC are involved in processing different attributes of the visual input. Furthermore, we found a progressive abstraction regime along the feed-forward direction in state-of-the-art deep learning models of vision. These results provide insights on critical questions in the field, including the purpose of information processing by ITC as well as the perceptual sequence and independence of object features.

Introduction

Primates possess a remarkable ability to recognize and categorize objects rapidly and accurately, a function essential for survival, navigation through complex environments, and

social interactions [1-3]. Psychophysical studies have long established that object recognition is not a singular, instantaneous process, but rather unfolds across multiple stages [4,5]. Early in this sequence, the brain extracts basic visual features—such as shapes, textures, and basic configurations—that allow for rapid, coarse distinctions between objects [3,4,6]. However, it is the ability to categorize objects at more abstract, conceptual levels, that makes the uniqueness of this system. These high-level perceptual judgments, often delayed relative to recognition of basic visual features, point toward a hierarchical and interactive system of visual processing, where distinct neural circuits contribute to progressively refine object representations [3,4].

Behaviorally, it is shown that the brain can swiftly extract mid-level semantic features, such as distinguishing faces from bodies, while the formation of abstract categories, like animate versus inanimate, typically requires more time and likely engages additional cognitive resources [4,5]. It is not clear whether this basic level precession is critical for perception of higher-level abstractions or it is just simply indicative of a temporal ordering due to evolutionary importance or experience [4]. At the neural level, this hierarchy is thought to involve the inferior temporal (ITC) and the prefrontal (PFC) cortices, two regions critical for object representation and decision-making [2,7-11]. The ITC has long been associated with the encoding of physical features, which are necessary for differentiating visually similar objects [12-15], whereas the PFC has been implicated in abstract, context-dependent categorizations [10,16]. However, we are still lacking theories that provide a holistic understanding of how semantic information is derived from visual input.

Despite advances in understanding the roles of the ITC-PFC circuit in object recognition, key questions remain unresolved. A major one is “how these regions interact to support semantic processing?” Early perception of mid-level semantic information is believed to stem from basic level advantage of ITC [7]. In this view, mid-level semantic information is solved by ITC, which is probably fed forward to higher order cognitive areas like PFC [7]. In contrast, an alternative hypothesis proposes that ITC conveys primarily visual, non-semantic information to downstream regions, which are responsible for making abstraction by incorporating these physical features and probably benefiting prior knowledge [9,10,12-20]. Overall, although current models suggest a bidirectional flow of information between ITC and PFC [21,22], the content of information in each of these directions is a matter of debate. To our knowledge, there is no experimental evidence on how abstract representations emerge in this circuit.

Here, we attempted to address how abstract semantic information is represented in the primate brain. To that aim, we recorded the neuronal activities of macaque ITC and PFC, while viewing a series of visual objects. We found that, mid-level semantic information is represented earlier in both ITC and PFC, compared to high-level abstractions. Interestingly, we show that this information is solved by the recurrent circuitries in the ITC population, while high-level categorization is performed by PFC. Furthermore, we observed that the functional connectivity of two neuronal populations is enhanced following exposure to the visual object, with bi-directional transfer of information. Additionally, orthogonal alignment of the high-level and mid-level encoding axes in the PFC neural space suggests a mechanism for the primates’ independent perception of different features of the same visual input. In line with

these findings, we also observed progressive abstraction along the feed-forward direction in the state-of-the-art deep learning models of vision. These findings will provide insights on several existing questions of the primate visual system (see Discussion).

Results

Neurophysiology and behavior

We simultaneously recorded the neuronal activity of ITC and PFC while macaque monkeys viewed a series of visual stimuli containing isolated grayscale natural and artificial objects of several categories (Fig. 1A). Since spatial frequency profile, as a basic visual feature, can affect category processing in ITC [23], we matched it between different categories of our stimulus set (see Methods). After adjustment, there was no significant difference between categories in none of the spatial frequencies (all comparisons showed a $p > 0.05$; Supplementary Fig. 1). The stimulus set was designed to elicit both mid-level and high-level semantic distinctions, as it spanned several conceptual categories, including animate versus inanimate objects and faces versus bodies. This vastness of object categories is critical for object recognition tasks, especially in this case, where we wanted to cover a broad range of categories for each class of high-level abstractions. To achieve this goal, we leveraged an experimental setting that allows us to record the neuronal responses in thousands of trials in a technically feasible amount of time, namely rapid serial visual presentation (RSVP); in every trial, the stimulus appeared in the central 7^{o} of the animal's visual field for 80 ms, which was followed by 400 ms of blank screen (Fig. 1A). The animal received a juice reward in random periods ranging 2-4 secs for continuous fixation at the center of the screen.

Fig. 1B shows the hypothetical semantic hierarchy, similar to a previous study [7]. At the highest level of abstraction in this hierarchy, stimuli are broadly divided into two primary categories: animate and inanimate objects. Below this level, that is mid-level semantic, the stimuli can be either primate or non-primate and face or body for the category of animates and artificial or natural for inanimates. Importantly, since some of these classifications are related to human-level knowledge, and might not necessarily be perceived by macaques, we focused on animate vs. inanimate distinction for high-level abstraction and face vs. body distinction for mid-level abstraction, throughout the rest of this study. Using this clearly defined stimulus set, we aimed to disentangle the temporal dynamics of object representation in the ITC-PFC circuit, exploring how these regions process different levels of semantic and conceptual information.

Basic level advantage of ITC and PFC

Fig. 1C-F presents sample units that show systematic selectivity for faces (Fig. 1C,D, upper panels), bodies (Fig. 1C,D, lower panels), animates (Fig. 1E,F, upper panels), and inanimates (Fig. 1E,F, lower panels) in the two regions (Fig. 1C,E for ITC and Fig. 1D,F for PFC). First, we sought to figure out the temporal dynamics of semantic representation in each of these

regions. For that, we trained linear classifiers, specifically using linear discriminant analysis (LDA), to distinguish faces from bodies and animates from inanimate objects using the population neuronal activity of each region (Fig. 1G,I). We observed that mid-level semantic information was represented earlier than high-level information in ITC (median ± SD in ms; onset times: mid-level = 83.00 ± 8.80, high-level = 82.00 ± 14.44, p = 0.002; peak times: mid-level = 103.50 ± 7.62, high-level = 182.00 ± 19.40, p < 1e-4; permutation test; Fig. 1G,H), similar to our previous study [7]. Furthermore, this early representation of mid-level information was also observed in PFC population activity (median ± SD in ms; onset times: mid-level = 83.00 ± 12.27, high-level = 112.00 ± 20.12, p < 1e-4; peak times: mid-level = 115.00 ± 19.70, high-level = 149.00 ± 20.81, p < 1e-4; permutation test; Fig. 1I,J). These results are consistent with previous findings on early representation of mid-level information in ITC [7]. Additionally, when considering PFC as the main brain area that engages in psychophysical tasks, where the subject must provide perceptual responses [16], these observations provide the rationale for faster categorization of basic level information at the behavioral or perceptual level, in line with previous reports [4,5].

ITC and PFC process different semantic attributes of the visual input

We were curious to figure out how the brain solves these two different levels of information, i.e., mid- and high-level semantics. For that, we need to consider that perceptual information, especially in the context of visual object's feature recognition, is not necessarily reflected in onset or peak times alone [4]. In this view, it is the speed of accumulation of perceptual evidence over time that defines the time-course of information representation [4,24-26], which is clearly defined in exemplar-based random walk model (Fig. 2A; see ref.[4] for a thorough discussion on this). To probe the issue, we used a metric aiming to measure the speed of information solution, not only representation, for a given problem in a brain area, namely solution time, which is defined as the difference of onset and peak times of information representation (Fig. 2B). Since it does not directly depend on any point estimate of representational times, it can capture the time required for a problem to be fully represented in a region, which can inversely show the ability of solution (Fig. 2B). Therefore, the combination of solution time with either one of onset or peak times can reliably point to resolution of information.

We looked at how these two levels of information evolve in the ITC-PFC neural circuit (Fig. 2C,G) using representational similarity analysis (RSA; see Methods) [27]. For that, we first created the ground truth representational similarity matrices (RSMs) for face vs. body (Fig. 2C) and animate vs. inanimate (Fig. 2G) categorizations. Next, instantaneous RSMs were derived from ITC and PFC population activities, and their similarities were computed to the related ground truth RSMs (Fig. 2C,G; see Methods). Despite negligible onset difference (median ± SD in ms; ITC = 81.00 ± 5.64, PFC = 80.00 ± 11.65, p = 0.004, permutation test; Fig. 2C,D), the mid-level information peaked earlier (median ± SD in ms; ITC = 135.00 ± 21.73, PFC = 157.00 ± 7.51, p < 1e-4, permutation test; Fig. 2C,E) and was solved faster (median ± SD in ms; ITC = 57.00 ± 22.12, PFC = 78.00 ± 12.44, p < 1e-4, permutation test; Fig. 2C,F) in the ITC, compared to PFC, which aligns well with previous evidence [7]. This is in line with a major theory of ITC processing, which states that ITC makes the semantic perception of a face in macaques [28-30] and humans

[30] (see Discussion). On the other hand, this relation was reversed in the PFC population activity; while high-level representations initiated later in the PFC (median ± SD in ms; ITC = 104.00 ± 15.62, PFC = 131.00 ± 21.17, p < 1e-4, permutation test; Fig. 2G,H), they reached full representation earlier than the ITC (median ± SD in ms; ITC = 195.00 ± 15.59, PFC = 167.00 ± 6.65, p < 1e-4, permutation test; Fig. 2G,I). These results point to greater capability of PFC population in solving the high-level abstractions, i.e., shorter solution times in PFC compared to ITC (median ± SD in ms; ITC = 87.50 ± 23.02, PFC = 35.00 ± 21.35, p < 1e-4, permutation test; Fig. 2G,J). This observation can fit in the vast body of literature depicting the roles of PFC for high-level categorizations and perceptions [9,10,16,19,31] (see Discussion). This notion also explains that since mid-level information is received in the feed-forward direction, they appear earlier in the population activity of PFC (i.e., earlier onset and peak times), compared to high-level information that does not exist in the feed-forward information (Fig. 1I,J), while PFC does not solve this information (Fig. 2G-J).

Confirmation of ITC and PFC roles in visual object's feature recognition with encoding models

By visual inspection, objects in the two categories at both semantic levels look different in terms of physical features. Regardless semantic processing, either region might be involved in processing of physical features [12-15,20,32]. Therefore, we quantified several low-level physical features of each object, namely circularity, elongation, spikiness, contrast, luminance, object area as well as the first and second principal components of the image in a grayscale (see Methods). Supplementary Fig. 2A shows the 20 most circular objects, which contains a greater number of inanimates; also, there are some faces, but no bodies. On the other hand, the 20 most spiky objects are predominantly animate (Supplementary Fig. 2B), with several bodies and no faces. Observing this striking difference in objects' physical features prompted us to search for the isolated semantic processing in the circuit; thus, we asked whether the temporal evolution of neural response could be explained by statistical regularities of objects. For that, we used generalized linear models (GLMs) to predict the instantaneous neural response using a set of semantic properties and the mentioned physical features of each object. Once the models were formed, we computed each object's residual for either level of semantic information, similar to a previous study [33] (see the full list of regressors and the implementation details in Methods).

