

1 Title: **Dynamics of scene representations in the human brain revealed by**
2 **magnetoencephalography and deep neural networks**

3

4 Radoslaw Martin Cichy¹, Aditya Khosla¹, Dimitrios Pantazis², Aude Oliva¹

5

6 ¹ Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA, USA

7

8 ² McGovern Institute for Brain Research, MIT, Cambridge, MA, USA

9

10

11

12

13

14 **CORRESPONDING AUTHOR**

15 Radoslaw Martin Cichy

16 Computer Science and Artificial Intelligence Laboratory

17 MIT

18 32-D430

19 Cambridge, MA, USA

20 Phone: +1 617 253 1428

21 Email: rmcichy@mit.edu

22

23 **ABSTRACT**

24

25 Human scene recognition is a rapid multistep process evolving over time from single
26 scene image to spatial layout processing. We used multivariate pattern analyses on
27 magnetoencephalography (MEG) data to unravel the time course of this cortical process.
28 Following an early signal for lower-level visual analysis of single scenes at ~100ms, we
29 found a marker of real-world scene size, i.e. spatial layout processing, at ~250ms
30 indexing neural representations robust to changes in unrelated scene properties and
31 viewing conditions. For a quantitative explanation that captures the complexity of scene
32 recognition, we compared MEG data to a deep neural network model trained on scene
33 classification. Representations of scene size emerged intrinsically in the model, and
34 resolved emerging neural scene size representation. Together our data provide a first
35 description of an electrophysiological signal for layout processing in humans, and a novel
36 quantitative model of how spatial layout representations may emerge in the human brain.

37

38

39

40

41

42 **KEY WORDS**

43

44 Scene perception, spatial layout, magnetoencephalography, deep neural network,
45 representational similarity analysis

46

47 **1 INTRODUCTION**

48 Perceiving the geometry of space is a core ability shared by all animals, with brain
49 structures for spatial layout perception and navigation preserved across rodents, monkeys
50 and humans (Epstein and Kanwisher, 1998, 1998; Doeller et al., 2008, 2010; Moser et al.,
51 2008; Epstein, 2011; Jacobs et al., 2013; Kornblith et al., 2013, 2013; Vaziri et al., 2014).
52 Spatial layout perception, the demarcation of the boundaries and size of real-world visual
53 space, plays a crucial mediating role in spatial cognition (Bird et al., 2010; Epstein, 2011;
54 Kravitz et al., 2011a; Wolbers et al., 2011a; Park et al., 2014) between image-specific
55 processing of individual scenes and navigation-related processing. Although the cortical
56 loci of spatial layout perception in humans have been well described (Aguirre et al.,
57 1998; Kravitz et al., 2011b; MacEvoy and Epstein, 2011; Mullally and Maguire, 2011;
58 Park et al., 2011; Bonnici et al., 2012), the dynamics of spatial cognition remain
59 unexplained, partly because neuronal markers indexing spatial processing remain
60 unknown.

61

62 Operationalizing spatial layout as scene size, that is the size of the space a scene subtends
63 in the real-world (Kravitz et al., 2011a; Park et al., 2011, 2014), we report here an
64 electrophysiological signal of spatial layout perception in the human brain. Using
65 multivariate pattern classification (Carlson et al., 2013; Cichy et al., 2014; Isik et al.,
66 2014) and representational similarity analysis (Kriegeskorte, 2008; Kriegeskorte and
67 Kievit, 2013; Cichy et al., 2014) on millisecond-resolved magnetoencephalography data
68 (MEG), we identified a marker of scene size around 250ms, preceded by and distinct
69 from an early signal for lower-level visual analysis of scene images at ~100ms.

70 Furthermore, we demonstrated that the scene size marker was independent of both low-
71 level image features (i.e. luminance, contrast, clutter) and semantic properties (the
72 category of the scene, i.e. kitchen, ballroom), thus indexing neural representations robust
73 to changes in viewing conditions as encountered in real-world settings.

74

75 To provide a quantitative explanation how space size representations emerge in cortical
76 circuits, we compared brain data to a deep neural network model trained to perform scene
77 categorization (Zhou et al., 2014, 2015), termed deep scene network. The deep scene
78 network *intrinsically* exhibited receptive fields specialized for layout analysis, such as
79 textures and surface layout information, without ever having been explicitly taught any of
80 those features. We showed that the deep scene neural network model predicted the human
81 neural representation of single scenes and scene space size better than a deep object
82 model and standard models of scene and object perception (Riesenhuber and Poggio,
83 1999; Oliva and Torralba, 2001). This demonstrates the ability of the deep scene model to
84 approximate human neural representations at successive levels of processing as they
85 emerge over time.

86

87 Together our findings provide a first description of an electrophysiological signal for
88 scene space processing in humans, and offer a novel quantitative and computational
89 model of the dynamics of visual scene space representation in the cortex. Our results
90 suggest that spatial layout representations naturally emerge in cortical circuits learning to
91 differentiate visual environments (Oliva and Torralba, 2001).

92 2 MATERIALS AND METHODS

93 2.1 Participants

94 Participants were 15 right-handed, healthy volunteers with normal or corrected-to-normal
95 vision (mean age \pm s.d. = 25.87 \pm 5.38 years, 11 female). The Committee on the Use of
96 Humans as Experimental Subjects (COUHES) at MIT approved the experiment and each
97 participant gave written informed consent for participation in the study, for data analysis
98 and publication of study results.

99 2.2 Stimulus material and experimental design

100 The image set consisted of 48 scene images differing in four factors with two levels each,
101 namely two scene properties: physical size (small, large) and clutter level (low, high);
102 and two image properties: contrast (low, high) and luminance (low, high) (Figure 1A).
103 There were 3 unique images for every level combination, for example 3 images of small
104 size, low clutter, low contrast and low luminance. The image set was based on
105 behaviorally validated images of scenes differing in size and clutter level, sub-sampling
106 the two highest and lowest levels of factors size and clutter (Park et al., 2014). Small
107 scenes were of size that would typically fit 2-8 people, whereas large scenes would fit
108 hundreds to thousands. Similarly, low clutter level scenes were empty or nearly empty
109 rooms, whereas high clutter scenes contained multiple objects throughout. The contrast
110 and luminance was adjusted to specific values for each image: images of low and high
111 contrast had root mean square values of 34% and 50% respectively; images of low and
112 high luminance had root mean square values of 34% and 51% respectively.

113

114 Participants viewed a series of scene images while MEG data was recorded (Figure 1B).
115 Images subtended 8° of visual angle in both width and height and were presented
116 centrally on a gray screen (42.5% luminance) for 0.5s in random order with an inter-
117 stimulus interval (ISI) of 1-1.2s, overlaid with a central red fixation cross. Every 4 trials
118 on average (range 3-5 trials, equally probable) a target image depicting concentric circles
119 was presented prompting participants to press a button and blink their eyes in response.
120 ISI between the concentric-circles and the next trial was 2s to allow time for eye blinks.
121 Target image trials were not included in analysis. Each participant completed 15 runs of
122 312s each. Every image was presented four times in a run, resulting in 60 trials per image
123 per participant in total.

124 **2.3 MEG recording**

125 We recorded continuous MEG signals from 306 channels (Elektra Neuromag TRIUX,
126 Elekta, Stockholm) at a sampling rate of 1000Hz. Raw data was band-pass filtered
127 between 0.03 and 330Hz, and pre-processed using spatiotemporal filters (maxfilter
128 software, Elekta, Stockholm). We used Brainstorm (Tadel et al., 2011) to extract peri-
129 stimulus MEG signals from -100 to +900ms with respect to stimulus onset, and then
130 normalized each channel by its baseline (-100 to 0ms) mean and standard deviation, and
131 temporally smoothed the time series with a 20ms sliding window.

132 **2.4 Multivariate pattern classification of MEG data**

133 *Single image classification:* To determine whether MEG signals can discriminate
134 experimental conditions (scene images), data were subjected to classification analyses
135 using linear support-vector machines (SVM) (Müller et al., 2001) in the libsvm
136 implementation (www.csie.ntu.edu.tw/~cjlin/libsvm) with a fixed regularization

137 parameter $C=1$. For each time point t , the processed MEG sensor measurements were
138 concatenated to 306-dimensional pattern vectors, resulting in $M=60$ raw pattern vectors
139 per condition (Figure 1B). To reduce computational load and improve signal-to-noise
140 ratio, we sub-averaged the M vectors in groups of $k = 5$ with random assignment, thus
141 obtaining M/k averaged pattern vectors. We then measured the performance of the SVM
142 classifier to discriminate between every pair (i,j) of conditions using a leave-one-out
143 approach: $M/k - 1$ vectors were randomly assigned to the training test, and 1 vector to the
144 testing set to evaluate the classifier decoding accuracy. The above procedure was
145 repeated 100 times, each with random assignment of the M raw pattern vectors to M/k
146 averaged pattern vectors, and the average decoding accuracy was assigned to the (i,j)
147 element of a 48×48 decoding matrix indexed by condition. The decoding matrix is
148 symmetric with an undefined diagonal. We obtained one decoding matrix
149 (representational dissimilarity matrix or RDM) for each time point t .

