Title:

Cortical feedback to V1 and V2 contains unique information about high-level scene structure.

Authors:

Andrew T. Morgan[1], Lucy S. Petro[1], Lars Muckli[1]*

Author Affiliations:

[1] Centre for Cognitive Neuroimaging, Institute of Neuroscience and Psychology, University of Glasgow, Glasgow G12 8QB, United Kingdom

Corresponding Author:

* Lars Muckli, 58 Hillhead St., Room 503, University of Glasgow, Glasgow G12 8QB, United Kingdom, +44 (0)141 330 6237, Lars.Muckli@glasgow.ac.uk

Abstract:

Early visual cortical neurons receive highly selective feedforward input, which is amplified or disamplified by contextual feedback and lateral connections. A significant challenge for systems neuroscience is to measure the feature space that drives these feedback channels. We occluded visual scenes and measured non-feedforward stimulated subregions of V1 and V2 using fMRI and multi-voxel pattern analyses. We found that response patterns in these subregions contain two high-level scene features, category and depth information. Responses in non-feedforward stimulated V1 and V2 differed from each other, suggesting that feedback to these two areas has unique information content. Further, we reveal that computational models of visual processing inadequately describe early visual cortex because they do not account for the brain's internal modeling of the world.

Introduction:

Cortical neurons receive two sources of input. The first, feedforward input, is sensitive to small portions of stimulus space as defined by classical receptive fields. The second consists of non-feedforward input from feedback and lateral interactions, which amplify and disamplify responses to the feedforward signals based on context[1,2]. The abundance of feedback connections in cortical circuits[3,4] and the diversity of responses to identical stimuli in different contexts[5] points out a challenge central to the understanding of neural computations. Improving our knowledge of feedback will provide insight to fundamental neuroscientific questions such as cognition, attention, perception, memory, action and mental disorders.

To study feedback, one must disentangle feedforward and non-feedforward sources of input. This involves independent stimulation or inactivation of feedback and feedforward pathways, which is achieved via single- or multiunit recordings paired with electrical stimulation, pharmacological intervention, cooling or optogenetics[4,6]. These methodologies provide detailed characterizations of neural activity, but are generally too invasive for studying the healthy human brain. Since the coordination of feedforward signals likely occurs through widely distributed feedback processes of contextual modulation[1], non-invasive modalities with large spatial coverage such as functional magnetic resonance imaging (fMRI) are advantageous for studying

feedback properties. Furthermore, fMRI is sensitive to energy-use rather than directly to spiking, and while feedback modulates feedforward responses, it does not necessarily generate spikes[1].

A non-invasive strategy to measure feedback is to homogenize feedforward input using visual occlusion. Functional brain imaging studies have measured feedback-specific effects in retinotopic visual cortex (V1) by occluding feedforward input across experimental conditions and observing that feedback signals are heterogeneous[4,7-9]. Smith and Muckli[8] blocked feedforward input to subsections of V1 using a uniform occluder and demonstrated that scenes can be decoded using pattern classification methods from responses in the occluded cortical area. Petro et al.[5] showed that decoding of faces in regions of V1 that process facial features is dependent on task. Vetter et al.[10] were able to decode different sounds from V1 activity in blindfolded subjects. These results provide evidence for feedback affecting V1 responses in a contextually relevant way, yet we can say little about the content of feedback signals. These signals may process specific feature predictions, and therefore produce response patterns similar to those observed when early visual cortex is stimulated by feedforward inputs[11]. Conversely, feedback may provide early visual cortex with general templates for expected scene structure based on high-level characteristics. In the current study, we aimed to investigate the specificity of feedback signals to human primary and secondary visual cortices (V1 and V2) during scene processing.

Results:

Investigating higher-level structure of cortical feedback to early visual cortex

We blocked feedforward input to subsections of retinotopic visual cortex with a uniform visual occluder covering one quarter of the visual field[8] while participants viewed 24 real-world scenes. To probe information characteristics of feedback, we identified two abstract features that modulate V1 responses: scene category and depth[12,13]. Scenes were balanced across six categories and two spatial depths. We localized subsections of left hemisphere V1 and V2 responding either to the occluded portion of the visual field (lower right image quadrant), or non-occluded visual field (upper right image quadrant; Figure 1). This process yielded four regions of interest (ROIs), hereafter referred to as Occluded and Non-Occluded V1 and V2. We also mapped

population receptive field (pRF) locations of individual voxels [14] to ensure that their response profiles were within regions of interest in the visual field.

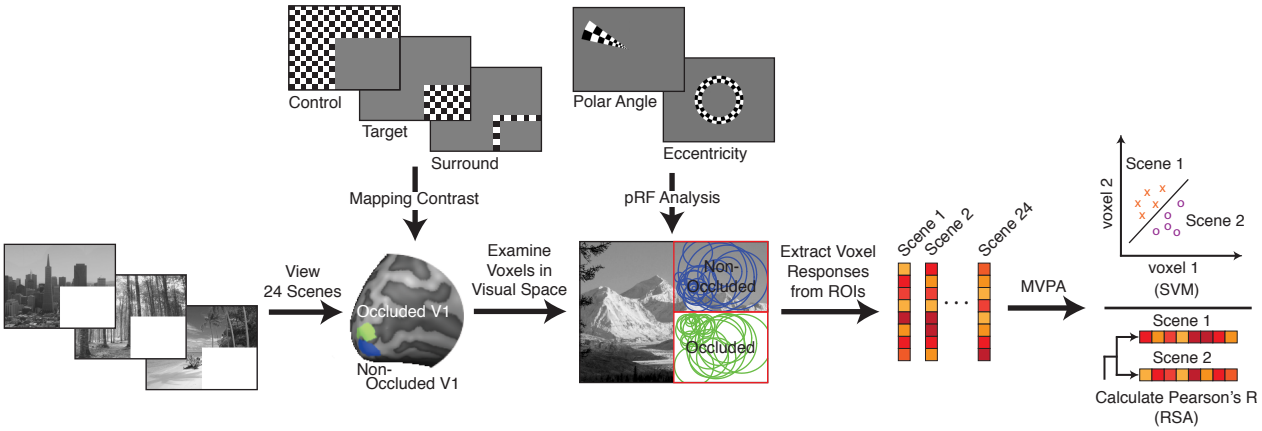

Figure 1. Experimental procedures. Participants viewed 24 scenes with lower right quadrants occluded. Scenes spanned 6 categories (Beaches, Buildings, Forests, Highways, Industry, and Mountains) and 2 depths (Near and Far). Occluded and Non-Occluded subsections of early visual cortex were localized using mapping contrasts. Retinotopic mapping data were used to separate V1 and V2 and for mapping population receptive fields (pRFs). Voxel pRFs not completely contained by the quadrant of interest (2σ from pRF center) were excluded from further analyses. Remaining voxels were included in multi-voxel pattern analyses.

Contextual Information in Occluded Early Visual Cortex

Using single-trial linear Support Vector Machine (SVM) classification we could decode individual scene, category, and depth information in occluded and non-occluded regions of V1 and V2 (Figure 2A, Table S1A; permutation tested against chance-level using 1000 samples, all p-values < 0.001). Consistent with previous results[8], we observed statistically higher classification performance in Occluded V1 than in Occluded V2, and this result held true for scene, category, and depth classification (bootstrap comparisons using 10000 samples, p-values = 0.0004, 0.0112, and 0.0024, respectively). Scene, category and depth decoding was also higher in Non-Occluded V1 than Non-Occluded V2, although differences were not statistically significant (all p-values ≥ 0.17). Our decoding results were reliable at the individual-subject level. Scene, category, and depth decoding were above chance-level in 11 of 12 subjects in nearly all regions tested. Only decoding of Depth in Occluded V2 had lower individual-subject classifier performances, which were above chance-level in 6 of 12 subjects. In

individual subjects, significance was determined by comparison to a null distribution of 1000 permutations, $p < 0.05$ considered significant.