Fig. 3A-D represents object residuals of sample units with systematic selectivity for faces (Fig. 1A,B, upper panels), bodies (Fig. 1A,B, lower panels), animates (Fig. 1C,D, upper panels), and inanimates (Fig. 1C,D, lower panels) in the two regions (Fig. 1C,E for ITC and Fig. 1D,F for PFC). As evidenced in Fig. 3A-D, some neurons in a given neuronal population increase/decrease their activity in category-selective manner. Subsequently, these object residuals, that near-exclusively show the contribution of a specific feature in forming the overall observed neural response, entered the RSA procedure (as in Fig. 2C-J). Interestingly the results were similar to the earlier observations. Specifically, face-body separation initiated earlier in PFC (median ± SD in ms; ITC = 80.00 ± 8.98, PFC = 75.50 ± 10.66, p < 1e-4, permutation test; Fig. 3E,F), while it peaked earlier (median ± SD in ms; ITC = 129.00 ± 21.98, PFC = 163.00 ± 12.99, p < 1e-4,

permutation test; Fig. 3E,G) and was solved faster (median ± SD in ms; ITC = 55.00 ± 23.46, PFC = 88.00 ± 16.01, $p < 1e\text{-}4$, permutation test; Fig. 3E,H) in the ITC population. Also, animacy information was has an earlier in onset in ITC (median ± SD in ms; ITC = 136.00 ± 40.00, PFC = 148.00 ± 39.92, $p = 0.04$, permutation test; Fig. 3I,J) as well as earlier peak representation (median ± SD in ms; ITC = 189.00 ± 32.37, PFC = 168.50 ± 39.28, $p = 0.07$, permutation test; Fig. 3I,K) and shorter solution time (median ± SD in ms; ITC = 49.00 ± 40.81, PFC = 23.00 ± 44.92, $p = 0.0007$, permutation test; Fig. 3I,L) in the PFC population. Here, we ruled-out the possible confounding effects due to several low-level visual features; spatial frequency profile and color were matched when preparing the stimuli, and the above-mentioned features are accounted for analytically. Needless to say, other low-level as well as mid- and high-level visual features are yet untouched, which suggests cautious interpretation of the results. However, overall, we provide near-confirmatory evidence that exclusive evaluation of semantic features further emphasizes the roles of ITC for processing mid-level rapidly-categorized semantic information and PFC for high-level perceptually-advanced abstractions in primate brain (see Discussion).

ITC and PFC populations transfer information for visual processing

To this point, we used the ITC and PFC population activities to find the evolution of semantic information in the two areas. As both regions are involved in object categorization and recognition tasks [1,2,7,8,11,12,19,21,30,32], and they dynamically transfer information to aid object recognition [21,34], we tried to see if the two neuronal populations are functionally connected while unfolding visual object. To this aim, similar to a previous study [35], we applied canonical correlation analysis (CCA) to the population neural representations. Theoretically, considering each area as a neural space where every dimension of this space is defined by the activity of a recorded neuron, CCA tries to find the dimensions of the two neural spaces, here ITC and PFC, that the activity in those dimensions will be maximally correlated (see Methods; Fig. 4A). Interestingly, we observed a pronounced and prolonged enhancement of the functional connectivity between ITC and PFC upon exposure to the object (statistical significance is measured by one-tailed permutation testing compared to baseline and is depicted in the Fig. 4B; thick and thin lines denote $p < 0.001$ and $p < 0.05$, respectively), potentially providing a route for inter-areal transfer of information.

If the two neuronal populations are connected, one will expect that they also share their representations. There is evidence that information flows from sensory regions, like ITC, to associative areas, like PFC, in the feed-forward direction [11,34,36-38]. Additionally, the prefrontal information is fed back to sensory systems in various conditions, including during object recognition [21,34,39,40]. Therefore, we implemented a previously defined approach similar to Granger causality, which measures how much a time-series is predictable from the past of another [41-43]. In short, we tried to predict a region's RSMs from earlier RSMs of the same or both regions over time. Specifically, at each timepoint t, we fitted regression models to predict the RSM_t from past RSMs of the same area or both areas; a reduction in the error term when using RSMs of both areas, suggests that the present moment population activity of a brain area (RSM_t) is predictable from the past of another (see Methods; Fig. 4C). Fig. 4D illustrates

the time-resolved dynamics of information transfer in the ITC-to-PFC and PFC-to-ITC directions. There is bidirectional information flow between ITC and PFC after stimulus onset. Importantly, the feed-forward transfer peaked earlier than the feedback (median ± SD in ms; ITC-to-PFC = 126.00 ± 10.32, PFC-to-ITC = 146.00 ± 21.07, $p < 1e\text{-}4$, permutation test; Fig. 4D, inlet panel), which is a significant observation due to the following: 1) PFC feedback to ITC is crucial for the late-resolved object information [21,22,34], and 2) animacy level information is solved/extracted at a later phase of visual object processing (for this, see Fig. 1G,H and ref.[7]). Therefore, this series of evidence strongly suggests that the feed-forward and feedback directions contain mid- and high-level information, respectively.

Semantic processing in ITC and PFC microcircuitries

Next, we tried to explore each region's neural architecture. To that aim, we used time-time decoding (TTD) to find the generalization of either representation across time in each region [44,45]. For a given timepoint, TTD measures how similar the current moment representation is to all other timepoints, giving valuable information about the functional architecture of the microcircuitry [44,45]. Interestingly, the TTD results of both semantic levels showed an off-diagonal generalization pattern in the two regions (statistical significance is measured by permutation testing and is depicted in the Fig. 5A,B; pixels with a color other than black, showed a $p < 0.05$ compared to baseline). This pattern suggests a functional structure containing recurrent interactions [44-46]. Earlier in this manuscript, we described the theory of accumulation of perceptual evidence over time as a competitive hypothesis for cognitive mechanisms of object feature processing and suggested that solution time can be an appropriate tool to approach the challenge (see Fig. 2A,B); presence of recurrent interactions in ITC and PFC for these two categorization tasks further strengthens the rationale behind these theoretical conceptualizations.

Up to now, we have shown that ITC and PFC have distinct, yet complementary, roles for processing object features; however, a significant question remains unresolved: how does this network support our perceptual independence for the two information types, i.e., mid- and high-level semantics? At the behavioral level, one can recognize whether a face (Fig. 5C, i vs. ii) or a body (Fig. 5C, iii vs. iv) is animate or not; at the same time, our face perception is intact when seeing both animate (Fig. 5C, i vs. iii) and animate (Fig. 5C, ii vs. iv) objects. Therefore, geometrically, we can consider that in the perceptual space, there are two orthogonal axes for these two categorization tasks. Next, we sought to find the neural underpinnings of this behavior in the ITC and PFC populations. Specifically, first, we tried to define the encoding axis for each categorization task in each region at every timepoint; subsequently, we measured the instantaneous cosine of the angle between the two axes in each region. If the two representations are independent in a neural space, the cosine of the angle between them will be zero, i.e., cosine of 90^o (Fig. 5D). To do this, we used LDA, as a linear supervised dimensionality reduction method similar to a previous study [47] (see Methods); LDA finds an encoding axis that maximizes the distance between the two groups of the data, while minimizing the within-group distances. Fig. 5E, upper panel demonstrates the temporal dynamics of alignment between the two encoding axes (for face-body and animacy) in ITC and

PFC. Interestingly, at the early phase of processing, the two axes are orthogonally aligned in the PFC neural space (statistical significance is measured by permutation testing against a uniform distribution with a mean at 0, i.e., U(-1,1), which is the null hypothesis for circular data; Fig. 5E, lower panel). Subsequently, at the late phase of processing, the same alignment appears in the ITC population (Fig. 5E). These observations point to similar encoding mechanisms in ITC and PFC, however with different timings and probably distinct purposes. In the early phase of processing when PFC is receiving the mid-level information from ITC, the configuration of its processing modules is set in a mode to represent the incoming feed-forward information and the in-run local processes in separate subspaces of the same neural space. This architecture can have profound neural and behavioral benefits, such as increasing the computational efficiency of a limited neural resource, reducing the chance of interference between parallel processing mechanisms, and providing perceptual independence. Additionally, the late phase independence in the ITC population could be analogous to how the PFC feedback (see Fig. 4D and refs.[21,34] for PFC effects on the late phase of object processing in ITC) interacts with the concurrent sensory input processing during working memory/attention [39,40] (see ref.[39] for a thorough discussion on the later issue).

Progressive abstraction along the feed-forward direction in deep learning models of vision

Thus far, we showed that as the visual input goes forward in the primate brain, more abstract information is derived from it. Subsequently, we asked “is this phenomenon only a feature of biological vision?” In other words, are deeper layers of artificial visual systems also more sensitive to more abstract information? To answer that, we tried to explore how semantic information is represented in state-of-the-art deep learning models of vision. We studied the models pretrained for object recognition task on various image datasets [48-56] (See Methods). Specifically, we studied CORnet-S, as the representative for CORnet family of networks [48], AlexNet [49], SqueezeNet [50], ResNet [51], DenseNet [52], Inceptionv3 [53], EfficientNet [54], VGG-16 [55], and MobileNet [56]. In each network, we extracted the activations of the layers of interest for our stimulus set (Fig. 1B), through the forward propagation, from which the layer-wise RSMs were constructed and were compared to the ground truth expectations (see Methods). We observed that, generally, as the input image reaches the deeper layers of the network, stronger abstract representations appear (statistical significance is measured by permutation testing, comparing every pair of layers; Fig. 6A,B). Importantly, this effect was true for both levels of abstractions, i.e., mid- and high-level semanticness. Also, VGG-16 showed the highest representational similarity to ground truth for both levels (mean ± SD of similarity; face-body = 0.62 ± 0.02, animate-inanimate = 0.21 ± 0.03; Fig. 6C,D). These results imply that, as we observed in primate brain, more advanced layers of the networks are relatively specialized for more semantic/abstract feature extractions.

