Unraveling Representations in Scene-selective Brain Regions Using Scene-Parsing Deep Neural Networks

Kshitij Dwivedi[1,2], Radoslaw Martin Cichy[1*], and Gemma Roig[2*]

Abstract

■ Visual scene perception is mediated by a set of cortical regions that respond preferentially to images of scenes, including the occipital place area (OPA) and parahippocampal place area (PPA). However, the differential contribution of OPA and PPA to scene perception remains an open research question. In this study, we take a deep neural network (DNN)-based computational approach to investigate the differences in OPA and PPA function. In a first step, we search for a computational model that predicts fMRI responses to scenes in OPA and PPA well. We find that DNNs trained to predict scene components (e.g., wall, ceiling, floor) explain higher variance uniquely in OPA and PPA than a DNN trained to predict scene category (e.g., bathroom, kitchen, office). This result is robust across several DNN architectures. On this basis, we then determine whether particular scene components predicted by DNNs differentially account for unique variance in OPA and PPA. We find that variance in OPA responses uniquely explained by the navigation-related floor component is higher compared to the variance explained by the wall and ceiling components. In contrast, PPA responses are better explained by the combination of wall and floor, that is, scene components that together contain the structure and texture of the scene. This differential sensitivity to scene components suggests differential functions of OPA and PPA in scene processing. Moreover, our results further highlight the potential of the proposed computational approach as a general tool in the investigation of the neural basis of human scene perception. ■

INTRODUCTION

Visual scene understanding is a fundamental cognitive ability that enables humans to interact with the components and objects present within the scene. Within the blink of an eye (Greene & Oliva, 2009; Fei-Fei, Iyer, Koch, & Perona, 2007; Thorpe, Fize, & Marlot, 1996; Potter, 1975), we know what type of scene we are in (e.g., kitchen or outdoors) as well as its spatial layout and the objects contained in it.

Research on the neural basis of scene understanding has revealed a set of cortical regions with a preferential response to images of scenes over images of objects. These regions are the parahippocampal place area (PPA; Epstein & Kanwisher, 1998), occipital place area (OPA; Dilks, Julian, Paunov, & Kanwisher, 2013; Hasson, Harel, Levy, & Malach, 2003), and retrosplenial cortex (O'Craven & Kanwisher, 2000). To investigate the distinct function each of these place regions has, subsequent research has begun to tease apart their commonalities and differences in activation profile and representational content (Bonner & Epstein, 2017; Silson, Chan, Reynolds, Kravitz, & Baker, 2015; Dilks et al., 2013; Hasson et al., 2003; O'Craven & Kanwisher, 2000; Epstein & Kanwisher, 1998). However, a complete picture of how scene-selective regions together orchestrate visual scene understanding is still missing.

To gain further insights, a promising but relatively less explored approach is computational modeling of brain activity. Recently, large advances have been made in modeling activity in visual cortex using deep neural networks (DNNs) trained on object categorization tasks (Krizhevsky, Sutskever, & Hinton, 2012) in both human and nonhuman primates (Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014). Inspired by this success, researchers have also begun to use DNNs trained on scene categorization to investigate scene-selective cortex (Bonner & Epstein, 2018; Groen et al., 2018; Cichy, Khosla, Pantazis, & Oliva, 2017).

In this process, two issues have emerged that need to be addressed. First, although DNNs trained on categorization tasks currently do best in predicting activity in scene-selective cortical regions, they do not account for all explainable variance. One particularly promising direction is the exploration of models trained on tasks different from categorization that might more closely resemble the brain region's functionality and thus predict brain activity better (Cichy & Kaiser, 2019; Yamins et al., 2014). Second, it remains unclear what is the nature of the representations in the DNNs that gives them their predictive power. Thus, additional effort is needed to clarify what these representations are.

To address the above issues, we investigated neural activity in the scene-selective cortex using DNNs trained on scene parsing instead of categorization. A scene parsing

This article is part of a Special Focus entitled, Human and Machine Cognition, presented at the 2019 annual meeting of the Cognitive Neuroscience Society Meeting.

[1]Freie Universität Berlin, [2]Goethe University Frankfurt
*Jointly directed work.

task requires the DNN to predict the location and category of each scene component in the image. Whereas the scene categorization task requires only recognizing the scene category, the scene parsing task requires deeper scene understanding involving categorization as well as a grasp of the spatial organization of components and objects within the scene. To help interact with different objects and navigate within the scene, scene-selective brain regions should also encode the spatial organization of components within the scene. Therefore, we hypothesize that the scene parsing task is closer to the task the brain has to solve, and a DNN trained on scene parsing will predict brain activity better than a DNN trained on scene categorization.

To evaluate our hypothesis, we compared the power of DNNs trained on scene parsing versus categorization to predict activity in scene-selective cortical regions. For this, we used an existing fMRI data set of brain responses elicited by viewing scene images (Bonner & Epstein, 2017) and applied representational similarity analysis (RSA) to compare brain responses with DNNs. We found that scene-parsing DNNs explain significantly more variance in brain responses uniquely in scene-selective regions than scene-classification DNNs.

We next investigated what representations in the DNNs trained on scene parsing gave the model its predictive power. For this, we queried the DNN's representations of different scene components, considering components that were present in all stimulus images: wall, floor, and ceiling. We showed that different scene components predict responses in OPA and PPA differently: Floor explained more variance in OPA than wall and ceiling, whereas wall explained more variance in PPA than floor and ceiling. Importantly, results were consistent across three different DNN architectures, showing the generalizability of our claims across architectures.

In summary, our results reveal differential representational content in scene-selective regions OPA and PPA and highlight DNNs trained on scene parsing as a promising model class for modeling human visual cortex with well-interpretable output.

METHODS

fMRI Data

We used fMRI data from a previously published study by Bonner and Epstein (2017) where all experimental details can be found, as well as instructions on how to download the data. The fMRI data were collected from 16 participants on a Siemens 3.0-T PRISMA scanner with a 64-channel head coil. The participants were presented with images of indoor environments. The images were presented for 1.5 sec on the screen followed by a 2.5-sec ISI. The images presented in the experiment were from a stimulus set of 50 color images depicting indoor environments. During the fMRI scan, participants were asked to fixate on a cross all the time and press a button if the image presented to them was a bathroom. The task required participants to attend to each image and categorize it. Voxel-wise (voxel size = $2 \times 2 \times 2$ mm) responses to each image during each scan run were extracted using a standard linear model.

We here focus on two scene-selective ROIs: PPA and OPA. PPA and OPA were identified from separate functional localizer scans using a contrast of brain responses to scenes larger than to objects and additional anatomical constraints. For both ROIs and all the subjects, each voxel's responses in a given ROI were z-scored across images in a given run and then averaged across runs. The responses to a particular image were further z-scored across voxels.

Behavioral Data

We used scene-related behavioral data representing navigational affordances assessed on the same stimulus set as used for recording the fMRI data described above (Bonner & Epstein, 2017). To represent navigational affordances, a behavioral experiment was conducted in which 11 participants (different from the participants in the fMRI experiment) indicated the path to walk through each image of the indoor environment used in the fMRI study using a computer mouse. The probabilistic maps of paths for each image were created, followed by a histogram construction of navigational probability in 1° angular bins radiating from the bottom center of the image. These histograms represent a probabilistic map of potential navigation routes from the viewer's perspective. The resultant histogram is referred to as the Navigational Affordance Model (NAM).