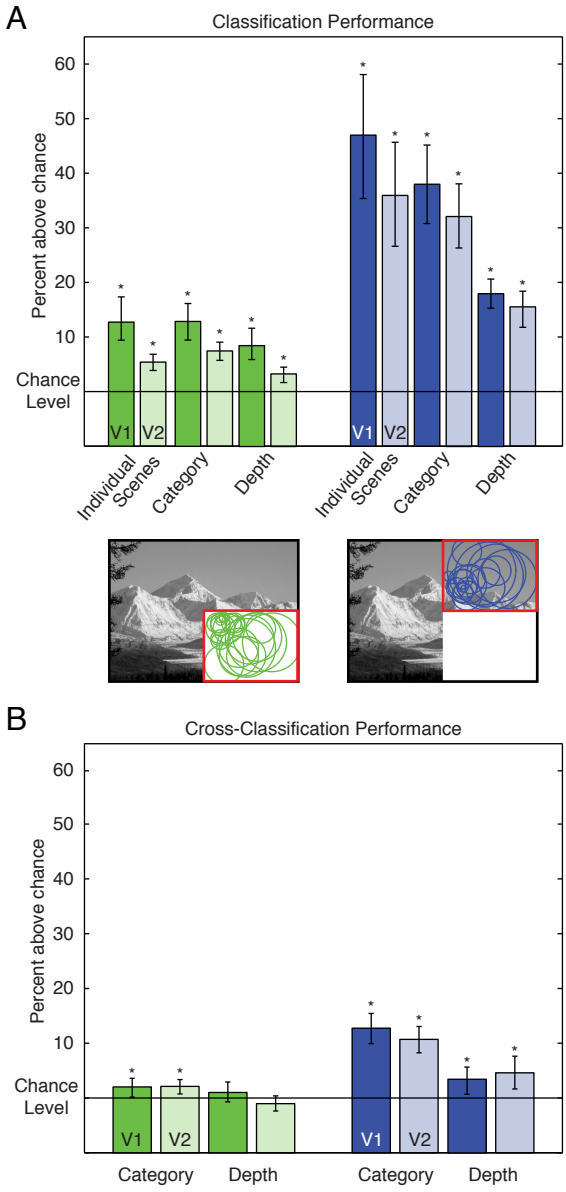

Figure 2. Classification and cross-classification performance. (A) Average classifier performances (12 subjects) are shown for each visual area with 95% confidence intervals (calculated via 1000 bootstrap samples of individual subject performances). Occluded analyses are shown in green; Non-Occluded analyses are shown in blue. Asterisks indicate greater than chance-level decoding accuracy, $p < 0.05$. Chance level is 4.17% for individual scenes, 16.67% for categories, and 50% for depth. (B) Cross-classification performance. Training occurred on 18 and 22 [of 24] randomly chosen scenes in category and depth analyses, respectively, and testing occurred on scenes not used for classifier training. This was repeated 100 times per subject, and 95% confidence intervals were calculated via 1000 bootstrap samples of individual subject performances. See also Table S1 and Figure S2.

To further test whether non-stimulated V1 and V2 represent higher-level properties of scenes, we performed cross-classification analyses for category and depth information (Figure 2B, Table S1B). We trained SVM models using responses to a subset of our scenes, leaving out a test-set for later cross-classification. For analysis of category, 18 (of 24) scenes were selected, leaving out one scene per category. For depth, we selected 22 (of 24) scenes, leaving out one scene per depth. We tested the classifier on the remaining scenes in a cross-classification approach. Due to the large number of possible image permutations in these analyses, we randomly assigned scenes to training and testing sets 100 times in each subject. Cross-classification of category was successful in Occluded and Non-Occluded areas (permutation tested against chance-level using 1000 samples, p-values = 0.012, 0.003 for Occluded V1 and V2; p-values < 0.001 for Non-Occluded V1 and V2). Cross-classification of depth was only successful in Non-Occluded areas (p-values ≥ 0.12 for Occluded V1 and V2; p-values ≤ 0.004 for Non-Occluded V1 and V2). Occluded and Non-Occluded V1 and V2 displayed comparable performance levels. Overall, these results demonstrate that category information in Occluded responses is generalizable across scenes, while depth information is not.

Visualizing Retintopic Response Patterns

Our decoding results show differences in response patterns to individual images in Occluded V1 and V2. The SVM algorithm detects these differences, but cannot indicate which scene features elicit them. To compare with scene features, we projected response patterns to each scene into visual space[15]. pRF estimates (2D isotropic Gaussian functions) were weighted by each voxel's response with the mean response to all 24 scenes removed. We plotted the sum of these functions as a heat map, with warm and cool colors indicating activations above and below the mean response in each voxel, respectively (Figure 3). In Non-Occluded portions of scenes, clutter and texture elicited above-average responses, while homogenous visual features such as sky evoked lower levels of activation. This observation is in line with previous studies that have shown stimulus contrast energy drives early visual cortical responses[16,17]. We observed continuations of textured objects into Occluded areas; examples include the beach debris in the fourth Beach and the tower in the second Industry scene. These projections highlight response differences between V1 and V2 when

receiving only feedback input (Figure 3A versus 3B). In Non-Occluded portions of scenes, V1 and V2 respond similarly, matching previous reports of responses to natural stimuli[18]. However, Occluded V1 and V2 do not display similar response patterns.

