Gabor and Non-Gabor Neural Representations Are Shared between Visual Perception and Mental Imagery

Yingying Huang[1,2], Frank Pollick[2], Ming Liu[1], and Delong Zhang[1,3]

Abstract

■ Visual perception and mental imagery have been shown to share a hierarchical topological visual structure of neural representation, despite the existence of dissociation of neural substrate between them in function and structure. However, we have limited knowledge about how the visual hierarchical cortex is involved in visual perception and visual imagery in a unique and shared fashion. In this study, a data set including a visual perception and an imagery experiment with human participants was used to train 2 types of voxel-wise encoding models. These models were based on Gabor features and voxel activity patterns of high-level visual cortex (i.e., fusiform face area, parahippocampal place area, and lateral occipital complex) to predict activity in the early visual cortex (EVC, i.e., V1, V2, V3) during perception, and then tested with respect to the generalization of these models to mental imagery. Our results showed that during perception and imagery, activities in the EVC could be independently predicted by the Gabor features and activity of high-level visual cortex via voxel-wise encoding models, which suggested that perception and imagery might share neural representation in the EVC. We further found Gabor-specific and non-Gabor-specific patterns of neural response to stimuli in the EVC, which were shared by perception and imagery. These findings provide insight into the mechanisms of how visual perception and imagery share representation in the EVC. ■

> *Whilst part of what we perceive comes through our senses from the object before us, another part (and it may be the larger part) always comes out of our own head.*
>
> William James (1890), *The Principles of Psychology*

INTRODUCTION

Every day, we are bombarded with an abundance of visual stimuli, such as colors, textures, and objects, and our brain then selectively processes the information to generate visual perception. Therefore, visual perception is a kind of reflection of the interaction between feedforward sensory input, externally driven by "bottom–up" pathways, and feedback signals, internally generated from "top–down" pathways (Hsieh, Vul, & Kanwisher, 2010; Kastner, De Weerd, Desimone, & Ungerleider, 1998). The typical internally generated mental process is mental imagery, which refers to the generation and representation of a visual image without corresponding feedforward stimuli from the real world (Andersson, Ragni, & Lingnau, 2019; Kosslyn & Thompson, 2003; Kosslyn, 1996). What is the relationship between visual perception and mental imagery? What are the differences and similarities between the visual experiences generated from both mental processes?

Considerable empirical evidence from behavioral and neurobehavioral studies supports the claim that visual perception and mental imagery have similar functions during sensory processing. For example, when participants were required to complete a mental image scanning task, a free view scanning task, and an iconic image scanning task with measured RT and error rate for each task, it was found that participants could achieve similar performance between scanning mental images and visual perception (Borst & Kosslyn, 2008). Kosslyn and Pearson have proposed that mental imagery resembles weak perception (Pearson, 2019; Pearson & Kosslyn, 2015), based on the similar neural mechanism of sensory processing between them (Maier, Frömer, Rost, Sommer, & Rahman, 2021; Xie, Kaiser, & Cichy, 2020; Dijkstra, Bosch, & van Gerven, 2019; Cichy, Heinzle, & Haynes, 2012; Reddy, Tsuchiya, & Serre, 2010; Stokes, Thompson, Cusack, & Duncan, 2009; Borst & Kosslyn, 2008; Ishai & Sagi, 1995). According to perception anticipation theory (Aitken, Turner, & Kok, 2020; Kok, Jehee, & de Lange, 2012; Sohoglu, Peelle, Carlyon, & Davis, 2012; Kosslyn & Thompson, 2003), expectation or prediction may have some impact on the top–down modulation process of perception (Berger & Ehrsson, 2013; Diekhof et al., 2011; Pearson, Clifford, & Tong, 2008). Some fMRI studies have provided evidence that mental imagery and visual perception share brain activity patterns in the early visual cortex (EVC, i.e., V1,

[1] South China Normal University, Guangzhou, China, [2] University of Glasgow, United Kingdom, [3] Kashi University, Kashgar, China

Journal of Cognitive Neuroscience 35:6, pp. 1045–1060
https://doi.org/10.1162/jocn_a_01992

V2, and V3) encoding low-level visual features (Maier et al., 2021; Kosslyn & Thompson, 2003) as well as the high-level visual areas encoding category information (Lee, Kravitz, & Baker, 2012; Stokes et al., 2009; O'Craven & Kanwisher, 2000). In earlier years, researchers used positron emission topography and repetitive TMS to demonstrate that EVC is involved in visual imagery (Kosslyn et al., 1999), and recent research with a multi-method approach (fMRI, TMS, and transcranial direct current stimulation) suggests that EVC serves a causative role in visual imagery (Keogh, Bergmann, & Pearson, 2020).

EVC plays an important role in many cognitive functions such as perception, memory, attention, and imagination. The neurons in the EVC not only receive sensory input and passively transfer to high areas, but also receive more feedback and lateral information from other areas (Muckli, Petro, & Smith, 2013; Budd, 1998). Moreover, the feedforward information from retina to primary visual cortex is only coordinated with 5% excitatory input from the LGN (Douglas & Martin, 2007), and only 20% of neural response in V1 could be explained by retinal input (Carandini et al., 2005). Consequently, the neurons from V1 receive more input from other cortical areas than retinal input.

Advanced machine learning approaches have been used in processing fMRI data to provide a novel viewpoint for exploring the specific shared neural representation between visual perception and mental imagery (Horikawa & Kamitani, 2017; Albers, Kok, Toni, Dijkerman, & de Lange, 2013; Lee et al., 2012; Reddy et al., 2010; Pearson et al., 2008). Gallant and colleagues first decoded visual perception by using Gabor features (i.e., spatial frequency, orientation, position, and phase) of natural images in the EVC (Kay, Naselaris, Prenger, & Gallant, 2008), and then they conducted the same decoding analysis of the content evoked upon imagining specific famous images (Naselaris, Olman, Stansbury, Ugurbil, & Gallant, 2015), which further provided insight into the similarity between visual perception and mental imagery in the EVC. Horikawa and Kamitani (2017) used features extracted by a deep convolutional neural network to decode brain activity in the ventral visual stream when participants were observing and imagining natural images, and these results supported the claim that visual perception and mental imagery share neural representation in the hierarchy of the visual system.

Overall, these studies have collectively shown that neural representation of the ventral visual cortex during perception has a hierarchical topological structure, where the brain activity from high-level visual areas represents the semantic category of natural images (Naselaris, Prenger, Kay, Oliver, & Gallant, 2009) and the activity from the EVC represents low-level visual features (Albers et al., 2013; Cichy et al., 2012; Lee et al., 2012). The hierarchical topological structure was also observed during mental imagery (Reddy et al., 2010). Despite these findings, it should be noted that many previous studies have also reported a dissociation of function and structure of neural substrate between perception and mental imagery (Spagna, Hajhajate, Liu, & Bartolomeo, 2021; Sirigu & Duhamel, 2001; Butter, Kosslyn, Mijovic-Prelec, & Riffle, 1997).

In the present study, we attempted to depict the internal structure of the neural representation during perception and mental imagery in the EVC. To this end, we obtained fMRI data from previous human research involving visual perception and mental imagery (Horikawa & Kamitani, 2017) and evaluated the neural representation of the EVC using two types of fMRI-based voxel-wise encoding models, that is, a stim-to-voxel encoding model trained on Gabor features of input stimuli and a voxel-to-voxel encoding model based on the voxel activity in the high-level visual cortex (HVC; which contains fusiform face area [FFA], parahippocampal place area [PPA], and lateral occipital complex [LOC]; Figure 1; see more details in the Methods section). We first examined whether the stim-to-voxel encoding model based on Gabor features could capture the linear mappings between Gabor feature patterns and neural activity in the EVC. Then, we tested whether the linear mappings also exist between neural activity in the high visual cortex and that in the EVC by using the voxel-to-voxel encoding model. Moreover, we evaluated the combination of stim-to-voxel and voxel-to-voxel encoding models by integrating Gabor features of visual stimuli and high visual information to measure brain activity in the EVC during visual perception. Based on this, we further generalized these trained voxel-wise encoding models from perception to mental imagery. After that, we investigated the neural relationship between perception and imagery via combination and separation of the predicted neural representation patterns in the EVC with the two types of encoding models. Finally, we divided the voxels in the EVC into two groups according to the specificity and nonspecificity of Gabor features to refine the neural substrate in both the perception and imagery conditions.

METHODS

Sample Size Justification

The fMRI data sets were provided by the Kamitani Lab at Kyoto University and Advanced Telecommunications Research Institute International (ATR) (https://github.com/KamitaniLab). Five healthy human participants (four men) joined in all experiments. The sample size of the current study is based on previous work about encoding and decoding human brain activity on a voxel-wise level (Mell, St-Yves, & Naselaris, 2021; Horikawa & Kamitani, 2017; Naselaris et al., 2015; Kay et al., 2008).

Description of Data Sets and Their Experimental Acquisition

For the original experiment where the data were acquired, there were 1250 different natural images ($j = 1, 2, 3, 4, \ldots, 1250$), which were selected from 200 representative

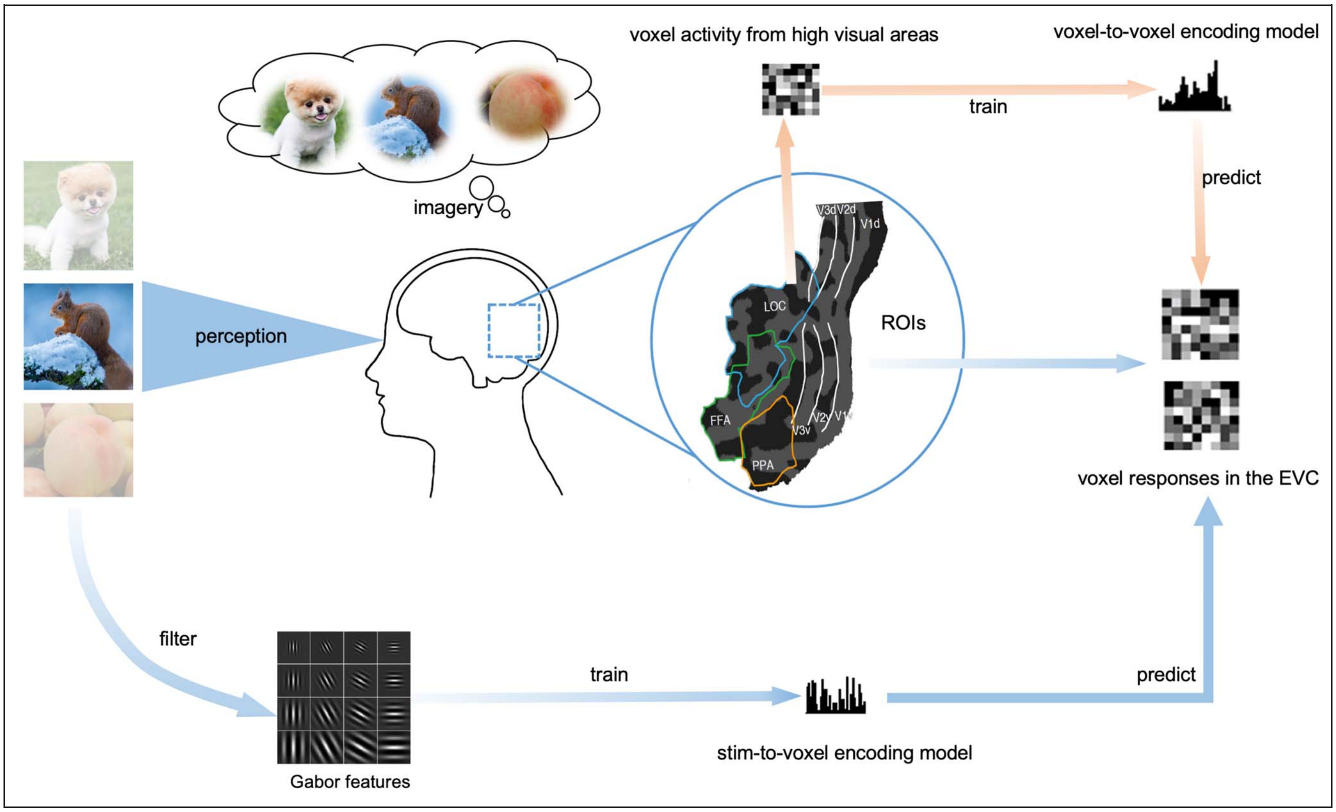
voxel activity from high visual areas
voxel-to-voxel encoding model
train
predict
imagery
perception
V3d V2d
V1d
LOC
ROIs
FFA
V2v V1v
V3v
PPA
voxel responses in the EVC
filter
train
predict
stim-to-voxel encoding model
Gabor features