DNN Models

We selected DNNs optimized on two different scene-related tasks: scene classification and scene parsing. We describe both types of models in detail below.

Scene Classification Models

For solving a scene classification task, a DNN model is optimized to predict the probabilities of the input image belonging to a particular scene category. For comparison with neural and behavioral data, we considered DNNs pretrained on the scene classification task on the Places-365 data set (Zhou, Lapedriza, Khosla, Oliva, & Torralba, 2017). Places-365 is a large-scale scene classification data set consisting of 1.8 million training images from 365 scene categories. We selected multiple scene-classification DNN architectures to investigate if our results generalize across different architectures. For this purpose, we considered three standard architectures, namely, Alexnet (Krizhevsky et al., 2012), Resnet18 (He, Zhang, Ren, & Sun, 2016), and Resnet50 (He et al., 2016), and downloaded pretrained models from github.com/CSAILVision/places365.

Alexnet consists of five convolutional layers (conv1–conv5) followed by three fully connected layers (fc6, fc7, and fc8). Both Resnet18 and Resnet50 consist of a convolutional layer followed by four residual blocks (block1–block4), each consisting of several convolutional layers

with skip connections leading to a final classification layer (fc). Resnet18 consists of 18 layers, and Resnet50 consists of 50 layers in total; they differ in the number of layers within each block.

Scene Parsing Models

We used scene parsing models trained on ADE20k scene parsing data set (Zhou et al., 2019). The ADE20k data set (publicly available at groups.csail.mit.edu/vision/datasets/ADE20K/) is a densely annotated data set consisting of 25k images of complex everyday scenes with pixel-level annotations of objects (chair, bed, bag, lamp, etc.) and components (wall, floor, sky, ceiling, water, etc.). The images were annotated using the LabelMe interface (Russell, Torralba, Murphy, & Freeman, 2008) by a single expert human annotator.

For the first set of experiments, where we compare the predictive power of scene parsing models to scene classification models for explaining the neuronal responses, we design scene parsing models such that their encoder architecture is taken from scene classification models while their decoder architecture is task specific. The encoder of the scene parsing models consists of the convolutional part (conv1–conv5 of Alexnet; block1–block4 of Resnet18 and Resnet50) of scene classification models. The decoder of scene parsing models is adapted to the scene parsing task following the architecture proposed by Zhao, Shi, Qi, Wang, and Jia (2017). It consists of a Pyramid Pooling module with deep supervision (Zhao et al., 2017; d1), followed by a layer (d2) that predicts several spatial maps, one spatial map per scene component predicted, that represent the probability of the presence of that component at a given spatial location. The encoder weights of scene parsing models are initialized with the weights learned on the scene classification task, and decoder weights are initialized randomly. The scene-parsing DNNs are then trained on ADE20k training data using a per-pixel cross-entropy loss, which measures the performance of the classifier at each pixel whether the correct component is assigned the highest probability or not. The above procedure ensures that gain/drop in explaining neural/behavioral responses could only be because of additional supervision on the scene parsing task.

The aforementioned scene-parsing DNNs are well suited for a direct comparison with the scene categorization DNNs as they have the same encoder architecture and were initialized with weights learned on the scene categorization task. However, they are not comparable to state-of-the-art models in terms of accuracy on the scene parsing task. Because our aim is to reveal differences in representations of the scene areas in the brain by comparing scene components, for the second set of experiments, it is crucial to select components detected with DNNs from the literature that achieve the highest accuracy in scene parsing. For this reason, we selected three state-of-the-art models on the scene parsing task, namely, Resnet101-PPM (Zhou et al., 2019), Upernet101 (Xiao, Liu, Zhou, Jiang, & Sun, 2018), and HR-Netv2 (Sun et al., 2019). All the three state-of-the-art models were trained on the ADE20k data set. We selected multiple models to investigate if the results we obtain are consistent across different models. Resnet101-PPM consists of a dilated version of the Resnet101 model (a deeper version of Resnet50 that consists of 101 layers) trained on Imagenet as the encoder and a Pyramid Pooling module with deep supervision (Zhao et al., 2017) as the decoder. Because of the small receptive field in the feature maps, scene-parsing DNNs fail to correctly segment larger objects/components. The Pyramid Pooling module (Zhao et al., 2017) tackles this issue by fusing the feature maps that have different receptive field sizes to merge high-spatial-resolution information with low-spatial-resolution information for a better local- and global-level scene understanding. Upernet101 (Xiao et al., 2018) is based on the feature pyramid network by Lin et al. (2017) that uses multilevel feature representations via a top–down architecture to fuse high-level semantic features with mid- and low-level lateral connections. Upernet101 also has a Pyramid Pooling Module before the top–down architecture to overcome the small receptive-field issue. HRNetv2 (Sun et al., 2019) relies on the importance of high-resolution feature maps for pixel labeling maps by maintaining high-resolution feature representations throughout the architecture and by merging information from both high- and low-resolution convolutions in parallel and overcomes the small receptive field issue mentioned above. We downloaded all abovementioned models from github.com/CSAILVision/semantic-segmentation-pytorch.

To reveal performance differences between different models on the scene parsing task, we compared the performance of state-of-the-art models with the scene parsing models used for comparison with scene-classification DNNs (see above: Alexnet, Resnet18, and Resnet50). For the comparison, we calculated the mean intersection-over-union (mIoU) score of detecting all components for all the images from the ADE20k validation data set. The intersection-over-union (IoU) score is calculated by dividing the intersection between a predicted and corresponding ground-truth pixel-level segmentation mask by their union. IoU is a standard metric to evaluate the overlap of a predicted and corresponding pixel-level mask of a particular component. Mean IoU is calculated by taking the mean of IoU scores across all images in the validation data set for all components.

As illustrated in Figure 1A, a scene parsing model decomposes an image into its constituent components. This decomposition allows investigating which scene components are more relevant to explaining the representations in scene-selective brain regions. We first identified which scene components are present in all the images from the stimulus set used for obtaining fMRI responses. To achieve this, we feedforwarded all the 50 images in the stimulus set of the fMRI data set through the models and checked the