150

151 *Representational clustering analysis for size:* Interpreting decoding accuracy as a
152 measure of dissimilarity between patterns, and thus as a distance measure in
153 representational space (Kriegeskorte and Kievit, 2013; Cichy et al., 2014), we partitioned
154 the RDM decoding matrix into within- and between-level segments for the factor scene
155 size (Figure 2A). The average of between-size minus within-size matrix elements
156 produced representational distances (percent decoding accuracy difference) indicative of
157 clustering of visual representations by scene size.

158

159 *Cross-classification*: To assess whether scene size representations were robust to changes
160 of other factors, we used SVM cross-classification assigning different levels of
161 experimental factors to the training and testing set. For example, Figure 2C shows the
162 cross-classification of scene size (small vs. large) across clutter, implemented by limiting
163 the training set to high clutter scenes and the testing set to low clutter scenes. The
164 procedure was repeated with reverse assignment (low clutter for training set and high
165 clutter for testing set) and decoding results were averaged. The training set was 12 times
166 larger ($M = 720$ raw pattern vectors) than for single-image decoding, as we pooled trials
167 across single images that had the same level of clutter and size. We averaged pattern
168 vectors by sub-averaging groups of $k = 60$ raw pattern vectors before the leave-one-out
169 SVM classification. Cross-classification analysis was performed for the cross-
170 classification of the factors scene size (Figure 2D) and scene clutter (Supplementary
171 Figure 3) with respect to changes across all other factors.

172 **2.5 Low and high-level computational models of image statistics**

173 We assessed whether computational models of object and scene recognition predicted
174 scene size from our image material. For this we compared four models: two deep
175 convolutional neural networks that were either trained to perform (1) scene or (2) object
176 classification; (3) the GIST descriptor (Oliva and Torralba, 2001), i.e. a model
177 summarizing the distribution of orientation and spatial frequency in an image that has
178 been shown to predict scene properties, among them size; and (4) HMAX model (Serre et
179 al., 2005), a model of object recognition most akin in structure to low-level visual areas
180 V1/V2. We computed the output of each of these models for each image as described
181 below.

182

183 *Deep neural networks*

184 The deep neural network architecture was implemented following Krizhevsky et al.,
185 2012. We chose this particular architecture because it was the best performing model in
186 object classification in the ImageNet 2012 competition (Russakovsky et al., 2014), uses
187 biologically-inspired local operations (convolution, normalization, max-pooling), and has
188 been compared to human and monkey brain activity successfully (Güçlü and van Gerven,
189 2014; Khaligh-Razavi and Kriegeskorte, 2014; Khaligh-Razavi et al., 2014). The network
190 architecture had 8 layers with the first 5 layers being convolutional and the last 3 fully
191 connected. For an enumeration of units and features for each layer see Table 3. We used
192 the convolution stage of each layer as model output for further analysis.

193

194 We constructed two deep neural networks that differed in the visual categorization task
195 and visual material they were trained on. A deep scene model was trained on 216 scene
196 categories from the Places dataset (available online at: <http://places.csail.mit.edu/>) (Zhou
197 et al., 2015) with 1300 images per category. A deep object model was trained on 683
198 different objects with 900,000 images from the ImageNet dataset (available online at:
199 <http://www.image-net.org/>) (Deng et al., 2009) with similar number of images per object
200 category (~1300). Both deep neural networks were trained on GPUs using the Caffe
201 toolbox (Jia et al., 2014). In detail, the networks were trained for 450,000 iterations, with
202 an initial learning rate of 0.01 and a step multiple of 0.1 every 100,000 iterations.
203 Momentum and weight decay were kept constant at 0.9 and 0.0005 respectively.

204

205 To visualize receptive fields (RFs) of model neurons in the deep scene network (Figure
206 3B) we used a reduction method (Zhou et al., 2015). In short, for a particular neuron we
207 determined the K images activating the neuron most strongly. To determine the empirical
208 size of the RF, we replicated the K images many times with small random occluders at
209 different positions in the image. We then passed the occluded images into the deep scene
210 network and compared the output to the original image, constructing the discrepancy map
211 that indicates which part of the image drives the neuron. We then recentered discrepancy
212 maps and averaged, generating the final RF. To illustrate the RFs tuning we further plot
213 the image patches corresponding to the top activation regions inside the RFs (Figure 3B).

214

215 *GIST*

216 For the GIST descriptor (Oliva and Torralba, 2001), each image was filtered by a bank of
217 Gabor filters with 8 orientations and 4 spatial frequencies (32 filters). Filter outputs were
218 averaged in a 4×4 grid, resulting in a 512-dimensional feature vector. The GIST
219 descriptor represents images in terms of spatial frequencies and orientations by position,
220 (code available: <http://people.csail.mit.edu/torralba/code/spatialenvelope/>).

221

222 *HMAX*

223 We used the HMAX model as applied and described by Serre et al (Serre et al., 2005), a
224 model inspired by the hierarchical organization of the visual cortex. In short, HMAX
225 consists of two sets of alternating S and C layers, i.e. in total 4 layers. The S-layers
226 convolve the input with pre-defined filters, and the C layers perform a max operation.

227 **2.6 Linking computational models of vision to brain data**

228 We used representational dissimilarity analysis to compare the output of computational
229 models to brain data. First, we recorded the output of each model for each of the 48
230 images of the image set. Then, to compare to human brain data, we calculated the pair-
231 wise dissimilarities between model outputs by 1- Spearman's rank order correlation R .
232 This formed 48x48 model dissimilarity matrices (RDMs), one for each layer of each
233 model: 8 for the deep scene and deep object network, 1 for GIST, and 4 for HMAX.

234

235 To compare models and brains, we determined whether images that were similarly
236 represented in a computational network were also similarly represented in the brain. This
237 was achieved by computing the similarity (Spearman's R) of layer-specific model
238 dissimilarity matrix with the time-point specific MEG decoding matrix for every subject
239 and time point and averaging results.

240

241 We then determined whether the computational models predicted the size of a scene. We
242 formulated an explicit size model, i.e. a 48×48 matrix with entries of 1 where images
243 differed in size and 0 otherwise. Equivalent matrices were produced for scene clutter,
244 contrast and luminance (Supplementary Figure 1). Correlation of the explicit size model
245 with any computational model RDM yielded a measure of how well computational
246 models predicted scene size.

247 Finally, we determined whether the above computational models accounted for neural
248 representations of scene size observed in MEG data. For this, we reformulated the
249 representational clustering analysis in a correlation framework. The two measures are

250 equivalent except that the correlation analysis takes into account the variability of the
251 data, which the clustering analysis does not for the benefit of clear interpretability as
252 percent change in decoding accuracy. The procedure had two steps. First, we calculated
253 the similarity (Spearman's R) of the MEG decoding accuracy matrix with the explicit size
254 model for each time point and each participant. Second, we re-calculated the similarity
255 (Spearman's R) of the MEG decoding accuracy matrix with the explicit size model after
256 partialling out all of the layer-specific RDMs of a given computational model.

257 2.7 *Statistical testing*

258 We used permutation tests for cluster-size inference, and bootstrap tests to determine
259 confidence intervals of onset times for maxima, cluster onsets and peak-to-peak latency
260 differences (Nichols and Holmes, 2002; Pantazis et al., 2005; Cichy et al., 2014).

261

262 *Sign permutation tests*

263 For the permutation tests, depending on the statistic of interest our null hypothesis was
264 that the MEG decoding time series were equal to 50% chance level, or that the decoding
265 accuracy difference of between- minus within-level segments of the MEG decoding
266 matrix was equal to 0, or that the correlation values were equal to 0. In all cases, under
267 the null hypothesis the sign of the observed effect in the MEG data is randomly
268 permutable, corresponding to a sign-permutation test that randomly multiplies the
269 participant-specific data with +1 or -1. We created 1,000 permutation samples, every
270 time re-computing the statistic of interest. This resulted in an empirical distribution of the
271 data, allowing us to convert our original data, as well as the permutation samples, into P -
272 values. We then performed cluster-size inference by setting a $P = 0.05$ cluster-definition

273 threshold on the original data and permutation samples, and computing a $P = 0.05$ cluster
274 size threshold from the empirical distribution of the resampled data.