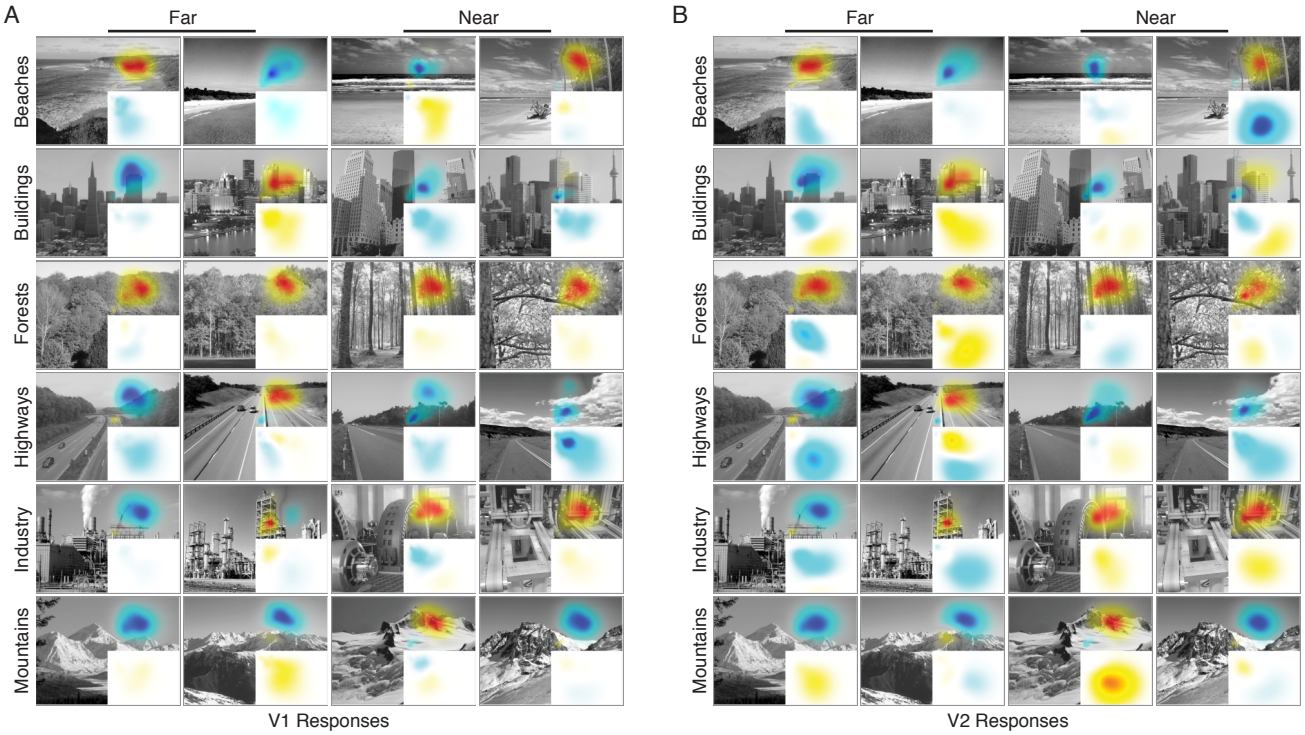

Figure 3. Projections of (A) V1 and (B) V2 response patterns into visual space. Projections of voxel responses versus baseline were averaged over subjects and scaled per scene. Projections were obtained by calculating a weighted average of all voxels' pRFs, where weights were each voxel's response amplitude with the mean response to all scenes removed. Warm colors above mean responses, and cool colors are below.

Relating Scene Representations in V1 and V2

Our visual field projections suggest that Non-Occluded V1 and V2 represent scenes similarly, while Occluded areas do not. To quantify these differences, we conducted Representational Similarity Analyses (RSA). By applying an iterative split-half cross-correlation procedure[13], we compiled matrices representing correlations between individual scene response pairs in each region (Figure 4). To assure reliability between

analysis methods, we first conducted individual scene classification using RSA. Diagonal points of similarity matrices indicate consistency in response patterns to individual scenes between the two halves of the data, and off-diagonal points are correlations between pairs of responses to different scenes. As such, higher correlations in the diagonal indicate the ability to distinguish individual scenes within a region of interest. Here, comparison of values along the diagonal against those off the diagonal showed significant differences in both occluded (V1: $t(25.93) = 5.29$, $p < 0.001$; V2: $t(25.55) = 6.32$, $p < 0.001$) and non-occluded regions (V1: $t(36.10) = 18.25$, $p < 0.001$; V2: $t(34.03) = 16.23$, $p < 0.001$), tested against null distributions of 5000 permutations. These results reaffirm that responses in Occluded subregions of V1 and V2 contain information about individual scenes.

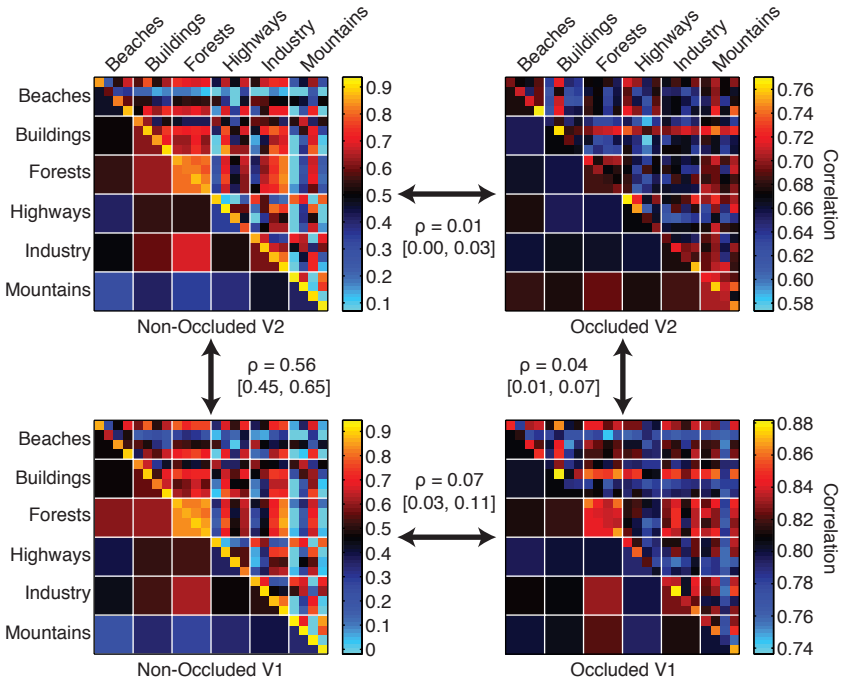

Figure 4. Representational Similarity Analyses for cortical regions. Similarity matrices comprise of correlations in the patterns of response between pairs of scenes, averaged amongst subjects. Diagonals are correlations of scene responses in each half of the data; a measure of signal reliability. For ease of interpretation, the lower triangle of each matrix displays average category correlations of off-diagonal points. Second-order correlations between matrices were calculated using Spearman's rho for each possible concatenation of the data to determine 95% confidence intervals (in brackets).

Having shown above-chance decoding, we then compared how each region represents the set of scenes by correlating the off-diagonal points of their similarity matrices (see Supplemental Methods). We used multidimensional scaling to visualize these comparisons in two dimensions[19,20]. In this scaled space, regions representing scenes similarly appear close together while those representing them differently appear far apart. We also compared representational structures of cortical responses to those of computational models. Since we know which image features models respond to, these comparisons provide insight to the information content of cortical responses. Figure 5 displays the representational space with confidence areas for the four cortical regions, idealized Category and Depth models, and three popular biologically-inspired computational models: the Weibull contrast model, inspired by lateral geniculate processing[17,21,22]; the Gist algorithm, similar to Gabor filters in V1[23]; and the H-MAX model (Layer C2), matched to tuning properties of intermediate ventral stream areas such as V4 or posterior IT[24].

Non-Occluded V1 and V2 occupy the same portion of the representational space and therefore represent scenes similarly, consistent with visual field projections. Likewise, Occluded V1 and V2 are separated, showing that they respond to scenes differently from one another. Interestingly, the Depth model is closer to Non-Occluded cortical regions than Occluded ones, while the Category model is more similar to Occluded V2. Further, no computational model overlaps with the representational space of cortical responses, consistent with previous findings[19]. These results suggest that models might be improved by incorporating feedback.

