# Modality-independent coding of concepts in prefrontal cortex

Yaelan Jung[1], Bart Larsen[2], & Dirk B. Walther[1]

[1]Department of Psychology, University of Toronto.
100 St. George Street, Toronto, ON, M5S 3G3 Canada.

[2]Department of Psychology, University of Pittsburgh.
Sennott Square, 210 South Bouquet Street, Pittsburgh, PA, 15260 USA

Correspondence should be addressed to Yaelan Jung,

100 St. George Street, Toronto, ON, M5S 3G3 Canada.

Phone: +1-416-946-3200

Email: yaelan.jung@mail.utoronto.ca

Classifications: Social Sciences/ Psychology and Cognitive Sciences

Keywords: Cognitive neuroscience, cross-modal integration, abstract representation, scene perception, prefrontal cortex

## Abstract

Natural environments convey information through multiple sensory modalities, all of which contribute to people's percepts. Although it has been shown that neural representations of visual content can be decoded from the visual cortex, it remains unclear where and how humans represent perceptual information at a conceptual level, not limited to a specific sensory modality. To address this question, we investigated how categories of scene images and sounds are represented in several brain regions. We found that both visual and auditory scene categories can be decoded not only from modality-specific areas, but also from several brain regions in the temporal, parietal, and prefrontal cortex. Intriguingly, only in the prefrontal cortex, but not in any other regions, categories of scene images and sounds appear to be represented in similar activation patterns, suggesting that scene representations in prefrontal cortex are modality-independent. Furthermore, the error patterns of neural decoders indicate that the category-specific neural activity patterns in the middle and superior frontal gyri are tightly linked to categorization behavior. Our findings suggest that complex visual information is represented at a conceptual level in prefrontal cortex, regardless of the sensory modality of the stimulus.

## Introduction

Imagine taking a walk on the beach. Your sensory experience would include the sparkle of the sun's reflection on the water, the calls of seagulls mixed with the calming sound of the crushing waves, the smell of salty ocean air, and the cold touch of wet sand between your toes. Even though the brain has separate, clearly delineated processing channels for all of these sensory modalities, the subjective perceptual experience will still be that of the integral concept "beach". What are the neural systems underlying this convergence, allowing us to form an abstract representation of concepts beyond the sensory modality domain? Here, we report finding neural representations of concepts that transcend sensory modalities in human prefrontal cortex. We characterize the properties of these representations in detail and conclusively link them to behavioral categorization performance.

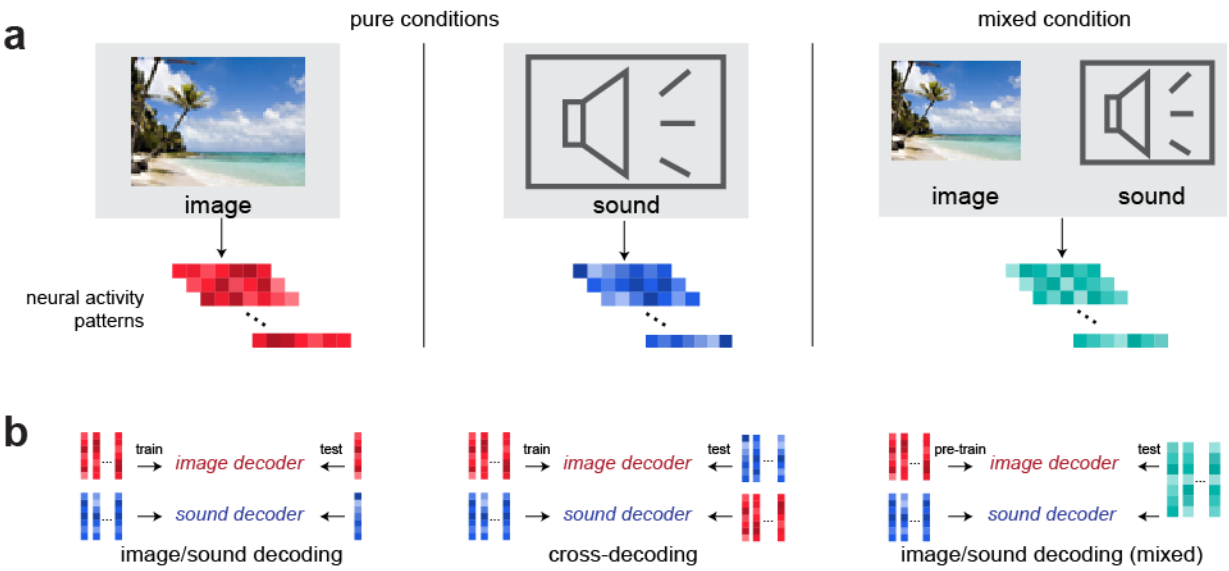
Specifically, we focused on the categorization of real-world scenes, which has been studied extensively in the visual domain. Three high-level brain regions are known to preferentially respond to scenes versus other visual stimuli: the Parahippocampal place area (PPA), Retrosplenial cortex (RSC), and the occipital place area (OPA) (Dilks, Julian, Paunov, & Kanwisher, 2013; Epstein & Kanwisher, 1998; Maguire, 2001). Furthermore, natural scene categories include a rich tapestry of sensory information from multiple sources, which requires not only one-to-one matching between visual (i.e. the shape of seagulls) and auditory (i.e. the calls of seagulls) objects, but generalization over various visual (i.e. the sun's reflection on the water) and auditory (i.e. the calls of seagulls) objects to conceptualize the scene category. Thus, scene category representations on an amodal level will reveal how a concept, which includes multiple instances, is abstracted and represented in the brain beyond a mere superposition of visual and auditory inputs.

In order to study scene representation beyond the sensory modality level, we presented participants with images and sounds of four scene categories (beaches, forests, cities, and offices) while their neural activity was measured in an fMRI scanner (Fig. 1a). We posit four different models of how visual and auditory information can be processed within a brain region (Fig. 1b). First, we expected that primary visual and auditory regions will contain neurons dedicated to the processing of specific modality information (image or sound). In these regions, we expected to decode scene categories from only the corresponding modality condition, but not across modalities. Also, conflicting information from a different modality (i.e. forest sounds) should not interfere with the processing of information from the dedicated modality (i.e. city image). In multi-modal regions, both visual and auditory information should be processed in anatomically collocated but functionally separate neural populations. Therefore, both image and sound categories can be decoded, but decoding across modalities should not be possible. Conflicting information from different modalities should not interfere with image and sound

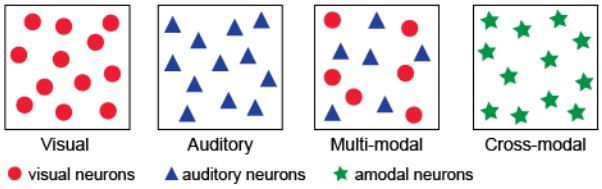
processing, as the information from each modality is processed separately. Finally, cross-modal regions are expected to be equally activated by scene information from different sensory modalities, as long as it pertains to the same category. So, both image and sound categories should be decodable, even across modalities. However, conflicting information from one modality will interfere with processing of information from the other modality in this region, so that decoding of scene categories will be degraded when the image and sound information are conceptually inconsistent. Based on previous neuroimaging and neurophysiology studies (Freedman, Riesenhuber, Poggio, & Miller, 2001; Mack, Preston, & Love, 2013; Meyers, Freedman, Kreiman, Miller, & Poggio, 2008; Miller & Cohen, 2001; Wood & Grafman, 2003), we hypothesized that the prefrontal cortex (PFC) should be involved in this cross-modal neural coding.

To test these hypotheses, we examined several ROIs from visual, auditory, temporal, parietal, and prefrontal cortex (summarized in Fig. 1d). Supporting our predictions, we found that different brain areas can be characterized according to our models' predictions. Specifically, middle frontal gyrus (MFG) and inferior frontal gyrus (IFG) show the same patterns of results as the cross-modal model, suggesting in these areas, scene information is not limited to a specific sensory modality, but rather, processed at a conceptual level.

We further explored how scene category information is classified in each brain area by analyzing the error patterns of the neural decoder. In the early stages of sensory processing, we expected neural representations of scene categories to reflect the physical properties of scene information in the respective sensory domain, whereas in later stages, it should be related to people's behavioral judgment of scene category (Walther, Caddigan, Fei-Fei, & Beck, 2009). To test this idea, we compared decoding error patterns with the error patterns of participants engaged in a separate behavioral scene categorization experiment, as well as with computational error patterns derived from categorizing the images and sounds based on their physical characteristics.