275

276 *Bootstrapping*

277 To calculate confidence intervals (95%) on cluster onset and peak latencies, we
278 bootstrapped the sample of participants 1,000 times with replacement. For each bootstrap
279 sample, we repeated the above permutation analysis yielding distributions of the cluster
280 onset and peak latency, allowing estimation of confidence intervals. In addition, for each
281 bootstrap sample, we determined the peak-to-peak latency difference for scene size
282 clustering and individual scene image classification. This yielded an empirical
283 distribution of peak-to-peak latencies. Setting $P < 0.05$, we rejected the null hypothesis of
284 a latency difference if the confidence interval did not include 0.

285

286 *Label permutation tests*

287 For testing the significance of correlation between the computational model RDMs and
288 the scene size model, we relied on a permutation test of image labels. This effectively
289 corresponded to randomly permuting the columns (and accordingly the rows) of the
290 computational model RDMs 1,000 times, and then calculating the correlation between the
291 permuted matrix and the explicit size model matrix. This yielded an empirical
292 distribution of the data, allowing us to convert our statistic into P -values. Effects were
293 reported as significant when passing a $P = 0.05$ threshold. Results were FDR-corrected
294 for multiple comparisons.

295

296 3 RESULTS

297 Human participants ($n = 15$) viewed images of 48 real-world indoor scenes that differed
298 in the layout property size, as well as in the level of clutter, contrast and luminance
299 (Figure 1A) while brain activity was recorded with MEG. While often real-world scene
300 size and clutter level correlate, here we de-correlated those stimulus properties explicitly
301 by experimental design, based on independent behavioral validation (Park et al., 2014) to
302 allow independent assessment. Images were presented for 0.5s with an inter-trial interval
303 of 1-1.2s (Figure 1B). Participants performed an orthogonal object-detection task on an
304 image of concentric circles appearing every four trials on average. Concentric circle trials
305 were excluded from further analysis.

306

307 To determine the timing of cortical scene processing we used a decoding approach: we
308 determined the time course with which experimental conditions (scene images) were
309 discriminated by visual representations in MEG data. For this, we extracted peri-stimulus
310 MEG time series in 1ms resolution from -100 to +900ms with respect to stimulus onset
311 for each subject. For each time point independently we classified scene images pair-wise
312 by MEG sensor patterns (support vector classification, Figure 1C). Time-point specific
313 classification results (percentage decoding accuracy, 50% chance level) were stored in a
314 48×48 decoding accuracy matrix, indexed by image conditions in rows and columns
315 (Figure 1C, inset). This matrix is symmetric with undefined diagonal. Repeating this
316 procedure for every time point yielded a set of decoding matrices (for a movie of
317 decoding accuracy matrices over time, averaged across subjects, see Supplementary
318 Movie 1). Interpreting decoding accuracies as a representational dissimilarity measure,

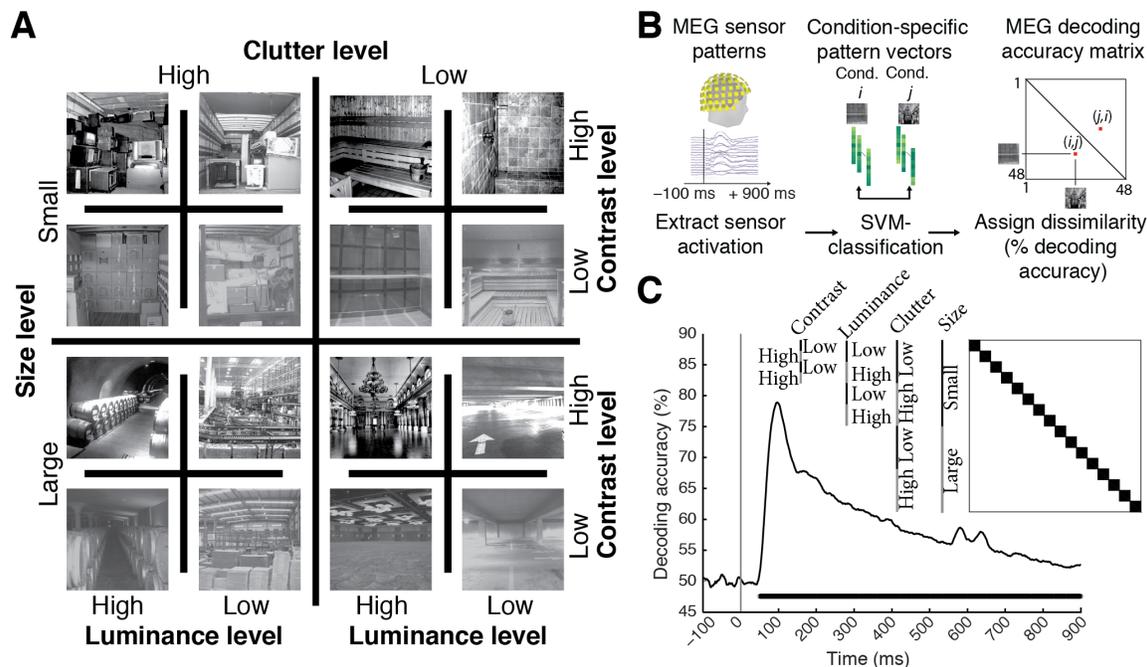
319 each 48x48 matrix summarized, for a given time point, which conditions were
320 represented similarly (low decoding accuracy) or dissimilarly (high decoding accuracy).
321 The matrix was thus termed MEG representational dissimilarity matrix (RDM) (Cichy et
322 al., 2014; Nili et al., 2014).

323

324 Throughout, we determined random-effects significance non-parametrically using a
325 cluster-based randomization approach (cluster-definition threshold $P < 0.05$, corrected
326 significance level $P < 0.05$) (Nichols and Holmes, 2002; Pantazis et al., 2005; Maris and
327 Oostenveld, 2007). 95% confidence intervals for mean peak latencies and onsets
328 (reported in parentheses throughout the results) were determined by bootstrapping the
329 participant sample.

330 **3.1 Neural representations of single scene images emerged early in cortical** 331 **processing**

332 We first investigated the temporal dynamics of image-specific individual scene
333 information in the brain. To determine the time course with which individual scene
334 images were discriminated by visual representations in MEG data, we averaged the
335 elements of each RDM matrix representing pairwise comparisons with matched
336 experimental factors (luminance, contrast, clutter level and scene size) (Figure 1C). We
337 found that the time course rose sharply after image onset, reaching significance at 50ms
338 (45-52ms) and a peak at 97ms (94-102ms). This indicates that single scene images were
339 discriminated early by visual representations, similar to single images with other visual
340 content (Thorpe et al., 1996; Carlson et al., 2013; Cichy et al., 2014; Isik et al., 2014),
341 suggesting a common source in early visual areas (Cichy et al., 2014).

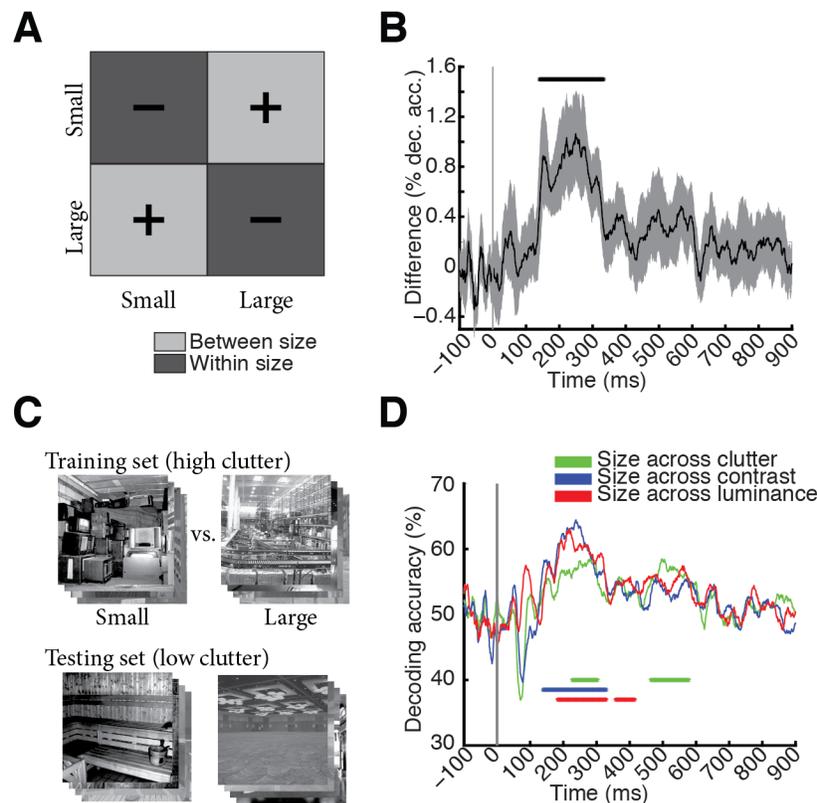


342

343 **Figure 1. Image set and single-image decoding.** **A)** The stimulus set comprised 48 indoor scene
 344 images differing in the size of the space depicted (small vs. large), as well as clutter, contrast, and
 345 luminance level; here each experimental factor combination is exemplified by one image. The
 346 image set was based on behaviorally validated images of scenes differing in size and clutter level,
 347 de-correlating factors size and clutter explicitly by experimental design (Park et al., 2014). Note
 348 that size refers to the size of the real-world space depicted on the image, not the stimulus
 349 parameters; all images subtended 8 visual angle during the experiment. **B)** Time-resolved (1ms
 350 steps from -100 to +900ms with respect to stimulus onset) pair-wise support vector machine
 351 classification of experimental conditions based on MEG sensor level patterns. Classification
 352 results were stored in time-resolved 48×48 MEG decoding matrices. **C)** Decoding results for
 353 single scene classification independent of other experimental factors. Decoding results were
 354 averaged across the dark blocks (matrix inset), to control for luminance, contrast, clutter level and
 355 scene size differences. Inset shows indexing of matrix by image conditions. Horizontal line below
 356 curve indicates significant time points ($n = 15$, cluster-definition threshold $P < 0.05$, corrected
 357 significance level $P < 0.05$); gray vertical line indicates image onset.