Unlike many other networks of object recognition, CORnet family models are designed based on the primate visual system, with modules corresponding to primate cortical areas, including V1, V2, V4, and ITC [48]. In fact, for the above results (Fig. 6A,B), we used the activations of the output layer of these four modules in the CORnet-S. CORnet-S was chosen since it has the highest behavioral and ITC neural predictivity in the family [48]. Next, we were curious to see

how much brain-like are these representations. Specifically, we asked "are these representations similar to those formed by the brain for each recognition task?" To that aim, we used the CORnet-S-extracted representation from the layer corresponding to ITC output, where the semantic information culminates, as the new ground truth, for each case of face vs. body and animate vs. inanimate distinction (see Methods). Supplementary Fig. 3A,C depicts the similarity of CORnet-S ITC layer population representation with the ITC- and PFC-derived representations for both cases over time. We found that CORnet-S ITC layer information was more similar to primate brain's ITC for both face-body (median ± SD of similarity; ITC = 0.44 ± 0.02, PFC = 0.36 ± 0.03, $p < 1e-3$, permutation test; Supplementary Fig. 3B) and animate-inanimate (median ± SD of similarity; ITC = 0.18 ± 0.02, PFC = 0.15 ± 0.02, $p < 1e-3$, permutation test; Supplementary Fig. 3D) separations. Subsequently, we tried to expand the same idea to the other networks; here, we also added CORnet-Z, CORnet-RT, and VGG-19. We performed the same comparison (as in Supplementary Fig. 3) while every time using the latest layer of each of these networks (as depicted in Fig. 6A,B) as the ground truth matrix. Fig. 6E,G demonstrates the temporal evolution for the similarity of ITC and PFC mid- (Fig. 6E) and high-level (Fig. 6G) semantic representations to those of deep models. We found that the primate ITC-derived representations were more similar to these networks, compared to PFC representations, for both face-body (mean ± SD of similarity; NN~ITC = 0.46 ± 0.05, NN~PFC = 0.35 ± 0.04, n = 12 networks, p = 0.0005, one-sample Wilcoxon signed rank test from a theoretical null value of 0 for similarity difference; Fig. 6F) and animate-inanimate (mean ± SD of similarity; NN~ITC = 0.21 ± 0.04, NN~PFC = 0.16 ± 0.02, n = 12 networks, p = 0.0005, one-sample Wilcoxon signed rank test from a theoretical null value of 0 for similarity difference; Fig. 6H) cases. Also, VGG-16 showed highest similarity to ITC population activity (mean ± SD of similarity = 0.52 ± 0.02; Fig. 6I, left panel), while AlextNet was the most similar network to PFC representation (mean ± SD of similarity = 0.42 ± 0.03; Fig. 6I, right panel) for face-body discrimination. For animacy separation, AlexNet was the most similar model to ITC (mean ± SD of similarity = 0.30 ± 0.02; Fig. 6J, left panel) and MobileNet formed the most PFC-like (mean ± SD of similarity = 0.20 ± 0.03; Fig. 6J, right panel) representation. Overall, we can infer that in line with previous reports [21,22], the visual function of PFC, which is subject to growing interest [32], is most probably absent, at least partially, in currently available deep models of vision. Also, while VGG-16 has more visual cortex-like properties, which is similar to a previous study [57], AlexNet and MobileNet produce more abstract and PFC-like representations.

Discussion

Here, we showed that the primate brain solves mid-level semantic information in higher visual cortex during feed-forward pathway, while the more abstract information remains to be extracted by more cognitive areas of neocortex, which then feedback a copy of that information to upstream regions. Our work addresses critical gaps in the literature by linking behavioral evidence of hierarchical visual processing to the neural circuits that mediate these functions. While previous studies have focused on either the ITC or PFC in isolation, our

investigation of their interaction over time will provide a more comprehensive view of how the primate brain integrates sensory information and abstract knowledge to achieve robust, flexible object recognition. We also show that this progressive abstraction regime can be generalized to artificial models of vision.

The purpose of information processing by ITC is strongly controversial [12-15,28-30]. While the vast body of literature in humans [30] and monkeys [28-30] proposes that ITC makes the semantic perception of certain evolutionarily important attributes, such as being a face or a body, there is an alternative theory suggesting that ITC processes physical, and not semantic, aspects of the visual input [12-15,20]; the latter, suggests that these visual properties are combined by more cognitively developed areas, like PFC, to form abstract perceptions [8,16]. The observation that mid-level information is represented earlier in ITC and travels in the forward direction to PFC, is consistent with the former theory. Furthermore, there is evidence that forward propagation is not enough for solving object recognition [22]. On the other hand, PFC has established roles in more cognitively advanced behaviors, which includes abstractions [9,10,16,31]. Also, PFC sends feedback signals to ITC which helps most for processing late-stage difficult-to-recognize objects [21]. Here, we show that PFC precedes ITC for representing high-level semantic information, and feedbacks this representation to visual cortex, which fits quite well in the mentioned literature. Therefore, we suggest that the dynamic interactions between ITC and PFC are required for a complete perception of visual input.

From another perspective, our results explain the behavioral observations on perceptual sequence of object features [4,5]. Object recognition is not a one-stage process, but rather a sequential one [4,5]. Certain features of an object are recognized earlier, while others take longer to be perceived [4,5]; the more abstract, such as animate or not, and more memory-dependent, like identity, attributes are typically processed later, while salient characteristics are detected rapidly [4,5]. Parallel to behavior, the same sequence is represented in ITC population activity [7]. While the early representation of mid-level information in ITC is a wide belief [7,28-30] and solves part of the behavioral hierarchy [7], the late perception of more abstract information remained unresolved. Considering the PFC as the area responsible for this degree of abstraction, fills the mentioned gap. It is reasonable to assume that this delayed perception could be due to the time required for information to reach PFC and the internal cognitive processing mechanisms within PFC to solve the problem.

Is this sequential progressive abstraction only a property of biological vision? Computer vision systems are far less capable than primates for categorization and image recognition tasks [58-60]; besides the etiological bases of these phenomena, we are also lacking methodological knowledge to improve their performance. One reasonable approach to construct efficient networks for such problems is to simulate biological vision, for which primate visual system is a remarkable candidate [48]. But a major problem in this case would be our very limited knowledge of the primate vision itself. Object recognition has long been attributed to the ventral visual stream [2,11], while a number of studies point to substantial roles of PFC as a major contributor to these processes [8,19,21]. Interestingly in this case, the roles of PFC become crucial for more difficult categorization problems [21], which could most probably overlap with the situations in which current deep models of vision fail. With the data presented here, we

suggest that since obviously PFC activity is crucial for object recognition, semantic processing, and abstract categorizations in primates [8,19,21], considering a stage for accomplishing PFC duties will probably improve the performance of computer vision systems.

In conclusion, we provide mechanistic insights for why and how the collaboration of visual and prefrontal cortices is required to form robust semantic perceptions of the visual input in primates. These results suggest that recurrent neural processes in ITC and PFC solve diverse attributes of the visual input and explain several previously introduced behavioral observations. From a general perspective, these results lay the foundations for a more networked view of visual processing, contrary to the traditional modular processing theories. Also, we suggest approaches to improve the artificial visual systems in object recognition tasks.

Methods

Animals and surgery

Two male rhesus macaque monkeys (monkey F/V; Macaca mulatta, weight: 9.4/8.8 kg, age: 10/9 years old) entered the study. Experiments were performed in accordance with the National Institutes of Health Guide for the Care and Use of Laboratory Animals and were approved by the internal ethics committee at Royan Institute (code: ---). In the beginning, monkeys underwent MRI imaging to help design recording chambers and head posts, which were subsequently implanted through surgery. The head post was located midline and monkey F/V had one recording chamber on left/right hemisphere. In a second surgery, the craniotomy was performed over the area covering both ITC and PFC. Finally, a CT scan was acquired to help correctly localize the regions of interest, in combination with the MRI.

Stimuli, task, and behavior

Visual stimuli were isolated natural and artificial objects in grayscale and were shown on gray background (Fig. 1A,B). The stimulus set comprised of several basic categories required to capture the diversity of real-world objects underneath the most abstract level, that is animacy (Fig. 1B). Specifically, it contained faces and bodies of humans and monkeys, four-limb animals, reptiles, fishes, birds, and insects for the animate category and flowers, fruits, chairs, cars, houses, clocks, and tools as inanimates. Spatial frequency profile was matched among different categories, using SHINE toolbox [61].

Monkeys were trained to perform a fixation task, to receive juice reward following periods of continuous fixation. As factors like prior experience to categorizations affect the timing of perceptual information [4,17], subjects were kept naïve to any categorization or identification tasks. This is ideal for the present study purpose, since it helps purely capture the basic representational sequences within the brain. All experiments, including training, were

performed in one experimental rig, and the task was run in PsychToolbox v3.0.18. Stimuli were presented in the central 7° of the animal's visual field on a BenQ monitor with a resolution of 1920 × 1080 and a refresh rate of 144 Hz. Monkeys were positioned 50 cm distant from the center of the monitor. Simultaneously, the eye position was tracked using an infrared eye-tracking device (Zist Kankash Toos, Mashhad, Iran) with a sampling frequency of 200 Hz. Each visual stimulus was shown 5-10 times, in different recording sessions (same number of repetitions for all stimuli in every recording session).

Visual feature extraction

We used OpevCV v4.10.0 library in Python to extract physical features of objects. After preprocessing, *contourArea*, *arcLength*, and *boundingRect* functions were used to extract surface area (A_{obj}), perimeter (P_{obj}), and bounding rectangle (i.e., the smallest upright rectangle that fully encloses the object) of non-background pixels of each object's contour, respectively. Subsequently, the circularity and elongation were calculated as the following:

$$Circularity = \frac{4\pi \times A_{obj}}{P_{obj}^2}$$

$$Elongation = 1 - \frac{d_{min}}{d_{max}}$$

where d_{min} and d_{max} are the shorter and longer dimensions the object's bounding rectangle, respectively. Spikiness was computed using the A_{obj} and the area of the object's convex hull (A_{conv}; computed by OpenCV *convexHull* function), as the following:

$$Spikiness = 1 - \frac{A_{obj}}{A_{conv}}$$

Luminance and contrast and were defined as the mean and standard deviation, respectively, of the object's pixel values in grayscale. Also, PC1 and PC2 were computed after performing principal component analysis (using scikit-learn *PCA* function) on the pixel values in grayscale.

Electrophysiological recording

In each session, head-fixed animal sat in the monkey chair and viewed the visual stimuli at the center of the screen. Neural recordings were performed through grids uniquely designed for each subject's chamber with 1.5/1 mm spacing between centers of the neighboring holes for monkey F/V. Tungsten electrodes (FHC, 130 mm length; Bowdoin, ME, USA) and the covering stainless steel guide tubes were mounted on a Motorized Electrode Manipulator (MEM)™ (Thomas Recording; Gießen, Hessen, Germany) and were lowered to cross the dura, at AP and ML coordinates related to ITC and ventrolateral PFC. After passing the dura, the electrodes were cautiously inserted into the brain using the mentioned micro-driver. Neural data was recorded using a recording device (Blackrock Neurotech; Salt Lake City, UT, USA) in a sampling rate of 30 kHz. A total of 88/68 recording sessions were performed from monkey F/V. Most of

the sessions were dually recorded, from both ITC and PFC, while a few sessions contained the neural response of one region. Thus, we had 78/59 ITC neural sites and 57/63 PFC neural sites for monkey F/V. Data from online-detected neural sites with auto-thresholding were used for subsequent analyses.

Neural data analysis

Offline data analyses were performed in MATLAB 2022b and Python v3.11.7. Neuronal responses were time-locked to the stimulus presentation onset. For all analyses, each unit's response was z-scored to 80 ms time window prior to stimulus onset, which was subsequently smoothed with averaging in consecutive 20 ms-long windows (with step size of 1 ms) to form the final peri-stimulus time histogram (PSTH). Responses were averaged for all trials of the same stimulus. For all population analyses, each stimulus was considered as a point in an N-dimensional space, where N is the total number of the recorded neurons in a region. This procedure was true for all timepoints; therefore, we have:

$$S_i(\mathrm{t}) = [r_1(\mathrm{t}), r_2(\mathrm{t}), \ldots, r_N(\mathrm{t})]$$

where, at timepoint t, $S_i(\mathrm{t})$ is the vector of neural response defining the representation of $stimulus_i$ and $r_j(\mathrm{t})$ is response of neuron j to the $stimulus_i$.

All onset times were considered as the first moment of time that the time-course of response passed the following threshold:

$$Value_{threshold} = \text{baseline average} + 3 \times \text{baseline std}$$

where baseline was [-50,50] ms relative to stimulus presentation onset. Repetitions (for population data analyses) and units (for encoding models) with either onset or peak time outside the window of [50,300] ms relative to stimulus presentation onset were considered unreliable and excluded from subsequent analysis. Solution time was calculated as the time difference between onset and peak times.