Figure 1. The stim-to-voxel and voxel-to-voxel encoding models. The neural activity of the visual areas when observing and imagining an object. We used the GWP to filter images to obtain low visual features and then trained the stim-to-voxel encoding models to capture the linear association of these low visual features and the activity of each voxel in the early visual cortex (EVC, i.e., V1, V2, and V3). At the same time, we trained the analogous voxel-to-voxel encoding models with neural activity in the HVC (i.e., FFA, LOC, and PPA; take the place of the low visual features) to predict voxel activity in the EVC. Moreover, we further investigated the neural activity during perception and imagery according to the combination of stim-to-voxel and voxel-to-voxel encoding models.

object categories (c = 1, 2, 3, …, 200). During BOLD-fMRI scanning, each participant was instructed to view natural images at the center of the projection screen (visual perception experiment, two sessions) and to imagine corresponding objects according to the word cues presented on the screen (visual imagery experiment, one test session). There was a training image session (24 runs) and a test image session (35 runs) in the perception experiment with the same procedures. A total of 1200 different images (j = 1, 2, 3, …, 1200) from 150 object categories (eight images used in each category, c = 1, 2, 3, …, 150) were presented only once in the training image session. Fifty different images (j = 1201, 1202, …, 1250) from 50 new object categories (c = 151, 152, …, 200) were presented 35 times each in the test image session. In the imagery experiment, each participant was instructed to read the red word cue presented on the screen and to then imagine objects visually with closed eyes after hearing a beep sound. The word cues were the corresponding names of 50 object categories (c = 151, 152, …, 200) that were matched to the images presented in the perception test session, but participants were asked to freely imagine as many objects from the same category as they could during each trial. The imagery session consisted of 20 runs, each containing 25 imagery trials, and each duration was 10 min 39 sec.

The detailed procedures of MRI data preprocessing and functional brain region localization can be found in the original article (Horikawa & Kamitani, 2017). We used six ROIs extracted from the original article: the EVC (i.e., V1, V2, and V3) and three high-level visual areas (i.e., FFA, LOC, and PPA). The original study was approved by the ethics committee of ATR, and the present data reanalysis was approved by the authors of the original article.

Gabor Features

We used Gabor features extracted from the 1250 natural images by a Gabor wavelet pyramid (GWP) model to encode the activity in the EVC (Ringach, 2002; Jones & Palmer, 1987). The GWP model could be viewed as an appropriate method to describe voxel activity in the EVC (Rainer, Augath, Trinath, & Logothetis, 2001; Lee, 1996; Jones & Palmer, 1987; Daugman, 1985), which has been used to describe the dimensions of space (DeYoe et al., 1996; Sereno et al., 1995; Engel et al., 1994), orientation (Sasaki et al., 2006; Haynes & Rees, 2005; Kamitani & Tong, 2005), and spatial frequency (Olman, Ugurbil, Schrater, & Kersten, 2004; Singh, Smith, & Greenlee, 2000) of natural images. According to a prior study (Kay et al., 2008), these Gabor features extracted from natural images by a GWP model could be fitted to voxel responses in the EVC, especially in the primary visual cortex (V1).

In the present study, we first constructed the GWP model with a series of Gabor filters. Based on Kay et al.'s (2008) study, we defined the same wavelets with six spatial frequencies: 1, 2, 4, 8, 16, and 32 cycles per field of view (192 × 192 mm^2). At each frequency, f cycles per field of view, wavelets were positioned on an $f \times f$ grid. At each grid position, wavelets occurred at eight orientations: 0, 22.5°, 45°, 67.5°, 90°, 112.5°, 135°, and 157.5°, and the two quadrature phases were 0° and 90°. Therefore, there were 10,920 ($[1^2 + 2^2 + 4^2 + 8^2 + 16^2 + 32^2] \times 8 =$ 10,920) Gabor filters in the GWP model.

We then applied the constructed GWP model to extract Gabor features of the stimuli images. The original 500 px × 500 px images were downsampled to 128 px × 128 px image resolution for the analysis. Accordingly, the features were defined as

$$\boldsymbol{f}(\boldsymbol{s}) = \log\left(\left|W^T\boldsymbol{s}\right| + \mathbf{1}\right) \tag{1}$$

where $\boldsymbol{f}$ is an $F \times 1$ vector containing the features ($F =$ 10,920, the number of filters used for the model), and W is a matrix of complex Gabor filters. The variable $\boldsymbol{s}$ indicates the matrix of each image used in the whole experiment, W includes as many rows as the pixels in $\boldsymbol{s}$, and each column contains a different Gabor filter; thus, the dimensions of Gabor features were $128^2 \times$ 10,920. The features represent the log of the magnitudes derived from filtering the image by each filter. These parameters correspond to those used by Gallant and colleagues (Naselaris et al., 2015; Kay et al., 2008).

Encoding Models

The neural activity in the EVC when people observe something could be explained by two sources, namely, the external low-level visual features and internally high visual activity. To simulate these two sources of information processing, we constructed two types of voxel-wise encoding models, one from the Gabor features aspect (stim-to-voxel encoding model) and one from the aspect of neural activity of HVC (voxel-to-voxel encoding model). These two sources formed an encoding model of the neural activity of each voxel in the EVC.

For training the stim-to-voxel encoding model, we extracted the BOLD-fMRI signals of each voxel in the EVC related to the 1200 training images (j = 1, 2, …, 1200) during perception. For each voxel, we regarded the Gabor features as the input variable and the fMRI signal activity of the EVC as the output response to train the voxel-wise encoding models for perception.

According to a prior study, p was defined as the number of training images, and q was defined as the number of input channels. The neural activity of each voxel in the EVC could be modeled as

$$\boldsymbol{y} = \boldsymbol{X}h + c\mathbf{1} + n \tag{2}$$

where $\boldsymbol{y}$ is the set of neural activity (i.e., the response of each voxel in the EVC, $p \times 1$), $\boldsymbol{X}$ is the set of input channels ($p \times q$), h is the kernel ($q \times 1$), c is the DC offset (1×1), $\mathbf{1}$ is a vector of constant ones ($p \times 1$), and n is the noise ($p \times 1$). We used the functions from the STRFlab toolbox (Version 1.45, https://strflab.berkeley.edu/) to automatically estimate the model parameters. The model parameters were estimated with gradient descent based on the early stopping algorithm to prevent parameters from overfitting, and the stopping set consisted of 20% of randomly selected responses (Kay et al., 2008; Tugnait, 1994). There was a bootstrap sampling approach for iterative analysis. This procedure was conducted independently for each voxel in the EVC; thus, we ultimately obtained an encoding model for each voxel in V1, V2, and V3 during perception.

For training the voxel-to-voxel encoding model, we extracted the voxel activity in each high visual area when participants observed each image in the training session and regarded each voxel signal as an input feature. This model is also called a voxel-to-voxel model (Mell et al., 2021). We also used the same algorithm (i.e., Equation 2) from the stim-to-voxel encoding model to predict the voxel activity in the EVC during perception. Because there was no stimulus presented on the screen during the imagery experiment, we applied the stim-to-voxel and voxel-to-voxel encoding models trained by perception data to the imagery test session.

Additionally, to better demonstrate the neural connection between HVC and EVC, we trained an additional type of voxel-to-voxel encoding model, which had a reverse prediction direction (i.e., EVC → HVC). This model utilizes voxels from the EVC to predict the responses of voxels in HVC. We employed the same algorithm to train and test prediction performance as in the previous voxel-to-voxel encoding model.

Image Identification Analysis

The stim-to-voxel and voxel-to-voxel encoding models trained with the perception training data set were used in the identification of images from the testing data set based on brain activities during perception and mental imagery, respectively.

In the viewed image identification analysis (i.e., perception), we used the test session data set of 50 newly viewed images from 50 different categories (j = 1201, 1202, …, 1250, c = 151, 152, …, 200) to estimate the performance of the trained stim-to-voxel and voxel-to-voxel encoding models. For the stim-to-voxel encoding model, we extracted Gabor features of each test image and then input these low-level visual features into the model to predict the corresponding voxel activity in the EVC. In this way, each voxel-wise encoding model would produce a prediction value for each voxel in the EVC. Each region of the EVC (i.e., V1, V2, and V3) was viewed as a basic unit, and thus the predicted brain activity pattern in V1, V2, and V3 could be used for comparison with the real fMRI signal pattern in the EVC to estimate the performance of the encoding model. To this end, a Pearson's correlation coefficient between each pair of the two sets (predicted set and real set) of the test images ($n = 50$) was calculated, and then a 50 × 50 correlation matrix for each participant was obtained. If the diagonal value of the obtained matrix was the maximum in each column, it suggested that the predicted voxel activity pattern matched very well with the real one for the same image, and we then regarded it as the correct prediction. Finally, we calculated the ratio of the number of correct predictions to the total number of test images (correct predictions/50), which was regarded as the predictive accuracy.

We used the same procedure to calculate accuracy with the trained voxel-to-voxel encoding models. The only difference was the input feature. We regarded the voxel activity from HVC as the input feature and then predicted the brain activity of each region of the EVC when participants were observing each image. Finally, we compared the predicted responses with the real responses to obtain prediction accuracies.