358 **3.2 Neural representations of scene size emerged later in time and were robust to**
359 **changes in viewing conditions and other scene properties**

360 When is the spatial layout property scene size processed by the brain? To investigate, we
361 partitioned the decoding accuracy matrix into two subdivisions: images of different
362 (between subdivision light gray, +) and similar size level (within subdivision, dark gray,
363 -). The difference of mean between-size minus within-size decoding accuracy is a
364 measure of clustering of visual representations by size (Figure). Peaks in this measure
365 indicate time points at which MEG sensor patterns cluster maximally by scene size,
366 suggesting underlying neural visual representations allowing for explicit, linear readout
367 (DiCarlo and Cox, 2007) of scene size by the brain. Scene size (Figure 2B) was
368 discriminated first at 141ms (118 – 156ms) and peaked at 249ms (150 – 274ms), which
369 was significantly later than the peak in single image classification ($P = 0.001$, bootstrap
370 test of peak-latency differences).



371

372 **Figure 2. Scene size is discriminated by visual representations.** **A)** To determine the time
373 course of scene size processing we determined when visual representations clustered by scene
374 size. For this we subtracted mean within-size decoding accuracies (dark gray, -) from between-
375 size decoding accuracies (light gray, +). **B)** Scene size was discriminated by visual
376 representations late in time (onset of significance at 141ms (118-156ms), peak at 249ms (150-
377 274ms). Gray shaded area indicates 95% confidence intervals determined by bootstrapping
378 participants. **C)** Cross-classification analysis, exemplified for cross-classification of scene size
379 across clutter level. A classifier was trained to discriminate scene size on high clutter images, and
380 tested on low clutter images. Results were averaged following an opposite assignment of clutter
381 images to training and testing sets. Before entering cross-classification analysis, MEG trials were
382 grouped by clutter and size level respectively independent of image identity. A similar cross-
383 classification analysis was applied for other image and scene properties. **D)** Results of cross-
384 classification analysis indicated robustness of scene size visual representations to changes in other
385 scene and image properties (scene clutter, luminance, and contrast). Horizontal lines indicate
386 significant time points ($n = 15$, cluster-definition threshold $P < 0.05$, corrected significance level
387 $P < 0.05$); gray vertical line indicates image onset. For result curves with 95% confidence
388 intervals see Supplementary Figure 2.

389 Equivalent analyses for the experimental factors scene clutter, contrast, and luminance
390 level yielded diverse time courses (Supplementary Figure 1, Table 1A). Importantly,
391 representations of low-level image property contrast emerged significantly earlier than
392 scene size ($P = 0.004$) and clutter ($P = 0.006$, bootstrap test of peak-latency differences).
393 For the factor luminance, only a weak effect and thus no significant onset response was
394 observed, suggesting a pre-cortical luminance normalization mechanism.

395

396 To be of use in the real world, visual representations of scene size must be robust against
397 changes of other scene properties, such as clutter level (i.e. space filled by different types
398 and amounts of objects) and semantic category (i.e. the label by which we name it), and
399 changes in viewing conditions, such as luminance and contrast. We investigated the
400 robustness of scene size representations to all these factors using cross-classification
401 (Figure 2C; for 95% confidence intervals on curves see Supplementary Figure 2). For this
402 we determined how well a classifier trained to distinguish scenes at one clutter level
403 could distinguish scenes at the other level, while collapsing data across single image
404 conditions of same level in size and clutter. We found that scene size was robust to
405 changes in scene clutter, luminance and contrast (Figure 2D; onsets and peaks in Table
406 1B). Note that by experimental design, the scene category always differed across size
407 level, such that cross-classification also established that scene size was discriminated by
408 visual representations independent of the scene category.

409

410 An analogous analysis for clutter level yielded evidence for viewing-condition
411 independent clutter level representations (Supplementary Figure 3), reinforcing the notion

412 of clutter level as a robust and relevant dimension of scene representations in the human
413 brain (Park et al., 2014). Finally, an analysis revealing persistent and transient
414 components of scene representations indicated strong persistent components for scene
415 size and clutter representations, with little or no evidence for contrast and luminance
416 (Supplementary Figure 4). Persistence of scene size and clutter level representations
417 further reinforces the notion of size and clutter level representations being important end
418 products of visual computations kept online by the brain for further processing and
419 behavioral guidance.

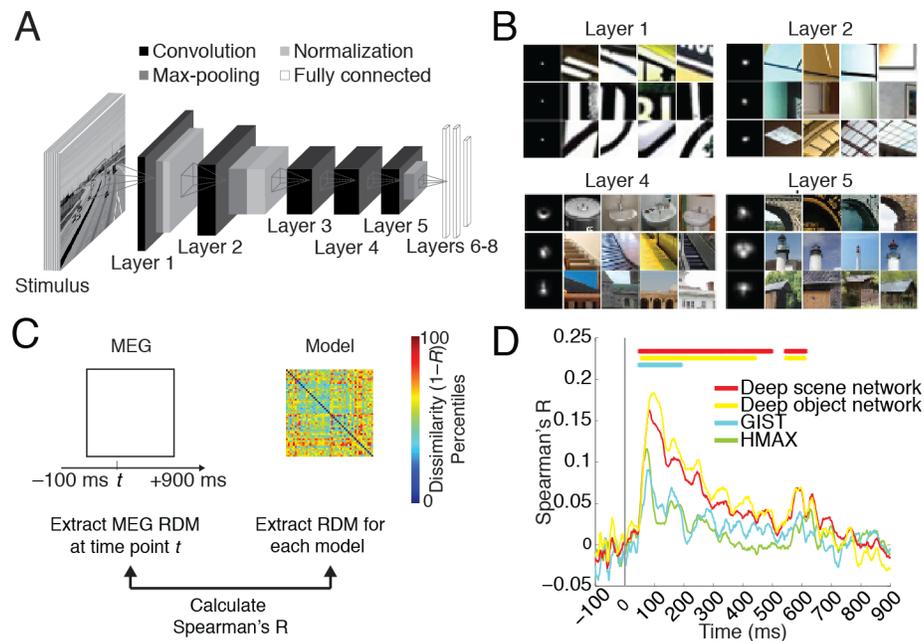
420

421 In sum, our results constitute evidence for representations of scene size in human brains
422 from non-invasive electrophysiology, apt to describe scene size discrimination under real
423 world changes in viewing conditions.

424 **3.3 Neural representations of single scene images were predicted by deep** 425 **convolutional neural networks trained on real world scene categorization**

426 Visual scene recognition in cortex is a complex hierarchical multi-step process, whose
427 understanding necessitates a quantitative model that captures this complexity. Here, we
428 evaluated whether an 8-layer deep neural network trained to perform scene classification
429 on 205 different scene categories (Zhou et al., 2014) predicted human scene
430 representations. We refer to this network as deep scene network (Figure 3A).
431 Investigation of the receptive fields (RFs) of model neurons using a reduction method
432 (Zhou et al., 2015) indicated a gradient of increasing complexity from low to high layers,
433 and selectivity to whole objects, texture, and surface layout information (Figure 3B). This

434 suggests that the network might be able to capture information about both single scenes
 435 and scene layout properties.