Classification and time-time decoding

All classification procedures were performed using LDA method (MATLAB *fitcdiscr.m* and *predict.m*) on ITC and PFC population neural data for different levels of the semantic hierarchy (Fig. 1 G-J and Fig. 5A,B). Specifically, for Fig. 1G,I, an LDA classifier was trained for every timepoint to either detect animates from inanimates (high-level abstraction) or faces from bodies (mid-level abstraction). In all cases, %70/%30 of the data was used to the train/test the model. After forming confusion matrices, the average of within-class accuracies was used as the representative accuracy. This procedure was repeated 200 times. Subsequently, critical times, i.e., onset and peak, and solution time were computed as described above.

For TTD (Fig. 5A,B), LDA classifiers were trained at every timepoint on mid- (Fig. 5A) and high-level (Fig. 5B) abstractions. Subsequently, to test the generalizability of the trained classifiers, and thus the information at that specific timepoint, across time, they were tested on all

timepoints [44,45]. In every run of 100 iterations, %70/%30 of the samples was used to train/test the models.

Representational similarity analysis

For RSA [27], first the ground truth RSMs of each case were created, which theoretically is 0 for no similarity and 1 when perfect similarity is expected. At every timepoint in each region, the cosine-similarity (using scikit-learn *cosine_similarity* function) was computed between the vectors of neural response to every possible pair of the stimuli, as the following:

$$\cos(\theta) = \frac{Resp_{stim_i} \cdot Resp_{stim_j}}{||Resp_{stim_i}||\ ||Resp_{stim_j}||}$$

where $Resp_{stim_i}$ and $Resp_{stim_j}$ are the vectors of neural response to $stim_i$ and $stim_j$, respectively, and θ is the angle between these two vectors in the high-dimensional neural space. The greater the $\cos(\theta)$, the more similar the two vectors are. Of note, only the neural responses to face and body objects were used to construct face-body RSMs. This process would create the instantaneous $N \times N$ regional RSMs for face-body and animate-inanimate conditions, which would have the following structure:

$$\begin{bmatrix} Similarity_{stim_{1,1}} & \cdots & Similarity_{stim_{1,N}} \\ \vdots & \ddots & \vdots \\ Similarity_{stim_{N,1}} & \cdots & Similarity_{stim_{N,N}} \end{bmatrix}$$

where N is total number of objects for both categories and $Similarity_{stim_{i,j}}$ is the cosine-similarity between vectors of neural response to $stim_i$ and $stim_j$. Next, the correlation between each data-derived RSM and the ground truth RSM was calculated with Kendall's tau correlation (using scipy *kendalltau* function). This procedure was repeated 200 times and, in every run, 20/50 stimuli per each class (a total of 40/100 objects) were randomly selected to form face-body/animate-inanimate RSMs. Subsequently, onset, peak, and solution times were computed as described above.

Generalized linear models

Encoding models, specifically GLMs, were formed to predict the neural response from a set of semantic and physical features of each object. Specifically, the semantic features were the following variables in binary format: animate, face, body, human, monkey; each regressor for each object was either True or False. Physical features were circularity, elongation, spikiness, contrast, luminance, object area as well as the first and second principal components of the image in a grayscale, as floating-point numbers. One model was formed for every unit at every timepoint (using Statsmodels *GLM* function). The full model was as the following:

$$y_t = \beta_1 \times X_{animacy} + \beta_2 \times X_{face} + \beta_3 \times X_{body} + \beta_4 \times X_{human} + \beta_5 \times X_{monkey} + \beta_6 \times X_{circularity} + \beta_7 \times X_{elongation} + \beta_8 \times X_{spikiness} + \beta_9 \times X_{contrast} + \beta_{10} \times X_{PC1} + \beta_{11} \times X_{PC2} + \beta_{12} \times X_{luminance} + \beta_{13} \times X_{object\_area}$$

where y_t is the baseline z-scored, stimulus-averaged, and smoothed (moving average with 20-ms window and 1-ms step) neural response at timepoint t. Next, similar to the notion of a recent study [33], which states:

$$\text{neuronal firing rate} = \text{firing rates explained by regressors} + residuals$$

and is because:

$$\text{total sum of squares (TSS)} = \text{explained sum of squares (ESS)} + \text{residual sum of squares (RSS)}$$

or

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

in which y_i and $\hat{y}_i$ are the true and model-predicted neural response to $stimulus_i$, respectively, and $\bar{y}$ is the average neural response to all stimuli, to extract object residuals when animacy is absent in the model (and therefore present and enhanced in the residuals), we formed models without the $\beta_1 \times X_{animacy}$ term and computed the residuals as:

$$residual_i(\mathrm{t}) = y_i(\mathrm{t}) - \hat{y}_i(\mathrm{t})$$

where, at timepoint t, $residual_i(\mathrm{t})$ is the model's residual for $stimulus_i$, $y_i(\mathrm{t})$ is the neural response of a given neuron to the same stimulus, and $\hat{y}_i(\mathrm{t})$ is the model prediction of the neural response of that particular neuron and stimulus. This procedure was performed for all neurons of the population. Next, the computed residuals, that are free of all other regressors and have enhanced animacy signal, were normalized to the baseline and entered the RSA for 500 times, similar to the process described above. The same method was applied to enhance the face and body signals; this time, the $\beta_2 \times X_{face}$ and $\beta_3 \times X_{body}$ terms were excluded from the initial predictive model. Subsequently, the onset, peak, and solution times were computed as described above.

Canonical correlation analysis

To measure the functional connectivity between two neuronal populations, we used CCA [62], similar to a recent study [35]. Briefly, CCA defines pairs of dimensions in the neural spaces of two areas (will be denoted as a, b), one in each area, that meets the following criterion: once the neural activities of the two regions are projected onto the corresponding dimensions, they give maximum possible correlation; specifically, at every timepoint we have:

$$(a, b) = arg_{a,b} \max corr(Xa, Yb)$$

where X and Y are the $n \times p_x$ and $n \times p_y$ matrices of residual activities in the two areas, respectively, in which n is the number of data points and p_x and p_y are the number of

recorded neurons in the two regions, respectively. The vectors a and b have dimensions $p_x \times 1$ and $p_y \times 1$, respectively.

Granger causality

To quantify information transfer rate over time, we used the RSMs created during the RSA (see above), in a way similar to a previous study [41-43]. First, the RSMs were created for each region separately over time; this would result in $N \times N \times T$ matrices for each region, where N is the number of stimuli and T is number of timepoints. Next, we formed two regression models to predict the RSM at timepoint t (RSM_t) of the theoretical receiver area from the RSMs of timepoints $[\mathrm{t}-60, \mathrm{t}-20]$ in the following manner: (1) the first regression model only had the past RSMs of the receiver area to predict the RSM_t, while (2) the second model had to predict the RSM_t from the past RSMs of both areas. Hypothetically, if the theoretical sender area transfers information to the theoretical receiver area the residual of the second model should be lower than that of the first model. To compare the residuals, we used the following formulation:

$$Predictivity = \log \frac{SSR_{model(2)}}{SSR_{model(1)}}$$

where $SSR_{model(1)}$ and $SSR_{model(2)}$ are the sum squared residuals for model (1) and model (2), respectively. PFC and ITC were the receiver areas for the feedforward and feedback directions, respectively. This procedure was repeated 200 times, and in every run, 100 stimuli were randomly selected to measure the information transfer. Subsequently, the peak information transfer was calculated for each direction.

Encoding axes alignment

Similar to a recent study [47], LDA classifier (scikit-learn library's *LinearDiscriminantAnalysis* function) was used to find the animacy (or face-body) encoding axis in the ITC or PFC neural spaces. LDA is also a supervised dimensionality reduction method; specifically, it tries to maximize the following objective function:

$$J(w) = \frac{w^T S_B w}{w^T S_W w}$$

where S_B is the between-class scatter matrix representing the squared distance between class means and S_w is the within-class scatter matrix quantifying the variance within each class; w is the weight vector; in this context, in contains the weights corresponding to different neurons; therefore, each item in the vector w indicates the contribution of a single neuron to the discrimination task. Maximizing the $J(w)$ ensures that the projected class means are maximally separated relative to the spread within each class, yielding an axis in the high-dimensional neural space that captures the direction of maximal class separability, i.e., task-specific encoding axis. We used the eigenvalue decomposition solver to fit the model and the

shrinkage parameter was computed automatically with the Ledoit-Wolf lemma algorithm [63]. Specifically, at every timepoint and in each region, after training an LDA classifier to decode animate objects from inanimates (or faces from bodies), the neuronal weights were extracted from the model and were considered as the animacy (or face-body) encoding axis. Subsequently, the cosine of the angle between the two axes, i.e., animacy and face-body, in each region was computed, as described above. This procedure was repeated 200 times, and in every run, %70 of the samples was used to fit the model.

Deep learning models of vision

Neural network evaluations were performed in PyTorch v2.3.0 and Torchvision v0.18.0. Layers of interest for each network are listed in Table 1. For each network, first, the activations of each layer of interest were extracted in the feedforward direction. Next, the layer-wise RSMs were created and compared to the related ground truth matrices similar to the procedure employed for neural data (Fig. 6A-D). To compute the brain-network similarities (Fig. 6E-J and Supplementary Fig. 3), the RSMs of the last layer of interest of each network was used instead of the ground truth matrices used in Fig. 2C,G. Both procedures were repeated 200 times and, in every run, 20/50 stimuli per each class (a total of 40/100 objects) were randomly selected for face-body/animate-inanimate condition. Peak similarity value of every repetition entered the statistical comparisons.

Table 1. Layers of interest for deep models of vision

Network	Layers
CORnet-S CORnet-RT CORnet-Z	“module.V1.output”, “module.V2.output”, “module.V4.output”, “module.IT.output”
AlexNet	“features.1”, “features.4”, “features.7”, “features.9”, “features.12”
SqueezeNet	“features.0”, “features.1”, “features.2”, “features.3.cat”, “features.4.cat”, “features.5”, “features.6.cat”, “features.7.cat”, “features.8”, “features.9.cat”, “features.10.cat”, “features.11.cat”, “features.12.cat”
ResNet	“layer1.2.relu_2”, “layer2.3.relu_2”, “layer3.5.relu_2”, “layer4.2.relu_2”
DenseNet	“features.transition1.pool”, “features.transition2.pool”, “features.transition3.pool”, “flatten”
Inceptionv3	“maxpool1”, “maxpool2”, “Mixed_5d.cat”, “Mixed_6e.cat”, “avgpool”
EfficientNet	“features.1.1.add”, “features.2.2.add”, “features.3.2.add”, “features.4.3.add”, “features.5.3.add”, “features.6.4.add”, “features.7.1.add”, “avgpool”
VGG-16	“features.4”, “features.9”, “features.16”, “features.23”, “features.30”
VGG-19	“features.4”, “features.9”, “features.18”, “features.27”, “features.36”
MobileNet	“features.0”, “features.1.add”, “features.2.block.2”, “features.3.add”, “features.4.block.3”, “features.5.add”, “features.6.add”, “features.7.block.2”, “features.8.add”, “features.9.add”, “features.10.add”, “features.11.block.3”, “features.12.add”, “features.13.block.3”, “features.14.add”, “features.15.add”, “features.16”

Statistical analyses

Statistical and machine learning analyses were performed in Python v3.11.7, using scikit-learn v1.2.2, Statsmodels v0.14.0, and SciPy v1.13.0 libraries, and MATLAB 2022b. Details of the statistical tests used for each comparison are described wherever appropriate throughout the text. All permutations were repeated 100001 times, except for Fig. 5A,B (1001 times). All tests were two-tailed, unless mentioned otherwise, and p-values less than 0.05 were considered as statistically significant.