In the imagined image identification analysis (i.e., mental imagery), there was no stimulus input during mental imagery, so we adopted the same Gabor features of the viewed images to estimate the performance of the stim-to-voxel encoding model. We then compared the predicted activity in the EVC with the real activity of the imagery test data. The specific calculation procedures of prediction accuracy were consistent with those described above in the perception experiment. The prediction of the voxel-to-voxel encoding model in the imagery experiment was conducted in the same way to obtain corresponding prediction accuracies.

Linear and Nonlinear Combination with Two Voxel-wise Encoding Models

The neural activity in the EVC can be explained by different contributions, for example, in the current study it might be derived from stimuli-relevant low visual features and corresponding activity in the HVC. Based on the comprehensive framework proposed by Op de Beeck, Haushofer, and Kanwisher (2008), there are several possibilities (i.e., additive combination and non-additive combination) to explore how different functional properties can generate interaction. To examine how these two sources might interact with one another in the EVC, we applied two mathematical approaches (i.e., linear and nonlinear combination) to the predicted voxel responses by the stim-to-voxel and voxel-to-voxel encoding models (Zhang et al., 2014). The linear combination was conducted by calculating the arithmetic mean of the predicted neural activity patterns from two types of voxel-wise encoding models. The nonlinear combination was obtained by calculating the geometric mean of the predicted neural activity patterns from two types of encoding models. Taking voxel activity in V1 as an example, the trained stim-to-voxel and voxel-to-voxel encoding models could separately predict voxel activity and each voxel in V1 got two predictions (i.e., one from stim-to-voxel model and another from voxel-to-voxel model). For each observed image or imagined stimulus, there were two different predicted voxel patterns in V1. Next, we combined these two-sources patterns linearly or nonlinearly to get the combined voxel activity patterns for each stimulus in the test session, which were further compared with the real V1 fMRI signal pattern to show the combined effect of two sources via stim-to-voxel and voxel-to-voxel encoding models. We also evaluated the combination effect within V2 and V3 via the same approach.

Separation of Imagery from Perception

Imagery is considered as a weak perception (Pearson & Kosslyn, 2015), which means that visual imagery is similar to visual perception. Based on this view, we further refined the relationship between the two mental processes by comparing the neural association of high visual activity to the EVC. For each test image/category, we acquired the predicted voxel activity patterns in V1, V2, and V3 during perception and imagery, separately, by the image identification analysis of the trained voxel-to-voxel encoding model. We then regarded the predicted imagery activity pattern as a covariate of the predicted perception activity pattern to remove. Finally, the remaining predicted activity pattern in perception was compared with the real fMRI signal pattern in the EVC to make a prediction. To illustrate the effect of this separation, we examined two other prediction performances for comparison. One pattern was the shuffled predicted activity pattern of imagery, and the other was a random activity pattern consisting of white noise. We regarded them as covariates to remove from the predicted perception pattern and then calculated prediction accuracies.

Predictions of Gabor Specificity and Non-Gabor Specificity

To further elaborate the information content in the activity of EVC, we labeled the voxels in each region of the EVC with two opposite names. The names were defined by the weight values obtained from the process of training the stim-to-voxel encoding model: Gabor specificity, which refers to the voxels with useful (nonzero) weight values, and non-Gabor specificity, which refers to the voxels with zero weight value. Here, the useful or zero weight value denoted whether the voxels could encode the Gabor features. Accordingly, we divided the voxels in each region of the EVC into two groups: Gabor-specificity group (i.e., G group) and non-Gabor specificity group (i.e., N group). After that, we conducted the same prediction procedures (image identification analysis and linear combination analysis) via the voxel-to-voxel encoding model for each group in V1, V2, and V3.

In addition, as a control analysis, we examined the prediction performance for the G group and the N group via the other type of voxel-to-voxel encoding model (i.e., direction: EVC → HVC).

RESULTS

Hierarchical Structure Exists in the EVC

Stimuli-relevant Response to Perception and Imagery

During perception, the performance of the trained stim-to-voxel encoding model showed that Gabor features could remarkably predict brain activity patterns in the EVC. The average accuracies of the five participants were 29% ± 0.13% (chi-square test, $\chi^2(1, n = 50) = 11.76, p = .0006$), 22% ± 0.17% ($\chi^2(1, n = 50) = 8.33, p = .004$), and 18% ± 0.11% ($\chi^2(1, n = 50) = 6.40, p = .01$) in V1, V2, and V3, respectively (see the first feature "Gabor" in Figure 2A), and all of them were significantly higher than the change level of 2% (1/50; see the red dash line in Figure 2A). Interestingly, there was a decreasing trend of prediction accuracies in the EVC (i.e., V1 > V2 > V3), except for Participant 1, which had the highest accuracy for V3 (V3 was 10%, 5/50; V1 was 6%, V2 was 4%). The predicted accuracies attenuated from V1 to V3, showing that Gabor feature could explain more voxel activity in V1 and there might be a hierarchical organization of the EVC with respect to the representation of Gabor features.

During imagery, the performance of the trained stim-to-voxel encoding model showed that Gabor features could not effectively predict brain activities in the EVC under the

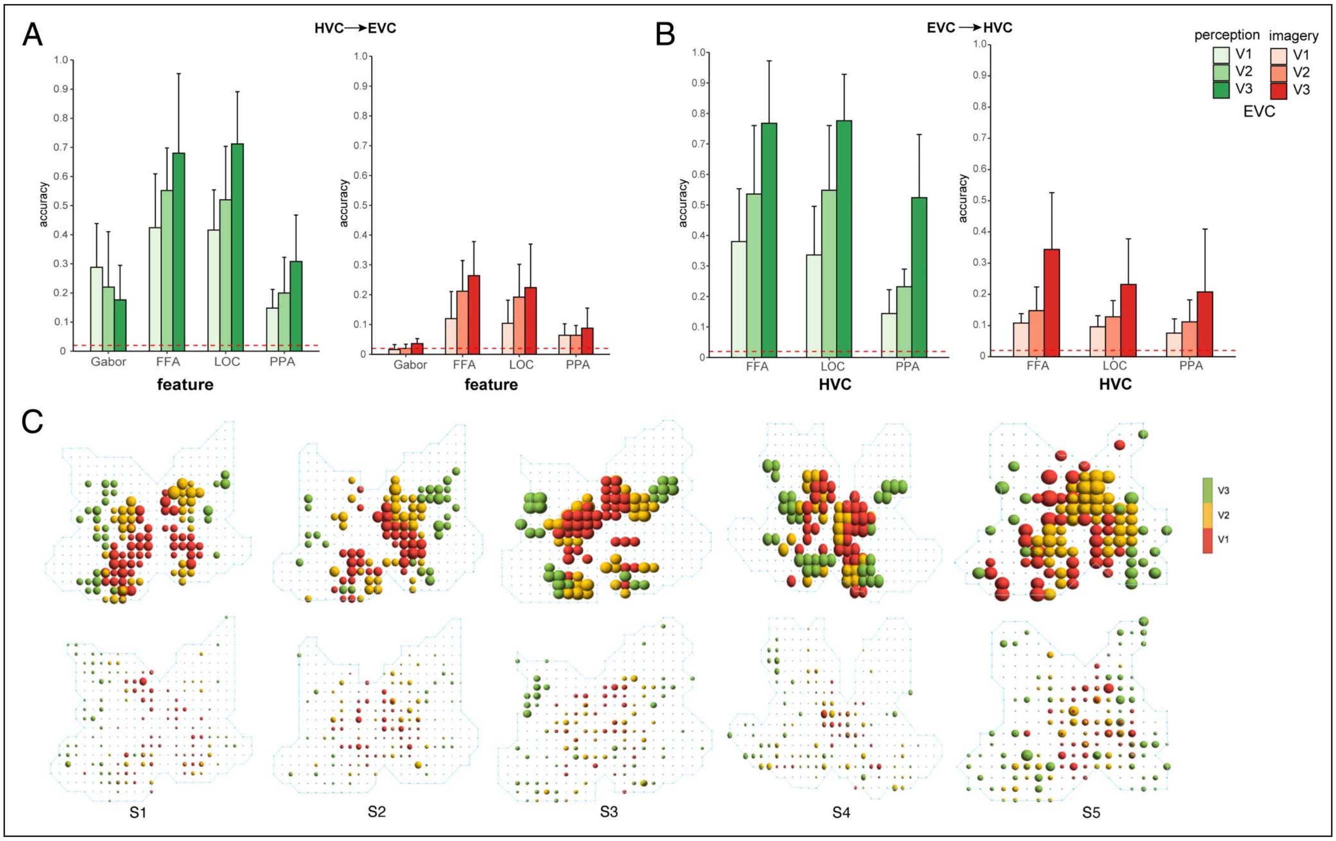

Figure 2. Prediction performance with stim-to-voxel and voxel-to-voxel encoding models in the EVC and HVC. (A) Average prediction accuracy of five participants with two types of voxel-wise encoding models during perception (green bar) and imagery (red bar). The three different levels of green bars represent the prediction accuracy in the EVC during perception. Another three different levels of red bars represent the prediction accuracy in the EVC during imagery. The first attributes of the abscissa denote the prediction of Gabor features via the stim-to-voxel encoding model, and the other three attributes denote the predictions of three different high-level visual areas via the voxel-to-voxel encoding models (direction: HVC → EVC). The red dashed line refers to chance level (2%). (B) Average prediction accuracy of five participants with the other voxel-to-voxel encoding models via EVC predicting HVC during perception (green bar) and imagery (red bar). This result can be compared with the results from the previous voxel-to-voxel encoding model. The three different levels of green bars represent the prediction accuracy of the EVC predicting high brain areas during perception. Another three different levels of red bars represent the prediction accuracy of the EVC predicting high areas during imagery. The three attributes denote the predictions of three different HVC via the voxel-to-voxel encoding models (direction: EVC → HVC). The red dashed line refers to chance level (2%). (C) The spatial distribution of 1000 voxels with the order of Pearson's correlation coefficients in the EVC. The upper row shows the voxel distribution during perception of five participants, and the row beneath shows the voxel distribution during imagery. The size of the colored dots was determined by the weight (Pearson's correlation coefficients). The red dots are voxels in V1; the orange dots are voxels in V2; and the green dots are voxels in V3. The voxel overlap proportions between perception and imagery for five participants were: 96%, 95%, 76.3%, 84.5%, and 100%.

same statistical standard with the analysis of perception data. The average V1 accuracy of the five participants was 1.6% ± 0.01%, $\chi^2(1, n = 50) = 0.02, p = .88$; V2 was 2% ± 0.01%, $\chi^2(1, n = 50) = 0, p = 1.00$; and V3 was 3.2% ± 0.02%, $\chi^2(1, n = 50) = 0.14, p = .71$ (see the first feature "Gabor" in Figure 2B). These results revealed that almost all of the prediction accuracies using the exogenous encoding model with Gabor features in the EVC during mental imagery were around the chance level (2%), with poor prediction performances. Some prior studies (Naselaris et al., 2015; Cichy et al., 2012) demonstrated that brain activity in the EVC generated during mental imagery could be explained by Gabor features. The insignificant results in the present study were likely caused largely by task-dependent factors. In the imagery experiment, participants were required to freely imagine as many object images matched with the word cue as possible (Horikawa & Kamitani, 2017), but we assessed the prediction performance based on a specific image that was the same as in the perception procedures. Thus, when using the Gabor features of a specific image as input values of the model to predict the brain activity triggered by the many imagined images from the same category, we could not obtain high prediction performance.