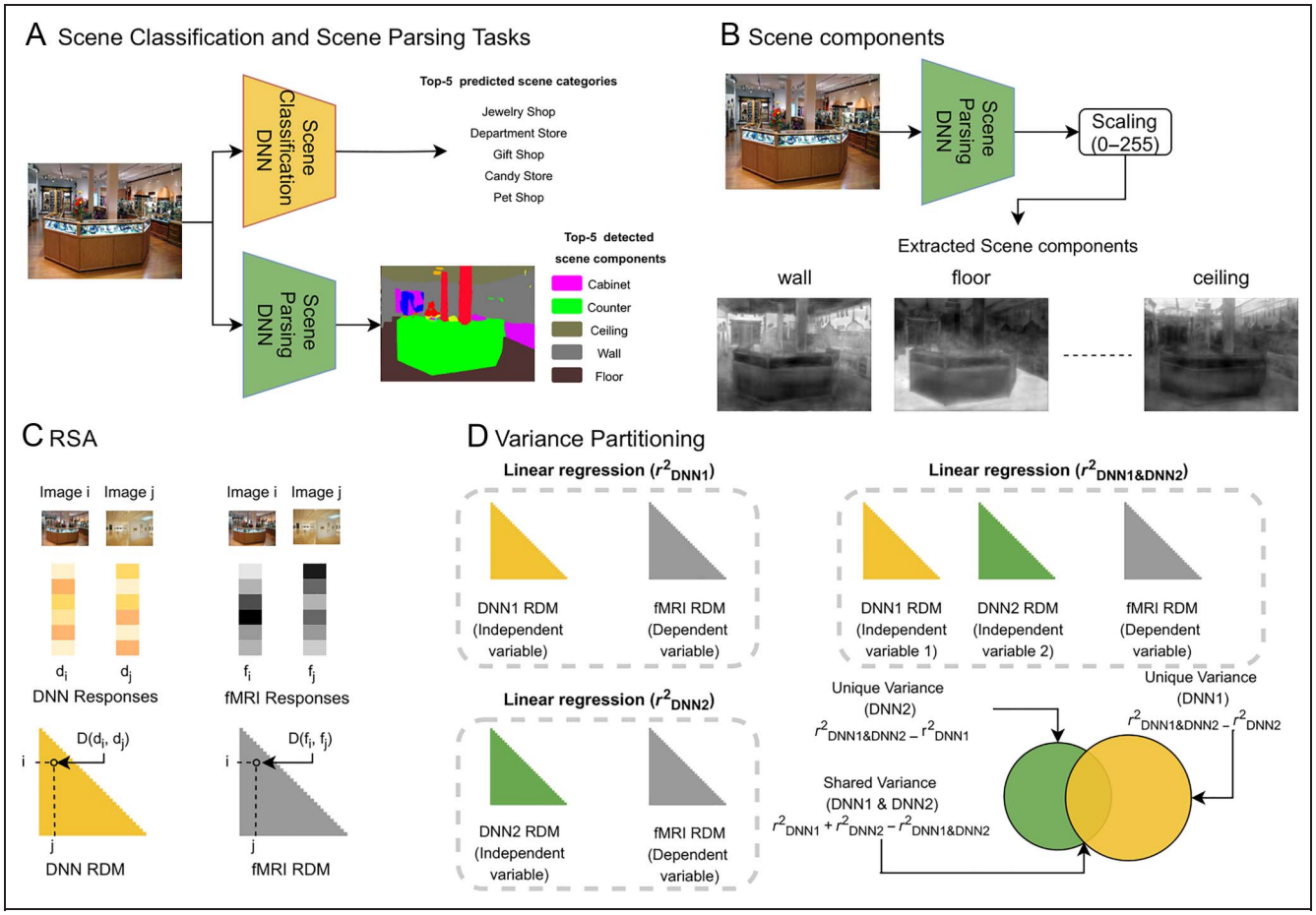

Figure 1. Outline of our approach. (A) In the scene classification task, the model outputs the probability of an image belonging to a particular class. In the scene parsing task, the model outputs a spatial map for each component. The pixel value of the spatial map corresponding to a component represents the probability of that pixel belonging to that component. (B) We use DNNs trained on scene parsing to extract responses corresponding to individual scene components. (C) RSA: We first compute RDMs for a DNN model and a brain ROI by computing pairwise distance (D) between DNN (d_i, d_j)/fMRI(f_i, f_j) responses corresponding to each pair (i, j) of images in the stimulus set. We next compute the correlation of a DNN RDM with fMRI RDM to determine the similarity between the brain and the DNN. (D) Variance partitioning: We conduct three multiple linear regressions with DNN RDMs as the independent variables and fMRI RDM as the dependent variable to estimate unique and shared variance of fMRI RDM explained by DNN RDMs.

presence of all the components in the image. Because the DNN has been trained on an image data set that is different from the set of stimuli used for the fMRI data, not all scene components predicted by the DNN appear in the stimulus set. In this particular set, we found that wall, floor, and ceiling were core scene components present in all images.

A scene parsing model outputs a spatial probability map for each component. To scale the spatial probability maps corresponding to different components in the same range, we normalized the spatial probability map for each component independently such that each pixel value lies in the range [0, 255]. We show the extracted normalized scene components corresponding to the wall, floor, and ceiling components for an example stimulus in Figure 1B.

RSA

We applied RSA (Kriegeskorte, Mur, & Bandettini, 2008) to compare DNN activations and scene components with neural and behavioral responses. RSA enables relating signals from different source spaces (such as here behavior, neural responses, and DNN activations) by abstracting signals from separate source spaces into a common similarity space. For this, in each source space, condition-specific responses are compared to each other for dissimilarity (e.g., by calculating Euclidean distances between signals), and the values are aggregated in so-called representational dissimilarity matrices (RDMs) indexed in rows and columns by the conditions compared. RDMs thus summarize the representational geometry of the source space signals. Different from source space signals themselves, RDMs from different sources spaces are directly comparable to each other for similarity and thus can relate signals from different spaces. We describe the construction of RDMs for different modalities and the procedure by which they were compared in detail below.

fMRI ROI RDMs

First, for each ROI (OPA and PPA), individual participant RDMs were constructed using Euclidean distances between

the voxel response patterns for all pairwise comparisons of images. Then, subject-averaged RDMs were constructed by calculating the mean across all individual subject RDMs. We downloaded the participant-averaged RDMs of OPA and PPA from the link (figshare.com/s/5ff0a04c2872e1e1f416) provided in Bonner and Epstein (2018).

NAM RDMs

NAM RDMs were constructed using Euclidean distances between the navigational affordance histograms for all pairwise comparisons of images. We downloaded the NAM RDM from figshare.com/s/5ff0a04c2872e1e1f416.

DNN RDMs

For all the DNNs we investigated in this work, we constructed the RDM for a particular layer using $1 - \rho$, where ρ is the Pearson's correlation coefficient, as the distance between layer activations for all pairwise comparisons of images. For scene-classification DNN RDMs, we created one RDM for each of the five convolutional layers (conv1–conv5) and for the three fully connected layers (fc6, fc7, and fc8) for Alexnet as well as the last layer of each block (block1–block4) and the final classification layer (fc) of Resnet18/Resnet50 to compare with neural/behavioral RDMs. For scene-parsing DNN RDMs, we created one RDM for each of the five convolutional layers (conv1–conv5) and for the two decoder layers (d1 and d2) for Alexnet as well as the last layer of each block (block1–block4) and two decoder layers (d1 and d2) of Resnet18/Resnet50 to compare with neural/behavioral RDMs.

Scene Component RDMs

For each of the scene components investigated, we constructed RDM for it using $1 - \rho$ as the distance between normalized spatial probability maps of that scene component, based on all pairwise comparisons of images.

Comparing DNN and Scene Component RDMs with Behavioral and Neural RDMs

In this work, we pose two questions: first, whether scene parsing models can better explain scene-selective neural responses and navigational affordance behavioral responses better than scene classification models and, second, whether the scene components detected by scene parsing models reveal differences in representations of scene-selective ROIs.

To investigate the first question, we calculated the Spearman's correlation between the RDMs of different layers of a scene-classification DNN with a particular behavioral/neural RDM and selected the layer RDM that showed the highest correlation with the behavioral/neural RDM. We used the selected layer RDM as the representative RDM for that architecture. We repeated the same procedure to select the representative RDM from a scene parsing model. As a baseline for comparison, we also considered a randomly initialized model and selected the representative RDM from it.