**c**




| | Pure | | | Mixed | |
|---|---|---|---|---|---|
| | image | sound | across | image | sound |
| Visual | + | - | - | + | - |
| Auditory | - | + | - | - | + |
| Multi-modal | + | + | - | +/- | +/- |
| Cross-modal | + | + | + | - | - |

**d**

| ROI | | Pure I | Pure S | Pure C | Mixed I | Mixed S |
|---|---|---|---|---|---|---|
| Visual | V1-V4 | +++ | - | - | +++ | - |
| | PPA | +++ | - | - | +++ | - |
| | OPA | +++ | + | - | +++ | - |
| | RSC | +++ | + | - | +++ | - |
| | LOC | +++ | - | - | +++ | - |
| Auditory | A1 | + | +++ | - | - | +++ |
| | A1 sub | - | +++ | - | - | +++ |

| ROI | | Pure I | Pure S | Pure C | Mixed I | Mixed S |
|---|---|---|---|---|---|---|
| Temporal | STG | ++ | ++ | - | - | + |
| | STS | + | - | - | + | - |
| | MTG | ++ | ++ | - | - | - |
| Parietal | AG | ++ | + | - | + | - |
| | SPG | +++ | ++ | - | +++ | - |
| | IPS | - | - | - | - | - |

| ROI | | Pure I | Pure S | Pure C | Mixed I | Mixed S |
|---|---|---|---|---|---|---|
| Prefrontal | $IFG_{orb}$ | - | - | + | - | - |
| | $IFG_{tri}$ | ++ | - | - | - | - |
| | $IFG_{opr}$ | + | + | + | - | - |
| | MFG | +++ | +++ | + | + | - |
| | SFG | + | ++ | - | - | - |

**Figure 1 (a)** Illustration of the image, sound, and mixed conditions. In the pure image and sound conditions, either images or sounds from four different scene categories were presented while neural activity patterns of participants were recorded. In the mixed condition, both images and sounds were presented simultaneously, but they were always from different categories (i.e. a beach images with city sounds). (**b**) Multivariate analysis of fMRI data for decoding image and sound category, cross-decoding across image and sound, and decoding image and sound from the mixed condition. (**c**) Models for separate brain areas dedicated to visual, auditory, multi-modal, and cross-modal processing, and the prediction for decoding performance of each model in different conditions (+: decodable, -: not decodable). (**d**) ROIs included in the analysis with their decoding results in different conditions (+++: decodable with $q < .001$, ++ decodable with $q < .01$, + decodable with $q < .05$, - not-decodable). ROIs with the same patterns of results as our models in (**c**) are color-coded accordingly.

## Results

*Decoding scene categories of images and sounds*

To assess neural representations of scene categories from images and sounds, we performed multivoxel pattern analysis on each ROI. A linear support vector machine (SVM; using LIBSVM, Chang & Lin, 2001) was trained using neural activity patterns with category labels and then tested to determine if a trained classifier can predict the scene category in a leave-one-run-out cross validation.

As shown in previous studies (Choo & Walther, 2016; Park, Brady, Greene, & Oliva, 2011; Walther et al., 2009; Walther, Chai, Caddigan, Beck, & Fei-Fei, 2011; Kravitz, Peng, & Baker, 2011), both early visual areas V1 through V4 and high-level visual areas, including the parahippocampal place area (PPA), the retrosplenial cortex (RSC), the occipital place area (OPA), and lateral occipital complex (LOC), show category-specific scene representations (Fig. 2a). We were also able to decode the scene categories from activity elicited by sounds of the respective natural environments in primary auditory cortex (A1) as well as its anatomical sub-divisions (Fig. 2d).

Unlike previous reports showing that auditory content can be decoded from early visual cortex (Paton, Petro, & Muckli, 2016; Vetter et al., 2014), we did not find representations of auditory information in V1 to V4. However, we were able to decode auditory scene categories in higher visual areas, the OPA (30.5%; chance: 25%) and the RSC (31.3%; Fig. 2b). Intriguingly, we could also decode scene categories from images in A1 with a decoding accuracy of 29.8% (Fig. 2c).

Having found modality-specific representations of scene categories in visual and auditory cortices, we aimed to identify scene representations in areas which are not limited to a specific sensory modality. We could decode categories of both visual and auditory scenes in several temporal and parietal regions (Fig. 2e,f): the middle temporal gyrus (MTG), the superior temporal gyrus (STG), the superior temporal sulcus (STS) and the superior parietal gyrus (SPG). In the angular gyrus (AG), we could decode visual but not auditory scene categories. Although previous studies have suggested that the intraparietal sulcus (IPS) is involved in audiovisual processing (Calvert et al., 2001), we could not decode either visual or auditory scene categories in the IPS.

Next, we examined whether the prefrontal cortex (PFC) also showed category-specific representations for both visual and auditory scene information. Previous studies have found strong hemispheric specialization in PFC (Gaffan & Harrison,1991; Goel, et al., 2007; Slotnick & Moo, 2006). We therefore analyzed functional activity in PFC separately by hemisphere. We were able to decode visual scene categories significantly above chance from the left inferior frontal gyrus (IFG), pars opercularis, the right IFG, pars triangularis, and in both hemispheres

from the middle frontal gyrus (MFG) and the superior frontal gyrus (SFG; Fig. 2g). The categories of scene sounds were decodable in the right IFG, pars triangularis, as well as the MFG and SFG in both hemispheres (Fig. 2h).

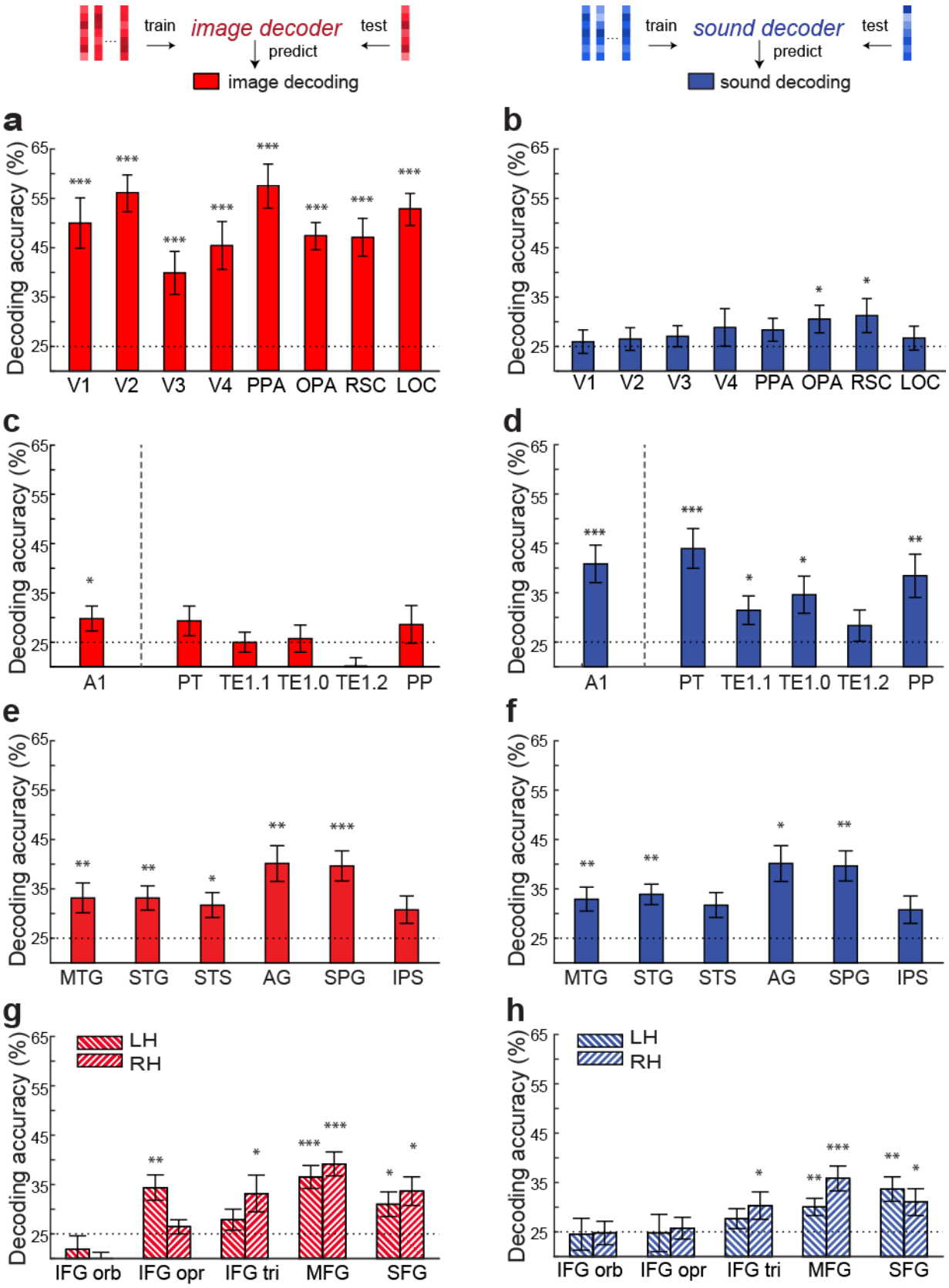

train
image decoder
test
predict
image decoding
train
sound decoder
test
predict
sound decoding
a
b
c
d
e
f
g
h
Decoding accuracy (%)
65
55
45
35
25
V1
V2
V3
V4
PPA
OPA
RSC
LOC
A1
PT
TE1.1
TE1.0
TE1.2
PP
MTG
STG
STS
AG
SPG
IPS
LH
RH
IFG orb
IFG opr
IFG tri
MFG
SFG