To further explore the extent of brain activity in the EVC during imagery that was explained by Gabor features, we extracted 1000 voxels from the whole EVC. First, the values of Pearson's correlation coefficients between the predicted voxel activity and the real activity of each participant were sorted in descending order, and then we analyzed the top 1000 values (except for those of the fifth participant, who had only approximately 700 voxels conforming to the condition). After that, we depicted the voxel spatial distribution pattern in the EVC, in which the Pearson's correlation coefficients were regarded as weights of nodes via the BrainNetViewer toolbox (https://www.nitrc.org/projects/bnv/). All the above calculations were performed with both perception and imagery data. We found that overlapping distributions of the extracted voxels between perception and mental imagery were very large: 96% for Participant 1, 95% for Participant 2, 76.3% for Participant 3, 84.5% for Participant 4, 100% for Participant 5 (where the number of corresponding voxels in the EVC was 734), which suggested that spatial distributions were similar between the two mental processes for each participant. The median of the average weight of participants during perception was 0.28 (ranging from −0.09 to 0.81), and the median during imagery was 0.02 (ranging from −0.21 to 0.44), which determined the size of the color nodes in Figure 2C.

Internal Neural Correlation of Visual Cortex during Perception and Imagery

From the prediction accuracies of the trained voxel-to-voxel encoding models, we found that the brain activity patterns in the three high-level visual areas could be used to successfully predict the brain activity patterns in the EVC (see the other three features "FFA, LOC, and PPA" in Figure 2A). FFA and LOC both obtained higher prediction accuracies of more than 40% of V1 during perception (more than 10% during imagery). Critically, the voxel-to-voxel encoding models were trained only with the perception training data, which could be generalized to predict the brain activity under the mental imagery condition (e.g., the average accuracies of FFA predictions were 12% ± 0.08%, $\chi^2(1, n = 50) = 3.57, p = .06$, in V1; 21% ± 0.09%, $\chi^2(1, n = 50) = 7.85, p = .005$, in V2; and 26% ± 0.10%, $\chi^2(1, n = 50) = 10.29, p = .001$, in V3. Comparing the prediction accuracies in the EVC during perception and imagery, we could find that there was an increasing trend of the prediction performances (i.e., accuracies: V1 < V2 < V3) in both mental processes. This therefore provided further evidence for the existence of a gradient hierarchical structure in the EVC, which was similar between perception and imagery (Figure 2A).

To investigate whether the voxel-to-voxel encoding model via brain activity from the HVC prediction of voxel responses in the EVC was potentially influenced by correlated neural activity between high- and low-level visual regions, we extracted the voxels from the EVC to train another voxel-to-voxel encoding model and predict the brain activity in the HVC (Mell et al., 2021; Zhang et al., 2014). Consistent with previous findings by Zhang and his colleagues, this voxel-wise encoding model with the reverse prediction direction successfully predicted brain activities in the HVC using the voxel responses from V1, V2, and V3 during both perception (see green bars in Figure 2B) and imagery (see red bars in Figure 2B) conditions.

We compared the prediction performance of these two voxel-to-voxel encoding models with completely reverse prediction directions, that is, HVC → EVC and EVC → HVC, using a paired t test. In the perception condition, the prediction accuracy of V1 predicting LOC was significantly lower than the prediction performance of LOC predicting V1, $t(4) = -3.27, p = .031$. Moreover, the prediction accuracy of V3 predicting LOC and PPA was higher than the prediction performance of LOC, $t(4) = 4.35, p = .012$, and PPA, $t(4) = 3.53, p = .024$, predicting V3. However, no significant prediction difference was observed between the two reverse prediction directions during the imagery condition (see Table 1 for more statistical details).

Linear Combination Effect between Gabor Features and High Visual Activity

To better understand how sensory input and high visual activity interact in the EVC, we explored the combined effect using the stim-to-voxel and voxel-to-voxel encoding models. We conducted linear and nonlinear combination analysis for both perception and imagery data. The results are shown in Figure 3. In Figure 3A, the different categories on the x axis denote linear and nonlinear combinations. For example, FFA + Gabor denote a linear combination

Table 1. Summary Statistics and Significance Results for the Prediction Accuracy of the Bidirectional Voxel-to-Voxel Encoding Models

Conditions	*Prediction Direction*	*t*	*df*	*p Value*
Perception	V1 ⇔ FFA	−1.11	4	0.33
Perception	V2 ⇔ FFA	−0.32	4	0.77
Perception	V3 ⇔ FFA	2.08	4	0.11
Perception	V1 ⇔ LOC	***−3.27***	4	***0.031***
Perception	V2 ⇔ LOC	1.00	4	0.37
Perception	V3 ⇔ LOC	***4.35***	4	***0.012***
Perception	V1 ⇔ PPA	−0.25	4	0.81
Perception	V2 ⇔ PPA	0.68	4	0.53
Perception	V3 ⇔ PPA	***3.53***	4	***0.024***
Imagery	V1 ⇔ FFA	−0.26	4	0.81
Imagery	V2 ⇔ FFA	−1.27	4	0.27
Imagery	V3 ⇔ FFA	1.48	4	0.21
Imagery	V1 ⇔ LOC	−0.23	4	0.83
Imagery	V2 ⇔ LOC	−1.50	4	0.21
Imagery	V3 ⇔ LOC	0.17	4	0.87
Imagery	V1 ⇔ PPA	0.35	4	0.74
Imagery	V2 ⇔ PPA	1.35	4	0.24
Imagery	V3 ⇔ PPA	1.46	4	0.21

The results of the paired *t* test between EVC predicting high visual areas and high visual areas predicting EVC showed significant differences only in the perception condition (especially between LOC and V1, V3) but not in the imagery condition. The bidirectional arrow shows the two reverse prediction directions (i.e., EVC → HVC and HVC → EVC). The negative value in the *t* value column means the accuracy of the EVC → HVC voxel-to-voxel encoding model is lower than the accuracy of the HVC → EVC voxel-to-voxel encoding model. The **bold font** in the *t* value and *p* value columns denotes a statistically significant result.

between FFA features (brain activity was regarded as the encoding model feature) and Gabor features, and FFA * Gabor denote a nonlinear combination between FFA features and Gabor features. For perception, the results showed that the linear combination of the two types of encoding models could effectively improve the prediction power, and the average accuracies were increased by more than 15%. However, the average accuracies were mostly reduced to lower than 10% under the nonlinear combination condition (the green lines in Figure 3A). There was no evident linear or nonlinear combination effect in the mental imagery experiment (the red lines in Figure 3A).

Shared Low–High Visual Neural Association between Perception and Imagery

Visual imagery is a kind of weak visual perception. To support the claim, we hypothesized that two mental processes shared the low–high visual neural association. We separated the voxel activity of mental imagery from that of perception prediction analysis via removing covariance during perception. The results are shown in the Figure 3B. The four different categories on the *x* axis in Figure 3B present four different ways to do the separation analysis. *Original* means that we used the perception data in the HVC to predict the voxel activities in the EVC via the voxel-to-voxel encoding model without removing anything (the same as the above image identification analysis). *Imagery* means that we removed the predicted voxel activity of imagery from that of perception and then conducted the prediction analysis via the voxel-to-voxel encoding model. *Noise* means removing random white noise from perception data, and then we conducted the prediction analysis. *Scatter* means removing scattered imagery activity pattern from perception data, and then we conducted the prediction analysis. We found that the average prediction accuracies in the EVC via removing voxel responses of imagery from that of perception fell within a range from 10% to 50%, significantly higher than the chance level (2%). However, after we removed the scattered imagery pattern and random white noise from the perception part, the prediction accuracies of voxel activity in the EVC were close to the original prediction performances of high visual activity prediction via the voxel-to-voxel encoding model. Therefore, these comparisons indicated that imagery shared information representation with perception in the part of neural association from HVC to EVC.

Shared Neural Representation in the EVC

To further elaborate the information content of the voxel activity in the EVC, we divided these voxels into two groups with Gabor specificity and with non-Gabor specificity (for more details, see the Methods section). We then conducted the prediction analysis via the voxel-to-voxel encoding model with voxel activity from the HVC to predict voxel activity of these two groups of voxels in the EVC.

Interestingly, during the procedure of labeling the voxels in the EVC, we found a stable topological distribution of the rates between Gabor specificity and non-Gabor specificity. Regardless of whether the total number of selected voxels was 1000, 800, 600, 400, or 200, the ratio of the two labeled groups was almost 1:2 (see Figure 4).

Comparing the prediction performances of the voxels from Gabor specificity group (the G group in Figure 5), non-Gabor specificity group (the N group in Figure 5), we found that accuracies obtained via the voxel-to-voxel encoding model with the brain activity from each high visual area were almost equal in each area of the EVC, in both the perception and the mental imagery conditions. In addition, the prediction performances of the two groups were similar to the original performance (i.e., without grouping, the results from O group in the Figure 5 are