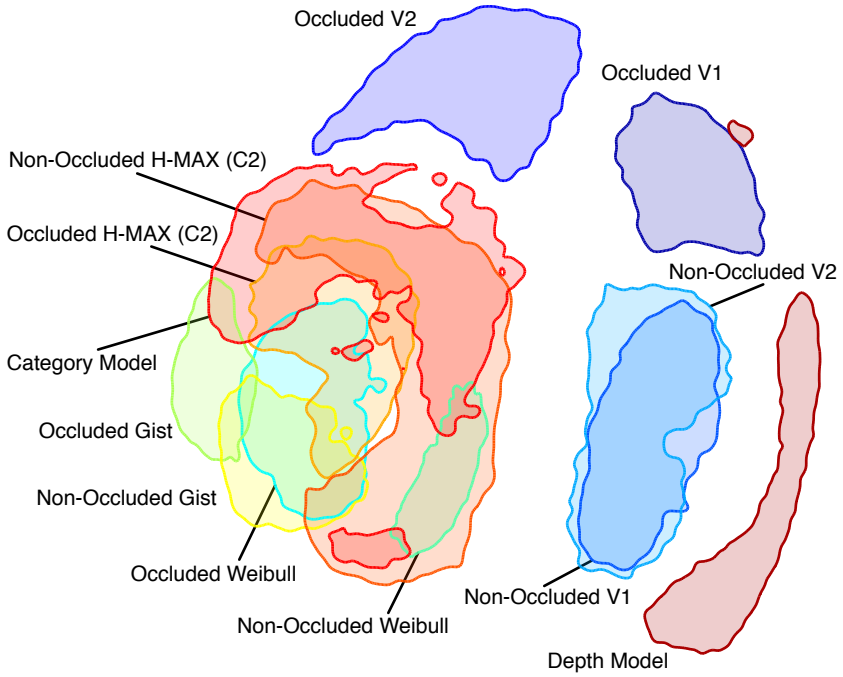

Figure 5. Multidimensional scaling configurations of cortical regions and model representations. 95% confidence areas are shown for cortical regions, idealized Category and Depth models, and three computational models: Weibull, Gist, and H-MAX C2 (run separately on Non-Occluded and Occluded portions of scenes). Multidimensional scaling was performed on 10,000 bootstrap samples of the data and aligned using Procrustes transformations without scaling. Kernel density estimation (KDE) determined the distribution of each cortical area or model in the representational space, and confidence areas were defined as a contour that contained 95% of the KDE mass. See also Figure S1.

Discussion:

Our data challenge models of visual processing in early visual cortex; V1 and V2 response patterns to visual stimulation are different than would be predicted by common computational models. This discrepancy is even more pronounced in Occluded regions, where models would predict identical responses to the white occluder because they do not account for contributions from contextual feedback. However, in Occluded regions, we observed activation patterns that were informative for determining individual scene, category, and depth information about the surrounding images, thus indicating that contextual feedback to early visual cortex is scene-specific, yet exhibits high-level structure. Moreover, we found that responses in Occluded V1 and V2 differed from each other, suggesting that feedback to these two areas has unique information content.

Our findings are consistent with descriptions of the visual system as a hierarchical inference network, with V1 acting as a high-resolution geometric buffer or blackboard[25,26]. In other words, V1 preserves scene information for reference in calculations where high-resolution image details or spatial precision are required. At first glance, this description seems incongruent with the idea that V1 can depict visual information that is missing or hidden from view, such as that masked by the occluder in our study. Importantly though, feedback to V1 depicts detailed internal models that share characteristics with expected or predictable feedforward input[27]. Our visual field projections illustrate this by revealing a number of response patterns that match predictable occluded features. The current results and a number of other studies lend support to inferential filling-in of expected information in early visual cortex (i.e. during perception of absent visual input[8,9,25,26,28-30].

These data go beyond the inferential filling-in of expected information in early visual cortex. Our cross-classification results suggest that internal models are constrained by feedforward inputs but are based on generalized scene templates. The visual system first attends to coarse-scale scene information in order to quickly estimate input and activate scene templates in memory[31,32]. Feedback loops connecting higher areas to early visual cortex may incorporate scene-specific information into the initial template structure. Since fMRI is associated with late components of neural response time courses[18], Occluded responses may display characteristics of initial templates as well as additional scene-specific details. This could explain our ability to classify individual scenes and cross-classify category information; Non-Occluded portions of scenes activate global scene templates that also cover Occluded portions of scenes. Scene-specific additions to these templates would allow for individual scene classification, even between scenes of the same category.

Sensory processing outside of the visual domain can also transmit contextual information to early visual cortex. A recent study found categorical responses in early visual cortices of blindfolded participants presented with complex natural sounds, yet these responses did not represent individual exemplars[10]. Together with the current study these results support the idea that sensory processing occurs via a hierarchical inference framework. Feedback predicts potential feedforward input based on available sensory input from any modality. For sounds, generalized category templates may be fed back to V1, but not the scene-specific additions that

allowed us to decode individual images. Since contextual sensory input in our study comes from surrounding visual information, visual features are predicted with greater spatial acuity.

In cross-classification analyses, generalization of depth information was weaker than that of category. We could not train on one set of images and use depth information to generalize to a new set of scenes, which was possible with category. While Occluded V1 and V2 responses may not actually contain depth information, an alternative explanation is that depth is not sufficiently defined in our stimulus set. We included scenes that were Near or Far within scene categories, as in[13]. We found generalizable depth information in Non-Occluded regions, but not in Occluded regions. Scene depth is highly correlated with category[33], and it is possible that our stimuli did not effectively span depth ranges in order to detect such depth-category interactions.

Occluded V1 and V2 responses differ, thus suggesting that feedback to these two areas has unique information content. Both V1 and V2 receive feedback from many brain areas[34,35]. Feedback to each area therefore likely plays a specialized role in predicting aspects of visual scenes. Despite V2's interconnectedness with extrastriate cortex, it is highly dependent on input from V1 due to its role in summarizing natural patterns[18,24,36]. In the current study, input to V2 from V1 could be a disparate reconciliation of non-meaningful feedforward input and contextual expectations from feedback[25].

To determine the specific aspects of scene processing which feedback signals to V1 and V2 specialize in, feedforward and feedback signals need to be effectively separated – a significant challenge to systems neuroscience. Feedforward and feedback signals opportunely input to different layers of visual cortex, with feedforward connections terminating in middle layers and feedback in superficial and deep layers[37]. Ultra-high-field fMRI (7T) provides a high-resolution tool to examine activity in different layers of human cortex. Indeed, Muckli et al.[28] recently found that three occluded scenes (similar to those in the current study) could be decoded using the response patterns of superficial layers (but not middle layers) of V1 and V2. However, the number of scenes used in the current study allowed for direct comparison of how different areas represent scenes through RSA[19]. This was not possible in Muckli et al. because they used a smaller stimulus set. High-resolution fMRI along with experiments utilizing larger stimulus sets will enable researchers to fully investigate the information content of feedback to these areas and the relationships between different cortical areas.

The computational models that we compared to cortical representations are inspired by biological visual systems, but they process scene statistics in a strictly feedforward manner. These models differ from any of our early visual cortical regions, strikingly even from Non-Occluded regions (Figure 5). These results indicate that models that neglect the role of feedback oversimplify early visual cortical processing.

Compared models also used relatively low-level image features. High-level feature detection has enabled deep learning networks to achieve incredibly high performance on a number of natural signal processing tasks including visual object and speech recognition[38]. Still, such networks predominately use feedforward architectures, and would therefore perform sub-optimally when presented with partially occluded visual scenes because they are unable to recognize occluded portions of the visual field as missing information. Occluded areas would be integrated identically to the rest of the scene, but without contributing any useful information. Reichert and Serre[39] show that networks utilizing synchronization (a form of feedback) by allowing complex weights between network nodes can disentangle occluded shapes. Further development of such networks will explain aspects of cortical responses not yet captured by current computational models, and may provide insight to mechanisms of feedback processing in early visual cortex.

Our results are in line with theoretical views explaining the visual system as a hierarchical inference network. Moreover, feedback to Occluded V1 & V2 has unique information content. We found that early visual cortex represents scenes differently from three popular biologically inspired computational models. Together these results highlight that a true understanding of the neural computations of early visual cortex will involve understanding how and what information is conveyed by feedback.

Methods:

Participants

Twelve healthy individuals (8 female, age 26.42 ± 5.76, mean ± SD) with normal or corrected-to-normal vision gave written informed consent to participate in this study, in accordance with the institutional guidelines of the local ethics committee of the College of Science & Engineering at the University of Glasgow (#CSE01127).