436

437 **Figure 3. Predicting emerging neural representations of single scene images by**
 438 **computational models.** **A)** Architecture of deep convolutional neural network trained on scene
 439 categorization (deep scene network). **B)** Receptive field (RF) of example deep scene neurons in
 440 layers 1, 2, 4, and 5. Each row represents one neuron. The left column indicates size of RF, and
 441 the remaining columns indicate image patches most strongly activating these neurons. Lower
 442 layers had small RFs with simple Gabor filter-like sensitivity, whereas higher layers had
 443 increasingly large RFs sensitive to complex forms. RFs for whole objects, texture, and surface
 444 layout information emerged although these features were not explicitly taught to the deep scene
 445 model. **C)** We used representational dissimilarity analysis to compare visual representations in
 446 brains with models. For every time point, we compared subject-specific MEG RDMs
 447 (Spearman's R) to model RDMs and results were averaged across subjects. **D)** All investigated
 448 models significantly predicted emerging visual representations in the brain, with superior
 449 performance for the deep neural networks compared to HMAX and GIST. Horizontal lines
 450 indicate significant time points ($n = 15$, cluster-definition threshold $P < 0.05$, corrected
 451 significance level $P < 0.05$); gray vertical line indicates image onset.
 452

453 To determine the extent to which visual representations learned by the deep scene model
454 and the human brain are comparable, we used representational similarity analysis
455 (Kriegeskorte, 2008; Cichy et al., 2014). The key idea is that if two images evoke similar
456 responses in the model, they should evoke similar responses in the brain, too.

457

458 For the deep neural network, we first estimated image response patterns by computing the
459 output of each model layer to each of the 48 images. We then constructed layer-resolved
460 48×48 representational dissimilarity matrices (RDMs) by calculating the pairwise
461 dissimilarity (1-Spearman's R) across all model response patterns for each layer output.

462

463 We then compared (Spearman's R) the layer-specific deep scene model RDMs with the
464 time-resolved MEG RDMs and averaged results over layers, yielding a time course
465 indicating how well the deep scene model predicted and thus explained scene
466 representations (Figure 3D). To compare against other models, we performed equivalent
467 analyses to a deep neural network trained on object-categorization (termed deep object
468 network) and standard models of object (HMAX) and scene-recognition (GIST) (Oliva
469 and Torralba, 2001; Serre et al., 2007).

470

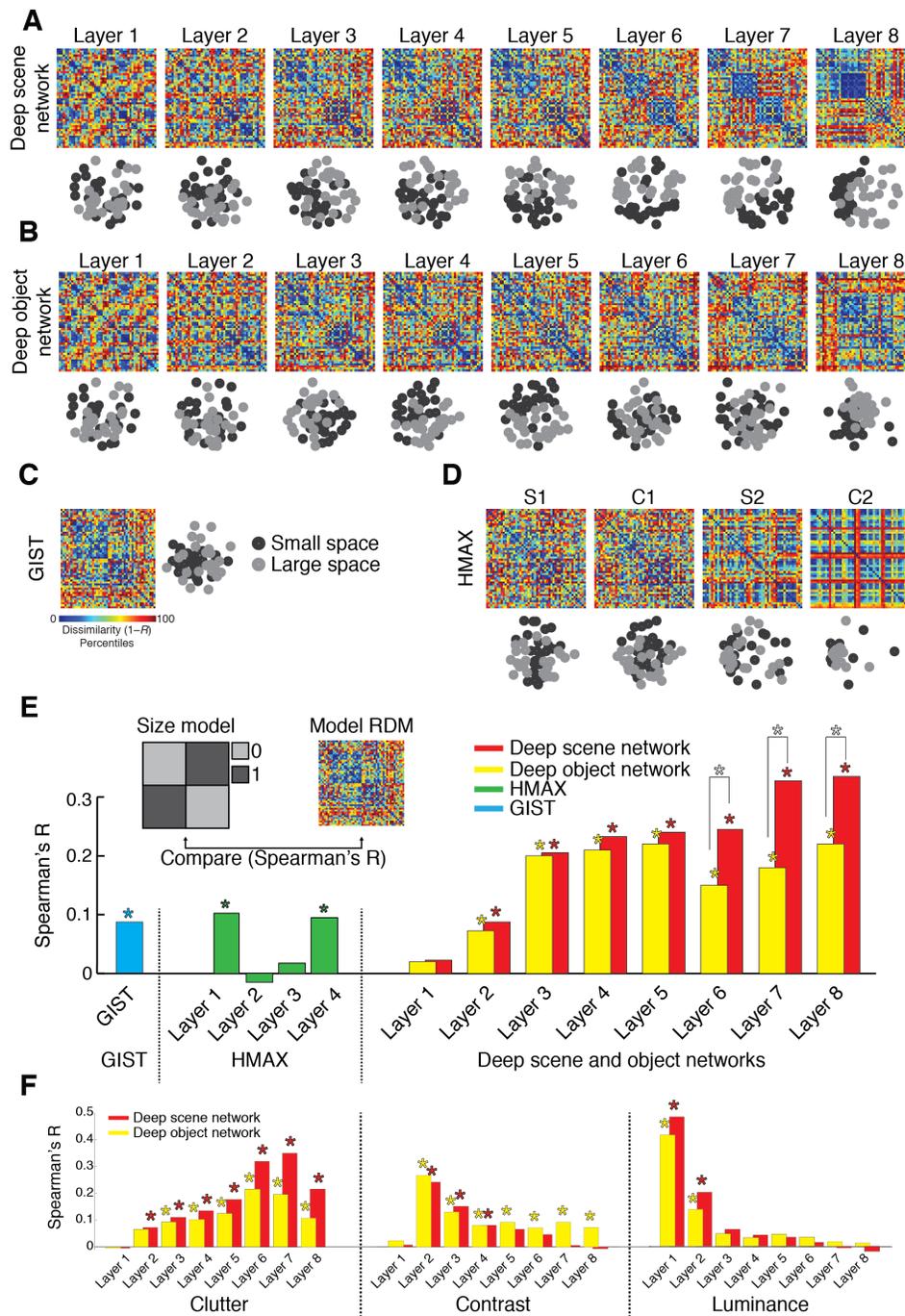
471 We found that the deep object and scene network performed similarly at predicting visual
472 representations over time (Figure 3D, for details see Table 2A; for layer-resolved results
473 see Supplementary Figure 5), and better than the HMAX and GIST models (for direct
474 quantitative comparison see Supplementary Figure 6).

475

476 In sum, our results show that brain representations of single scene images were best
477 predicted by deep neural network models trained on real-world categorization tasks,
478 demonstrating the ability of the models to capture the complexity of scene recognition,
479 and their semblance to the human brain representations.

480 **3.4 Representations of scene size emerged in the deep scene model**

481 Beyond prediction of neural representations of single scene images, does the deep scene
482 neural network indicate the spatial layout property scene size? To visualize, we used
483 multidimensional scaling (MDS) on layer-specific model RDMs, and plotted the 48 scene
484 images into the resulting 2D arrangement color-coded for scene size (black= small, gray
485 = large). We found a progression in the representation of scene size in the deep scene
486 network: low layers showed no structure, whereas high layers displayed a progressively
487 clearer representation of scene size (A). A similar, but weaker progression, was visible
488 for the deep object network (Figure 4B). Comparable analysis for HMAX and GIST
489 (Figure 4C,D) found no prominent representation of size.



490

491 **Figure 4. Representation of scene size in computational models of object and scene**
 492 **category.** **A-D)** Layer-specific RDMs and corresponding 2D multidimensional scaling
 493 (MDS) plots for a deep scene network, deep object network, GIST, and HMAX. MDS plots are
 494 color-coded by scene size (small = black; large = gray). **E)** Quantifying the representation of
 495 scene size in computational models. We compared (Spearman's R) each model's RDMs with an
 496 explicit size model (RDM with entries 0 for images of similar size, 1 for images of dissimilar

497 size). Results are color-coded for each model. **F)** Similar to (E) for clutter, contrast and luminance
498 (results shown only for deep scene and object networks). While representations of the abstract
499 scene properties size and clutter emerged with increasing layer number, the low-level image
500 properties contrast and luminance successively abstracted away. Stars above bars indicate
501 statistical significance. Stars between bars indicate significant differences between the
502 corresponding layers of the deep scene vs. object network. Complete layer-wise comparisons
503 available in Supplementary Figure 7. ($n = 48$; label permutation tests for statistical inference, $P <$
504 0.05 , FDR-corrected for multiple comparisons).

505

506 We quantified this descriptive finding by computing the similarity of model RDMs with
507 an explicit size model (an RDM with entries 0 for images of similar size, 1 for images of
508 dissimilar size; Figure 4E inset). We found a significant effect of size in all models ($P <$
509 0.05 , FDR-corrected, stars above bars indicate significance). The size effect was larger in
510 the deep neural networks than in GIST and HMAX, it was more pronounced in the high
511 layers, and the deep scene network displayed a significantly stronger effect of scene size
512 than the deep object network in layers 6-8 (stars between bars; for all pair-wise layer-
513 specific comparisons see Supplementary Figure 7). A supplementary partial correlation
514 analysis confirmed that the effect of size in the deep scene network was not explained by
515 correlation with the other experimental factors (Supplementary Figure 8).