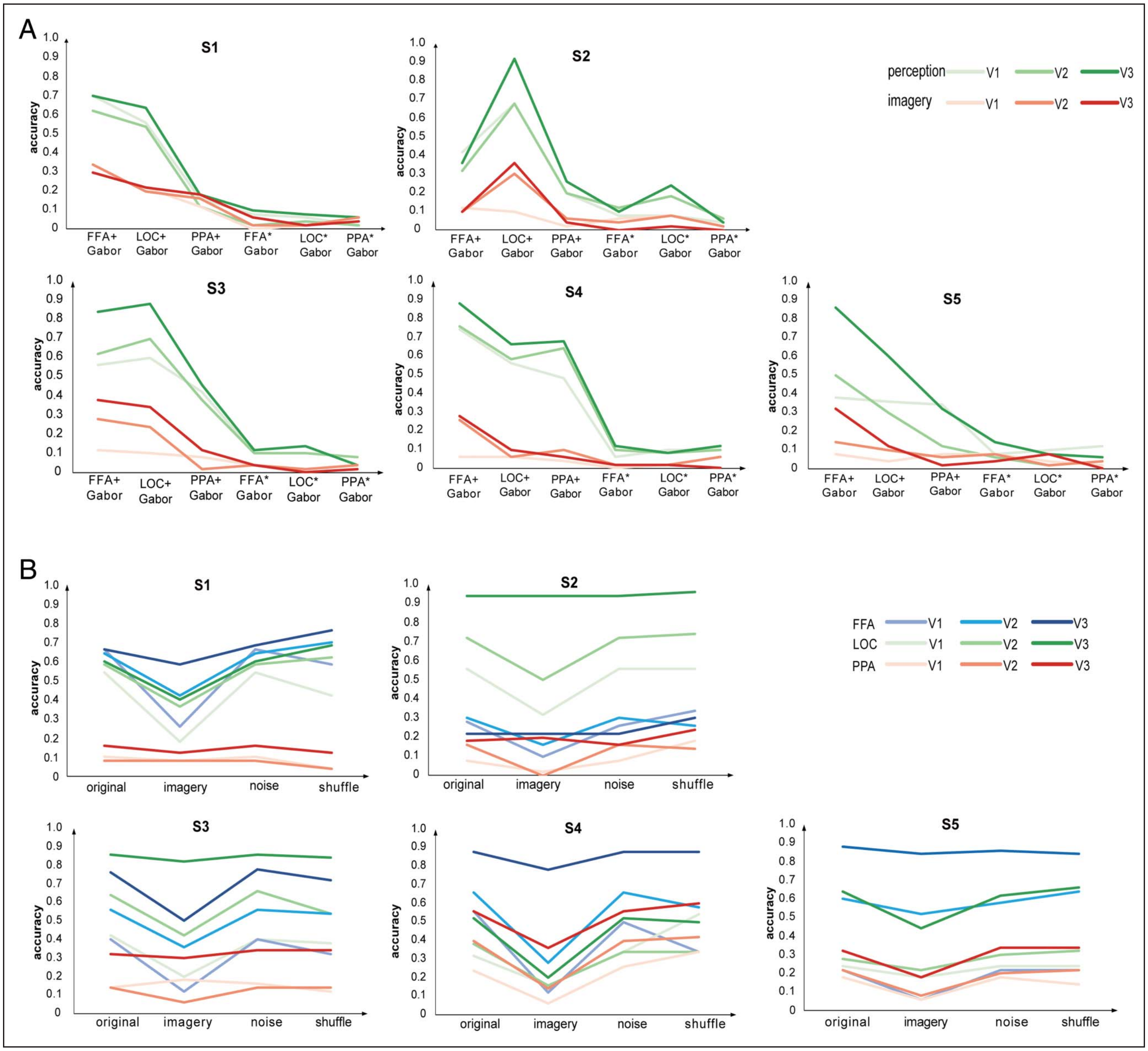

Figure 3. Neural prediction performance during perception and mental imagery. (A) The combination effect of the stim-to-voxel and voxel-to-voxel encoding models via linear and nonlinear combinations, respectively. The left three attributes of the abscissa in each participant line chart represent the linear combination, and the right counterpart represents the nonlinear combination, respectively. The green line represents the prediction performance during perception in the EVC, and the red line represents the counterpart during imagery in the EVC. (B) Performance in the separation of the predicted imagery pattern from the predicted perception pattern. The blue line represents the performance of FFA predicting the EVC; the green line represents the performance of LOC predicting the EVC; and the red line represents the performance of PPA predicting the EVC. We conducted four comparisons: original, the voxel activity from HVC during perception to predict voxel responses in the EVC; imagery, the voxel activity from HVC during perception with removal of the predicted voxel activity of imagery to predict voxel responses in the EVC; noise, the voxel activity from HVC during perception with removal of the random white noise to predict voxel responses in the EVC; shuffle, the voxel activity from HVC during perception with removal of the shuffled imagery to predict voxel responses in the EVC.

the same results of FFA, LOC, and PPA features in Figure 2A) achieved by the voxel-to-voxel encoding model. For example, during perception condition, the average prediction accuracies of FFA features in V1 are 32% ± 0.18% (G group), 39% ± 0.18% (N group), and 42% ± 0.19% (O group); in V2, the accuracies are 49% ± 0.16% (G group), 51% ± 0.11% (N group), and 55% ± 0.15% (O group); in V3, the accuracies are 51% ± 0.21% (G group), 68% ± 0.28% (N group), and 68% ± 0.27% (O group). See the average prediction results from five participants in the Figure 5. Likewise, we explored the combined effect of the stim-to-voxel and voxel-to-voxel voxel-wise encoding models in the EVC for the two groups of voxels. The results revealed that the prediction performances for each group in V1, V2, and V3 via the voxel-to-voxel encoding model during perception and imagery

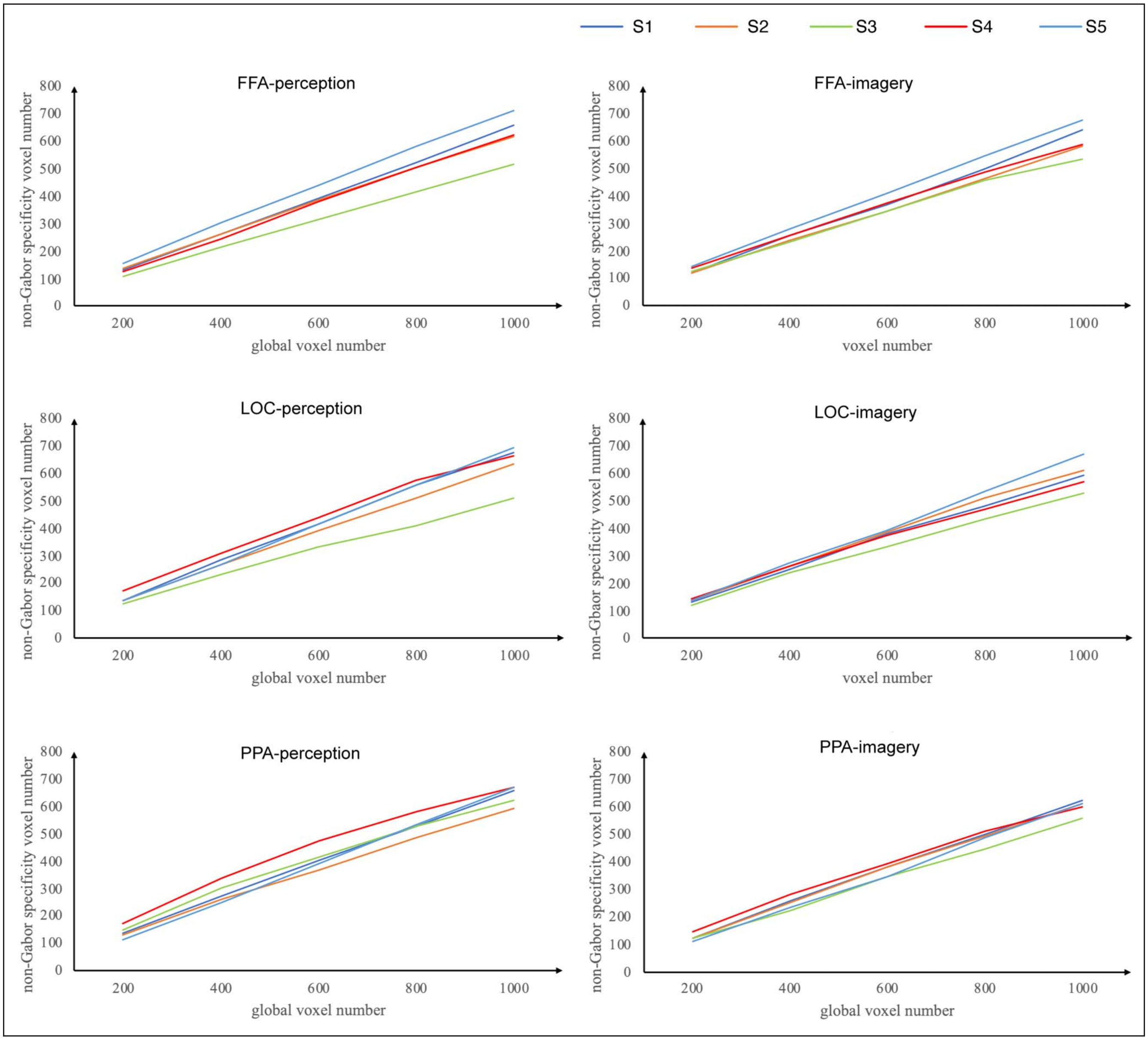

Figure 4. Rate of Gabor nonspecificity and global selected voxel number in the EVC. The abscissa number is selected voxel number from the whole EVC according to the prediction performance rated from high to low and the ordinate number is Gabor nonspecificity voxel number selected by the weight of by the feature weight with zero value from the stim-to-voxel voxel-wise encoding model. The left column is the selected voxels of three HVC predicting EVC in perception of five participants, whereas the right column is the corresponding performance in imagery.

were similar and also showed the hierarchical structure in the EVC (i.e., accuracies: V1 < V2 < V3). These results further indicated that the independent representation contents corresponding to Gabor and non-Gabor features in the EVC were equally predicted by the voxel activity from the HVC during perception and imagery.

Moreover, as a control analysis, we used the other voxel-to-voxel encoding model (direction: EVC → HVC) to perform prediction analysis with the G group and the N group of voxels from the EVC predicting each high-level visual area. We then compared the prediction performance of the two groups using a paired *t* test. The prediction accuracies of both groups during the perception and imagery conditions are shown in Figure 6, whereas the statistical results are presented in Table 2.

Our findings indicated that the prediction accuracies of the G group were lower than those of the N group, as evidenced by the prediction performance shown in Figure 6 and statistical results (see all *t* values in Table 2; negative values indicate lower accuracies for the G group). Furthermore, in the perception condition, the accuracies of V3 predicting FFA, $t(4) = -3.76$, $p = .020$; and LOC, $t(4) = -4.66$, $p = .010$; and of V1 predicting LOC, $t(4) = -2.80$, $p = .049$, in the G group were significantly lower than those in the N group. In addition, in the perception condition, the prediction accuracies of EVC (i.e., V1, V2,