**Figure 2** Decoding accuracy for images (**a**, **c**, **e**, **g**; plotted in red) and sounds (**b**, **d**, **f**, **h**; plotted in blue). (**a**,**b**) ROIs from visual cortex. (**c**,**d**) primary auditory cortex (A1) with its anatomical subdivisions: Planum Temporale (PT), Posteromedial Heschl's gyrus (TE1.1), Middle Heschl's gyrus (TE1.0), Anterolateral Heschl's gyrus (TE1.2), Planum Polare (PP). (**e**, **f**) temporal and parietal ROIs: middle temporal gyrus (MTG), superior temporal gyrus (STG), superior temporal sulcus (STS), angular gyrus (AG), superior parietal gyrus (SPG), intraparietal sulcus (IPS) (**g**, **h**) Prefrontal ROIs, separately for left (LH) and right hemisphere (RH): inferior frontal gyrus (IFG), pars orbitalis (orb), pars opercularis (opr), and pars triangularis (tri), as well as middle (MFG) and superior frontal gyrus (SFG). Significance of the one-sample t-tests (one-tailed) was adjusted for multiple comparisons using false discovery rate, *q < .05, **q < .01, ***q < .001. The error bars are SEM.

Although the temporal, parietal, and prefrontal cortex all showed both visual and auditory scene representations, this does not necessarily imply that these areas process scene information at an abstract and conceptual level that transcends sensory modalities. A conceptual representation of scene categories in the brain should not merely consist of co-existing populations of neurons with visually and auditorily triggered activation patterns; the voxels in these ROIs should be activated equally by visual and auditory inputs if they represent the same category. In other words, if the neural activity pattern elicited by watching a picture of a forest reflects the general concept of *forest*, then this neural representation should be similar to that elicited by listening to forest sounds. Intriguingly, we found supporting evidence for this kind of abstract representations in the prefrontal areas when we examined the error patterns from neural decoder. When the error patterns of the image category decoder were compared to those from the sound decoder in the same ROI, we found considerable similarities only in prefrontal ROIs but not in any other ROIs where both image and sound scene categories can be decoded (Fig. S1). Based on this finding, we aimed to explicitly examine whether scene category information in the prefrontal areas transcends sensory modalities using cross-decoding analysis between the image and sound conditions.

*Cross-modal decoding*

For the cross-decoding analysis, we trained the decoder using the image labels from the image runs and then tested whether it could correctly predict the categories of scenes presented as sounds in the sound runs. We also performed the reverse analysis, training the decoder on the sound runs and testing it on the image runs (Fig. 3a).

Cross-decoding from images to sounds succeeded in the MFG in both hemispheres and in the right IFG, pars orbitalis. The right MFG and the right IFG, pars triangularis, showed significant decoding accuracy for cross-decoding from sounds to images (Fig. 3c). However, cross-decoding was not possible in either direction anywhere in sensory cortices or temporal and parietal cortices (Fig. 3b). These results suggest that temporal and parietal cortices do not contain abstract representations of categories even though they represent both visual and auditory scene information.

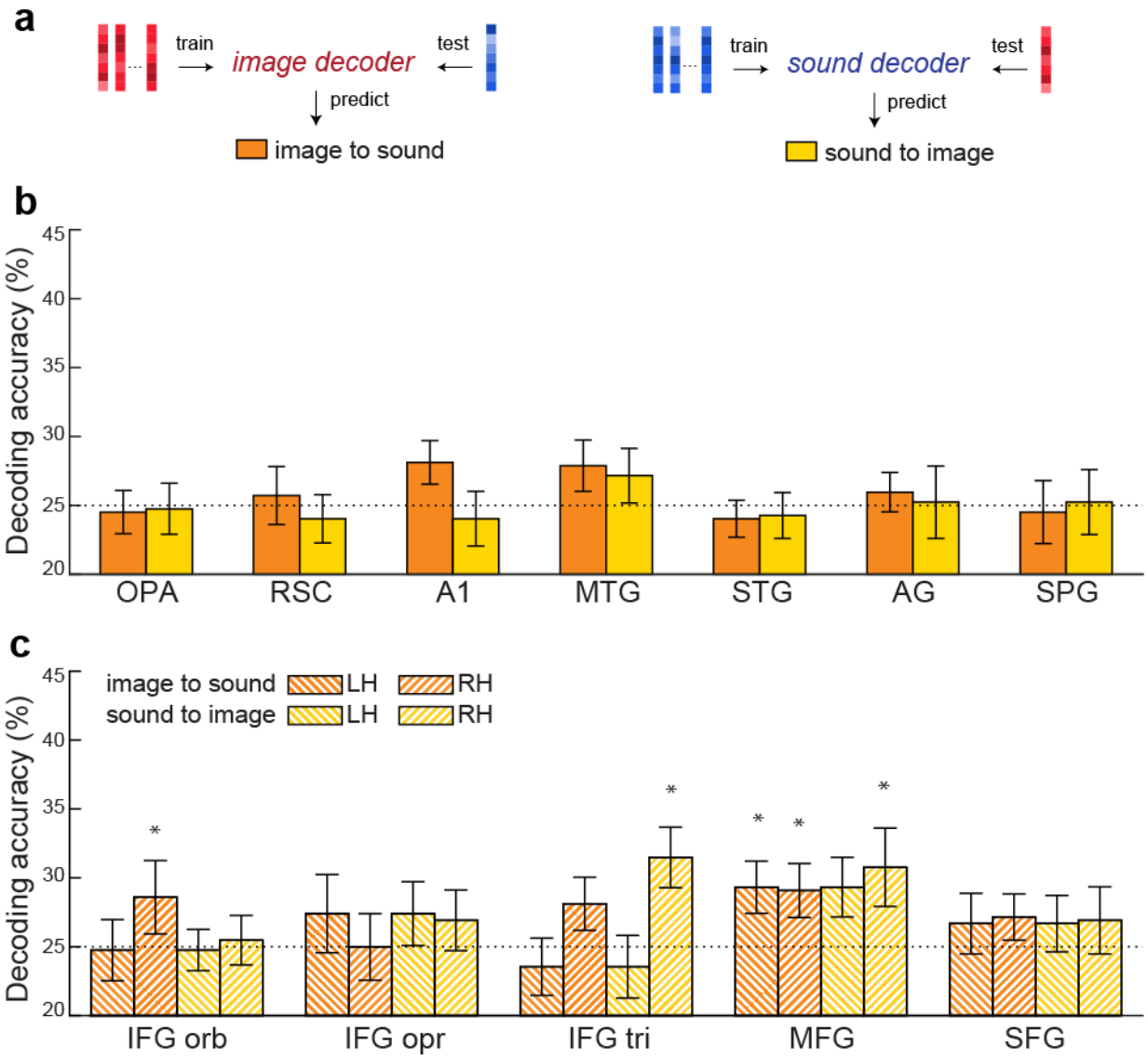



**Figure 3** Accuracy of cross-decoding between images and sounds. (**a**) Illustration of cross-decoding analysis. (**b**) Non-prefrontal ROIs that allow for decoding of images and sounds. (**c**) prefrontal ROIs in PFC. The significance of the one-sample t-tests (one-tailed) was adjusted for multiple comparisons (FDR) and marked above each bar, *q < .05. The error bars are SEM.

*Presenting images and sounds concurrently*

As a further test to examine the cross-modal nature of scene category representations in PFC, we used an interference condition, in which we presented images and sounds from incongruous categories simultaneously. If a population of neurons encodes a scene category independently of the sensory modality, then we should see a degradation of the category representation in the presence of a competing signal from the other modality. If, on the other

hand, two separate but intermixed populations of neurons encode the visual and auditory categories, respectively, then we should be able to still decode the category from at least one of the two senses.

To decode scene categories from this mixed condition, we created an image and a sound decoder by training separate classifiers with data from the image-only and the sound-only conditions. We then tested these decoders with neural activity patterns from the mixed condition, using either image or sound labels as ground truth (Fig. 4a). As the training and the test data are from separate sets of runs, cross-validation was not needed for this analysis.