516

517 Together, these results indicate the deep scene network captured scene size better than all
518 other models, and that scene size representations emerge gradually in the deep neural
519 network hierarchy. Thus representations of visual space can emerge intrinsically in neural
520 networks constrained to perform visual scene categorization without being trained to do
521 so directly.

522 **3.5 Neural representations of scene size emerged in the deep scene model**

523 The previous sections demonstrated that representations of scene size emerged in both
524 neural signals (Figure 2) and computational models (Figure 4). To evaluate the overlap
525 between these two representations, we combined representational similarity analysis with
526 partial correlation analysis (Clarke and Tyler, 2014) (Figure 5A).

527

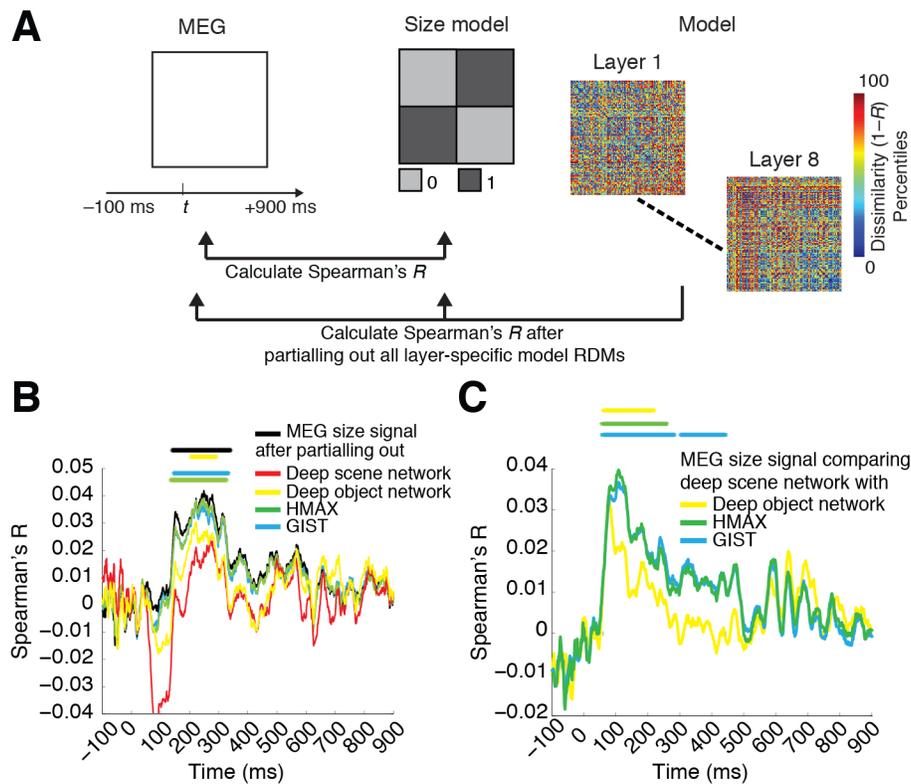
528 We first computed the neural representations of scene size by correlating (Spearman's R)
529 the MEG RDMs with the explicit size model (black curve). We then repeated the process,
530 but only after partialling out all layer-specific RDMs of a model from the explicit size
531 model (color-coded by model) (Figure 5B). The reasoning is that if neural signals and
532 computational models carry the same scene size information, the scene size effect will
533 vanish in the latter case.

534

535 When partialling out the effect of the deep scene network, the scene size effect was
536 considerably reduced and was no longer statistically significant. In all other models, the
537 effect was reduced but was still statistically significant (Figure 5B). Further, the reduction
538 of the size effect was higher for the deep scene network than all other models (Figure
539 5C). Equivalent analyses for scene clutter, contrast and luminance indicated that the deep
540 scene and object networks abolished all effects, while other models did not
541 (Supplementary Figure 9).

542

543 Together, these results show that only the deep scene model captured the neural
 544 representation of scene size in the human brain, singling it out as the best of the scene
 545 representation models tested here.
 546



547

548 **Figure 5. The deep scene model accounts for more of the MEG size signal than other**
 549 **models. A)** We combined representational similarity with partial correlation analysis to determine
 550 which computational models explained emerging representations of scene size in the brain. **B)**
 551 MEG representations of scene size (termed MEG size signal) before (black) and after (color-
 552 coded by model) partialling out the effect of different computational models. Only partialling out
 553 the effect of the deep scene network abolished the MEG size signal. **C)** Difference in amount of
 554 variance partialled out from the size signal: comparing all models to the deep scene network. The
 555 deep scene network accounted for more MEG size signal than all other models ($n = 15$; cluster-
 556 definition threshold $P < 0.05$, significance threshold $P < 0.05$; results corrected for multiple
 557 comparisons by 5 for panel B and 3 for panel C).

558 4 DISCUSSION

559 We characterized the emerging representation of scenes in the human brain using
560 multivariate pattern classification methods (Carlson et al., 2013; Cichy et al., 2014) and
561 representational similarity analysis (Kriegeskorte, 2008; Kriegeskorte and Kievit, 2013)
562 on combined MEG and computational model data. We found that neural representations
563 of individual scenes and the low-level image property contrast emerged early, followed
564 by the scene layout property scene size at around 250 ms. The neural representation of
565 scene size was robust to changes in viewing conditions and scene properties such as
566 contrast, luminance, clutter level and category. Our results provide novel evidence for an
567 electrophysiological signal of scene processing in humans that remained stable under
568 real-world viewing conditions. To capture the complexity of scene processing in the brain
569 by a computational model, we trained a deep convolutional neural network on scene
570 classification. We found that the deep scene model predicted representations of scenes in
571 the brain and accounted for abstract properties such as scene size and clutter level better
572 than alternative computational models, while abstracting away low-level image properties
573 such as luminance and contrast level.

574 4.1 A multivariate pattern classification signal for the processing of scene layout 575 property scene size

576 A large body of evidence from neuropsychology, neuroimaging and invasive work in
577 humans and monkeys has identified locally circumscribed cortical regions of the brain
578 dedicated to the processing of three fundamental visual categories: faces, bodies and
579 scenes (Allison et al., 1994; Kanwisher et al., 1997; Aguirre et al., 1998; Downing et al.,
580 2001; Tsao et al., 2006; Kornblith et al., 2013). For faces and bodies, respective

581 electrophysiological signals in humans have been identified (Allison et al., 1994; Bentin
582 et al., 1996; Jeffreys, 1996; Liu et al., 2002; Stekelenburg and de Gelder, 2004; Thierry et
583 al., 2006). However, electrophysiological markers for scene-specific processing have
584 been identified for the auditory modality only (Fujiki et al., 2002; Tiitinen et al., 2006),
585 and a visual scene-specific electrophysiological signal had not been described until now.

586

587 Our results provide the first evidence for an electrophysiological signal of visual scene
588 size processing in humans. Multivariate pattern classification analysis on MEG data
589 revealed early discrimination of single scene images (peak at 97ms) and the low-level
590 image property contrast (peak at 74ms), whereas the abstract property of space size was
591 discriminated later (peak at 249ms). While early scene-specific information in the MEG
592 likely emerged from low-level visual areas such as V1 (Cichy et al., 2014), the
593 subsequent scene size signal had properties commonly ascribed to higher stages of visual
594 processing in ventral visual cortex: the representation of scene size was tolerant to
595 changes occurring in real world viewing conditions, such as luminance, contrast, clutter
596 level and category. The electrophysiological signal thus reflected scene size
597 representations that could reliably be used for scene recognition in real world settings
598 under changing viewing conditions (Poggio and Bizzi, 2004; DiCarlo and Cox, 2007;
599 DiCarlo et al., 2012). This result paves the way to further studies of the representational
600 format of scenes in the brain, e.g. by measuring the modulation of the scene-specific
601 signal by other experimental factors.

602

603 The magnitude of the scene size effect, although consistent across subjects and
604 statistically robust to multiple comparison correction, is small with a maximum of ~1%.
605 Note however that the size effect, in contrast to single image decoding (peak decodability
606 at ~79%), is not a measure of how well single images differing in size can be
607 discriminated, but a difference measure of how much better images of different size can
608 be discriminated rather than images of the same size. Thus, it is a measure of information
609 about scene size over-and-above information distinguishing between any two single
610 scenes. The magnitude of the size effect is comparable to effects reported for abstract
611 visual properties such as animacy (1.9 and 1.1% respectively, Cichy et al., 2014).