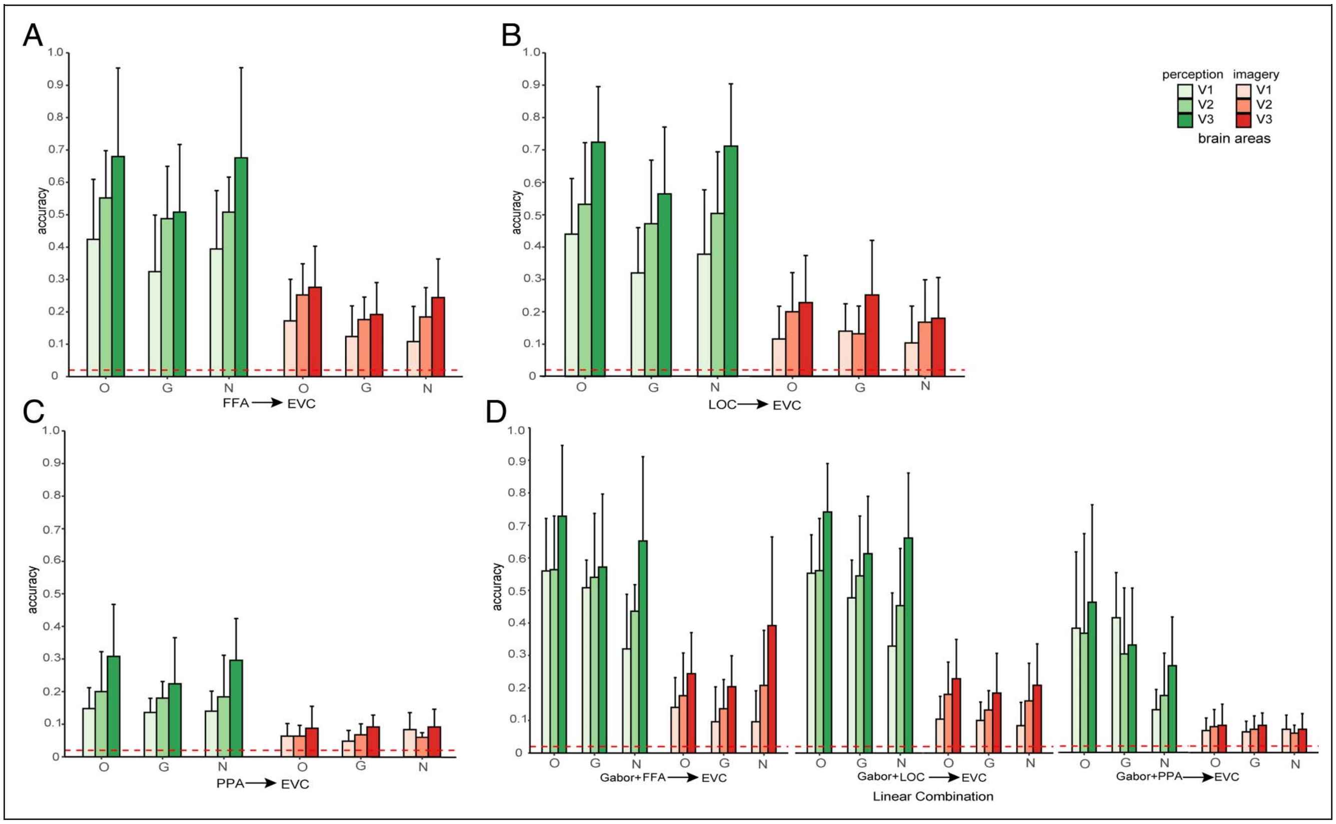

Figure 5. Shared neural representation in the EVC. Here, we present five participants' average results of three high visual area predictions in Gabor specificity and non-Gabor specificity voxel groups via the voxel-to-voxel encoding models. The categories on the *x* axis with the letter P refer to the perception condition (green bars), and those with the letter I refer to the imagery condition (red bars). GP = Gabor-specificity prediction during perception (G group); NP = non-Gabor specificity prediction during perception (N group); OP = original high visual area prediction during perception (O group, i.e., the prediction performance corresponding with Figure 2); GI = Gabor-specificity prediction during imagery (G group); NI = non-Gabor specificity prediction during imagery (N group); OI = original high visual area prediction during imagery (O group). (A, B, C) Prediction performance for the HVC in three groups (Gabor specificity, non-Gabor specificity, and original group) with respect to perception and imagery conditions, respectively. (D) Prediction performances of linear combination in the HVC in three groups with respect to perception and imagery conditions, respectively. OF = original prediction in the FFA during perception (green) and imagery (red); GF = Gabor specificity prediction in the FFA during perception (green) and imagery (red); NF = non-Gabor specificity prediction in the FFA during perception (green) and imagery (red). We extract letter "L" and "P" to refer to "LOC" and "PPA" separately, so other categories on the *x* axis have the same meaning as those in the FFA.

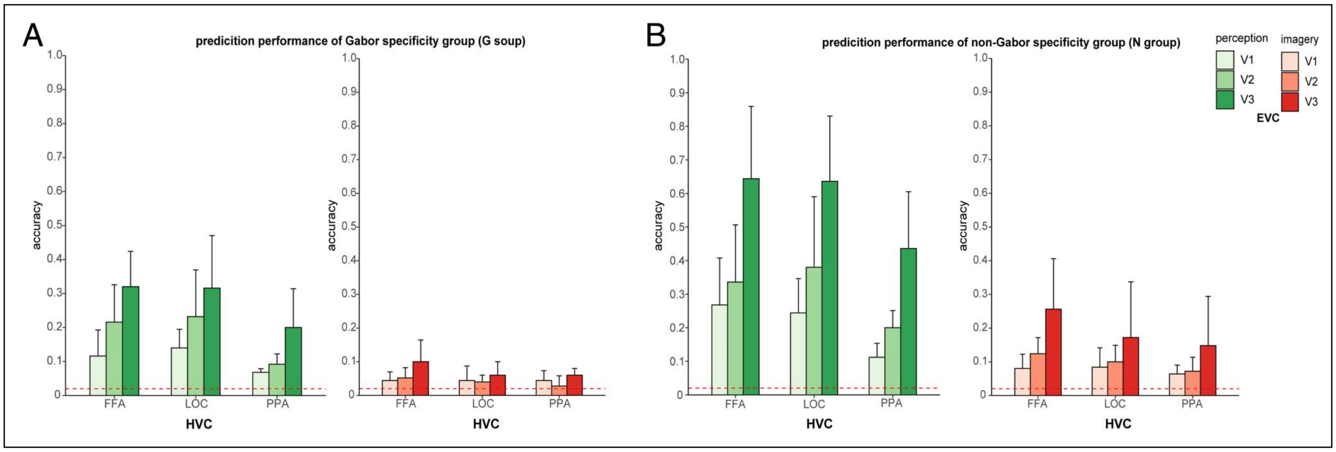

Figure 6. Prediction performance of Gabor specificity and non-Gabor specificity groups via the voxel-to-voxel encoding model (direction: EVC → HVC). (A) Prediction performance of the Gabor specificity group during perception (green bars) and imagery (red bars) via the voxel-to-voxel encoding model. (B) Prediction performance of the non-Gabor specificity group during perception (green bars) and imagery (red bars) via the voxel-to-voxel encoding model.

Table 2. Summary Statistics and Significance Results for the Prediction Accuracy between the Gabor Specificity Group and the Non-Gabor Specificity Group via the Voxel-to-Voxel Encoding Models (Direction: EVC → HVC)

Conditions	*Prediction Direction*	*t*	*df*	*p Value*
Perception	V1 → FFA	−2.06	4	0.11
Perception	V2 → FFA	−1.70	4	0.16
Perception	**V3** → FFA	***−3.76***	4	***0.020***
Perception	**V1** → LOC	***−2.80***	4	***0.049***
Perception	V2 → LOC	−1.81	4	0.14
Perception	**V3** → LOC	***−4.66***	4	***0.010***
Perception	**V1** → PPA	***−2.99***	4	***0.040***
Perception	**V2** → PPA	***−3.96***	4	***0.017***
Perception	**V3** → PPA	***−2.83***	4	***0.047***
Imagery	V1 → FFA	−1.29	4	0.27
Imagery	V2 → FFA	−2.45	4	0.070
Imagery	V3 → FFA	−2.03	4	0.11
Imagery	V1 → LOC	−1.22	4	0.29
Imagery	V2 → LOC	−2.18	4	0.10
Imagery	V3 → LOC	−1.73	4	0.16
Imagery	V1 → PPA	−1.58	4	0.19
Imagery	V2 → PPA	−2.16	4	0.10
Imagery	V3 → PPA	−1.43	4	0.23

The arrow direction shows the prediction direction (i.e., EVC → HVC). The negative value in the *t* value column means the accuracy of the G group is lower than the accuracy of the N group. The **bold font** in the *t* value and *p* value columns denotes a statistically significant result.

and V3) predicting PPA in the G group were significantly lower than those in the N group (V1 → PPA: $t(4) = -2.99$, $p = .040$; V2 → PPA: $t(4) = -3.96$, $p = .017$; V3 → PPA: $t(4) = -2.83$, $p = .047$).

DISCUSSION

Using the trained stim-to-voxel and voxel-to-voxel encoding models to predict the fMRI signal in the EVC, this study compared the encoded information content related to visual perception and mental imagery in the visual system. The results revealed the existence of neural representations with Gabor specificity and non-Gabor specificity that were shared between perception and visual imagery, respectively, which helped us to understand the neural underpinnings between the two similar mental processes.

Before this current work, research has applied multivariate pattern analysis to investigate the relationship between perception and imagery (Albers et al., 2013; Tong, 2013; Lee et al., 2012; Harrison & Tong, 2009) and to show the complexity of interaction within the visual cortex. However, MVPA is limited in its ability to decompose brain activity patterns into distinct sources of variation, and how to measure the represented content remains largely unclear. This study used voxel-wise encoding models (Naselaris et al., 2015; Thirion et al., 2006) to refine the shared neural representation between visual perception and mental imagery. The results provided excellent prediction performance with using the voxel activity in the HVC to predict the voxel responses in the EVC in both perception and imagery conditions via the voxel-to-voxel encoding model, in which the voxel-wise relationship among visual areas cannot be observed with other mathematical models, such as the convolutional neural networks (Mell et al., 2021). Moreover, the prediction performance of another voxel-to-voxel encoding model with the reverse direction (i.e., EVC → HVC) showed that the neural representation of visual stimuli in the high-level visual areas could be predicted by the voxel activity in the EVC, further supporting the idea that the voxel-to-voxel encoding models reflect and emphasize the neural correlation of the HVC and the EVC. The statistical comparisons of the bidirectional voxel-to-voxel encoding models indicated that, in the perception condition, the voxel activity of LOC was found to be better suited to predict the voxel responses in V1, rather than the other way around. This result might indicate that some additionally different information flows from LOC to primary visual cortex. However, there is no such significant difference in the imagery condition. This distinction is an important dimension for differentiating perception and imagery and warrants further attention in the future research.

In addition, we found a linear combination effect of the interaction between low-level visual features and internal high visual activity in the EVC, which was consistent with a previous finding (Zhang et al., 2014). Furthermore, we regarded imagery activity as a covariate and removed it from the internal activation part of perception to conduct an in-depth exploration of the representation mechanism of perception. The results showed that the neural association from HVC to the EVC during perception could be partly explained by that during imagery, but another unexplained part is unique for perception. More importantly, we could easily divide the voxels in the EVC into a Gabor specificity group and a non-Gabor specificity group based on the strategy of voxel-wise encoding model. These results indicated that the Gabor specificity and non-Gabor specificity neural representations were shared between visual perception and imagery. Importantly, in the control analysis, we observed differences between the G and the N groups, with the N group demonstrating superior performance in predicting brain activity in the HVC using the other voxel-to-voxel encoding model (i.e., direction: EVC → HVC). Specifically, the voxels of the N group in V3 could better explain voxel response in the HVC than the prediction performance of the voxels from the G group, supporting the existence of a hierarchical structure in the EVC and suggesting that V3 is closer in function to the high-level visual areas.