Stimuli

Twenty-four real-world scenes from six categories were chosen from a dataset compiled by Walther et al.[12]. Images were displayed in gray-scale (matched for global luminance) on a rear-projection screen using a projector system (1024 x 768 resolution, 60 Hz refresh rate). Stimuli spanned 19.5 x 14.7˚ of visual angle, and were presented with the lower-right quadrant occluded by a white box (occluded region spanned ≈ 9 x 7˚). A centralized fixation checkerboard (9 x 9 pixels) marked the center of the scene images.

Experimental Design

Each of the eight runs consisted of six blocks of eight sequences of stimulation with intervening fixation periods, plus two mapping blocks (total scanning time per run was 804s). Each stimulation sequence lasted 120s, with 12s fixation at the beginning and the end of each series. Stimuli were flashed at a rate of 5Hz in order to maximize the signal-to-noise ratio of the BOLD response[40]. Each sequence was presented in a pseudo-randomized order where individual images were not shown repeatedly. Over the course of the experiment, each scene was presented 16 times (two times per run). To ensure fixation, we instructed participants to keep fixation on the central checkerboard and respond via a button press to randomized color changes of the fixation (from black/white to red/green) according to the category of the scene during the color change.

We used mapping blocks to localize the cortical representation of the occluded region[9]. In a block design, subjects viewed contrast-reversing checkerboard stimuli (5Hz) at three visual locations: Target (lower-right quadrant), Surround (of the target), and Control (remaining three quadrants). Each condition was

displayed for 12s with a 12s fixation period following, and mapping blocks were randomly inserted between experimental blocks, once per run. We conducted retinotopic mapping (polar-angle and eccentricity) runs separately from the main experiment.

fMRI Acquisition

MRI data were collected at the Centre for Cognitive Neuroimaging, University of Glasgow. T1-weighted anatomical and echo-planar (EPI) images were acquired using a research-dedicated 3T Tim Trio MRI system (Siemens, Erlangen, Germany) with a 32-channel head coil and integrated parallel imaging techniques (IPAT factor: 2). Functional scanning used EPI sequences to acquire partial brain volumes aligned to maximize coverage of early visual areas (18 slices; voxel size: 3mm, isotropic; 0.3mm interslice gap; TR = 1000ms; TE = 30ms; matrix size = 70x64; FOV = 210x192mm). Four runs of the experimental task (804 vol.), one run of retinotopic mapping [session 1: polar angle (808 vol.); session 2: eccentricity (648 vol.)], and a high-resolution anatomical scan (3D MPRAGE, voxel size: 1mm, isotropic) were performed during each of two scanning sessions.

fMRI Data Preprocessing

Functional data for each run were corrected for slice time and 3D motion, temporally filtered (high-pass filter with Fourier basis set [6 cycles], linearly detrended), and spatially normalized to Talairach space using Brain Voyager QX 2.8 (Brain Innovation, Maastricht, Netherlands). No spatial smoothing was performed. These functional data were then overlaid onto their respective anatomical data in the form of an inflated surface. Retinotopic mapping runs were used to define early visual areas V1 and V2 using linear cross-correlation of eight polar angle conditions.

A general linear model (GLM) with one predictor for each condition (Target > Surround; mapping conditions from experimental runs) was used to define regions of interest (ROI) that responded to the visual target region (lower-right quadrant) and the control region (upper-right quadrant), within V1 and V2. We then performed pRF analyses[14] on all ROI voxels, and excluded those voxels whose response profiles were not fully contained (within 2σ of their pRF center) by the respective visual ROI. Lastly, a conjunction of two GLM contrasts (Target > Surround & Target > Control for Occluded ROIs, and Control > Surround & Control > Target

for Non-Occluded ROIs) was used to exclude any voxels responding to stimuli presentation outside their respective visual ROI (see Figure 1). Time courses from each selected vertex were then extracted independently per run and a GLM was applied to estimate response amplitudes on a single-block basis. The resulting beta weights estimated peak activation for each single block assuming a standard 2γ hemodynamic response function.

Multivariate Pattern Analyses

For SVM classification analyses, a separate regressor modeled each experimental trial. This procedure yielded a pattern of voxel activations for each single trial, and parameter estimates (β values) were obtained for each voxel and then z-scored. A Linear SVM classifier was trained to learn the mapping between a set of all available multivariate observations of brain activity and the particular scene presented, and the classifier was tested on an independent set of test data. Classification analyses were performed using a pairwise multiclass method. Classifier performance was assessed using an n-fold leave-one-run-out cross-validation procedure where models were built on n – 1 runs and were tested on the independent nth run (repeated for the eight runs of the experiment). In analyses of category and depth-based classification, individual scene presentation labels were combined based on these distinctions before training and testing of the SVM classifiers.

Cross-classification analyses were performed similarly to those of our cross-validated classification, but our scene set was split up prior to model training. Training set sizes were 18 and 22 scenes for category and depth analyses, respectively, and testing sets consisted of the remaining scenes. Due to the large number of possible scene permutations, we conducted 100 iterations of our analyses in each subject. For these analyses, training and testing sets were defined in a pseudo-random manner where each category or depth was evenly represented within both sets. As in our cross-validated classification, training and testing of models occurred on independent data sets using a leave-one-run-out procedure.

For RSA, an iterative split-half correlation method[13,19] was applied to each subject's data where runs were split into two halves (four runs in each) and concatenated. GLM analyses were then conducted to estimate condition-based responses to each scene in the respective half of the data and repeated for the 35 possible combinations of data splits. Cross-correlation was then used to establish the similarity between the

reponse patterns of each pair of scenes. A Fisher transformation was applied to each correlation value before combining correlations into group analyses, and the transformation was reversed for results presentation. To calculate second-order correlations of similarity matrices, subjects' data were concatenated by quarters (two runs in each) and analyzed similarly to first-order correlations with each half producing a similarity matrix. Off-diagonal points were then correlated using Spearman rank correlation in 210 different combinations of runs to create a distribution of correlation values. Here, we used Spearman correlations instead of Kendall's Tau-a because we were not comparing response patterns to models that predict rank-ties between scenes[20].

Multidimensional scaling was conducted on Occluded and Non-Occluded V1 and V2, idealized Category and Depth models, and three computational models: Weibull[17,22], Gist[23], and H-MAX C2[24] were included. Each computational model was run on the non-occluded portion of scenes and separately on the occluded portion of scenes, and dissimilarity of model outputs was calculated using normalized Euclidean distance (see details on the dimensionality of outputs below). Next, 10,000 bootstrap samples were drawn from previously calculated data-splits for each pair-wise image comparison. Here, multidimensional scaling was performed on second-order dissimilarity matrices (Figure S1; 1- Kendall's Tau-a was used because Category and Depth models predicted rank-ties) with non-metric scaling to two dimensions (Stress normalized by inter-point sum of squares). Each configuration was aligned using a Procrustes transformation without scaling, and we used kernel density estimation (KDE) for each cortical area or model to determine its distribution in the representational space. Confidence areas were defined as contours that contained 95% of the KDE mass.

Weibull Model

The Weibull image contrast model measures the distribution of contrast values for an image and seeks to emulate the X and Y cells in the Lateral Geniculate. It has a two-dimensional output. We used the Weibull image contrast model outlined in Groen et al., (2012; 2013) with the field of view for estimation of the beta and gamma parameters set to 1.5 and 5 degrees, respectively.

Gist Model

The Gist algorithm[23] measures the distribution of oriented bandpass Gabor filter responses in localized portions of images. Our model consisted of 16 locations (4 x 4 grid), 8 orientations, and 4 spatial frequencies. This model had a 512 dimensional output.