References

1 DiCarlo, J. J. & Cox, D. D. Untangling invariant object recognition. *Trends in cognitive sciences* **11**, 333-341 (2007).

2 DiCarlo, J. J., Zoccolan, D. & Rust, N. C. How does the brain solve visual object recognition? *Neuron* **73**, 415-434 (2012).

3 Serre, T., Oliva, A. & Poggio, T. A feedforward architecture accounts for rapid categorization. *Proceedings of the national academy of sciences* **104**, 6424-6429 (2007).

4 Mack, M. L. & Palmeri, T. J. The timing of visual object categorization. *Frontiers in Psychology* **2**, 165 (2011).

5 Mack, M. L. & Palmeri, T. J. The dynamics of categorization: Unraveling rapid categorization. *Journal of Experimental Psychology: General* **144**, 551 (2015).

6 Riesenhuber, M. & Poggio, T. Models of object recognition. *Nature neuroscience* **3**, 1199-1204 (2000).

7 Dehaqani, M.-R. A. *et al.* Temporal dynamics of visual category representation in the macaque inferior temporal cortex. *Journal of neurophysiology* **116**, 587-601 (2016).

8 Meyers, E. M., Freedman, D. J., Kreiman, G., Miller, E. K. & Poggio, T. Dynamic population coding of category information in inferior temporal and prefrontal cortex. *Journal of neurophysiology* **100**, 1407-1419 (2008).

9 Rainer, G., Rao, S. C. & Miller, E. K. Prospective coding for objects in primate prefrontal cortex. *Journal of Neuroscience* **19**, 5493-5505 (1999).

10 Wallis, J. D., Anderson, K. C. & Miller, E. K. Single neurons in prefrontal cortex encode abstract rules. *Nature* **411**, 953-956 (2001).

11 Lehky, S. R. & Tanaka, K. Neural representation for object recognition in inferotemporal cortex. *Current opinion in neurobiology* **37**, 23-35 (2016).

12 Vinken, K., Prince, J. S., Konkle, T. & Livingstone, M. S. The neural code for “face cells” is not face-specific. *Science Advances* **9**, eadg1736 (2023).

13 Bardon, A., Xiao, W., Ponce, C. R., Livingstone, M. S. & Kreiman, G. Face neurons encode nonsemantic features. *Proceedings of the national academy of sciences* **119**, e2118705119 (2022).

14 Arcaro, M. J., Ponce, C. & Livingstone, M. The neurons that mistook a hat for a face. *Elife* **9**, e53798 (2020).

15 Baldassi, C. *et al.* Shape similarity, better than semantic membership, accounts for the structure of visual object representations in a population of monkey inferotemporal neurons. *PLoS computational biology* **9**, e1003167 (2013).

16 Miller, E. K. & Cohen, J. D. An integrative theory of prefrontal cortex function. *Annual review of neuroscience* **24**, 167-202 (2001).

17 Tanaka, J. W. & Taylor, M. Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive psychology* **23**, 457-482 (1991).
18 Johnson, K. E. & Mervis, C. B. Effects of varying levels of expertise on the basic level of categorization. *Journal of experimental psychology: General* **126**, 248 (1997).
19 Freedman, D. J., Riesenhuber, M., Poggio, T. & Miller, E. K. Categorical representation of visual stimuli in the primate prefrontal cortex. *Science* **291**, 312-316 (2001).
20 Sharma, S., Vinken, K., Jagadeesh, A. V. & Livingstone, M. S. Face cells encode object parts more than facial configuration of illusory faces. *Nature Communications* **15**, 9879 (2024).
21 Kar, K. & DiCarlo, J. J. Fast recurrent processing via ventrolateral prefrontal cortex is needed by the primate ventral stream for robust core visual object recognition. *Neuron* **109**, 164-176. e165 (2021).
22 Kar, K., Kubilius, J., Schmidt, K., Issa, E. B. & DiCarlo, J. J. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature neuroscience* **22**, 974-983 (2019).
23 Toosi, R. *et al.* The Spatial Frequency Representation Predicts Category Coding in the Inferior Temporal Cortex. *bioRxiv*, 2023.2011. 2007.566068 (2023).
24 Emadi, N. & Esteky, H. Neural representation of ambiguous visual objects in the inferior temporal cortex. *PloS one* **8**, e76856 (2013).
25 Nosofsky, R. M. & Palmeri, T. J. Comparing exemplar-retrieval and decision-bound models of speeded perceptual classification. *Perception & Psychophysics* **59**, 1027-1048 (1997).
26 Nosofsky, R. M. & Palmeri, T. J. An exemplar-based random walk model of speeded classification. *Psychological review* **104**, 266 (1997).
27 Kriegeskorte, N., Mur, M. & Bandettini, P. A. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience* **2**, 249 (2008).
28 Shi, Y. *et al.* Rapid, concerted switching of the neural code in inferotemporal cortex. *bioRxiv*, 2023.2012. 2006.570341 (2023).
29 Landi, S. M., Viswanathan, P., Serene, S. & Freiwald, W. A. A fast link between face perception and memory in the temporal pole. *Science* **373**, 581-585 (2021).
30 Kriegeskorte, N. *et al.* Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* **60**, 1126-1141 (2008).
31 Miller, E. K. The prefontral cortex and cognitive control. *Nature reviews neuroscience* **1**, 59-65 (2000).
32 Rose, O. & Ponce, C. R. A concentration of visual cortex-like neurons in prefrontal cortex. *Nature Communications* **15**, 7002 (2024).
33 Yu, G., Katz, L. N., Quaia, C., Messinger, A. & Krauzlis, R. J. Short-latency preference for faces in primate superior colliculus depends on visual cortex. *Neuron* **112**, 2814-2822. e2814 (2024).
34 Noroozi, J., Rezayat, E. & Dehaqani, M.-R. A. Frontotemporal network contribution to occluded face processing. *Proceedings of the National Academy of Sciences* **121**, e2407457121 (2024).
35 Semedo, J. D. *et al.* Feedforward and feedback interactions between visual cortical areas use different population activity patterns. *Nature communications* **13**, 1099 (2022).
36 Seger, C. A. & Miller, E. K. Category learning in the brain. *Annual review of neuroscience* **33**, 203-219 (2010).
37 Webster, M. J., Bachevalier, J. & Ungerleider, L. G. Connections of inferior temporal areas TEO and TE with parietal and frontal cortex in macaque monkeys. *Cerebral cortex* **4**, 470-483 (1994).
38 Ungerleider, L., Gaffan, D. & Pelak, V. Projections from inferior temporal cortex to prefrontal cortex via the uncinate fascicle in rhesus monkeys. *Experimental brain research* **76**, 473-484 (1989).
39 Buschman, T. J. Balancing flexibility and interference in working memory. *Annual review of vision science* **7**, 367-388 (2021).

40 Stokes, M. G. *et al.* Dynamic coding for cognitive control in prefrontal cortex. *Neuron* **78**, 364-375 (2013).

41 Granger, C. W. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, 424-438 (1969).

42 Bernasconi, C. & KoÈnig, P. On the directionality of cortical interactions studied by structural analysis of electrophysiological recordings. *Biological cybernetics* **81**, 199-210 (1999).

43 Kietzmann, T. C. *et al.* Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences* **116**, 21854-21863 (2019).

44 King, J.-R. & Dehaene, S. Characterizing the dynamics of mental representations: the temporal generalization method. *Trends in cognitive sciences* **18**, 203-210 (2014).

45 Rajaei, K., Mohsenzadeh, Y., Ebrahimpour, R. & Khaligh-Razavi, S.-M. Beyond core object recognition: Recurrent processes account for object recognition under occlusion. *PLoS computational biology* **15**, e1007001 (2019).

46 King, J.-R., Pescetelli, N. & Dehaene, S. Brain mechanisms underlying the brief maintenance of seen and unseen sensory information. *Neuron* **92**, 1122-1134 (2016).

47 Dehaqani, A. A. *et al.* A mechanosensory feedback that uncouples external and self-generated sensory responses in the olfactory cortex. *Cell Reports* **43** (2024).

48 Kubilius, J. *et al.* Brain-like object recognition with high-performing shallow recurrent ANNs. *Advances in neural information processing systems* **32** (2019).

49 Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* **25** (2012).

50 Iandola, F. N. *et al.* SqueezeNet: AlexNet-level accuracy with 50× fewer parameters and. *arXiv preprint arXiv:1602.07360* (2016).

51 He, K., Zhang, X., Ren, S. & Sun, J. in *Proceedings of the IEEE conference on computer vision and pattern recognition.* 770-778.

52 Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. in *Proceedings of the IEEE conference on computer vision and pattern recognition.* 4700-4708.

53 Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. in *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2818-2826.

54 Tan, M. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946* (2019).

55 Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

56 Howard, A. G. *et al.* MobileNets: efficient convolutional neural networks for mobile vision applications (2017). *arXiv preprint arXiv:1704.04861* (2017).

57 Nonaka, S., Majima, K., Aoki, S. C. & Kamitani, Y. Brain hierarchy score: Which deep neural networks are hierarchically brain-like? *IScience* **24** (2021).

58 Rajalingham, R. *et al.* Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience* **38**, 7255-7269 (2018).

59 Schrimpf, M. *et al.* Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, 407007 (2018).

60 Yamins, D. L. & DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience* **19**, 356-365 (2016).

61 Willenbockel, V. *et al.* Controlling low-level image properties: the SHINE toolbox. *Behavior research methods* **42**, 671-684 (2010).

62 Hotelling, H. RELATIONS BETWEEN TWO SETS OF VARIATES. *Biometrika* **28**, 321-377, doi:10.1093/biomet/28.3-4.321 (1936).

63 Ledoit, O. & Wolf, M. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis* **88**, 365-411 (2004).

Acknowledgements

The authors would like to thank Mohammad Rabiei Ghahfarokhi for his technical assistance.

Author Contributions

MRAD & FS conceptualized the study. MM, MZ, & MQ collected the data. MM designed the analyses plan, analyzed the data, performed neural network evaluations, performed visualizations, and wrote the manuscript. MJ & MRAD supervised the study.

Data Availability

Data will be made available upon reasonable request to the corresponding author.

Code Availability

MATLAB scripts and functions as well as Python notebooks will be made publicly available at https://github.com/mooziri/Paper_VisualSemanticProcessing following publication of the study.

Competing Interests

None declared.

Funding

This work was funded by --- to MRAD (grant number: ---).