We first found that deeper layers of both scene-classification and scene-parsing models showed a higher correlation with neural responses than earlier layers. A possible explanation behind the observed trend could be that deeper layers are more task-specific whereas early layers learn low-level visual features irrespective of tasks and therefore do not represent task-specific information in the model. Moreover, the highest correlation with PPA and OPA was found in deeper layers of the network, further supporting this idea. We report the layer used to select the representative RDMs for each model in Table 1. To compare which model RDM (scene parsing/scene classification/random) explains behavioral/neural RDM better, we compared the correlation values of all three RDMs with behavioral/neural RDM (illustrated in Figure 1C for scene classification vs. scene parsing).

To investigate whether the scene components detected by scene parsing models reveal differences in representations of scene-selective ROIs, we computed the correlation

Table 1. Correlation Value and Layer Information of the Layer That Showed the Highest Correlation with a Particular Brain Area or Behavior for All the Models Considered (AlexNet, ResNet18, and ResNet50)

Task	*Models*	*OPA*	*PPA*	*NAM*
Scene classification	Alexnet	.395 (fc7)	.369 (fc7)	.122 (conv5)
	Resnet18	.391 (fc)	.397 (fc)	.114 (block4)
	Resnet50	.364 (fc)	.365 (fc)	.111 (block4)
Scene parsing	Alexnet	.393 (d2)	.393 (d1)	.162 (conv5)
	Resnet18	.426 (d1)	.415 (d1)	.198 (block4)
	Resnet50	.415 (d1)	.418 (d1)	.165 (d2)
Random	Alexnet	.169 (fc7)	.167 (fc6)	.015 (fc7)
	Resnet18	.194 (fc)	.176 (fc)	.041 (block4)
	Resnet50	.152 (fc)	.142 (fc)	.042 (block4)

between a scene component RDM and a neural RDM and compared which scene component explains better a particular ROI.

Variance Partitioning

Although, in its basic formulation, RSA provides insights about the degree of association between a DNN RDM and a behavioral/neural RDM, it does not provide a full picture of how multiple DNN RDMs together explain the behavioral/neural RDM. Therefore, we applied a variance partitioning analysis that determines the unique and shared contribution of individual DNN RDMs in explaining the behavioral/neural RDM when considered in conjunction with the other DNN RDMs. Furthermore, variance partitioning allows selection of multiple layers from a single model to explain the variance in neural and behavioral RDMs.

We illustrate the variance partitioning analysis in Figure 1D. We assigned a behavioral/neural RDM as the dependent variable (referred to as predictand). We then assigned two model (DNN/scene component) RDMs as the independent variables (referred to as predictors). Then, we performed three multiple regression analyses: one with both independent variables as predictors and two with individual independent variables as the predictors. Then, by comparing the explained variance (r^2) of a model used alone with the explained variance when it was used with other models, the amount of unique and shared variance between different predictors can be inferred (Figure 1D). In the case of three independent variables, we performed seven multiple regression analyses: one with all three independent variables as predictors, three with different combinations of two independent variables as predictors, and three with individual independent variables as the predictors.

To compare scene parsing and scene classification models with a randomly initialized model as the baseline, the predictors were the respective DNN RDMs and predictands were the behavioral and neural RDMs. We performed variance partitioning analysis first using the selected representative RDMs (Table 1) for each model using RSA. In a second analysis, we relax the criteria of representing a model by a single layer RDM and use multiple layer RDMs together to represent the model. We selected all the layer RDMs from each model (scene classification/scene parsing/random) and used them as predictors for variance partitioning.

To compare different scene components, the predictors were the respective scene component RDMs and predictands were the neural RDMs of scene-selective ROIs and behavioral RDM.

Statistical Testing

We applied nonparametric statistical tests to assess the statistical significance in a similar manner to a previous related study (Bonner & Epstein, 2018). We assessed the significance of the correlation between neural/behavioral responses with a DNN through a permutation test by permuting the conditions randomly 5000 times in either the neural ROI RDM or the DNN RDM.

From the distribution obtained using these permutations, we calculated p values as one-sided percentiles. We calculated the standard errors of these correlations by randomly resampling the conditions in the RDMs for 5000 iterations. We used resampling without replacement by subsampling 90% (45 of 50 conditions) of the conditions in the RDMs. We used an equivalent procedure for testing the statistical significance of the correlation difference and unique variance difference between the two models. The statistical outcomes were corrected for multiple comparisons using false discovery rate correction with a threshold equal to 0.05.

RESULTS

Are Scene Parsing Models Suitable to Account for Scene-selective Brain Responses and Scene-related Behavior?

We investigated the potential of DNNs trained on scene parsing to predict scene-related human brain activity focusing the analysis on scene-selective regions OPA and PPA. To put the result into context, we compared the predictive power of DNNs trained on scene parsing to DNNs trained on scene classification, which are currently the default choice in investigating scene-related brain responses and behavior (Bonner & Epstein, 2018; Groen et al., 2018; Cichy et al., 2017) and against a randomly initialized DNN as baseline. To ensure that the results can be attributed to differences in the task rather than being specific to a particular network architecture, we investigated three different network architectures: Alexnet, Resnet18, and Resnet50.

We applied RSA to relate DNN models (scene classification, scene parsing, and random) with the brain responses in OPA and PPA (Figure 2A). We found that DNNs trained on scene parsing significantly predicted brain activity in all investigated regions ($p = .0001$ for Alexnet, $p = .0001$ for Resnet18, and $p = .0001$ for Resnet50). This shows that they are suitable candidate models for the investigation of brain function. We further found that DNNs trained on scene parsing explain as much or more variance in scene-selective regions than DNNs trained on scene categorization. We note both scene parsing and scene-classification DNNs explain significantly higher variance in scene-selective regions than a randomly initialized DNN across different architectures ($p < .001$ for all the comparisons).

If scene parsing models are suitable models for predicting responses in scene-selective brain regions, and these regions underlie scene understanding, the models should predict scene-related behavior, too. We considered navigational affordance behavior operationalized as