Different from the nonlinear and modular representation of the human brain (Op de Beeck et al., 2008), our results (in Figure 3A) showed perceptual representation in the EVC with a linear way by combing the stim-to-voxel and voxel-to-voxel encoding models. This kind of linear combination might indicate that the neural representation related to the visual features and the high visual activity could be independent in the EVC. In addition, in the separation analysis, we found that mental imagery shared an underlying neural association from high visual activity to that in the EVC with visual perception, which was consistent with the assertion that cortical interaction is an essential part of brain processing (Morgan, 2018).

The further linear representation results (Figure 5D) showed the same prediction performances between Gabor specificity and non-Gabor specificity voxels via the voxel-to-voxel encoding model. Our results here may imply that the high visual activity effectively and equally measured the two groups of voxel activity in the EVC during perception and imagery using the encoding model. Furthermore, our results demonstrated that the neural association from high visual activity to EVC could be explained by Gabor features and non-Gabor features, independently. A previous study using encoding models and representational similarity analysis quantitatively showed that the semantic content of an image mainly predicted the activity of voxels in the EVC (V1, V2, and V3) and these features were essentially depictive but also propositional (Naselaris et al., 2009). These findings might imply that the activity of early visual areas could not only depict an object preserving visual features but also naturally reflect semantics of the object to some extent. However, it is still unclear how the representation of information in EVC is essentially depictive and can implement symbolic functions naturally. In the present study, we showed that there are two parallel activation modes in the EVC of the brain, one that corresponds to the visual features and the other that is nonvisual, which means that this mode may further reveal different information representation modes of the EVC to some extent. It should be noted that there have been debates about the essence of mental imagery, mainly focusing on the relationship between depictive representation and propositional representation in our brain. The dual-coding theory (Clark & Paivio, 1991) emphasized that there were two distinct subsystems in our brain specialized for dealing with different types of representations. The present findings might also imply that the two types of representations could be compatible in the early stages of the visual ventral stream with different neural activity. Of course, this problem needs to be further clarified. On the one hand, the types of voxels in EVC should be further differentiated, especially on aspect of the Gabor-specificity voxels, which were sensitive to visual characteristic changes of external visual stimuli. This problem can be further examined with the help of external stimuli, such as the changes in background brightness, so as to reveal its impact on mental imagery ability and its neural basis. On the other hand, the different EVC activity models observed in this study need to be further combined using technology with higher time resolution to further reveal the timing of their different processes (Xie et al., 2020; Dijkstra, Mostert, de Lange, Bosch, & van Gerven, 2018), so as to better understand these different activity modes and their interactions, which provides a new perspective for us to further understand the relationship between imagery and perception, as well as the individual differences of mental imagery.

The results of the current research may be applied to improve future technology. The clear and elaborate representation mechanism offers an alternative possibility concerning applying and updating an artificial intelligence system with consciousness. In addition, the shared and independent neural representation mechanism may lead us toward a deeper understanding and explanation of abnormal phenomena of brain information processing, such as those found in hallucinations and schizophrenia. These results could then in turn lead to more effective and reasonable treatments. The current study included some limitations that should be addressed in future work. First, we did not obtain significant results when using the stim-to-voxel encoding model with Gabor features to predict the voxel response in the EVC during visual imagery, which was presumably primarily caused by the imagery experimental design. In the future, one could revise the experimental design by using specific imagery items for participants (Naselaris et al., 2015) to improve the model performance. Second, some researchers argue that visual imagery engages the left fusiform gyrus instead of the EVC (Spagna et al., 2021); perhaps we could take other brain regions into consideration. This current research also leads to avenues for future research, for example, aphantasia (Keogh, Pearson, & Zeman, 2021), a new special research topic, should be investigated and compared with normal visual perception and imagery to further explore brain representation. In addition, application of the voxel-wise encoding models afforded us an opportunity to explore the shared and independent representation mechanism between visual perception and mental imagery; and thus, in the future, new techniques and methods need to be developed and the questions of dynamic neural representation (Dijkstra et al., 2018) and visualization of the independent representation need to be addressed.

In conclusion, the present study first showed that during perception, there is a linear combination of Gabor features and high visual activity in the EVC, which was shared by mental imagery. We further demonstrated that Gabor specificity and non-Gabor specificity neural activity were shared in the hierarchy of the visual system during both visual perception and mental imagery. These observations provide new insights into the underlying neural substrate between visual perception and imagery.

Code and Data Accessibility

The code for generating the GWP is freely available online (https://www.mathworks.cn/matlabcentral/fileexchange/60088-gabor-wavelet-pyramid). The code for the encoding models is based on the STRFlab toolbox, which is accessible at https://strflab.berkeley.edu/. The preprocessed data and relevant code are publicly available on the website of the Kamitani Lab (https://github.com/KamitaniLab/GenericObjectDecoding), and the raw fMRI data are available from OpenNeuro (https://openneuro.org/datasets/ds001246). All calculation processes were run on MATLAB R2017a on a Windows operating system.

Acknowledgments

We thank Professor Zhaoping Li, who is with Max Planck Institute for Biological Cybernetics, for giving writing suggestions and assistance.

Reprint requests should be sent to Delong Zhang, School of Psychology, Centre for Studies of Psychological Application, and Guangdong Key Laboratory of Mental Health and Cognitive Science, South China Normal University, Guangzhou 510631, China, or via e-mail: delong.zhang@m.scnu.edu.cn.

Data Availability Statement

Our data and materials can be retrieved from https://github.com/KamitaniLab.

Author Contributions

Yingying Huang: Conceptualization; Data curation; Formal analysis; Methodology; Resources; Validation; Visualization; Writing—Original draft; Writing—Review & editing. Frank Pollick: Writing—Review & editing. Ming Liu: Supervision. Delong Zhang: Conceptualization; Funding acquisition; Methodology; Project administration; Supervision; Writing—Original draft; Writing—Review & editing.

Funding Information

National Natural Science Foundation of China (https://dx.doi.org/10.13039/501100001809), grant number: 31600907.

Diversity in Citation Practices

Retrospective analysis of the citations in every article published in this journal from 2010 to 2021 reveals a persistent pattern of gender imbalance: Although the proportions of authorship teams (categorized by estimated gender identification of first author/last author) publishing in the *Journal of Cognitive Neuroscience* (*JoCN*) during this period were M(an)/M = .407, W(oman)/M = .32, M/W = .115, and W/W = .159, the comparable proportions for the articles that these authorship teams cited were M/M = .549, W/M = .257, M/W = .109, and W/W = .085 (Postle and Fulvio, *JoCN*, 34:1, pp. 1–3). Consequently, *JoCN* encourages all authors to consider gender balance explicitly when selecting which articles to cite and gives them the opportunity to report their article's gender citation balance. The authors of this article report its proportions of citations by gender category to be as follows: M/M = .650, W/M = .250, M/W = .075, and W/W = .025.

REFERENCES

Aitken, F., Turner, G., & Kok, P. (2020). Prior expectations of motion direction modulate early sensory processing. *Journal of Neuroscience*, *40*, 6389–6397. https://doi.org/10.1523/JNEUROSCI.0537-20.2020, PubMed: 32641404

Albers, A. M., Kok, P., Toni, I., Dijkerman, H. C., & de Lange, F. P. (2013). Shared representations for working memory and mental imagery in early visual cortex. *Current Biology*, *23*, 1427–1431. https://doi.org/10.1016/j.cub.2013.05.065, PubMed: 23871239

Andersson, P., Ragni, F., & Lingnau, A. (2019). Visual imagery during real-time fMRI neurofeedback from occipital and superior parietal cortex. *Neuroimage*, *200*, 332–343. https://doi.org/10.1016/j.neuroimage.2019.06.057, PubMed: 31247298

Berger, C. C., & Ehrsson, H. H. (2013). Mental imagery changes multisensory perception. *Current Biology*, *23*, 1367–1372. https://doi.org/10.1016/j.cub.2013.06.012, PubMed: 23810539

Borst, G., & Kosslyn, S. M. (2008). Visual mental imagery and visual perception: Structural equivalence revealed by scanning processes. *Memory & Cognition*, *36*, 849–862. https://doi.org/10.3758/mc.36.4.849, PubMed: 18604966

Budd, J. M. (1998). Extrastriate feedback to primary visual cortex in primates: A quantitative analysis of connectivity. *Proceedings of the Royal Society of London, Series B: Biological Sciences*, *265*, 1037–1044. https://doi.org/10.1098/rspb.1998.0396, PubMed: 9675911

Butter, C. M., Kosslyn, S., Mijovic-Prelec, D., & Riffle, A. (1997). Field-specific deficits in visual imagery following hemianopia due to unilateral occipital infarcts. *Brain*, *120*, 217–228. https://doi.org/10.1093/brain/120.2.217, PubMed: 9117370

Carandini, M., Demb, J. B., Mante, V., Tolhurst, D. J., Dan, Y., Olshausen, B. A., et al. (2005). Do we know what the early visual system does? *Journal of Neuroscience*, *25*, 10577–10597. https://doi.org/10.1523/JNEUROSCI.3726-05.2005, PubMed: 16291931

Cichy, R. M., Heinzle, J., & Haynes, J. D. (2012). Imagery and perception share cortical representations of content and location. *Cerebral Cortex*, *22*, 372–380. https://doi.org/10.1093/cercor/bhr106, PubMed: 21666128

Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review*, *3*, 149–210. https://doi.org/10.1007/BF01320076

Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A: Optics and Image Science*, *2*, 1160–1169. https://doi.org/10.1364/JOSAA.2.001160, PubMed: 4020513

DeYoe, E. A., Carman, G. J., Bandettini, P., Glickman, S., Wieser, J., Cox, R., et al. (1996). Mapping striate and extrastriate visual areas in human cerebral cortex. *Proceedings of the National Academy of Sciences, U.S.A.*, *93*, 2382–2386. https://doi.org/10.1073/pnas.93.6.2382, PubMed: 8637882

Diekhof, E. K., Kipshagen, H. E., Falkai, P., Dechent, P., Baudewig, J., & Gruber, O. (2011). The power of imagination—How anticipatory mental imagery alters perceptual processing of fearful facial expressions. *Neuroimage*, *54*, 1703–1714. https://doi.org/10.1016/j.neuroimage.2010.08.034, PubMed: 20797441

Dijkstra, N., Bosch, S. E., & van Gerven, M. A. J. (2019). Shared neural mechanisms of visual perception and imagery. *Trends in Cognitive Sciences*, *23*, 423–434. https://doi.org/10.1016/j.tics.2019.02.004, PubMed: 30876729

Dijkstra, N., Mostert, P., de Lange, F. P., Bosch, S., & van Gerven, M. A. J. (2018). Differential temporal dynamics during visual imagery and perception. *eLife*, *7*, e33904. https://doi.org/10.7554/eLife.33904, PubMed: 29807570

Douglas, R. J., & Martin, K. A. C. (2007). Mapping the matrix: The ways of neocortex. *Neuron*, *56*, 226–238. https://doi.org/10.1016/j.neuron.2007.10.017, PubMed: 17964242

Engel, S. A., Rumelhart, D. E., Wandell, B. A., Lee, A. T., Glover, G. H., Chichilnisky, E. J., et al. (1994). fMRI of human visual cortex. *Nature*, *369*, 525. https://doi.org/10.1038/369525a0, PubMed: 8031403