H-MAX Model

The H-MAX model is a hierarchical model which gradually combines visual features of higher complexity. Here, we used the output of its fourth sequential stage, C2. The first two stages (S1 and C1) correspond to the simple and complex cells or early visual cortex. Stages S2 and C2 use the same pooling mechanisms as S1 and C1, but pool from the C1 stage and respond most strongly to a particular prototype input pattern. Prototypes were learned from a database of natural images outside of this study[24]. The output of this model had 2000 dimensions.

Retinotopic projections

Using each voxel's two-dimensional Gaussian response, as estimated in our pRF analysis, we projected activity patterns into visual space. The projected stimuli consisted of the sum of the 2D-Gaussians of all voxels in a given visual ROI, weighted by each voxel's response, as calculated in our GLM analyses. The stimuli were plotted as heat maps, with red and blue colors indicating activations above and below the mean response in each voxel, respectively. Each subject's data were calculated individually, and then averaged to obtain a group projection.

References:

1 Phillips, W. A., Clark, A. & Silverstein, S. M. On the functions, mechanisms, and malfunctions of intracortical contextual modulation. *Neurosci Biobehav Rev* **52**, 1-20, doi:10.1016/j.neubiorev.2015.02.010 (2015).

2 Gilbert, C. D. & Li, W. Top-down influences on visual processing. *Nat Rev Neurosci* **14**, 350-363, doi:10.1038/nrn3476 (2013).

3 Budd, J. M. Extrastriate feedback to primary visual cortex in primates: a quantitative analysis of connectivity. *Proc Biol Sci* **265**, 1037-1044, doi:10.1098/rspb.1998.0396 (1998).

4 Muckli, L. & Petro, L. S. Network interactions: non-geniculate input to V1. *Current opinion in neurobiology* **23**, 195-201, doi:10.1016/j.conb.2013.01.020 (2013).

5 Petro, L. S., Smith, F. W., Schyns, P. G. & Muckli, L. Decoding face categories in diagnostic subregions of primary visual cortex. *The European journal of neuroscience* **37**, 1130-1139, doi:10.1111/ejn.12129 (2013).

6 Petro, L. S., Vizioli, L. & Muckli, L. Contributions of cortical feedback to sensory processing in primary visual cortex. *Front Psychol* **5**, 1223, doi:10.3389/fpsyg.2014.01223 (2014).

7 Williams, M. A. *et al.* Feedback of visual object information to foveal retinotopic cortex. *Nat Neurosci* **11**, 1439-1445, doi:10.1038/nn.2218 (2008).

8 Smith, F. W. & Muckli, L. Nonstimulated early visual areas carry information about surrounding context. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 20099-20103, doi:10.1073/pnas.1000233107 (2010).

9 Muckli, L., Kohler, A., Kriegeskorte, N. & Singer, W. Primary visual cortex activity along the apparent-motion trace reflects illusory perception. *PLoS biology* **3**, e265, doi:10.1371/journal.pbio.0030265 (2005).

10 Vetter, P., Smith, F. W. & Muckli, L. Decoding sound and imagery content in early visual cortex. *Current biology : CB* **24**, 1256-1262, doi:10.1016/j.cub.2014.04.020 (2014).

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

12 Walther, D. B., Caddigan, E., Fei-Fei, L. & Beck, D. M. Natural scene categories revealed in distributed patterns of activity in the human brain. *J Neurosci* **29**, 10573-10581, doi:10.1523/JNEUROSCI.0559-09.2009 (2009).

13 Kravitz, D. J., Peng, C. S. & Baker, C. I. Real-world scene representations in high-level visual cortex: it's the spaces more than the places. *J Neurosci* **31**, 7322-7333, doi:10.1523/JNEUROSCI.4588-10.2011 (2011).

14 Dumoulin, S. O. & Wandell, B. A. Population receptive field estimates in human visual cortex. *Neuroimage* **39**, 647-660, doi:10.1016/j.neuroimage.2007.09.034 (2008).

15 Thirion, B. *et al.* Inverse retinotopy: inferring the visual content of images from brain activation patterns. *Neuroimage* **33**, 1104-1116, doi:10.1016/j.neuroimage.2006.06.062 (2006).

16 Kay, K. N., Winawer, J., Rokem, A., Mezer, A. & Wandell, B. A. A two-stage cascade model of BOLD responses in human visual cortex. *PLoS Comput Biol* **9**, e1003079, doi:10.1371/journal.pcbi.1003079 (2013).

17 Groen, II, Ghebreab, S., Prins, H., Lamme, V. A. & Scholte, H. S. From image statistics to scene gist: evoked neural activity reveals transition from low-level natural image structure to scene category. *J Neurosci* **33**, 18814-18824, doi:10.1523/JNEUROSCI.3128-13.2013 (2013).

18 Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P. & Movshon, J. A. A functional and perceptual signature of the second visual area in primates. *Nat Neurosci* **16**, 974-981, doi:10.1038/nn.3402 (2013).

19 Kriegeskorte, N., Mur, M. & Bandettini, P. Representational similarity analysis - connecting the branches of systems neuroscience. *Front Syst Neurosci* **2**, 4, doi:10.3389/neuro.06.004.2008 (2008).

20 Nili, H. *et al.* A toolbox for representational similarity analysis. *PLoS Comput Biol* **10**, e1003553, doi:10.1371/journal.pcbi.1003553 (2014).

21 Scholte, H. S., Ghebreab, S., Waldorp, L., Smeulders, A. W. M. & Lamme, V. a. F. Brain responses strongly correlate with Weibull image statistics when processing natural images. *Journal of Vision* **9**, 1-15, doi:10.1167/9.4.29.Introduction (2009).

22 Groen, II, Ghebreab, S., Lamme, V. A. & Scholte, H. S. Spatially pooled contrast responses predict neural and perceptual similarity of naturalistic image categories. *PLoS Comput Biol* **8**, e1002726, doi:10.1371/journal.pcbi.1002726 (2012).

23 Oliva, A. & Torralba, A. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *Int. J. Comput. Vision* **42**, 145-175, doi:10.1023/a:1011139631724 (2001).

24 Serre, T., Oliva, A. & Poggio, T. A feedforward architecture accounts for rapid categorization. *Proc Natl Acad Sci U S A* **104**, 6424-6429, doi:10.1073/pnas.0700622104 (2007).

25 Lee, T. S. & Mumford, D. Hierarchical Bayesian inference in the visual cortex. *J Opt Soc Am A Opt Image Sci Vis* **20**, 1434-1448 (2003).

26 Lee, T. S., Mumford, D., Romero, R. & Lamme, V. A. The role of the primary visual cortex in higher level vision. *Vision Res* **38**, 2429-2454 (1998).

27 Mumford, D. On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biol Cybern* **66**, 241-251 (1992).

28 Muckli, L. *et al.* Contextual Feedback to Superficial Layers of V1. *Curr Biol* **25**, 2690-2695, doi:10.1016/j.cub.2015.08.057 (2015).

29 Alink, A., Schwiedrzik, C. M., Kohler, A., Singer, W. & Muckli, L. Stimulus predictability reduces responses in primary visual cortex. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **30**, 2960-2966, doi:10.1523/JNEUROSCI.3730-10.2010 (2010).

30 Sugita, Y. Grouping of image fragments in primary visual cortex. *Nature* **401**, 269-272, doi:10.1038/45785 (1999).

31 Schyns, P. G. & Oliva, A. From Blobs to Boundary Edges - Evidence for Time-Scale-Dependent and Spatial-Scale-Dependent Scene Recognition. *Psychol Sci* **5**, 195-200, doi:DOI 10.1111/j.1467-9280.1994.tb00500.x (1994).