612

613 What might be the exact locus of the observed scene size signal in the brain? Previous
614 research has indicated parametric encoding of scene size in parahippocampal place area
615 (PPA) and retrosplenial cortex (Park et al., 2014), corroborating numerous studies
616 showing that spatial properties of scenes such as boundaries and layout are represented in
617 these cortical regions (Epstein and Kanwisher, 1998; Epstein et al., 1999; Wolbers et al.,
618 2011b). Both onset and peak latency of the observed scene size signal concurred with
619 reported latencies for parahippocampal cortex (Mormann et al., 2008), suggesting that
620 one or several nodes of the human spatial navigation network might be the source of the
621 scene size signal.

622

623 Last, we found that not only scene size representations, but also scene clutter
624 representations were tolerant to changes in viewing conditions, and emerged later than
625 the low-level image contrast representations. These results complement previous findings

626 in object perception research that representations of single objects emerge earlier in time
627 than representations of more abstract properties such as category membership (Carlson et
628 al., 2013; Cichy et al., 2014).

629 **4.2 Neural representations of abstract scene properties such as scene size are**
630 **explained by a deep neural network model trained on scene classification**

631 Scene processing in the brain is a complex process necessitating a formal quantitative
632 model that addresses this complexity. Here, our study of several models of scene and
633 object recognition provided three novel results, each with fundamental theoretical
634 implications.

635

636 First, deep neural networks offered the best characterization of neural scene
637 representations compared to other models tested. The superiority of high performing deep
638 neural networks over simpler models indicates that hierarchical architectures might be
639 necessary to capture the structure of single scene representations in the human brain.

640 While previous research has established that deep neural networks capture object
641 representations in human and monkey inferior temporal cortex well, we demonstrated
642 that a deep neural network explained millisecond-resolved dynamics underlying scene
643 recognition from processing of low- to high-level properties, better than other models of
644 object and scene-processing tested. Concerning high-level abstract scene properties in
645 particular, our results shed lights into the black box of cortical scene processing,
646 providing novel insight both from the perspective of modeling, and of experimental brain
647 science. From a modeling perspective, the near monotonic relationship between the
648 representation of size and clutter level in the deep neural network and the network layer

649 number indicates that scene size is an abstract scene property emerging through complex
650 multi-step processing. From the perspective of experimental brain science, our results
651 provide an advance in understanding neural representations of the processing of abstract
652 scene properties such as spatial layout. Neuronal responses in high-level visual cortex are
653 often sparse and nonlinear, making a full explanation by simple mathematical models in
654 low-dimensional spaces or basic image statistics unlikely (Groen et al., 2013; Rice et al.,
655 2014; Watson et al., 2014; Rice et al., 2014). Instead, our result concurs with the finding
656 that complex deep neural networks performing well on visual categorization tasks
657 represent visual stimuli similar to the human brain (Cadieu et al., 2013; Yamins et al.,
658 2014), and extends the claim to abstract properties of visual stimuli.

659

660 The second novel finding is that a deep neural network trained specifically on scene
661 categorization had superior representation of scene size compared to a deep neural
662 network trained on objects. Importantly, it also offered the best account of neural
663 representations of scene size in the MEG, indicating that the underlying algorithmic
664 computations matched the neuronal computations in the human brain. This indicates that
665 the constraints imposed by the task the network is trained on, i.e. object or scene
666 categorization, critically influenced the represented features. This makes plausible the
667 notion that spatial representations emerge naturally and intrinsically in neural networks
668 performing scene categorization, such as in the human brain. It further suggests that
669 separate processing streams in the brain for different visual content, such as scenes,
670 objects or faces, might be the result of differential task constraints imposed by
671 classification of the respective visual input (DiCarlo et al., 2012; Yamins et al., 2014).

672

673 The third novel finding is that representations of abstract scene properties (size, clutter
674 level) emerged with increasing layers in deep neural networks, while low-level image
675 properties (contrast, luminance) were increasingly abstracted away, mirroring the
676 temporal processing sequence in the human brain: representations of low-level image
677 properties emerged first, followed by representations of scene size and clutter level. This
678 suggests common mechanisms in both and further strengthen the idea that deep neural
679 networks are a promising model of the processing hierarchies constituting the human
680 visual system, reinforcing the view of the visual brain as performing increasingly
681 complex feature extraction over time (Thorpe et al., 1996; Liu et al., 2002; Reddy and
682 Kanwisher, 2006; Serre et al., 2007; Kourtzi and Connor, 2011; DiCarlo et al., 2012).

683

684 However, we did not observe a relationship between layer-specific representations in the
685 deep neural networks and temporal dynamics in the human brain. Instead, the MEG
686 signal predominantly reflected representations in low neural network layers
687 (Supplementary Figure 5). One reason for this might be that our particular image set
688 differed strongly in low-level features, thus strongly activating early visual areas that are
689 best modeled by low neural network layers. Activity in low-level visual cortex was thus
690 very strong, potentially masking weaker activity in high-level visual cortex that is
691 invariant to changes in low level features. Another reason might be that while early visual
692 regions are close to the MEG sensors, creating strong MEG signals, scene-processing
693 cortical regions such as PPA are deeply harbored in the brain, creating weaker MEG
694 signals. Future studies using image sets optimized to drive low-and high level visual

695 cortex equally are necessary, to test whether layer-specific representations in deep neural
696 networks can be mapped in both time and in space onto processing stages in the human
697 brain.

698 **4.3 Conclusions**

699 Using a combination of multivariate pattern classification and computational models to
700 study the dynamics in neuronal representation of scenes, we identified a neural marker of
701 spatial layout processing in the human brain, and showed that a deep neural network
702 model of scene categorization explains representations of spatial layout better than other
703 models. Our results pave the way to future studies investigating the temporal dynamics of
704 spatial layout processing, and highlight deep hierarchical architectures as the best models
705 for understanding visual scene representations in the human brain.

706 **5 ACKNOWLEDGEMENTS**

707 This work was funded by National Eye Institute grant EY020484 (to A.O.), National
708 Science Foundation grant BCS-1134780 (to D.P.), McGovern Institute Neurotechnology
709 Program (to A.O. and D.P.), a Humboldt Scholarship (to R.M.C), and was conducted at
710 the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research,
711 Massachusetts Institute of Technology. We thank Santani Teng for helpful comments on
712 the manuscript.

713 **6 REFERENCES**

714 Aguirre GK, Zarahn E, D'Esposito M (1998) An area within human ventral cortex
715 sensitive to “building” stimuli: evidence and implications. *Neuron* 21:373–383.

- 716 Allison T, Ginter H, McCarthy G, Nobre AC, Puce A, Luby M, Spencer DD (1994) Face
717 recognition in human extrastriate cortex. *J Neurophysiol* 71:821–825.
- 718 Bentin S, Allison T, Puce A, Perez E, McCarthy G (1996) Electrophysiological Studies
719 of Face Perception in Humans. *J Cogn Neurosci* 8:551–565.
- 720 Bird CM, Capponi C, King JA, Doeller CF, Burgess N (2010) Establishing the
721 Boundaries: The Hippocampal Contribution to Imagining Scenes. *J Neurosci*
722 30:11688–11695.
- 723 Bonnici HM, Kumaran D, Chadwick MJ, Weiskopf N, Hassabis D, Maguire EA (2012)
724 Decoding representations of scenes in the medial temporal lobes. *Hippocampus*
725 22:1143–1153.
- 726 Cadieu CF, Hong H, Yamins D, Pinto N, Majaj NJ, DiCarlo JJ (2013) The Neural
727 Representation Benchmark and its Evaluation on Brain and Machine.
728 ArXiv13013530.
- 729 Carlson T, Tovar DA, Alink A, Kriegeskorte N (2013) Representational dynamics of
730 object vision: The first 1000 ms. *J Vis* 13.
- 731 Cichy RM, Pantazis D, Oliva A (2014) Resolving human object recognition in space and
732 time. *Nat Neurosci* 17:455–462.
- 733 Clarke A, Tyler LK (2014) Object-Specific Semantic Coding in Human Perirhinal
734 Cortex. *J Neurosci* 34:4766–4775.
- 735 Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) ImageNet: A large-scale
736 hierarchical image database. *CVPR*, pp 248–255.
- 737 DiCarlo JJ, Cox DD (2007) Untangling invariant object recognition. *Trends Cogn Sci*
738 11:333–341.
- 739 DiCarlo JJ, Zoccolan D, Rust NC (2012) How Does the Brain Solve Visual Object
740 Recognition? *Neuron* 73:415–434.