Figure Legends

Figure 1. Experimental design, theoretical framework, and regional information representation. (A) Schematic of the experiment; the animal was trained to watch the visual objects at the center of the screen, to receive juice rewards. In every trial, the stimulus appeared in the central 7^{o} of the animal's visual field for 80 ms, which was followed by 400 ms of blank screen. Simultaneously, the neuronal activities of ITC and vlPFC was recorded for offline analyses. Lower left panel shows a schematic of the recording locations. (B) Semantic hierarchy of visual stimuli used in this study, displaying the categorical distinctions: animate vs. inanimate represent high-level and face vs. body represent mid-level abstraction. (C-F)

PSTHs of sample units with greater response for face (C,D, upper panels), body (C,D, lower panels), animate (E,F, upper panels), inanimate (E,F, lower panels) objects in ITC (C,E) and PFC (D,F). Solid lines and shaded areas indicate the mean values and SEM of the instantaneous firing rates, respectively. Gray rectangle at 0-80 ms represents the time window of stimulus presentation. (G,I) Time-course of decoding accuracy for classifiers trained and tested on neural data from ITC (G) and PFC (I) to distinguish face vs. body or animate vs. inanimate. Solid lines and shaded areas indicate the mean values and SD of the classifiers' accuracies, respectively. Gray rectangle at 0-80 ms represents the time window of stimulus presentation. Dashed lines denoted with arrows and arrowheads are median values of onset and peak times, respectively. (H,J) Statistical comparison of onset (upper panels) and peak (lower panels) times for the regional classifiers in G,I for ITC (H) and PFC (J). Bars and error bars indicate median values and SD, respectively. Statistical significance measured by permutation test. **p < 0.01, ***p < 0.001. ITC, inferior temporal cortex; PSTH, peri-stimulus time histogram; vlPFC, ventrolateral prefrontal cortex.

Figure 2. Temporal dynamics of semantic information in the ITC-PFC circuit. (A,B) Schematic illustrations of the theoretical frameworks describing random walk model (A) and solution time. (C,G, bottom) Time-course of similarity to ground truth for the population activities of ITC and PFC for face vs. body (C) or animate vs. inanimate (G) distinctions. Solid lines and shaded areas indicate the mean values and SD of the similarities, respectively. Gray rectangle at 0-80 ms represents the time window of stimulus presentation. Dashed lines denoted with arrows and arrowheads are the median of onset times and the peak time of average time-course, respectively. (C,G, top) Peak similarity RSMs of ITC (leftmost panels) and PFC (rightmost panels) alongside the ground truth (middle panels) for face-body (C) and animate-inanimate (G) conditions. Warmer colors indicate greater values of cosine-similarity. (D-F,H-J) Statistical comparison of onset (D,H), peak (E,I), and solution (F,J) times for the similarities in C,G for face-body (D-F) and animate-inanimate (H-J) separations. Bars and error bars indicate median values and SD, respectively. Statistical significance measured by permutation test. ***p < 0.001. ITC, inferior temporal cortex; PFC, prefrontal cortex; RSM, representational similarity matrix.

Figure 3. Temporal dynamics of "enhanced" semantic information in the ITC-PFC circuit. (A-D) Object residuals of sample units with greater response for face (A,B, upper panels), body (A,B, lower panels), animate (C,D, upper panels), inanimate (C,D, lower panels) objects in ITC (A,C) and PFC (B,D). Solid lines and shaded areas indicate the mean values and SEM of the instantaneous object residuals, respectively. Gray rectangle at 0-80 ms represents the time window of stimulus presentation. (E,I) Time-course of similarity to ground truth for the population residuals of ITC and PFC for face vs. body (E) or animate vs. inanimate (I) distinctions. Solid lines and shaded areas indicate the mean values and SD of the similarities, respectively. Gray rectangle at 0-80 ms represents the time window of stimulus presentation. Dashed lines denoted with arrows and arrowheads are the median values of onset and peak times, respectively. (F-H,J-L) Statistical comparison of onset (F,J), peak (G,K), and solution (H,L)

times for the similarities in E,I for face-body (F-H) and animate-inanimate (J-L) separations. Bars and error bars indicate median values and %95 CI, respectively. Statistical significance measured by permutation test. *p < 0.05, ***p < 0.001. CI, confidence interval; ITC, inferior temporal cortex; PFC, prefrontal cortex.

Figure 4. Temporal dynamics of inter-regional neuronal population communication in the ITC-PFC circuit. (A) Schematic illustration of CCA. In this context, CCA concurrently finds one dimension each neural space (top panels), that are maximally correlated with each other (bottom panel). (B) Time-course of the population functional connectivity between ITC and PFC. Solid lines and shaded areas indicate the mean values and SD of the functional connectivity, respectively. Gray rectangle at 0-80 ms represents the time window of stimulus presentation. Statistical significance measured by one-tailed permutation test; thick and thin horizontal lines denote $p < 0.001$ and $p < 0.05$ compared to baseline, respectively. (C) Schematic illustration of the Granger causality approach: the RSM of timepoint t of a given region (RSM in yellow square) was predicted from earlier RSMs of the same (RSMs in green squares) or both (RSMs in green and red squares) regions using regression models. (D) Time-course of Granger causality in the ITC-to-PFC and PFC-to-ITC directions. Solid lines and shaded areas indicate the mean values and SD of the predictivity, respectively. Gray rectangle at 0-80 ms represents the time window of stimulus presentation. Dashed lines are the peak times of average information transfer in each direction. Inlet: statistical comparison of peak times of information transfer. Bars and error bars indicate median values and SD, respectively. Statistical significance measured by permutation test. ***p < 0.001. CCA, canonical correlation analysis; ITC, inferior temporal cortex; PFC, prefrontal cortex.

Figure 5. Neural architecture of ITC and PFC microcircuitries for semantic processing. (A,B) TTD for across-time generalization of face-body (A) and animacy (B) information in the ITC (left panels) and PFC (right panels) populations. Gray rectangles at 0-80 ms represent the time window of stimulus presentation. Statistical significance measured by permutation testing compared to baseline; non-significant pixels are colored in black. (C) Schematic illustration of the hypothetical geometrical relationship between mid- and high-level information types in the perceptual space; in this framework, since our perception of animacy (C-i/C-iii vs C-ii/C-iv) is independent of face-body status of an object (and vice-versa), we can theoretically consider that there are two axes in the perceptual space, one for each categorization task, that are orthogonally aligned. (D) Schematic illustration of possible alignments between the two arbitrary feature axes of in a given neuronal population. If the angle between two axes in a neural space is 90^{o} (D-i), then, unlike non-perpendicular conditions (e.g., the situation depicted in D-ii), changes in either direction does not alter the representation along the other one. (E, upper panel) Time-course of the cosine of the angle between encoding axes corresponding to mid- and high-level abstractions in the ITC and PFC neural spaces. Solid lines and shaded areas indicate the mean values and SD of the angle cosine, respectively. Gray rectangle at 0-80 ms represents the time window of stimulus presentation. (E, lower panel) Time-course of the p-value for statistical comparison of the between-axes angle cosine

compared to a theoretical null distribution for circular data, i.e., U(-1,1), using permutation test. Gray rectangles at 0-80 ms represent the time window of stimulus presentation. Purple/orange rectangle demonstrates the early/late phase of axes orthogonality in the PFC/ITC neural space. ITC, inferior temporal cortex; PFC, prefrontal cortex; TTD, time-time decoding.

Figure 6. Semantic processing in deep models of vision. (A,B, top) From left, RSMs of ground truth expectations and last layer of the studied networks for mid- (A) and high-level (B) semanticness. Warmer colors indicate greater values of cosine-similarity. (A,B, bottom) Each panel depicts the similarity of all studied layers of a network to ground truth RSM, shown in the upper rows, for face-body (A) and animate-inanimate (B) conditions. Solid lines and shaded areas indicate the mean values and SD of similarity, respectively. (C,D) Comparison of all networks' last layer similarity to ground truth RSMs for face-body (C) and animate-inanimate (D) distinctions. Bars and error bars indicate mean values and SD, respectively. (E,G) Time-course of similarity of ITC and PFC population representations to the last layer of all studied networks for face-body (E) or animate-inanimate (G) separations. Solid lines and shaded areas indicate the mean values and SD of the similarities, respectively. Gray rectangle at 0-80 ms represents the time window of stimulus presentation. (F,H) Scatter plot for the networks' peak similarities to ITC and PFC population representations shown in E,G. Each point is one network, and the bars along the x and y axes are the SD of the network similarity to PFC and ITC, respectively (n = 12 networks). Histograms on the top right corners illustrate the similarity difference between NN~ITC and NN~PFC raw values. Dashed red lines in the histograms are the median similarity difference in each case. Statistical significance measured by one-sample Wilcoxon signed rank test from a theoretical null value of 0 for similarity difference. (I,J) Comparison of all networks' last layer similarity peak to ITC (left panels) and PFC (right panels) population representations for face-body (I) and animate-inanimate (J) distinctions. Bars and error bars indicate mean values and SD, respectively. ITC, inferior temporal cortex; PFC, prefrontal cortex; RSM, representational similarity matrix.

Supplementary Figure 1. Matching spatial frequency among categories. Spatial frequency profile for the visual stimuli in each category. Solid lines and shaded areas indicate the mean values and SD of the amplitude, respectively. Statistical significance measured by Mann-Whitney test between category pairs at every frequency.

Supplementary Figure 2. Visual feature differences between high-level semantic categories. (A,B) Illustration of top 20 most circular (A) and spiky (B) objects in the entire stimulus-set.

Supplementary Figure 3. Semantic similarity of CORnet-S ITC to primate ITC and PFC. (A,C) Time-course of similarity to CORnet-S ITC layer for the population activities of ITC and PFC for face-body (A) or animate-inanimate (C) distinctions. Solid lines and shaded areas indicate the

mean values and SD of the similarities, respectively. Gray rectangle at 0-80 ms represents the time window of stimulus presentation. (B,D) Histogram of peak similarity values in A,C. Dashed lines show the median value of each distribution. Statistical significance measured by permutation test. ITC, inferior temporal cortex; PFC, prefrontal cortex.