We were able to decode visual and auditory scene categories from the respective sensory brain areas, even in the presence of conflicting information from the other modality (Fig. 4b, c). In temporal and parietal ROIs, we could decode scene categories, but these ROIs were no longer multimodal (Fig. 4d); they only represented scene categories in either the visual or auditory domain but no longer both. These findings suggest that these ROIs contain separate but intermixed neural populations for visual and auditory information. For ROIs in PFC, we found that conflicting audiovisual stimuli interfered heavily with representations of scene categories (Fig. 4e, f). Scene categories could no longer be decoded in PFC from either modality, except for visual scenes in the right MFG. Presumably, this break-down of the decoding of scene categories is due to the conflicting information from the two sensory modalities arriving at the same cross-modal populations of neurons. Directing participants' attention to one of the modalities did not result in a recovery of decoding (see Fig. S2).

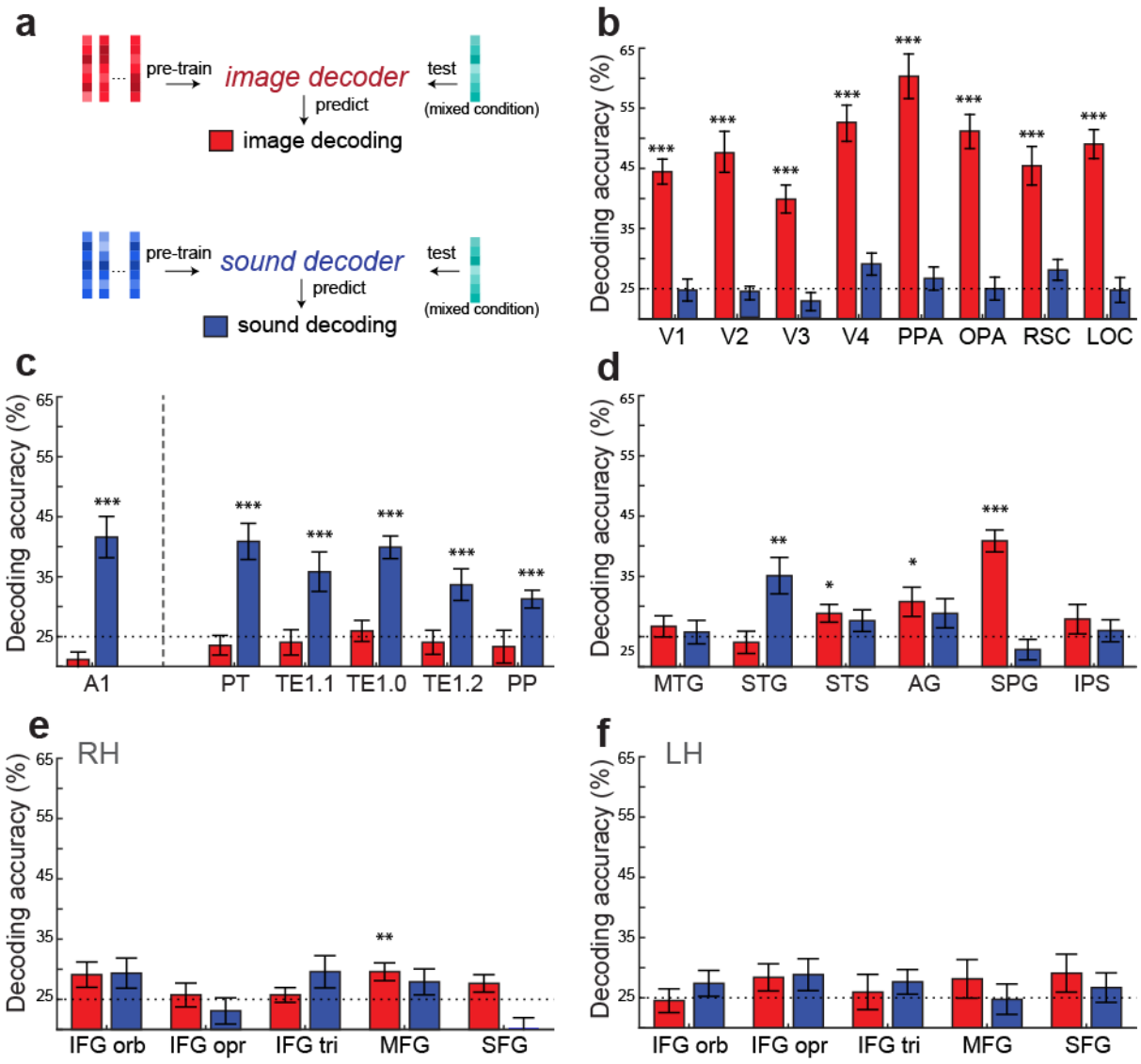



**Figure 4** Scene category decoding in the mixed condition. A classifier was pre-trained on the data from the pure image or sound conditions and then tested with the data from the mixed condition using either image or sound labels (**a**). It predicted the image categories (red) or sound categories (blue) from visual (**b**), auditory (**c**), temporal and parietal cortices (**d**), as well as PFC (**e, f**). The significance of the one-sample t-tests (one-tailed) was adjusted for multiple comparisons (FDR) and marked above each bar. The error bars are SEM.

*Error correlations*

To further explore the characteristics of the neural representations of scene categories, we compared the patterns of errors from the neural decoders with those from human behavior as well as with the physical attributes of the stimuli. If the neural representation of scenes in a

certain brain region is used directly for categorical decisions, then error patterns from this ROI should be similar to errors made in behavioral categorization (Walther et al., 2009). However, in early stages of neural processing, scene representations might reflect the physical properties of the scene images or sounds. In this case, the error patterns of the decoders should resemble the errors that a classifier would make solely based on low-level physical properties.

The errors from the neural decoders were recorded in a confusion matrix, whose rows indicate the true category of the stimulus, and whose columns indicate the prediction of the decoder. Images were characterized by the spatial distribution of oriented edges at four different scales. Sounds were characterized by the distribution of power over 128 frequency bands. These physical characteristics were used as features for a linear support vector machine classifier in a 16-fold leave-one-out cross validation, resulting in confusion matrices for image and sound properties (Fig. S3; top). The classifier accurately categorized 85.8% of the scene images and 57.8% of the scene sounds (chance: 25%).

We obtained behavioral error patterns from a separate behavioral experiment, in which participant classified images and sounds of the scenes into four categories (beaches, forests, cities, offices). We ensured that participants would make mistakes by restricting presentation duration of the images to 26.7ms, followed by a perceptual mask. The quality of the sound clips was degraded by superimposing pure tone noise at nine times the volume of the soundscapes themselves. Mean accuracy was 76.6% for the visual task (std = 12.35%, mean RT = 885.6 ms) and 78.1% for the auditory task (std = 6.86%, mean RT = 7.84 sec). Behavioral errors were recorded in confusion matrices, separately for images and sounds.

To assess similarity of representations, we correlated the patterns of errors (off-diagonal elements of the confusion matrices) between the neural decoders, physical structure of the stimuli, and human behavior (see Methods & Fig. S3). Statistical significance of the correlations was established with non-parametric permutation tests. Here we considered error correlations to be significant when none of the correlations in the null set exceeded the correlation of the correct ordering of the categories ($p < 0.0417$).

Behavioral errors from image categorization were not correlated with the errors derived from image properties ($r = -.458$, $p = .917$), suggesting that behavioral judgment of scene categories was not directly driven by low-level physical differences between the images. There was, however, a positive error correlation between the auditory task and physical properties of sounds ($r = .407$, $p < .0417$).

For the image condition, decoding errors were positively correlated with human behavior in the PPA and the RSC as well as in the right MFG and SFG in PFC (Fig. 5a), suggesting that scene representations in PFC reflects human categorization behaviors. Also, in the sound condition,

we found that error patterns from both sound structure and sound behavior are correlated with decoding errors in some of the sub-divisions of A1 and the right SFG (Fig. 5b).