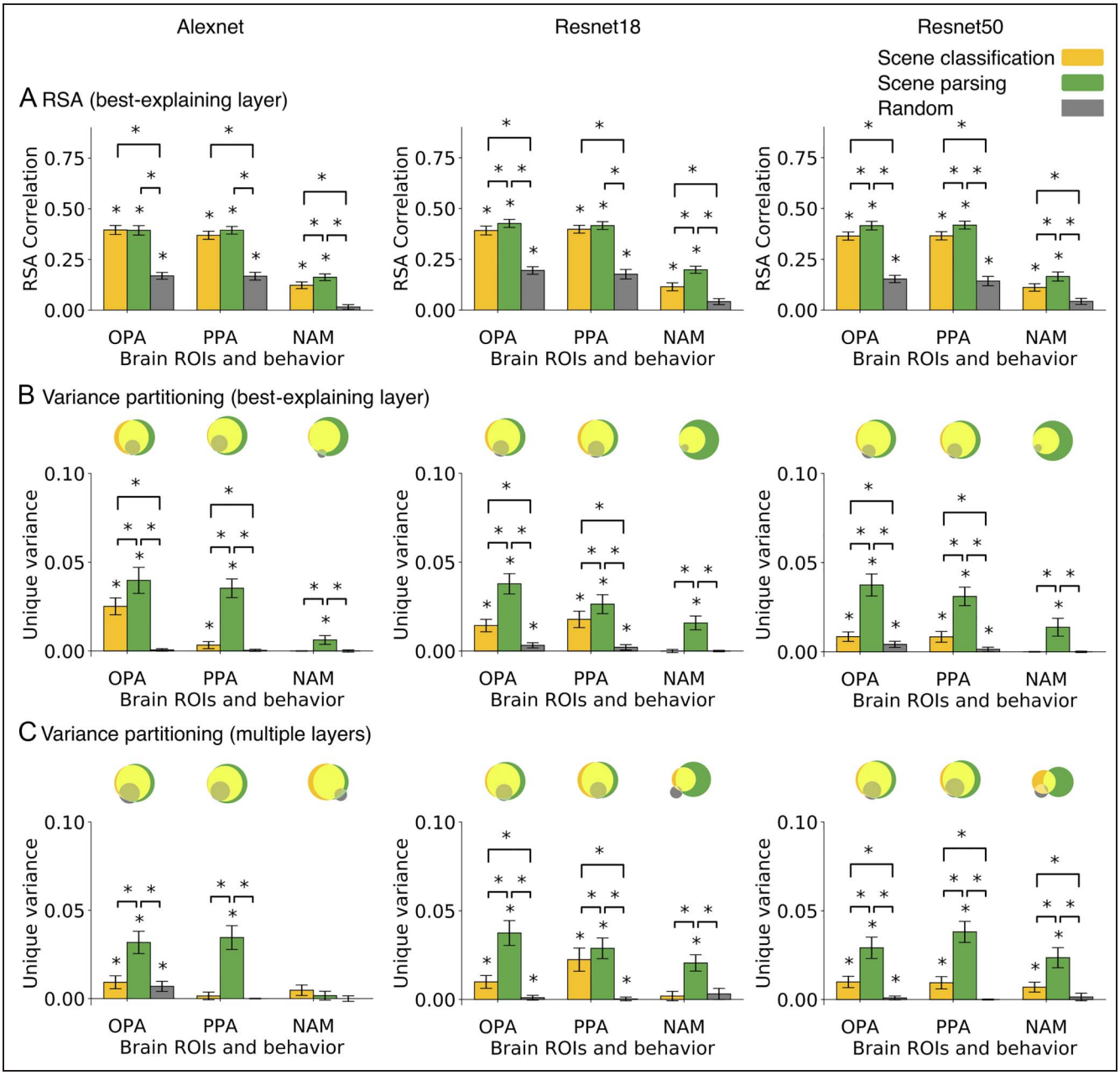

Figure 2. Model comparison in accounting for OPA and PPA as well as behavior. (A) RSA of scene-selective areas PPA and OPA as well as behavioral model NAM with scene-parsing, scene-classification, and random models (best-explaining layer) and (B) variance of scene-selective areas PPA and OPA as well as behavioral model NAM explained uniquely by scene-parsing, scene-classification, and random models (best-explaining layer) for the architecture Alexnet (left), Resnet18 (center), and Resnet50 (right). Venn diagram on top of each bar plot illustrates the unique and shared variance of ROIs and behavior explained by multiple models together. (C) Variance of scene-selective areas PPA and OPA as well as behavioral model NAM explained uniquely by scene-parsing, scene-classification, and random models (multiple layers) for the architecture Alexnet (left), Resnet18 (center), and Resnet50 (right). Venn diagram on top of each bar plot illustrates the unique and shared variance of ROIs and behavior explained by multiple models together. The asterisk at the top indicates the significance ($p < .05$) calculated by permuting the conditions 5000 times. * False discovery rate corrected with a threshold equal to 0.05.

the angular histogram of navigational trajectories that participants indicated for the stimulus set. Paralleling the results on brain function, the investigation of behavior showed that DNNs trained on scene parsing predicted behavior significantly ($p = .0005$ for Alexnet, $p = .0005$ for Resnet18, and $p = .0005$ for Resnet50) and also significantly better than DNNs trained on scene classification ($p = .01$ for Alexnet, $p = .0005$ for Resnet18, and $p = .03$ for Resnet50). Similar to results on brain function, we note that both scene-parsing and scene-classification DNNs explain significantly higher variance in behavior than a randomly initialized DNN across different architectures.

Although the RSA results above provided insights about the degree of association between a DNN RDM and a behavioral/neural RDM, it cannot tell how multiple DNN RDMs together predict the behavioral/neural RDM. For this more complete picture, we conducted variance partitioning to reveal the unique variance of neural/behavioral RDMs

explained by scene-classification and scene-parsing DNN RDMs (Figure 2B). We observe from Figure 2B that scene-parsing DNNs explain more variance uniquely (OPA: p = .0003 for Alexnet, p = .0002 for Resnet18, and p = .0002 for Resnet50; PPA: p = .0002 for Alexnet, p = .0003 for Resnet18, and p = .0002 for Resnet50) than scene-classification DNNs for both scene-selective ROIs. We further observe from Venn diagrams in Figure 2B that, for scene-selective neural RDMs, most of the variance explained is shared between scene-classification and scene-parsing DNNs across all three architectures. The results suggest that scene-parsing DNNs might be a better choice for investigating scene-selective neural responses than scene-classification DNNs.

We observe for behavior that the scene-classification DNNs explain nearly no unique variance, whereas on the other hand, scene-parsing DNNs explain significantly higher unique variance (p = .001 for Alexnet, p = .0002 for Resnet18, and p = .0005 for Resnet50) across all three architectures (see Figure 2B for unique variance and Venn diagrams illustrating both unique and shared variances). The results suggest that, because the scene parsing task takes into account the spatial arrangement of constituent components in the scene, a scene-parsing DNN explains behavioral affordance assessments better than a scene-classification DNN.

In both the above analyses, we selected the RDM of the layer of a model that showed the highest RSA correlation with a neural/behavioral RDM as the representative RDM for the model, but this begs the question how well a particular layer represents a model as a whole. To answer this question, we selected multiple layer RDMs from each model (scene classification/scene parsing/random) and applied variance partitioning to find out how much of the variance in neural/behavior RDMs is explained uniquely by a given model represented by multiple layer RDMs. We report the results in Figure 2C and observe a similar trend as observed in Figure 2B using layer RDMs that showed maximum correlation. The results suggest that layers that showed maximum correlation with a neural/behavior RDM were deeper in the network and therefore have more task-specific representation as opposed to earlier layers of the network. The variance explained by earlier layers contributes mostly to the shared variance of neural/behavioral RDMs explained by all models together.

Together, these results establish DNNs trained on scene parsing tasks as a promising model class for investigating scene-selective cortical regions in the human brain and for navigational behavior related to the spatial organization of scene components.

State-of-the-art Scene Parsing Models for Investigating Scene Components Represented in the Human Brain

Models trained on the scene parsing task offer the possibility to selectively investigate which of the scene components (such as wall, ceiling, or floor) they encode. However, what is the most suitable scene parsing model to compare to the brain? In the model comparison above, our choice was guided by making models as similar to each other as possible in complexity to rule out that observed differences in accounting for brain activity are simply because of differences in model complexity. However, for an in-depth investigation of scene-selective areas using scene components, it is crucial to choose models that detect the scene components with as high accuracy as possible. Therefore, we compared the performance of different scene parsing models qualitatively and quantitatively to select the most accurate ones to compare with the responses of scene-selective brain areas.

For performance comparison on the scene parsing task, we chose the three models used above (Alexnet, Resnet18, and Resnet50), plus three state-of-the-art models of scene parsing: HRNetv2, Upernet101, and Resnet101-PPM. The state-of-the-art models achieve high performance by merging low-resolution feature maps with high-resolution feature maps to generate results in high spatial resolution.