Harrison, S. A., & Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature*, *458*, 632–635. https://doi.org/10.1038/nature07832, PubMed: 19225460

Haynes, J. D., & Rees, G. (2005). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature Neuroscience*, *8*, 686–691. https://doi.org/10.1038/nn1445, PubMed: 15852013

Horikawa, T., & Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, *8*, 15037. https://doi.org/10.1038/ncomms15037, PubMed: 28530228

Hsieh, P. J., Vul, E., & Kanwisher, N. (2010). Recognition alters the spatial pattern of FMRI activation in early retinotopic cortex. *Journal of Neurophysiology*, *103*, 1501–1507. https://doi.org/10.1152/jn.00812.2009, PubMed: 20071627

Ishai, A., & Sagi, D. (1995). Common mechanisms of visual imagery and perception. *Science*, *268*, 1772–1774. https://doi.org/10.1126/science.7792605, PubMed: 7792605

James, W. (1890). *The principles of psychology* (Vol. 1). London: Macmillan. https://doi.org/10.1037/10538-000

Jones, J. P., & Palmer, L. A. (1987). An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, *58*, 1233–1258. https://doi.org/10.1152/jn.1987.58.6.1233, PubMed: 3437332

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, *8*, 679–685. https://doi.org/10.1038/nn1444, PubMed: 15852014

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*, 108–111. https://doi.org/10.1126/science.282.5386.108, PubMed: 9756472

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, *452*, 352–355. https://doi.org/10.1038/nature06713, PubMed: 18322462

Keogh, R., Bergmann, J., & Pearson, J. (2020). Cortical excitability controls the strength of mental imagery. *eLife*, *9*, e50232. https://doi.org/10.7554/eLife.50232, PubMed: 32369016

Keogh, R., Pearson, J., & Zeman, A. (2021). Aphantasia: The science of visual imagery extremes. In J. J. S. Barton & A. Leff (Eds.), *Handbook of clinical neurology* (Vol. 178, pp. 277–296). Elsevier. https://doi.org/10.1016/B978-0-12-821377-3.00012-X, PubMed: 33832681

Kok, P., Jehee, J. F. M., & de Lange, F. P. (2012). Less is more: Expectation sharpens representations in the primary visual cortex. *Neuron*, *75*, 265–270. https://doi.org/10.1016/j.neuron.2012.04.034, PubMed: 22841311

Kosslyn, S. M. (1996). Image and brain: The resolution of the imagery debate. *Journal of Cognitive Neuroscience*, *7*, 415–420. https://doi.org/10.1162/jocn.1995.7.3.415, PubMed: 23961870

Kosslyn, S. M., Pascual-Leone, A., Felician, O., Camposano, S., Keenan, J. P., Thompson, W. L., et al. (1999). The role of area 17 in visual imagery: Convergent evidence from PET and rTMS. *Science*, *284*, 167–170. https://doi.org/10.1126/science.284.5411.167, PubMed: 10102821

Kosslyn, S. M., & Thompson, W. L. (2003). When is early visual cortex activated during visual mental imagery? *Psychological Bulletin*, *129*, 723–746. https://doi.org/10.1037/0033-2909.129.5.723, PubMed: 12956541

Lee, S. H., Kravitz, D. J., & Baker, C. I. (2012). Disentangling visual imagery and perception of real-world objects. *Neuroimage*, *59*, 4064–4073. https://doi.org/10.1016/j.neuroimage.2011.10.055, PubMed: 22040738

Lee, T. S. (1996). Image representation using 2D Gabor wavelets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *18*, 959–971. https://doi.org/10.1109/34.541406

Maier, M., Frömer, R., Rost, J., Sommer, W., & Rahman, R. A. (2021). Conceptual knowledge affects early stages of visual mental imagery and object perception. *bioRxiv*. https://doi.org/10.1101/2020.01.14.905885

Mell, M. M., St-Yves, G., & Naselaris, T. (2021). Voxel-to-voxel predictive models reveal unexpected structure in unexplained variance. *Neuroimage*, *238*, 118266. https://doi.org/10.1016/j.neuroimage.2021.118266, PubMed: 34129949

Morgan, A. T. (2018). *Encoding and decoding of cortical feedback to human early visual cortex*. PhD Thesis. University of Glasgow.

Muckli, L., Petro, L. S., & Smith, F. W. (2013). Backwards is the way forward: Feedback in the cortical hierarchy predicts the expected future. *Behavioral and Brain Sciences*, *36*, 221. https://doi.org/10.1017/S0140525X12002361, PubMed: 23663531

Naselaris, T., Olman, C. A., Stansbury, D. E., Ugurbil, K., & Gallant, J. L. (2015). A voxel-wise encoding model for early visual areas decodes mental images of remembered scenes. *Neuroimage*, *105*, 215–228. https://doi.org/10.1016/j.neuroimage.2014.10.018, PubMed: 25451480

Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., & Gallant, J. L. (2009). Bayesian reconstruction of natural images from human brain activity. *Neuron*, *63*, 902–915. https://doi.org/10.1016/j.neuron.2009.09.006, PubMed: 19778517

O'Craven, K. M., & Kanwisher, N. (2000). Mental imagery of faces and places activates corresponding stimulus-specific brain regions. *Journal of Cognitive Neuroscience*, *12*, 1013–1023. https://doi.org/10.1162/08989290051137549, PubMed: 11177421

Olman, C. A., Ugurbil, K., Schrater, P., & Kersten, D. (2004). BOLD fMRI and psychophysical measurements of contrast response to broadband images. *Vision Research*, *44*, 669–683. https://doi.org/10.1016/j.visres.2003.10.022, PubMed: 14751552

Op de Beeck, H. P., Haushofer, J., & Kanwisher, N. G. (2008). Interpreting fMRI data: Maps, modules and dimensions. *Nature Reviews Neuroscience*, *9*, 123–135. https://doi.org/10.1038/nrn2314, PubMed: 18200027

Pearson, J. (2019). The human imagination: The cognitive neuroscience of visual mental imagery. *Nature Reviews Neuroscience*, *20*, 624–634. https://doi.org/10.1038/s41583-019-0202-9, PubMed: 31384033

Pearson, J., Clifford, C. W., & Tong, F. (2008). The functional impact of mental imagery on conscious perception. *Current Biology*, *18*, 982–986. https://doi.org/10.1016/j.cub.2008.05.048, PubMed: 18583132

Pearson, J., & Kosslyn, S. M. (2015). The heterogeneity of mental representation: Ending the imagery debate. *Proceedings of the National Academy of Sciences, U.S.A.*, *112*, 10089–10092. https://doi.org/10.1073/pnas.1504933112, PubMed: 26175024

Rainer, G., Augath, M., Trinath, T., & Logothetis, N. K. (2001). Nonmonotonic noise tuning of BOLD fMRI signal to natural images in the visual cortex of the anesthetized monkey. *Current Biology*, *11*, 846–854. https://doi.org/10.1016/S0960-9822(01)00242-1, PubMed: 11516645

Reddy, L., Tsuchiya, N., & Serre, T. (2010). Reading the mind's eye: Decoding category information during mental imagery. *Neuroimage*, *50*, 818–825. https://doi.org/10.1016/j.neuroimage.2009.11.084, PubMed: 20004247

Ringach, D. L. (2002). Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *Journal of Neurophysiology*, *88*, 455–463. https://doi.org/10.1152/jn.2002.88.1.455, PubMed: 12091567

Sasaki, Y., Rajimehr, R., Kim, B. W., Ekstrom, L. B., Vanduffel, W., & Tootell, R. B. H. (2006). The radial bias: A different slant on visual orientation sensitivity in human and nonhuman primates. *Neuron*, *51*, 661–670. https://doi.org/10.1016/j.neuron.2006.07.021, PubMed: 16950163

Sereno, M. I., Dale, A. M., Reppas, J. B., Kwong, K. K., Belliveau, J. W., Brady, T. J., et al. (1995). Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science*, *268*, 889–893. https://doi.org/10.1126/science.7754376, PubMed: 7754376

Singh, K. D., Smith, A., & Greenlee, M. W. (2000). Spatiotemporal frequency and direction sensitivities of human visual areas measured using fMRI. *Neuroimage*, *12*, 550–564. https://doi.org/10.1006/nimg.2000.0642, PubMed: 11034862

Sirigu, A., & Duhamel, J. R. (2001). Motor and visual imagery as two complementary but neurally dissociable mental processes. *Journal of Cognitive Neuroscience*, *13*, 910–919. https://doi.org/10.1162/089892901753165827, PubMed: 11595094

Sohoglu, E., Peelle, J. E., Carlyon, R. P., & Davis, M. H. (2012). Predictive top–down integration of prior knowledge during speech perception. *Journal of Neuroscience*, *32*, 8443–8453. https://doi.org/10.1523/JNEUROSCI.5069-11.2012, PubMed: 22723684

Spagna, A., Hajhajate, D., Liu, J., & Bartolomeo, P. (2021). Visual mental imagery engages the left fusiform gyrus, but not the early visual cortex: A meta-analysis of neuroimaging evidence. *Neuroscience & Biobehavioral Reviews*, *122*, 201–217. https://doi.org/10.1016/j.neubiorev.2020.12.029, PubMed: 33422567

Stokes, M., Thompson, R., Cusack, R., & Duncan, J. (2009). Top–down activation of shape-specific population codes in visual cortex during mental imagery. *Journal of Neuroscience*, *29*, 1565–1572. https://doi.org/10.1523/JNEUROSCI.4657-08.2009, PubMed: 19193903

Thirion, B., Duchesnay, E., Hubbard, E., Dubois, J., Poline, J. B., Lebihan, D., et al. (2006). Inverse retinotopy: Inferring the visual content of images from brain activation patterns. *Neuroimage*, *33*, 1104–1116. https://doi.org/10.1016/j.neuroimage.2006.06.062, PubMed: 17029988

Tong, F. (2013). Imagery and visual working memory: One and the same? *Trends in Cognitive Sciences*, *17*, 489–490. https://doi.org/10.1016/j.tics.2013.08.005, PubMed: 23958465

Tugnait, J. K. (1994). Estimation of linear parametric models of nonGaussian discrete random fields with application to texture synthesis. *IEEE Transactions on Image Processing*, *3*, 109–127. https://doi.org/10.1109/83.277894, PubMed: 18291913

Xie, S., Kaiser, D., & Cichy, R. M. (2020). Visual imagery and perception share neural representations in the alpha frequency band. *Current Biology*, *30*, 2621–2627. https://doi.org/10.1016/j.cub.2020.04.074, PubMed: 32531274

Zhang, D., Wen, X., Liang, B., Liu, B., Liu, M., & Huang, R. (2014). Neural associations of the early retinotopic cortex with the lateral occipital complex during visual perception. *PLoS One*, *9*, e108557. https://doi.org/10.1371/journal.pone.0108557, PubMed: 25251083