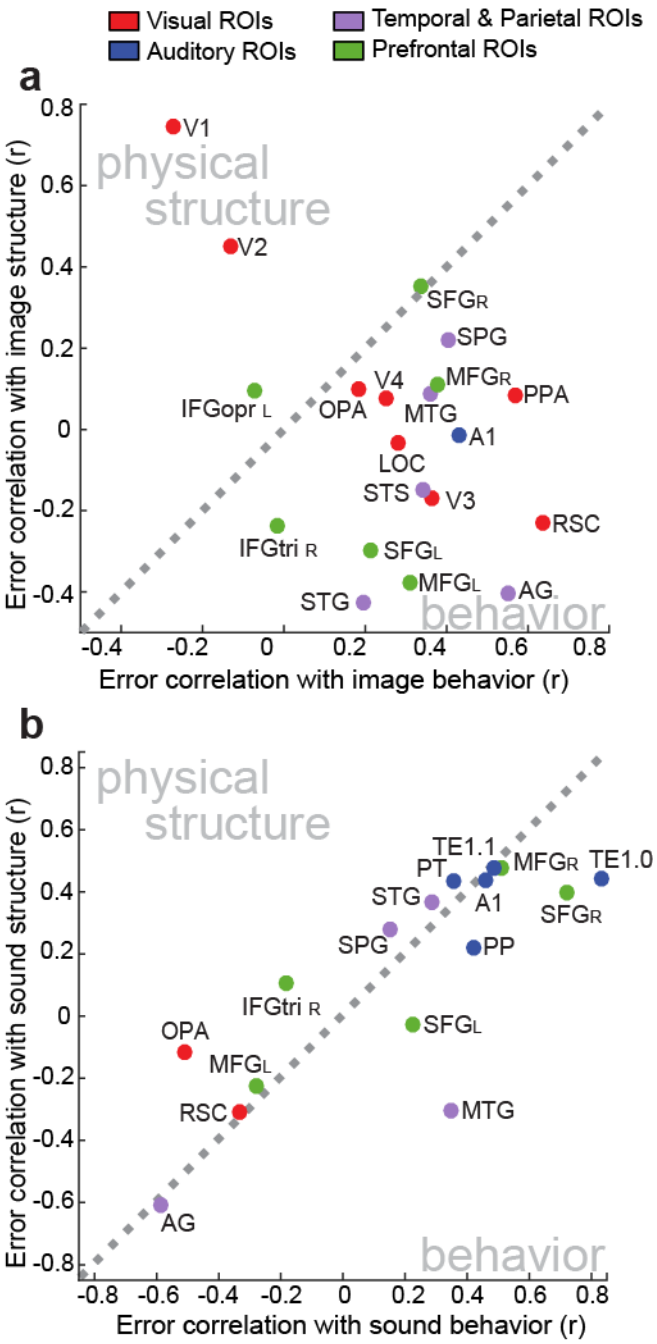



**Figure 5** Error correlations of the neural decoder with behavior and stimuli structure in the image (**a**) and sound (**b**) conditions. Diagonal axes indicate the points where the error correlations with behavior and with stimuli structure are equal. In the image condition, we see a clear progression from V1 through V2-4 to higher-level visual areas (in red), moving from strong error correlation with image structure to strong error correlation with visual behavior. Error patterns from decoding image categories from PFC (in

green) are most similar to visual behavior. In the sound condition, all ROIs are close to the diagonal, because sound structure and sound behavior error patterns are significantly correlated with each other. We see A1 and its subdivisions (in blue) high along the diagonal, indicating strong similarity with both sound structure and sound behavior. OP and RSC (in red) show low error correlation with either structure or behavior. Prefrontal ROIs (in green) are distributed along the diagonal, with right MFG and right SFG showing high and left MFG showing low error correlations.

*Whole-brain analysis*

In order to explore representations of scene categories beyond the pre-defined ROIs, we performed a whole-brain searchlight analysis with a size of 7x7x7 voxels (21x21x21 mm) cubic searchlight. The same LORO cross-validation analysis for image and sound conditions as well as the same two cross-decoding analyses as for the ROI-based analysis were performed at each searchlight location, followed by a cluster-level correction for multiple comparisons. For each decoding condition, we found several spatial clusters with significant decoding accuracy. Some of these clusters confirmed the pre-defined ROIs, others revealed scene representations in unexpected regions beyond the ROIs.

For the image and sound decoding, we found consistent results with our ROI-based approach: the respective sensory cortices as well as the right PFC areas showed clusters with significant decoding accuracy for the image or sound scene categories (Fig. 6& Fig. S4 a). Most of the clusters with significant decoding accuracy for cross-decoding were found in PFC, supporting the notion that only in PFC, scene information is computed beyond the modality level (Fig. 6& Fig. S4 b).

We performed the error pattern analysis on the clusters with significant decoding accuracy. For both image (Fig. 7a) and sound conditions (Fig. 7b), we observed a posterior-to-anterior (PA) trend; with voxels in the posterior (low-level) regions more closely matched to stimulus properties and with voxels more anterior (high-level) regions more closely related to behavior (see the sagittal view in Fig. 7& Fig S5).

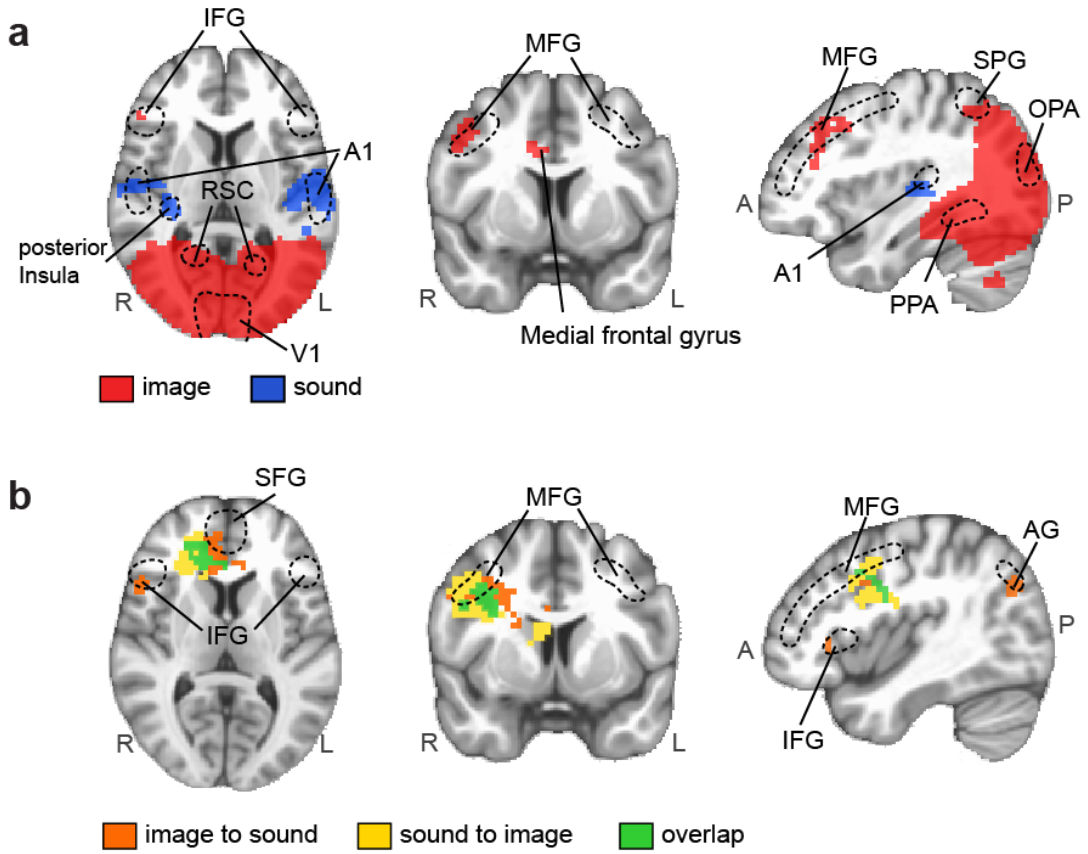



**Figure 6** Neural decoding accuracy maps for image and sound categories (**a**) and those for cross-decoding analysis (**b**). (**a**) Searchlight locations with above-chance decoding of images and sounds are highlighted in red and blue respectively (thresholded at $p < .01$ with a cluster-based multiple comparison correction). (**b**) Searchlight locations with above-chance level cross-decoding accuracy for image to sound (orange), sound to image (yellow), and both (green) (thresholded at $p < .01$ with a cluster-based multiple comparison correction). MNI coordinates of the sections shown: x = -42, y = -6, z = 7. For an exhaustive set of axial images see Fig. S4.