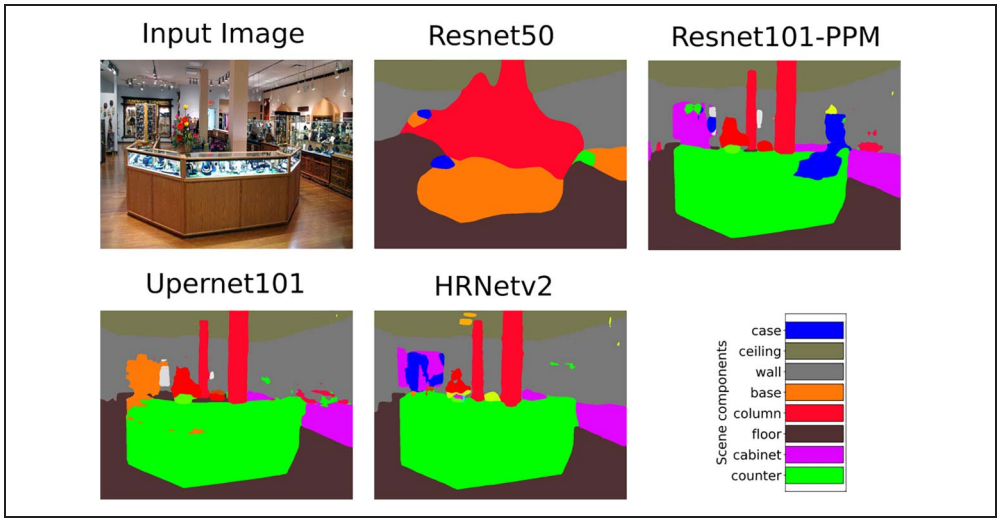

Figure 3. Qualitative comparison of scene parsing output for different models. Input image (top left) and corresponding scene parsing output of the different models investigated in this work.

Table 2. Scene Parsing Performance on ADE20k Validation Set

Model	*Wall*	*Ceiling*	*Floor*	*Overall Accuracy*
HRNetv2	0.7538	0.8278	0.7811	0.4320
Upernet101	0.7503	0.8265	0.7772	0.4276
Resnet101-PPM	0.7453	0.8195	0.7659	0.4257
Resnet50	0.6422	0.7356	0.6642	0.3020
Resnet18	0.6223	0.6997	0.6627	0.2741
AlexNet	0.5857	0.6810	0.6105	0.2306

The table shows the accuracy of detecting selected components along with overall accuracy for different scene parsing models in decreasing order.

We illustrate their parsing performance qualitatively by examining their output on an example image (Figure 3). We observe that the scene parsing output generated by Resnet50 had smooth and less precise boundaries of components whereas Resnet101-PPM, Upernet101, and HRNetv2 detected components accurately with precise boundaries in their outputs.

To quantitatively compare model performance, we evaluated the performance of all models on the ADE20k validation data set. For mIoU score of detecting all components for all the images from the ADE20k validation data set, we report mIoU scores of individual components that were present in all images in the stimulus set: wall, ceiling, and floor. The results are reported in Table 2. They indicate that state-of-the-art models beat the complexity-matched models by a margin of 12% accuracy. Therefore, for an in-depth investigation of representations in scene-selective brain areas, we used the top three models, namely, HRNetv2, Upernet101, and Resnet101-PPM.

Scene Parsing Networks Reveal a Differential Contribution of Wall, Floor, and Ceiling Components to Representations in Scene-selective Regions

We investigated whether the scene components detected by a scene-parsing DNN reveal a difference in the representational content of scene-selective ROIs. We focused on the three scene components—wall, floor, and ceiling—that were present in all the images of the stimulus set and compared them with scene-selective ROIs OPA and PPA and behavioral model NAM.

We first report the RSA results (Figure 4A) of comparing a scene component RDM with OPA, PPA, and NAM for three state-of-the-art architectures HRNetv2, Upernet101, and Resnet101-PPM. We found that the correlation of the OPA RDM with the floor RDM was significantly higher than that of the wall (p = .02 for HRNetv2, p = .05 for Upernet101, p = .0002 for Resnet101-PPM) and ceiling (p = .01 for HRNetv2, p = .01 for Upernet101, p = .0002 for Resnet101-PPM) RDMs, and the correlation of the PPA RDM with the wall (p = .006 for HRNetv2, p = .001 for Upernet101, p = .01 for Resnet101-PPM) and floor (p = .006 for HRNetv2, p = .003 for Upernet101, p = .001 for Resnet101-PPM) RDMs was significantly higher than that with the ceiling RDM. NAM, which represents the navigational paths in the scenes, showed the highest correlation with the floor RDM, which was significantly higher than the correlation with the wall (p = .0002 for HRNetv2, p = .0002 for Upernet101, p = .0002 for Resnet101-PPM) and ceiling (p = .0002 for HRNetv2, p = .0002 for Upernet101, p = .0002 for Resnet101-PPM) RDMs. The above results held consistently across all investigated models. Together, this suggests that OPA and PPA

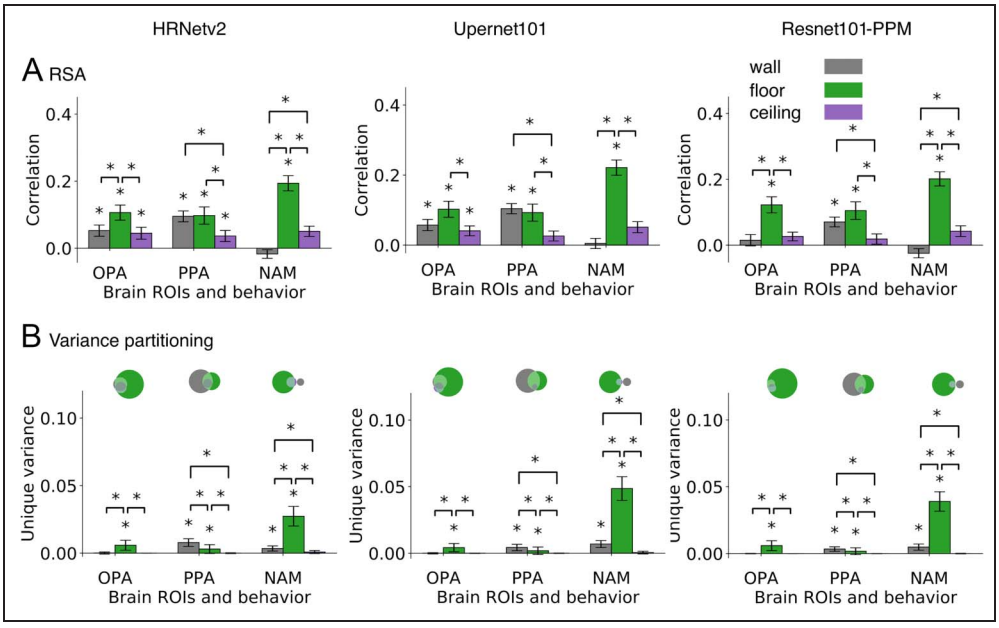

Figure 4. Scene components reveal the differences in representational content of OPA, PPA, and behavioral model NAM. (A) Results of RSA for OPA, PPA, and NAM with scene components of wall floor and ceiling for three state-of-the-art models on scene parsing task HRnetv2 (left), Upernet101 (center), and Resnet101 (right); (B) unique variance accounted for in OPA, PPA, and NAM by using components from the HRnetv2 (left), Upernet101 (center), and Resnet101 (right) models. Venn diagram on top of each bar plot illustrates the unique and shared variance of ROIs and behavior explained by multiple components together. The asterisk at the top indicates the significance (p < .05) calculated by permuting the conditions 5000 times, false discovery rate corrected with a threshold equal to 0.05.

have differential representational content with respect to scene components.