Figure 1

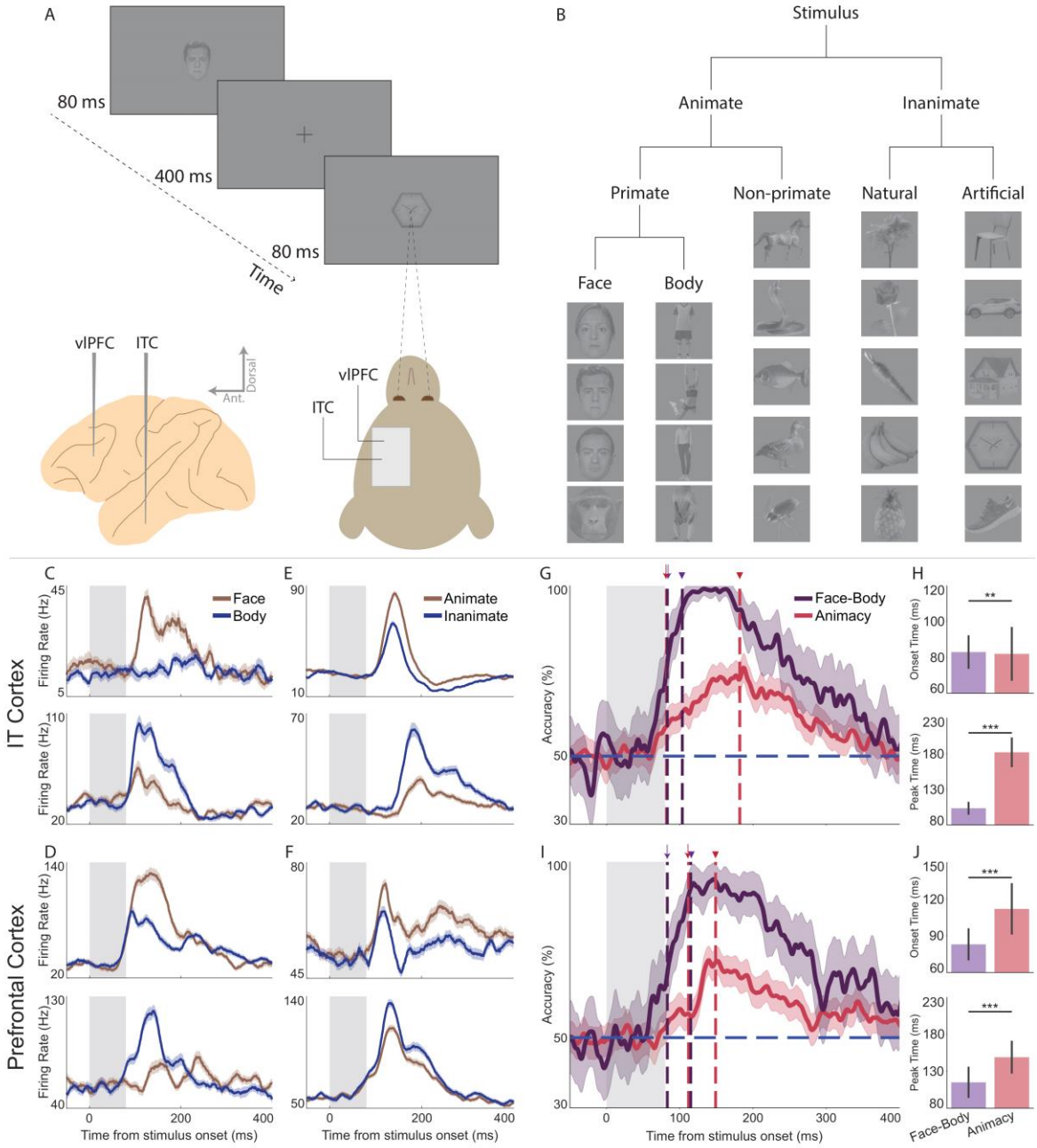
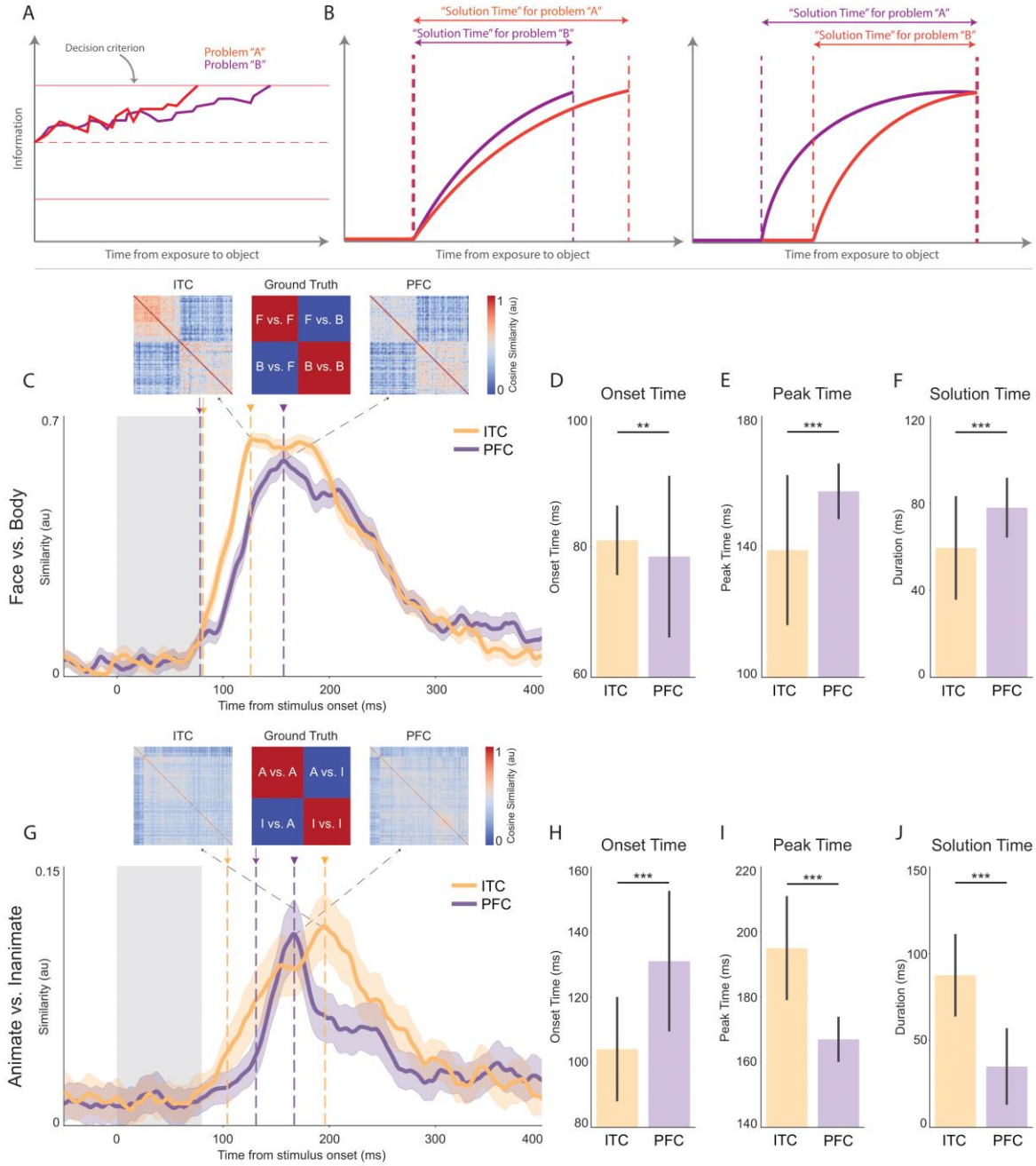

A
80 ms
400 ms
80 ms
Time
vlPFC
ITC
Ant.
Dorsal
vlPFC
ITC
B
Stimulus
Animate
Inanimate
Primate
Non-primate
Natural
Artificial
Face
Body
C
E
G
H
IT Cortex
Face
Body
Animate
Inanimate
Face-Body
Animacy
Firing Rate (Hz)
Accuracy (%)
Onset Time (ms)
Peak Time (ms)
D
F
I
J
Prefrontal Cortex
Time from stimulus onset (ms)
Face-Body
Animacy