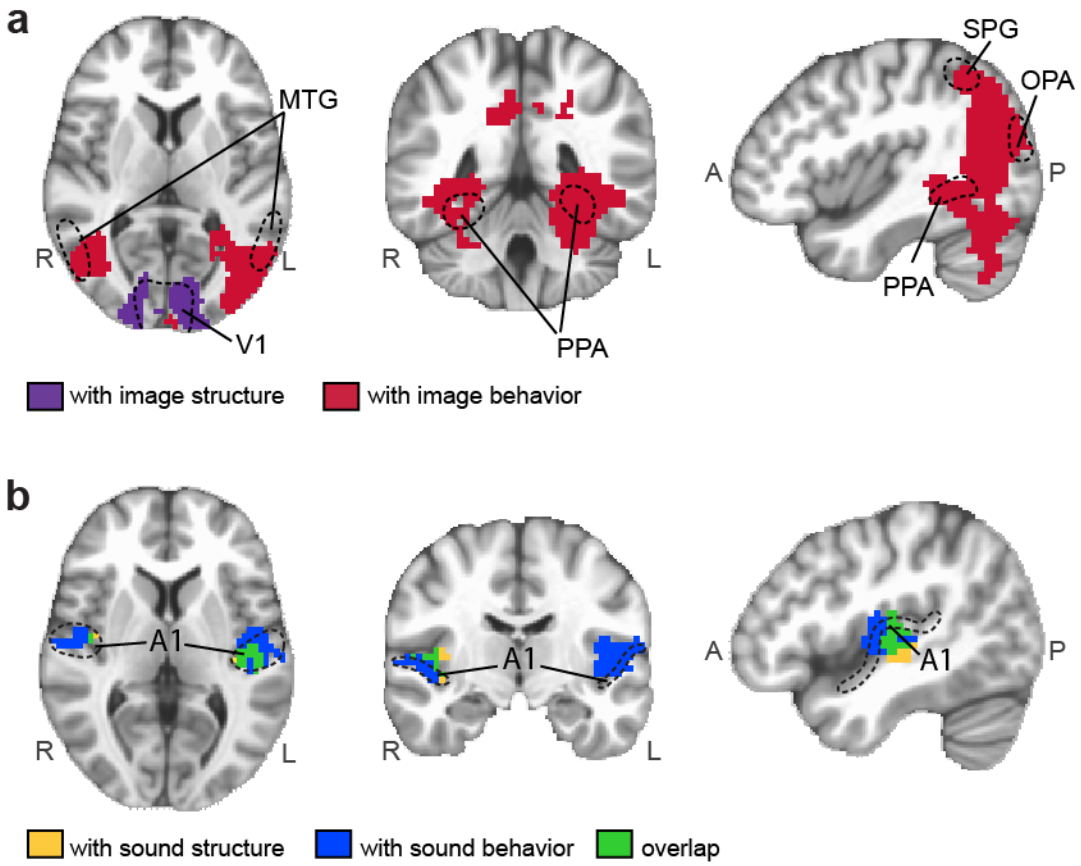



**Figure 7** Error correlation maps for the error correlation analysis in image (**a**) and sound (**b**) conditions. Only searchlight locations with significant decoding accuracy were included. (**a**) Searchlight locations with significant correlations with image structure and image behavior are highlighted in purple and magenta, respectively (thresholded at $p < .05$ with a cluster-based multiple comparison correction). MNI coordinates of the sections shown: x = -42, y = 47, z = 7. (**b**) Searchlight locations with significant correlations with sound structure and sound behavior are highlighted in yellow and blue, respectively (threshold at $p < .05$ with a cluster-based multiple comparison correction). Those locations with significant correlations for both sound structure and sound behavior are marked in green. MNI coordinates of the sections shown: x = 46, y = 11, z = 5. For an exhaustive set of axial images see Fig. S5.

## Discussion

The present study investigated where and how scene information from different sensory modalities forms categorical representations at the conceptual level. We have found that both visual and auditory stimuli of the natural environment elicit representations of scene categories in sub-regions of prefrontal cortex (PFC). These neural representations of scene categories generalize across sensory modalities and resemble human categorization behavior, suggesting that scene representations in PFC reflect scene categories at a conceptual level. To our knowledge, our study is the first to demonstrate a neural representation of scenes at such an abstract level.

There have been attempts to address how sensory information is integrated to form a concept, especially with audiovisual stimuli (Beauchamp, Lee, Argall, & Martin, 2004; Hsier, Colas, & Kanwisher, 2012). However, those studies employed univariate analysis, focusing on how the BOLD signal reflects the integration (Beauchamp et al., 2004; Downar et al., 2000) or correlations (Hsier et al., 2012) of content-specific visual and auditory information in certain brain area. The present study expands on those studies by showing that similar neural activity patterns are elicited by visual and auditory stimuli when they represent the same concept of a scene category. Our results support previous findings showing that PFC is involved in the computation of categorical representations (Freedman et al., 2001; Meyers et al., 2008; Miller & Cohen, 2001; Mack et al. 2013).

Three distinct characteristics support the idea that neural representations of scene categories in PFC are distinct from those in modality-specific areas such as the visual or the auditory cortices. First, both image and sound categories could be decoded from the same areas in PFC. Thus, it can be inferred that neural representations of scene categories in PFC are not limited to a specific sensory modality. Second, the representations in PFC could be cross-decoded from one modality to the other, showing that the category-specific neural activity patterns were similar across the sensory modalities. Third, when subjects were presented with incongruous visual and auditory scene information simultaneously, it was no longer possible to decode scene categories in PFC, whereas modality-specific areas still carried the category-specific neural activity patterns. This result shows that inconsistent information entering through the two sensory channels in the mixed condition interferes, preventing the formation of clear concepts of scene categories in PFC.

Although scene categories could be decoded from both images and sounds in several ROIs in the temporal and parietal lobes, cross-decoding across sensory modalities was not possible there, suggesting that neural representations elicited by visual inputs were not similar to those elicited by auditory inputs. Further supporting the idea that visual and auditory representations are separate but intermixed in these regions, decoding of scene categories from the visual or

auditory domain was still possible in the presence of a conflicting signal in the other domain. These findings suggest that even though information from both visual and auditory scenes is present in these regions (Beauchamp et al., 2004; Calvert, Hansen, Iversen, & Brammer, 2001), scene information is computed separately for each sensory modality, unlike in PFC. The discrimination between multi-modal and truly cross-modal representations is not possible with the univariate analysis techniques used in those studies.

Analysis of decoding errors demonstrated that the category representations in the visual areas have a hierarchical organization. In the early stage of processing, categorical representations are formed based on the physical properties of visual inputs, whereas in the later stage, the errors of neural decoders correlate with human behavior, confirming previous findings which mainly focused on the scene-selective areas (Walther et al., 2009; 2011; Choo & Walther, 2016). Significant error correlation between human behavior and the neural decoders in prefrontal areas confirms that this hierarchical organization is extended to PFC, beyond the modality-specific areas such as PPA, OPA, or RSC.

Intriguingly, no similar hierarchical structure of category representations was found in the auditory domain. Both types of the errors, the errors representing the physical properties and those from human behavior, were correlated to the errors of neural decoder in the A1. This difference between the visual and the auditory domain might reflect the fact that much of auditory processing occurs in sub-cortical regions, before the information arrives in auditory cortex. Further experiments using different manipulations on auditory features and their statistical structure will be needed to clarify what kind of auditory features contribute to human auditory categorization.

Previous fMRI studies have shown that auditory content can be decoded from early visual cortex, suggesting cross-modal interactions in the modality-specific areas (Paton et al., 2016; Vetter et al., 2014). Although we did not observe representations of auditory scenes in the early visual areas, our data show that auditory content can be decoded from high-level scene-selective areas (RSC and OPA). Visual content can be decoded from A1. This seeming inconsistency with previous studies might be driven by different levels of complexity and diversity of the auditory stimuli used in each experiment: Vetter et al. (2014) used few exemplars of object-level sounds. We used more complex sounds with multiple objects overlapping in time, recorded from real-world settings, whose statistics could only be classified at the category level. Thus, category-specific visual imagery caused by auditory stimulation may underlie the decoding of auditory scene categories in RSC and OPA. These findings lead to a host of further questions for future research, such as how these visual and auditory areas are functionally connected, whether the multisensory areas mediate this interaction between the

visual and auditory areas by sending feedback signals, or whether these cross-modal representations can influence or interfere with perceptual sensitivity in each sensory domain.

The whole-brain searchlight analysis confirmed the findings of our ROI-based analysis. In the image and sound decoding analyses, we found clusters with significant decoding accuracies in the visual and the auditory areas as well as in the temporal, the parietal, and the prefrontal regions. Furthermore, the clusters in the prefrontal areas showed significant accuracy in the cross-decoding analysis, while the clusters in other modality-specific or multimodal areas did not, supporting the view that only in PFC representations are computed that transcend sensory modalities. In the analysis of decoding errors, we observed that the errors of the image decoders were significantly correlated with human categorization behavior in scene-selective areas as well as in parietal regions, consistent with previous work by our group (Walther et al., 2009; 2011; Choo & Walther, 2016).

In a recent review, Grill-Spector and Weiner (2014) suggested that the ventral temporal cortex contains a hierarchical structure for visual categorization, which has the more exemplar-specific representations in posterior areas, but the more abstract representations in anterior areas of the ventral temporal cortex. In the present study, we show that the posterior-to-anterior hierarchy of levels of abstraction extends to the PFC, which represents concepts at an amodal level. The abstraction and generalization across sensory modalities is likely to contribute to the efficiency of cognition by representing similar concepts in a consistent manner, even when the physical signal might be delivered via different sensory channels (Huth, de Heer, Griffiths, Theunissen, & Gallant, 2016).