32 Oliva, A. & Torralba, A. Building the gist of a scene: the role of global image features in recognition. *Prog Brain Res* **155**, 23-36, doi:10.1016/S0079-6123(06)55002-2 (2006).

33 Torralba, A. & Oliva, A. Depth estimation from image structure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **24**, 1226-1238, doi:10.1109/TPAMI.2002.1033214 (2002).

34 Kennedy, H. & Bullier, J. A double-labeling investigation of the afferent connectivity to cortical areas V1 and V2 of the macaque monkey. *J Neurosci* **5**, 2815-2830 (1985).

35 Markov, N. T. *et al.* A weighted and directed interareal connectivity matrix for macaque cerebral cortex. *Cereb Cortex* **24**, 17-36, doi:10.1093/cercor/bhs270 (2014).

36 Sincich, L. C. & Horton, J. C. The circuitry of V1 and V2: integration of color, form, and motion. *Annu Rev Neurosci* **28**, 303-326, doi:10.1146/annurev.neuro.28.061604.135731 (2005).

37 Bastos, A. M. *et al.* Canonical Microcircuits for Predictive Coding. *Neuron* **76**, 695-711, doi:DOI 10.1016/j.neuron.2012.10.038 (2012).

38 LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436-444, doi:10.1038/nature14539 (2015).

39 Reichert, D. P. & Serre, T. Neuronal Synchrony in Complex-Valued Deep Networks. *arXiv preprint arXiv:1312.6115* (2013).

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

Author Contributions:

A.T.M., L.S.P., and L.M. designed the experiment; A.T.M. and L.S.P. performed the experiment; A.T.M. performed analyses; A.T.M., L.S.P., and L.M. wrote the paper.

Acknowledgments:

This work was supported by the European Research Council (ERC StG 2012_311751-BrainReadFBPredCode, awarded to L.M.). A.T.M. is supported by the College of Science & Engineering, University of Glasgow. We would like to thank Nikolaus Kriegeskorte and Luca Vizioli for useful discussions regarding Representational Similarity Analysis, Fraser Smith for discussion of SVM classification analyses, Steven Scholte for discussion and code for the Weibull model, and James Haxby for helpful discussion regarding Multidimensional Scaling. We would also like to thank Grace Edwards, Fiona McGruer, Matthew Bennett and Yulia Revina for useful comments and discussion on the manuscript.

Supplemental Information:

Table S1:

	Individual Scenes	Category	Depth
Occluded V1	12.70 [9.64 16.65]	12.91 [10.09 15.58]	8.33 [6.27 11.13]
Occluded V2	5.36 [4.10 6.68]	7.44 [5.92 8.81]	3.19 [1.93 4.38]
Non-Occluded V1	47.13 [38.02 57.27]	37.93 [31.86 43.66]	37.93 [31.86 43.66]
Non-Occluded V2	35.98 [28.19 43.95]	32.01 [27.41 36.85]	32.01 [27.41 36.85]

Table S1A. Related to Figure 2. Classification results, reported as percent above chance-level (chance-level = 4.17%, 16.67%, and 50% for individual scenes, category and depth, respectively). 1000 bootstrap samples of mean performance were drawn from individual subject results, and 95% confidence intervals are shown in brackets.

	Category	Depth
Occluded V1	1.99 [0.49 3.36]	1.04 [-0.42 2.66]
Occluded V2	2.11 [0.93 3.14]	-1.13 [-2.21 0.11]
Non-Occluded V1	12.72 [10.45 14.91]	3.33 [1.08 5.39]
Non-Occluded V2	10.66 [8.62 12.84]	4.67 [2.17 7.10]

Table S1B. Related to Figure 2. Cross-classification results, reported as percent above chance-level (chance-level = 16.67% for category and 50% for depth). For analysis of category, 18 (of 24) scenes were selected, leaving out one scene per category. For depth, 22 (of 24) were selected, leaving out one scene per depth. The classifier was tested on the remaining scenes in a cross-classification approach. Due to the large number of possible image permutations in these analyses, we randomly assigned scenes to training and testing sets 100 times in each subject. 1000 bootstrap samples of mean performance were drawn from individual subject results, and 95% confidence intervals are shown in brackets.

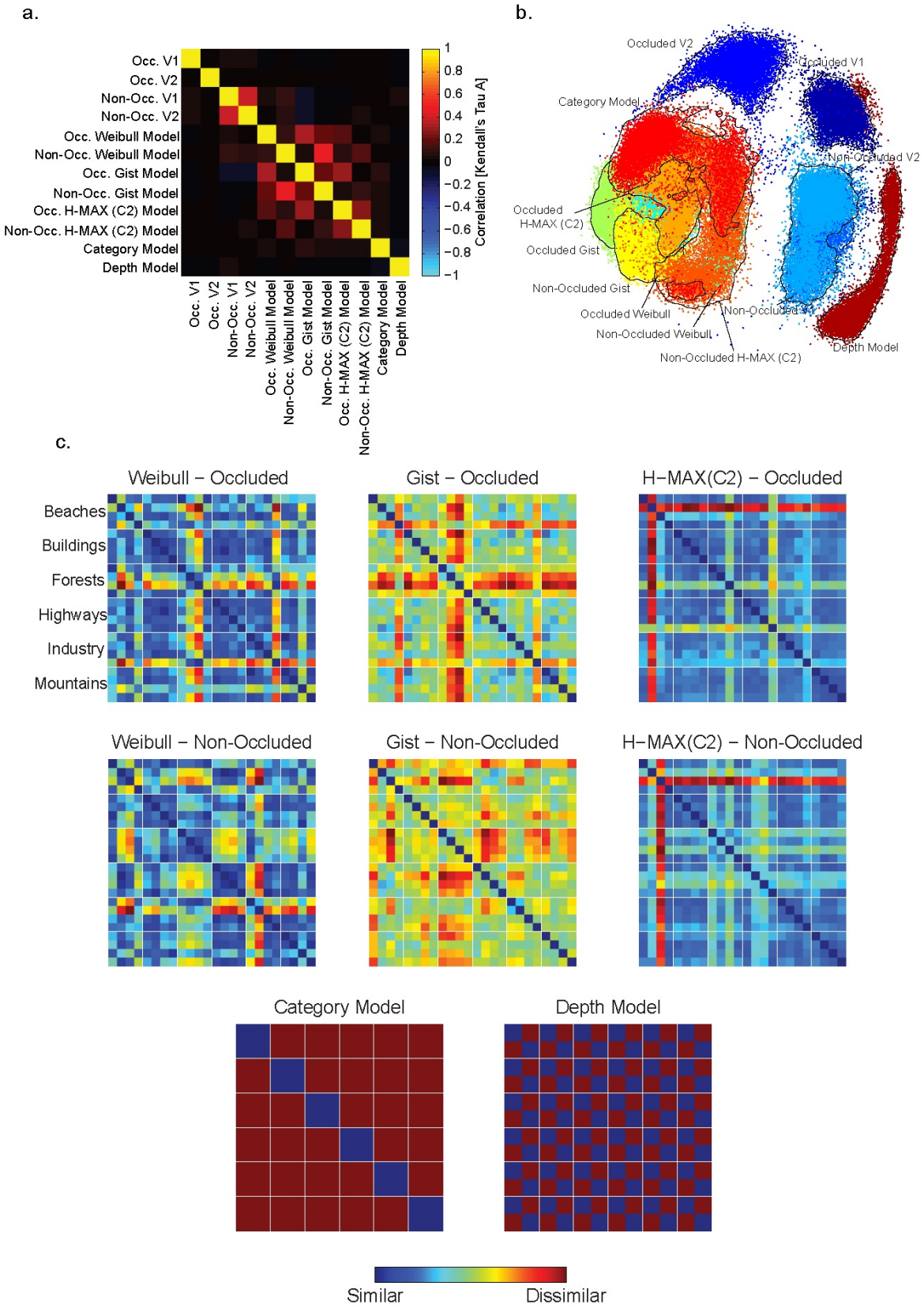

Fig. S1. Related to Fig. 5. Second-order relationships of cortical data, computational models, and idealized models. (A) One example of the matrix form of cortical area and model comparisons (distance is measured as

1- Kendall's Tau A). Individual points in (B) are cortical area or model representational structures, whose distances from each other are determined by multidimensional scaling of 10,000 bootstrap resamples. The black outlines in (B) correspond directly to Figure 5, representing 95% of the data for each condition. (C) First-order representational dissimilarity matrices (standardized Euclidean distance) for computational and idealized models.