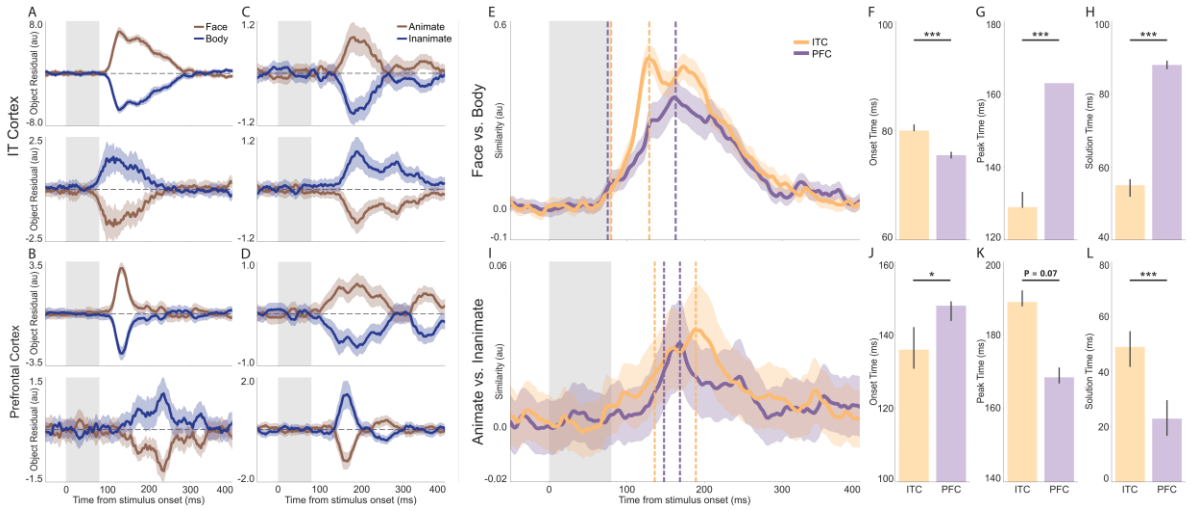
Figure 2



961

962

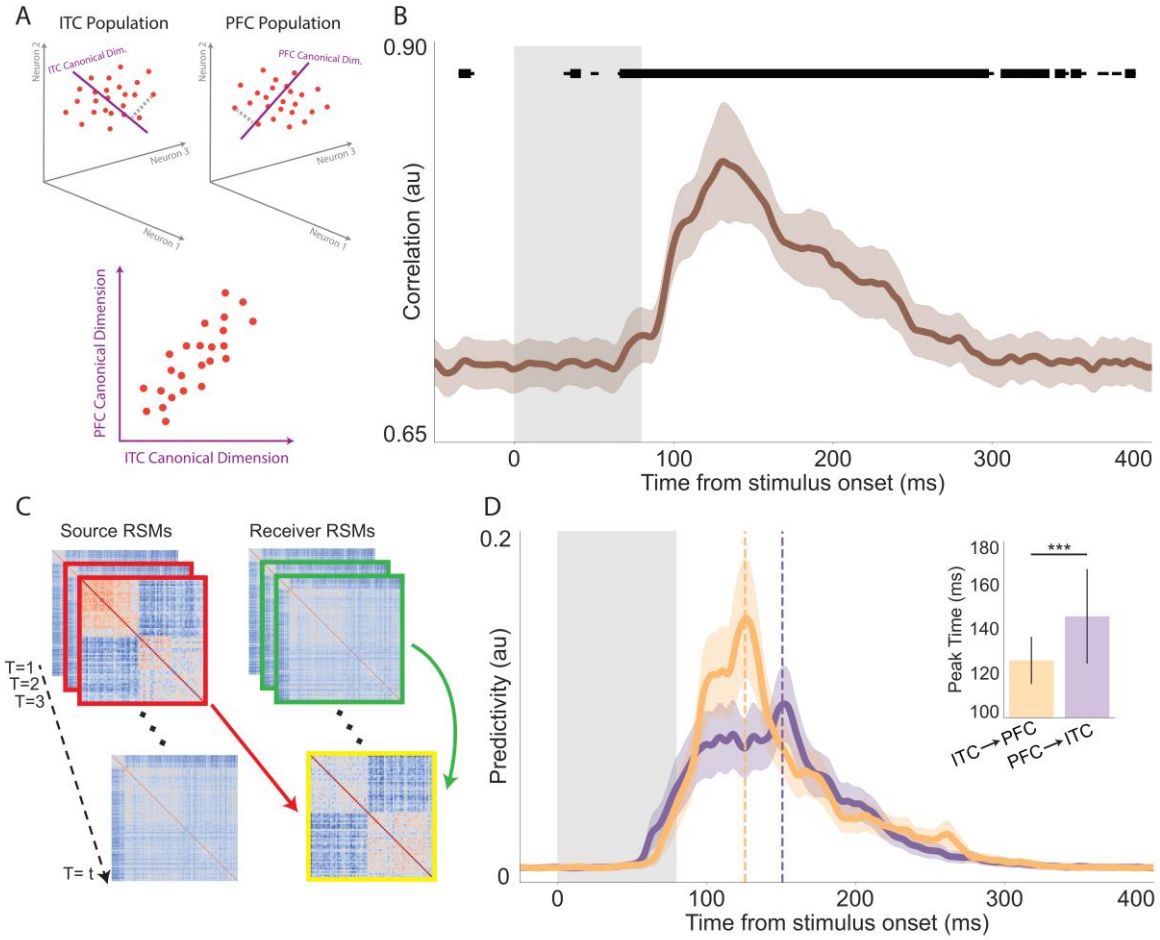
963 **Figure 3**



964

965

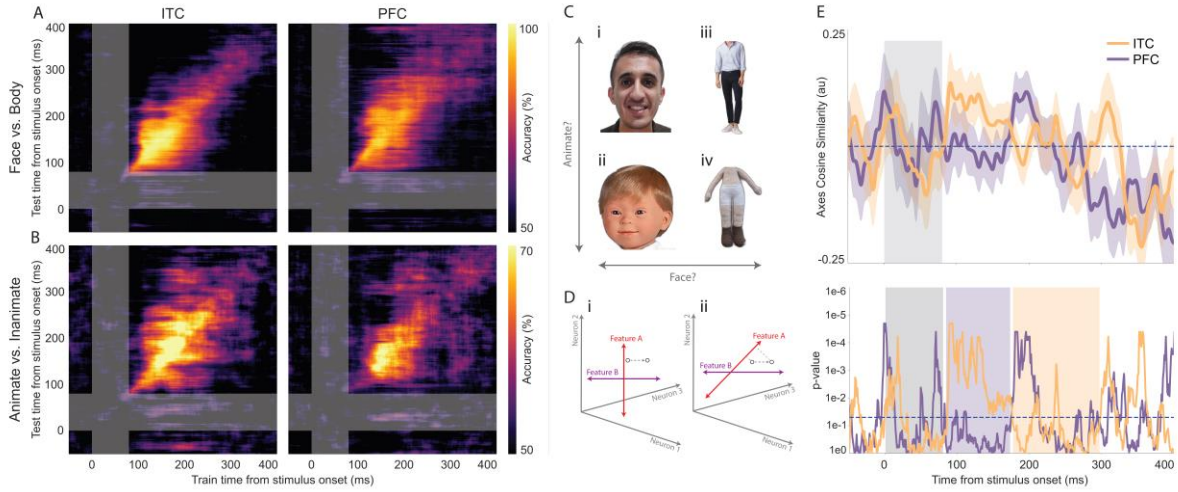
966 **Figure 4**



967

968

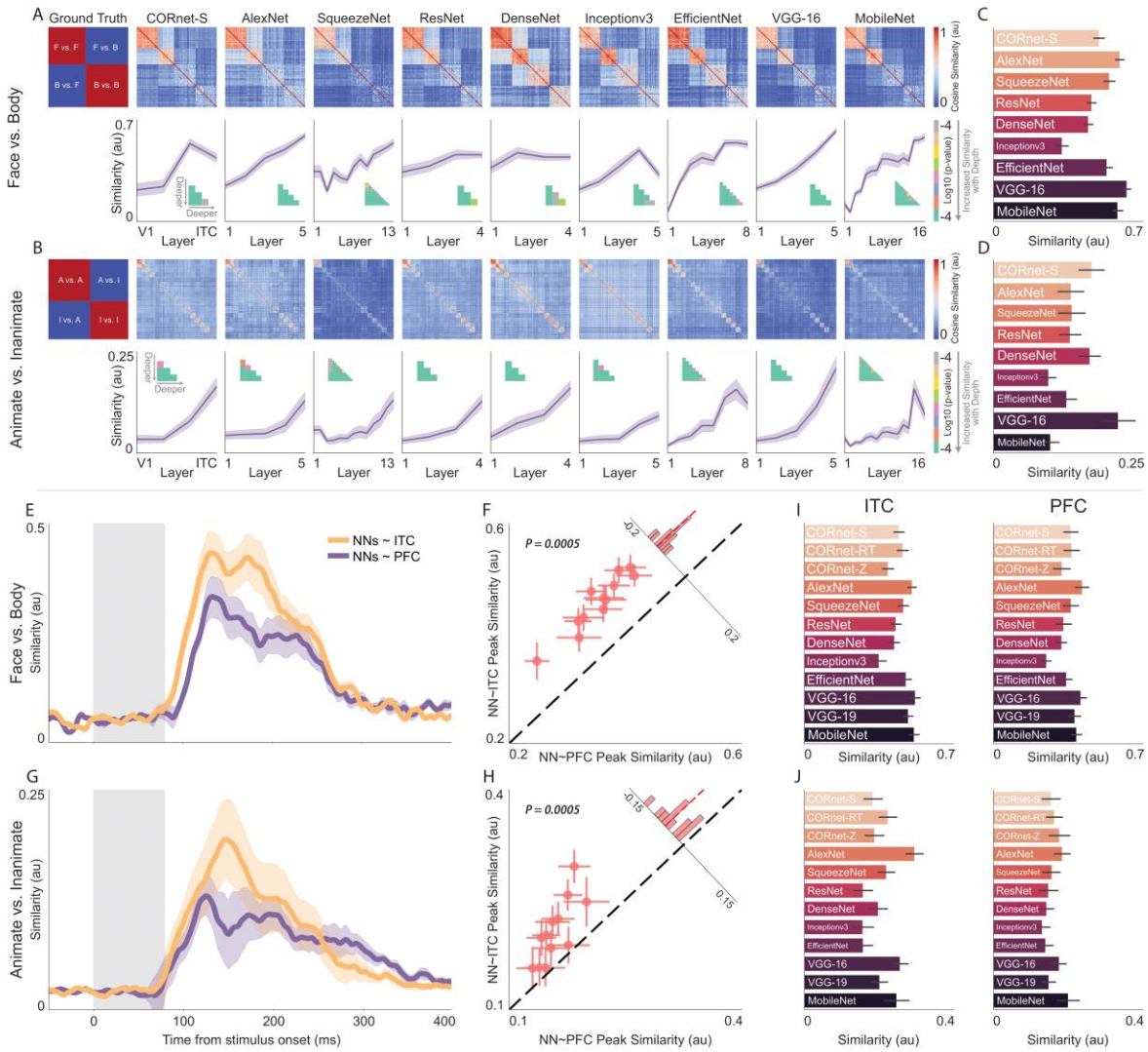
969 **Figure 5**



970

971

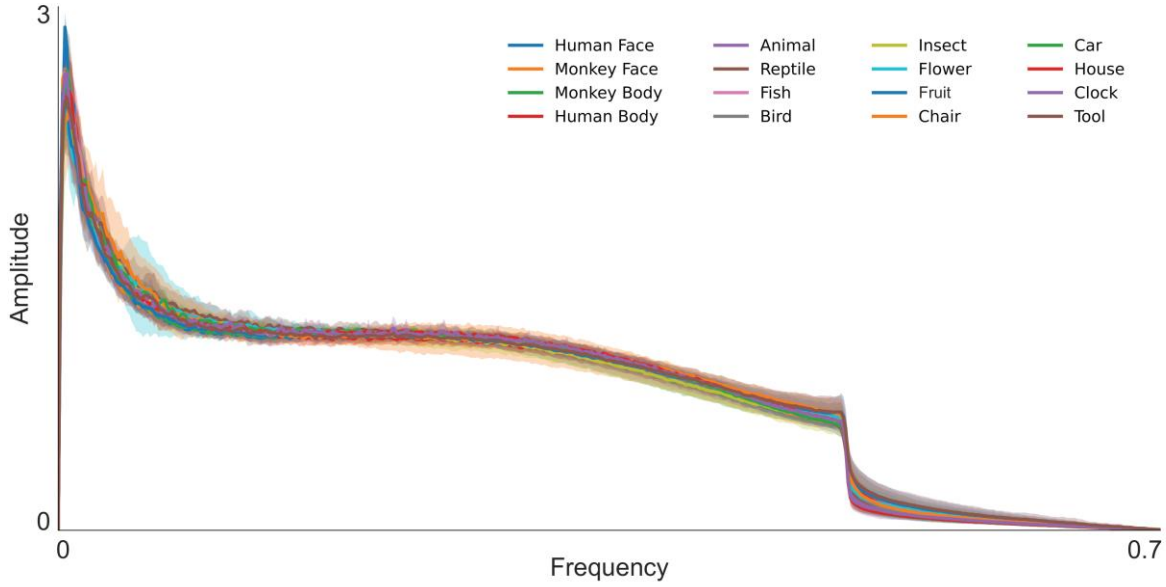
Figure 6



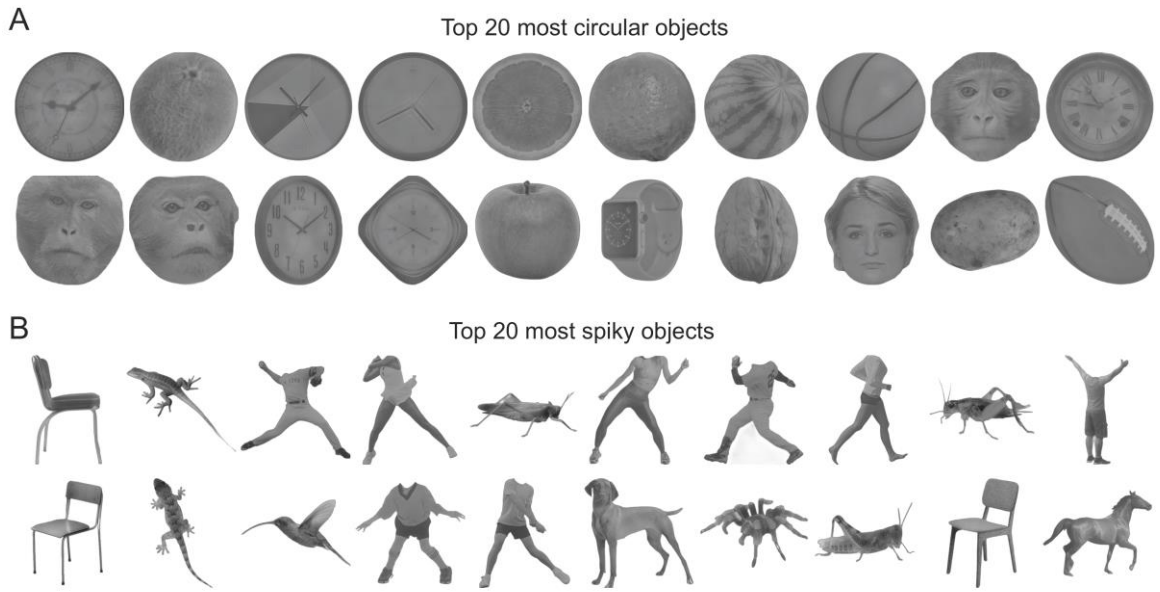
973

974

Supplementary Figure 1



978 **Supplementary Figure 2**



979

980

Supplementary Figure 3