Here we investigated the representation of abstract categories in the absence of a particular task. In fact, categorization at the basic level has been shown to be automatic and compulsory (Greene & Fei-Fei 2014). It is less clear how task context or intentions shape the representation of categories, which should reflect the utility of a particular type of scene in a given behavioral context (Greene, Baldassano, Esteva, Beck, & Fei-Fei, 2016). Further studies with diverse task settings are needed to answer this question.

## Materials and Methods

### fMRI experiment

*Participants*

Thirteen subjects (18 to 25 years old, 6 females; 7 males) participated in the fMRI experiment. All participants were in good health with no past history of psychiatric or neurological disorders and reported having normal hearing and normal or corrected-to-normal vision. They gave written informed consent before the experiment began according to the Institutional Review Board of The Ohio State University.

*Stimuli*

Participants were shown 640 color photographs of natural scenes of four different scene categories (beaches, forests, cities, offices). The images have previously been rated as the best exemplars of their categories from a data base of about 4000 images that were downloaded from the internet (Torralbo et al. 2013). Images were presented at a resolution of 800x600 pixels using a Christie DS+6K-M projector operating at a refresh rate 60Hz. Images subtended approximately 21x17 degrees of visual angle.

Sixty-four sound clips representing the same four scene categories (beaches, forests, cities, or offices) were used as auditory stimuli in both experiments. The sound clips were purchased from various commercial sound libraries and edited to 15 seconds of length. Perceived loudness was equated using Replay Gain as implemented in the Audacity sound editing software (Audacity Team, 2012). In a pilot experiment, the sound clips were correctly identified and rated as highly typical for their categories by 14 naïve subjects.

*Procedures*

Participants' brains were scanned during twelve experimental runs, four runs with images only, four runs with sounds only, and four runs with images and sounds presented concurrently. Each run started with the instruction asking participants to attend, for the duration of the run, to either images (image runs and half of the mixed runs) or sounds (sound runs and the other half of the mixed runs).

Runs contained eight blocks, two for each scene category, interleaved with 12.5 s fixation periods to allow for the hemodynamic response to return to baseline levels. The beginning and the end of a run also included a fixation period of 12.5 sec. The order of blocks within runs and the order of runs were counter-balanced across participants. Mixed runs were only presented after at least two pure image and sound runs. Stimuli were arranged into eight blocks of 15 seconds duration. During image blocks participants were shown ten color photographs of the same scene category for 1.5 seconds each. During sound blocks they were shown a blank screen with a fixation cross, and a 15-second sound clip was played using Sensimetrics S14 MR-

compatible in-ear noise-canceling headphones at approximately 70 dB. During mixed blocks participants were shown images and played a sound clip of a different scene category at the same time. A fixation cross was presented throughout each block, and subjects were instructed to maintain fixation. Each run lasted 3 min and 52.5 seconds.

*fMRI data acquisition and Preprocessing*

Imaging data were recorded on a 3 Tesla Siemens MAGNETOM Trio MRI scanner with a 12-channel head coil at the Center for Cognitive and Behavioral Brain Imaging (CCBBI) at The Ohio State University. High resolution anatomical images were acquired with a 3D-MPRAGE (magnetization-prepared rapid acquisition with gradient echo) sequence with sagittal slices covering the whole brain; inversion time = 930ms, repetition time (TR) = 1900ms, echo time (TE) = 4.68ms, flip angle = 9°, voxel size = 1 x 1 x 1 mm, matrix size = 224 x 256 x 160 mm. Functional images for the main experiment were recorded with a gradient echo, echo-planar imaging sequence with a volume repetition time (TR) of 2.5 s, an echo time (TE) of 28ms and a flip angle of 78 degree. 48 axial slices with 3 mm thickness were recorded without gap, resulting in an isotropic voxel size of 3 x 3 x 3 mm.

FMRI data were motion corrected to one EPI image (the 72nd volume of the 10th run), followed by spatial smoothing with a Gaussian kernel with 2 mm full width at half maximum (FWHM) and temporal filtering with a high-pass filter at 1/400 Hz. Data were normalized to percent signal change by subtracting the mean of the first fixation period in each run and dividing by the mean across all runs. The effects of head motion (6 motion parameters) and scanner drift (second degree polynomial) were regressed out using a general linear model (GLM). The residuals of this GLM analysis were averaged over the duration of individual blocks, resulting in 96 brain volumes that were used as input for a multi-voxel pattern analysis (MVPA). Preprocessing was performed using AFNI (Cox, 1996).

*Multi-voxel pattern analysis*

Neural representations of scene categories were assessed by decoding scene categories from the neural activity patterns in a range of regions of interest (ROI). The following ROIs were defined using separate localizer scans: V1-V4, parahippocampal place area (PPA), occipital place area (OPA), retrosplenial cortex (RSC), and lateral occipital complex (LOC). Other ROIs were defined using an anatomical atlas (Destrieux et al., 2010; Norman-Haignere, Kanwisher, & McDermott, 2013): A1 and its subdivisions (Planum Temporale, Posteromedial Heschl's gyrus, Middle Heschl's gyrus, Anterolateral Heschl's gyrus, & Planum Polare) middle temporal gyrus (MTG), superior temporal gyrus (STG), superior temporal sulcus (STS), angular gyrus (AG), superior parietal gyrus (SPG), intraparietal sulcus (IPS), middle frontal gyrus (MFG), superior

frontal gyrus (SFG), and inferior frontal gyrus (IFG) with pars opercularis, pars orbitalis, and pars triangularis.

For each participant, we trained a linear support vector machine (SVM; using LIBSVM, Chang & Lin, 2001) to assign the correct scene category labels to the voxel activations inside an ROI based on the fMRI data from all runs except one. The SVM decoder then produced predictions for the labels of the data in the left-out run. This leave-one-run-out (LORO) cross validation procedure was repeated with each run being left out in turn, thus producing predicted scene category labels for all runs. Decoding accuracy was assessed as the fraction of blocks with correct category labels. Group-level statistics was computed over all thirteen participants using one-tailed t tests, determining if decoding accuracy was significantly above chance level (0.25). Significance of the t-test was adjusted for multiple comparisons using false discovery rate (FDR) (Westfall & Young, 1993).

To curb over-fitting of the classifier to the training data, we reduced the dimensionality of the neural data by selecting a subset of voxels in each ROI. Voxel selection was performed by ranking voxels in the training data according to the F statistics of a one-way ANOVA of each voxel's activity with scene category as the main factor (Pereira et al. 2009). We determined the optimal number of voxels by cross validation within the training data. In the case of cross validation analysis of the pure image and sound conditions, voxel selection was performed in a nested cross-validation, using the training data of each cross validation fold. Optimal voxel numbers varied by ROI and participant but were generally between 100 and 1000 (mean voxel number averaged across all ROIs and subjects = 107.125).

*Error correlations*

Category label predictions of the decoder were recorded in a confusion matrix, whose rows indicate the category of the stimulus, and whose columns represent the category predictions by the decoder. Diagonal elements indicate correct predictions, and off-diagonal elements represent decoding errors. Neural representations of scene categories were compared with human behavior by correlating the error patterns (the off-diagonal elements of the confusion matrices) between neural decoding and behavioral responses (Walther, Beck, & Fei-Fei 2012). Statistical significance of the correlations was established non-parametrically against the null distribution of all error correlations that were obtained by jointly permuting the rows and columns of one of the confusion matrices in question (24 possible permutations of four labels). Error correlations were considered as significant when none of the correlations in the null set exceeded the correlation for the correct ordering of category labels ($p < 0.0417$).

To assess the similarity between neural representations and the physical characteristics of the stimuli, we constructed simple computational models of scene categorization based on low-level stimulus features. Scene images were filtered with a bank of Gabor filters with four

different orientations at four scales. Images were categorized based on the resulting feature vector in a 16-fold cross validation, using a linear SVM. Error patterns from the computational analysis were correlated with error patterns from the neural decoder.
Physical properties of the sounds were assessed using a cochleagram, which mimics the biomechanics of the human ear (Meddis, Hewitt, & Shackleton, 1990; Wang & Brown, 2006). The cochleagrams of individual sound clips were integrated over their duration and subsampled to 128 frequency bands, resulting in a biomechanically realistic frequency spectrum. The activation of the frequency bands was used as input to a linear SVM, which predicted scene categories of sounds in a 16-fold cross validation. Again, error patterns from this analysis were correlated with those obtained from the neural decoder.