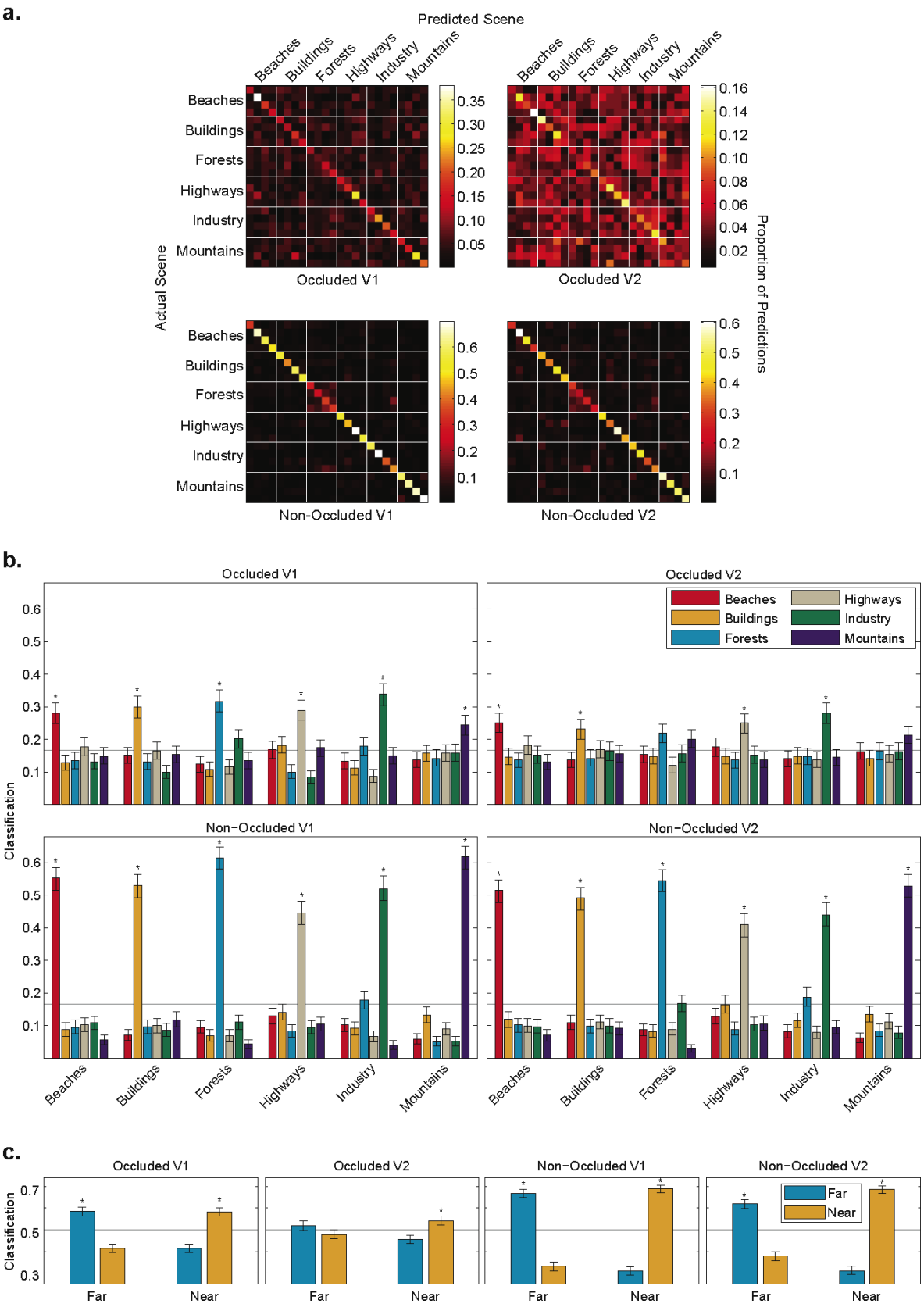

Figure S2. Related to Figure 2. Classification confusion for (A) individual scenes (B) category- and (C) depth-based analyses. (A) Classification confusion matrices for individual scene analyses. Actual scenes are

presented on the y-axis (within each category, the first two scenes are 'Far' and the last two are 'Near'); predicted scenes are presented on the x-axis. Correct predictions are displayed as the diagonal. Note that an ideal confusion matrix would have 1 everywhere on the diagonal (correct predictions) and 0 in all off-diagonal points (errors). (B) Actual categories (ground truth) are presented along the x-axes, while the proportion of total classifier predictions are displayed on the y-axes. Perfect classification would have a ground truth proportion of 1, and 0 for all other categories or distances. Chance-levels are displayed as grey lines. For each proportion, 95% confidence intervals are presented via 1000 bootstrapped samples. Asterisks indicate classifier predictions of true conditions being significantly greater than other conditions, $p < 0.05$. (C) Identical to (B), but corresponding to depth-based classification.

Discussion of Figure S2.

To better understand feedback properties, we examined classification errors from our initial classification analysis. In individual scene classification, correct scene predictions were significantly more common than misclassifications in Occluded V1, as indicated by the strong diagonal in its confusion matrix (Figure S2a). This was not the case with many scenes in Occluded V2. Also, relatively few misclassifications occurred in non-occluded areas, indicated by the strong diagonals in these matrices. We hypothesized that individual scene representations within category or depth groups should be more difficult to differentiate than those between categories or depths. Accordingly, we expected misclassifications to occur more frequently within categories or depths. To test for this bias, we calculated differences between individual scene misclassifications based on being within or between category or depth groups. We found that misclassifications were not biased by either grouping in non-stimulated areas, and were only biased for category in some Forest scenes in stimulated areas (Forest scenes 1, 2 and 4 in Non-occluded V1, and Forests 2, 3 and 4 in Non-occluded V2, $p < 0.05$). These results suggest that scene-specific properties, which are prominent in representations of individual images in early visual cortex during natural vision, are an important component of representations under visual occlusion as well.

Figure S2b shows the classifier predictions for each condition in category-based analyses. Predictions in all six categories significantly outnumber misclassifications in three of our examined regions, while in Occluded V2, Forest and Mountain categories suffer from greater numbers of classification errors. In addition, distinct misclassification patterns appear in each region. In Occluded V1, misclassification of Forest scenes as Industry scenes is significantly more common than other classification errors, while the opposite is not true. This situation is reversed in Non-Occluded V1, with Industry scenes being mistaken for Forest scenes, but not vice versa. In Non-Occluded V2, both of these confusions are more common than other errors, but neither error is present in Occluded V2 analyses. When examining depth-based classification (Figure S2c), proportions of correct predictions to errors do not differ between depths in V1 analyses. In both V2 regions, classifiers misclassify scenes of greater depth more often than those of lesser depth, and in Occluded V2, Far scenes do not classify above chance-level.