To tease out how much variance in OPA and PPA is explained by individual scene components, we apply variance partitioning to find the unique and shared variance of OPA and PPA RDMs explained by different scene component RDMs. We report the variance partitioning results showing unique variance explained by each component along with Venn diagrams illustrating both unique and shared variances in Figure 4B. We observed that, in the case of OPA, the floor RDM explains significantly higher variance of OPA RDM uniquely compared to wall (p = .0002 for HRNetv2, p = .002 for Upernet101, p = .0008 for Resnet101-PPM) and ceiling (p = .0002 for HRNetv2, p = .002 for Upernet101, p = .0008 for Resnet101-PPM) RDMs. For PPA, the wall RDM explains significantly higher variance of PPA RDM uniquely compared to the floor (p = .001 for HRNetv2, p = .006 for Upernet101, p = .015 for Resnet101-PPM) and ceiling (p = .0005 for HRNetv2, p = .002 for Upernet101, p = .007 for Resnet101-PPM) RDMs. Furthermore, for NAM, the floor RDM explains significantly higher variance as compared to the wall (p = .0002 for HRNetv2, p = .0002 for Upernet101, p = .0002 for Resnet101-PPM) and ceiling (p = .0002 for HRNetv2, p = .05 for Upernet101, p = .0002 for Resnet101-PPM) RDMs. Consistent with the RSA results above, this result reinforces the differences between OPA and PPA in the representation of scene components.

DISCUSSION

In this study, we investigated the potential of scene-parsing DNNs in predicting neural responses in scene-selective brain regions. We found that scene-parsing DNNs predicted responses in scene-selective ROIs OPA and PPA better than scene-classification DNNs. We further showed that scene components detected by scene-parsing DNNs revealed differences in representational content of OPA and PPA.

Previous work using DNNs to predict neural responses has emphasized the importance of the task for which the DNNs were optimized (Richards et al., 2019; Yamins & DiCarlo, 2016; Khaligh-Razavi & Kriegeskorte, 2014). We argue that the higher unique variance of scene-selective neural responses explained by scene-parsing DNNs over scene-classification DNNs is because of such a difference in tasks. The scene classification task aims at identifying the category of the scene irrespective of the spatial organization of different components and objects in the scene. In contrast, the scene parsing task requires pixelwise labeling of the whole image and thus a more comprehensive understanding of the scene in terms of how different objects and components are spatially organized within a given scene. Higher variance of the scene-selective neural responses explained by the scene-parsing DNNs that encode spatial structure suggests that scene-selective neural responses also encode spatial structure of the scene entailing the position of different objects and components. This view is supported further by evidence from neuroimaging literature (Kravitz, Peng, & Baker, 2011; Park, Brady, Greene, & Oliva, 2011) showing that scene-selective regions represent the spatial layout of scenes. The information about the spatial structure of the scene might be required by the brain to plan interaction within the scene, such as navigating to a target, reaching objects, or performing visual search.

Our in-depth analysis using scene components revealed differential representations in OPA and PPA. We observed that OPA had a significantly higher correlation with floor than ceiling and wall. A possible explanation for the observed difference could be because of OPA's involvement in detecting navigational affordances (Bonner & Epstein, 2017), for which the floor plays a major role, and could explain the high sensitivity of OPA to stimulation in the lower visual field (Silson et al., 2015). In contrast, we found that PPA shows a significantly higher correlation with wall as well as floor compared to ceiling. This could explain why PPA has sensitivity to the upper visual field (Silson et al., 2015). A plausible explanation could be that detecting the wall is relevant to identifying the type of room, its texture (Henriksson, Mur, & Kriegeskorte, 2019; Park & Park, 2017), and landmarks (Troiani, Stigliani, Smith, & Epstein, 2014).

Previous work has already aimed at determining the nature of OPA representations by computational modeling (Bonner & Epstein, 2018) on the same fMRI data set that was used in our study. For this, the authors determined for a DNN trained on scene categorization which individual DNN units most correlated with NAM and OPA and visualized those units using receptive field mapping and segmentation from Zhou, Khosla, Lapedriza, Oliva, and Torralba (2014). The units extracted corresponded mostly to uninterrupted portions of floor and wall or the junctions between floor and wall. The results align with our findings that floor components explain OPA responses uniquely whereas the wall units could be attributed to shared variance explained by floor and wall components in our study. However, arguably, segmentation maps extracted using the receptive field mapping method are less interpretable as they cannot be directly assigned to meaningful entities without additional human annotations (Zhou et al., 2014) or by comparing with ground-truth segmentation maps of meaningful entities (Bau, Zhou, Khosla, Oliva, & Torralba, 2017). Furthermore, assigning a segmentation map of a unit obtained using receptive field mapping to a meaningful entity using ground-truth segmentation maps leads to less accurate segmentation maps (Bau et al., 2017) compared to the segmentation from a scene-parsing DNN (Xiao et al., 2018) trained to segment components using ground-truth segmentation maps. Thus, we believe our approach using scene components generated by a scene-parsing DNN to be particularly well suited to reveal the representational content of scene-selective brain regions.

In our analysis investigating navigational affordance behavioral responses, we found that scene-parsing DNNs explained significantly higher variance of behavioral RDM than the scene-classification DNN. A plausible explanation for this finding is that determining navigational affordances requires spatial understanding of the scene, including floor and obstacle detection. Scene parsing tasks explicitly require such spatial understanding, whereas scene classification tasks can be performed without explicitly detecting the spatial organization of objects and components in the scene. Furthermore, the layers that show the highest correlation with NAM in scene classification models are later convolutional layers, which preserve spatial information of the image suggesting that spatial information is required to explain NAM. In the comparison of NAM with scene components, we found that NAM is best explained by the floor component, whereas other components (wall, ceiling) explained insignificant unique variance. The above finding further reinforces our argument that scene-parsing DNNs explain NAM responses better because of the task requirement of finding spatial layout of objects and components. Although, in this study, we focused on showing the advantage of scene-parsing DNNs over scene-classification DNNs in explaining scene-selective neural and navigational affordance-related behavioral responses, our results do not rule out scene-classification DNNs as useful models to explain semantic behavioral responses related to scene category or semantic similarity.

A limitation of our study is that the differences revealed between OPA and PPA are based on the analysis of only three scene components. This is because of the limitations of the stimulus set, which consistently had only three components that were present in all 50 images. Future work should exploit the full richness of scene components provided by DNNs trained on scene parsing. For this, a stimulus set would have to be designed that contains many components in all images of the stimulus set. Another possible direction would be to use stimuli with annotations and use these annotations directly to compare with the fMRI responses. The advantage of using a scene-parsing DNN over human annotations is that, once the DNN is trained on scene parsing, the components can be extracted for a new set of stimuli with zero cost as opposed to annotations, where human effort is required to annotate every new stimulus set.