*Searchlight analysis*

To explore representations of scene categories outside pre-defined ROIs, we performed a searchlight analysis. We defined a cubic "searchlight" of 7x7x7 voxels (21x21x21 mm). The searchlight was centered on each voxel in turn (Kriegeskorte, Göbel, & Bandettini, 2006), and LORO cross-validation analysis was performed within each searchlight location using a Gaussian Naïve Bayes classifier until all voxels served as the center of the searchlight (Searchmight Toolbox; Pereira & Botvinick 2011). Decoding accuracy as well as the full confusion matrix at a given searchlight location, were assigned to the central voxel.
For group-analysis, we first co-registered each participant's anatomical brain to the Montreal Neurological Institute (MNI) 152 template using a diffeomorphic transformation as calculated by AFNI's 3dQWarp. We then used the same transformation parameters to register individual decoding accuracy maps to MNI space using 3dNWarpApply, followed by spatial smoothing with a 4 mm FWHM Gaussian filter. To identify voxels with decodable categorical information at the group level, we performed one-tailed t-tests to test whether decoding accuracy at each searchlight location was above chance (0.25). After thresholding at $p < .01$ (one-tailed) we conducted a cluster-level correction for multiple comparisons, applying a minimum cluster size of 15 voxels, the average cluster size obtained from the $\alpha$ probability simulations conducted for each participant.

## Acknowledgments

We thank Michael Mack and Heeyoung Choo for their helpful comments on the early version of this manuscript. This work is supported by NSERC Discovery Grant (#498390) and Canadian Foundation for Innovation (#32896).

## References


1. Dilks, D. D., Julian, J. B., Paunov, A. M., & Kanwisher, N. (2013). The occipital place area is causally and selectively involved in scene perception. *Journal of Neuroscience*, 33(4), 1331-1336.
2. Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, 392(6676), 598-601.
3. Maguire, E. (2001). The retrosplenial contribution to human navigation: a review of lesion and neuroimaging findings. *Scandinavian Journal of Psychology*, 42(3), 225-238.
4. Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science*, 291(January), 312–317.
5. Mack, M. L., Preston, A. R., & Love, B. C. (2013). Decoding the brain's algorithm for categorization from its neural implementation. *Current Biology*, 23(20), 2023–2027.
6. Meyers, E. M., Freedman, D. J., Kreiman, G., Miller, E. K., & Poggio, T. (2008). Dynamic population coding of category information in inferior temporal and prefrontal cortex. *Journal of Neurophysiology*, 100(June 2008), 1407–1419. http://doi.org/10.1152/jn.90248.2008
7. Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, 24, 167–202. http://doi.org/10.1146/annurev.neuro.24.1.167
8. Wood, J. N., & Grafman, J. (2003). Human prefrontal cortex: processing and representational perspectives. *Nature Reviews Neuroscience*, 4(2), 139–147. http://doi.org/10.1038/nrn1033.
9. Walther, D. B., Caddigan, E., Fei-Fei, L., & Beck, D. M. (2009). Natural Scene Categories Revealed in Distributed Patterns of Activity in the Human Brain. *Journal of Neuroscience*, 29(34), 10573–10581.
10. Chang, C. C., & Lin, C. J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *2*(3), 27.
11. Choo, H., & Walther, D. B. (2016). Contour junctions underlie neural representations of scene categories in high-level human visual cortex. *NeuroImage*, 135, 32–44.
12. Park, S., Brady, T. F., Greene, M. R., & Oliva, A. (2011). Disentangling scene content from spatial boundary: complementary roles for the parahippocampal place area and lateral occipital complex in representing real-world scenes. *Journal of Neuroscience*, 31(4), 1333–1340.
13. Walther, D. B., Chai, B., Caddigan, E., Beck, D. M., & Fei-Fei, L. (2011). Simple line drawings suffice for functional MRI decoding of natural scene categories. *Proceedings of the National Academy of Sciences*, *108*(23), 9661-9666.
14. Kravitz, D. J., Peng, C. S., & Baker, C. I. (2011). Real-world scene representations in high-level visual cortex: it's the spaces more than the places. *Journal of Neuroscience*, *31*(20), 7322-7333.
15. Paton, A., Petro, L., & Muckli, L. (2016). An Investigation of Sound Content in Early Visual Areas. *Journal of Vision*, 16(12), 153-153.
16. Vetter, P., Smith, F. W., & Muckli, L. (2014). Decoding sound and imagery content in early visual cortex. *Current Biology*, 24(11), 1256–1262. http://doi.org/10.1016/j.cub.2014.04.020
17. Gaffan, D., & Harrison, S. (1991). Auditory-visual associations, hemispheric specialization and temporal-frontal interaction in the rhesus monkey. *Brain*, *114*(5), 2133-2144.

18. Goel, V., Tierney, M., Sheesley, L., Bartolo, A., Vartanian, O., & Grafman, J. (2007). Hemispheric specialization in human prefrontal cortex for resolving certain and uncertain inferences. *Cerebral cortex*, *17*(10), 2245-2250
19. Slotnick, S. D., & Moo, L. R. (2006). Prefrontal cortex hemispheric specialization for categorical and coordinate visual spatial memory. *Neuropsychologia*, *44*(9), 1560-1568.
20. Beauchamp, M., Lee, K., Argall, B., & Martin, A. (2004). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron*, 41, 809–823.
21. Hsieh, P.-J., Colas, J. T., & Kanwisher, N. (2012). Spatial pattern of BOLD fMRI activation reveals cross-modal information in auditory cortex. *Journal of Neurophysiology*, 107(12), 3428–3432.
22. Downar, J., Crowley, A. P., Mikulis, D. J., & Davis, K. D. (2000). A multimodal cortical network for the detection of changes in the sensory environment. *Nature Neuroscience*, 3(3), 277–283.
23. Calvert, G. A., Hansen, P. C., Iversen, S. D., & Brammer, M. J. (2001). Detection of audio-visual integration sites in humans by application of electrophysiological criteria to the BOLD effect. *NeuroImage*, 14(2), 427–38. http://doi.org/10.1006/nimg.2001.0812
24. Grill-Spector, K., & Weiner, K. S. (2014). The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, 15(8), 536–548.
25. Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453-458.
26. Greene, M. R., & Fei-Fei, L. (2014). Visual categorization is automatic and obligatory: Evidence from Stroop-like paradigm. *Journal of vision*, *14*(1), 1-14.
27. Greene, M. R., Baldassano, C., Esteva, A., Beck, D. M., & Fei-Fei, L. (2016). Visual scenes are categorized by function. *Journal of Experimental Psychology: General*, *145*(1), 82-94.
28. Torralbo, A., Walther, D. B., Chai, B., Caddigan, E., Fei-Fei, L., & Beck, D. M. (2013). Good exemplars of natural scene categories elicit clearer patterns than bad exemplars but not greater BOLD activity. *PloS one*, *8*(3), e58594.
29. Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical research*, *29*(3), 162-173.
30. Westfall, P. H., & Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment* (Vol. 279). John Wiley & Sons.
31. Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage*, *45*(1), S199-S209.
32. Walther, D. B., Beck, D. M., & Fei-Fei, L. (2012). To err is human: Correlating fMRI decoding and behavioral errors to probe the neural representation of natural scene categories. *Visual population codes–Toward a common multivariate framework for cell recording and functional imaging*, 391-416.
33. Meddis, R., Hewitt, M. J., & Shackleton, T. M. (1990). Implementation details of a computation model of the inner hair-cell auditory-nerve synapse. *The Journal of the Acoustical Society of America*, *87*(4), 1813-1816.
34. Wang, D., & Brown, G. J. (2006). *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE Press.

35. Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National academy of Sciences of the United States of America*, *103*(10), 3863-3868.
36. Pereira, F., & Botvinick, M. (2011). Information mapping with pattern classifiers: a comparative study. *Neuroimage*, *56*(2), 476-496.
37. Kastner, S., De Weerd, P., Desimone, R., & Ungerleider, L. G. (1998). Mechanisms of directed attention in the human extrastriate cortex as revealed by functional MRI. *Science*, *282*(5386), 108-111.
38. Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, *17*(11), 4302-4311.
39. Malach, R., Reppas, J. B., Benson, R. R., Kwong, K. K., Jiang, H., Kennedy, W. A., ... & Tootell, R. B. (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proceedings of the National Academy of Sciences*, *92*(18), 8135-8139.
40. Destrieux, C., Fischl, B., Dale, A., & Halgren, E. (2010). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *Neuroimage*, *53*(1), 1-15.
41. Norman-Haignere, S., Kanwisher, N., & McDermott, J. H. (2013). Cortical pitch regions in humans respond primarily to resolved harmonics and are located in specific tonotopic regions of anterior auditory cortex. *Journal of Neuroscience*, *33*(50), 19451-19469.
42. Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International journal of computer vision*, *40*(1), 49-70.