It is crucial to point out that our findings could be influenced by the task participants performed while inside the scanner. The participants were required to identify whether the presented image was a bathroom or not. Most of the studies do not look into the influence of tasks in the fMRI studies, and the opinion on whether the difference in tasks results in different representations is divided. For instance, some studies (Woolgar, Hampshire, Thompson, & Duncan, 2011; Duncan, 2010) suggest that the parietal cortex and pFC are involved in representing task context whereas the scene processing is attributed to the occipito-temporal cortex. Recently, some studies (Hebart, Bankson, Harel, Baker, & Cichy, 2018; Bracci, Daniels, & Op de Beeck, 2017; Bugatus, Weiner, & Grill-Spector, 2017; Vaziri-Pashkam & Xu, 2017; Lowe, Gallivan, Ferber, & Cant, 2016; Erez & Duncan, 2015; Harel, Kravitz, & Baker, 2014) have shown evidence of task influence in the occipito-temporal cortex. Therefore, a promising future direction of research might be to find out whether our findings are replicated or not on another fMRI study where participants performed a different task.

To summarize, our findings provide evidence supporting the use of DNNs trained on the scene parsing task as a promising tool to predict and understand activity in the visual brain. We believe that this approach has the potential to be applied widely, providing interpretable results that give insights into how the human visual cortex represents the visual world.

Acknowledgments

We thank Agnessa Karapetian and Greta Häberle for their valuable comments on the paper. G. R. thanks the support of the Alfons and Gertrud Kassel Foundation. R. M. C. is supported by Deutsche Forschungsgemeinschaft grants (CI241/1-1, CI241/3-1) and the European Research Council Starting Grant (ERC-2018-StG 803370).

Reprint requests should be sent to Gemma Roig, Department of Computer Science, Goethe University Frankfurt, Robert-Mayer-Str. 11-15, Frankfurt am Main, Germany 60325, or via e-mail: roig@cs.uni-frankfurt.edu.

REFERENCES

Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6541–6549). Honolulu, HI: IEEE.

Bonner, M. F., & Epstein, R. A. (2017). Coding of navigational affordances in the human visual system. *Proceedings of the National Academy of Sciences, U.S.A.*, *114*, 4793–4798.

Bonner, M. F., & Epstein, R. A. (2018). Computational mechanisms underlying cortical responses to the affordance properties of visual scenes. *PLoS Computational Biology*, *14*, e1006111.

Bracci, S., Daniels, N., & Op de Beeck, H. (2017). Task-context overrules object- and category-related representational content in the human parietal cortex. *Cerebral Cortex*, *27*, 310–321.

Bugatus, L., Weiner, K. S., & Grill-Spector, K. (2017). Task alters category representations in prefrontal but not high-level visual cortex. *Neuroimage*, *155*, 437–449.

Cichy, R. M., & Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in Cognitive Sciences*, *23*, 305–317.

Cichy, R. M., Khosla, A., Pantazis, D., & Oliva, A. (2017). Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. *Neuroimage*, *153*, 346–358.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*, 27755.

Dilks, D. D., Julian, J. B., Paunov, A. M., & Kanwisher, N. (2013). The occipital place area is causally and selectively involved in scene perception. *Journal of Neuroscience*, *33*, 1331–1336.

Duncan, J. (2010). The multiple-demand (MD) system of the primate brain: Mental programs for intelligent behaviour. *Trends in Cognitive Sciences*, *14*, 172–179.
Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, *392*, 598–601.
Erez, Y., & Duncan, J. (2015). Discrimination of visual categories based on behavioral relevance in widespread regions of frontoparietal cortex. *Journal of Neuroscience*, *35*, 12383–12393.
Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision*, *7*, 10.
Greene, M. R., & Oliva, A. (2009). The briefest of glances: The time course of natural scene understanding. *Psychological Science*, *20*, 464–472.
Groen, I. I. A., Greene, M. R., Baldassano, C., Fei-Fei, L., Beck, D. M., & Baker, C. I. (2018). Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *eLife*, *7*, e32962.
Harel, A., Kravitz, D. J., & Baker, C. I. (2014). Task context impacts visual object processing differentially across the cortex. *Proceedings of the National Academy of Sciences, U.S.A.*, *111*, E962–E971.
Hasson, U., Harel, M., Levy, I., & Malach, R. (2003). Large-scale mirror-symmetry organization of human occipito-temporal object areas. *Neuron*, *37*, 1027–1041.
He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778). Las Vegas, NV: IEEE.
Hebart, M. N., Bankson, B. B., Harel, A., Baker, C. I., & Cichy, R. M. (2018). The representational dynamics of task and object processing in humans. *eLife*, *7*, e32816.
Henriksson, L., Mur, M., & Kriegeskorte, N. (2019). Rapid invariant encoding of scene layout in human OPA. *Neuron*, *103*, 161–171.
Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, *10*, e1003915.
Kravitz, D. J., Peng, C. S., & Baker, C. I. (2011). Real-world scene representations in high-level visual cortex: It's the spaces more than the places. *Journal of Neuroscience*, *31*, 7322–7333.
Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4.
Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*, 1097–1105.
Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2117–2125). Honolulu, HI: IEEE.
Lowe, M. X., Gallivan, J. P., Ferber, S., & Cant, J. S. (2016). Feature diagnosticity and task context shape activity in human scene-selective cortex. *Neuroimage*, *125*, 681–692.
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.
Park, J., & Park, S. (2017). Conjoint representation of texture ensemble and location in the parahippocampal place area. *Journal of Neurophysiology*, *117*, 1595–1607.
Park, S., Brady, T. F., Greene, M. R., & Oliva, A. (2011). Disentangling scene content from spatial boundary: Complementary roles for the parahippocampal place area and lateral occipital complex in representing real-world scenes. *Journal of Neuroscience*, *31*, 1333–1340.
Potter, M. C. (1975). Meaning in visual search. *Science*, *187*, 965–966.
Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., et al. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, *22*, 1761–1770.
Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). LabelMe: A database and web-based tool for image annotation. *International Journal of Computer Vision*, *77*, 157–173.
Silson, E. H., Chan, A. W.-Y., Reynolds, R. C., Kravitz, D. J., & Baker, C. I. (2015). A retinotopic basis for the division of high-level scene processing between lateral and ventral human occipitotemporal cortex. *Journal of Neuroscience*, *35*, 11921–11935.
Sun, K., Zhao, Y., Jiang, B., Cheng, T., Xiao, B., Liu, D., et al. (2019). High-resolution representations for labeling pixels and regions. arXiv preprint arXiv:1904.04514.
Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*, 520–522.
Troiani, V., Stigliani, A., Smith, M. E., & Epstein, R. A. (2014). Multiple object properties drive scene-selective regions. *Cerebral Cortex*, *24*, 883–897.
Vaziri-Pashkam, M., & Xu, Y. (2017). Goal-directed visual processing differentially impacts human ventral and dorsal visual representations. *Journal of Neuroscience*, *37*, 8767–8782.
Woolgar, A., Hampshire, A., Thompson, R., & Duncan, J. (2011). Adaptive coding of task-relevant information in human frontoparietal cortex. *Journal of Neuroscience*, *31*, 14592–14599.
Xiao, T., Liu, Y., Zhou, B., Jiang, Y., & Sun, J. (2018). Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision* (pp. 418–434). Munich, Germany: Springer.
Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*, 356–365.
Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences, U.S.A.*, *111*, 8619–8624.
Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2881–2890). Honolulu, HI: IEEE.
Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2014). Object detectors emerge in deep scene CNNs. arXiv preprint arXiv:1412.6856..
Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017). Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*, 1452–1464.
Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., et al. (2019). Semantic understanding of scenes through the ADE20K dataset. *International Journal of Computer Vision*, *127*, 302–321.