- 741 Doeller CF, Barry C, Burgess N (2010) Evidence for grid cells in a human memory
742 network. *Nature* 463:657–661.
- 743 Doeller CF, King JA, Burgess N (2008) Parallel striatal and hippocampal systems for
744 landmarks and boundaries in spatial memory. *Proc Natl Acad Sci* 105:5915–5920.
- 745 Downing PE, Jiang Y, Shuman M, Kanwisher N (2001) A Cortical Area Selective for
746 Visual Processing of the Human Body. *Science* 293:2470–2473.
- 747 Epstein RA (2011) Cognitive Neuroscience: Scene Layout from Vision and Touch. *Curr*
748 *Biol* 21:R437–R438.
- 749 Epstein R, Harris A, Stanley D, Kanwisher N (1999) The Parahippocampal Place Area:
750 Recognition, Navigation, or Encoding? *Neuron* 23:115–125.
- 751 Epstein R, Kanwisher N (1998) A cortical representation of the local visual environment.
752 *Nature* 392:598–601.
- 753 Groen IIA, Ghebreab S, Prins H, Lamme VAF, Scholte HS (2013) From Image Statistics
754 to Scene Gist: Evoked Neural Activity Reveals Transition from Low-Level
755 Natural Image Structure to Scene Category. *J Neurosci* 33:18814–18824.
- 756 Güçlü U, van Gerven MAJ (2014) Deep Neural Networks Reveal a Gradient in the
757 Complexity of Neural Representations across the Brain’s Ventral Visual Pathway.
758 *ArXiv14116422 Q-Bio* Available at: <http://arxiv.org/abs/1411>.
- 759 Isik L, Meyers EM, Leibo JZ, Poggio TA (2014) The dynamics of invariant object
760 recognition in the human visual system. *J Neurophysiol* 111:91–102.
- 761 Jacobs J, Weidemann CT, Miller JF, Solway A, Burke JF, Wei X-X, Suthana N, Sperling
762 MR, Sharan AD, Fried I, Kahana MJ (2013) Direct recordings of grid-like
763 neuronal activity in human spatial navigation. *Nat Neurosci* 16:1188–1190.
- 764 Jeffreys DA (1996) Evoked Potential Studies of Face and Object Processing. *Vis Cogn*
765 3:1–38.

- 766 Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T
767 (2014) Caffe: Convolutional Architecture for Fast Feature Embedding.
768 ArXiv14085093.
- 769 Kanwisher N, McDermott J, Chun MM (1997) The Fusiform Face Area: A Module in
770 Human Extrastriate Cortex Specialized for Face Perception. *J Neurosci* 17:4302–
771 4311.
- 772 Khaligh-Razavi S-M, Henriksson L, Kay K, Kriegeskorte N (2014) Explaining the
773 hierarchy of visual representational geometries by remixing of features from
774 many computational vision models. bioRxiv:009936.
- 775 Khaligh-Razavi S-M, Kriegeskorte N (2014) Deep Supervised, but Not Unsupervised,
776 Models May Explain IT Cortical Representation. *PLoS Comput Biol*
777 10:e1003915.
- 778 King J-R, Dehaene S (2014) Characterizing the dynamics of mental representations: the
779 temporal generalization method. *Trends Cogn Sci* 18:203–210.
- 780 Kornblith S, Cheng X, Ohayon S, Tsao DY (2013) A Network for Scene Processing in
781 the Macaque Temporal Lobe. *Neuron* 79:766–781.
- 782 Kourtzi Z, Connor CE (2011) Neural Representations for Object Perception: Structure,
783 Category, and Adaptive Coding. *Annu Rev Neurosci* 34:45–67.
- 784 Kravitz DJ, Peng CS, Baker CI (2011a) Real-World Scene Representations in High-Level
785 Visual Cortex: It's the Spaces More Than the Places. *J Neurosci* 31:7322–7333.
- 786 Kravitz DJ, Saleem KS, Baker CI, Mishkin M (2011b) A new neural framework for
787 visuospatial processing. *Nat Rev Neurosci* 12:217–230.
- 788 Kriegeskorte N (2008) Representational similarity analysis – connecting the branches of
789 systems neuroscience. *Front Syst Neurosci* 2:4.

- 790 Kriegeskorte N, Kievit RA (2013) Representational geometry: integrating cognition,
791 computation, and the brain. *Trends Cogn Sci* 17:401–412.
- 792 Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep
793 convolutional neural networks. In: *Advances in Neural Information Processing*
794 *Systems*.
- 795 Liu J, Harris A, Kanwisher N (2002) Stages of processing in face perception: an MEG
796 study. *Nat Neurosci* 5:910–916.
- 797 MacEvoy SP, Epstein RA (2011) Constructing scenes from objects in human
798 occipitotemporal cortex. *Nat Neurosci* 14:1323–1329.
- 799 Maris E, Oostenveld R (2007) Nonparametric statistical testing of EEG- and MEG-data. *J*
800 *Neurosci Methods* 164:177–190.
- 801 Mormann F, Kornblith S, Quiroga RQ, Kraskov A, Cerf M, Fried I, Koch C (2008)
802 Latency and Selectivity of Single Neurons Indicate Hierarchical Processing in the
803 Human Medial Temporal Lobe. *J Neurosci* 28:8865–8872.
- 804 Moser EI, Kropff E, Moser M-B (2008) Place Cells, Grid Cells, and the Brain’s Spatial
805 Representation System. *Annu Rev Neurosci* 31:69–89.
- 806 Mullally SL, Maguire EA (2011) A New Role for the Parahippocampal Cortex in
807 Representing Space. *J Neurosci* 31:7441–7449.
- 808 Nichols TE, Holmes AP (2002) Nonparametric permutation tests for functional
809 neuroimaging: A primer with examples. *Hum Brain Mapp* 15:1–25.
- 810 Nili H, Wingfield C, Walther A, Su L, Marslen-Wilson W, Kriegeskorte N (2014) A
811 Toolbox for Representational Similarity Analysis. *PLoS Comput Biol*
812 10:e1003553.
- 813 Oliva A, Torralba A (2001) Modeling the Shape of the Scene: A Holistic Representation
814 of the Spatial Envelope. *Int J Comput Vis* 42:145–175.

- 815 Pantazis D, Nichols TE, Baillet S, Leahy RM (2005) A comparison of random field
816 theory and permutation methods for the statistical analysis of MEG data.
817 *NeuroImage* 25:383–394.
- 818 Park S, Brady TF, Greene MR, Oliva A (2011) Disentangling Scene Content from Spatial
819 Boundary: Complementary Roles for the Parahippocampal Place Area and Lateral
820 Occipital Complex in Representing Real-World Scenes. *J Neurosci* 31:1333–
821 1340.
- 822 Park S, Konkle T, Oliva A (2014) Parametric Coding of the Size and Clutter of Natural
823 Scenes in the Human Brain. *Cereb Cortex*.
- 824 Poggio T, Bizzi E (2004) Generalization in vision and motor control. *Nature* 431:768–
825 774.
- 826 Reddy L, Kanwisher N (2006) Coding of visual objects in the ventral stream. *Curr Opin*
827 *Neurobiol* 16:408–414.
- 828 Rice GE, Watson DM, Hartley T, Andrews TJ (2014) Low-Level Image Properties of
829 Visual Objects Predict Patterns of Neural Response across Category-Selective
830 Regions of the Ventral Visual Pathway. *J Neurosci* 34:8837–8844.
- 831 Riesenhuber M, Poggio T (1999) Hierarchical models of object recognition in cortex. *Nat*
832 *Neurosci* 2:1019–1025.
- 833 Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A,
834 Khosla A, Bernstein M, Berg AC, Fei-Fei L (2014) ImageNet Large Scale Visual
835 Recognition Challenge. *ArXiv14090575 Cs*.
- 836 Serre T, Oliva A, Poggio T (2007) A feedforward architecture accounts for rapid
837 categorization. *Proc Natl Acad Sci* 104:6424–6429.
- 838 Serre T, Wolf L, Poggio T (2005) Object recognition with features inspired by visual
839 cortex. In: *IEEE CVPR*, pp 994–1000.

- 840 Stekelenburg JJ, de Gelder B (2004) The neural correlates of perceiving human bodies:
841 an ERP study on the body-inversion effect. *Neuroreport* 15:777–780.
- 842 Thierry G, Pegna AJ, Dodds C, Roberts M, Basan S, Downing P (2006) An event-related
843 potential component sensitive to images of the human body. *Neuroimage* 32:871–
844 879.
- 845 Thorpe S, Fize D, Marlot C (1996) Speed of processing in the human visual system.
846 *Nature* 381:520–522.
- 847 Tsao DY, Freiwald WA, Tootell RBH, Livingstone MS (2006) A Cortical Region
848 Consisting Entirely of Face-Selective Cells. *Science* 311:670–674.
- 849 Vaziri S, Carlson ET, Wang Z, Connor CE (2014) A Channel for 3D Environmental
850 Shape in Anterior Inferotemporal Cortex. *Neuron* 84:55–62.
- 851 Watson DM, Hartley T, Andrews TJ (2014) Patterns of response to visual scenes are
852 linked to the low-level properties of the image. *NeuroImage* 99:402–410.
- 853 Wolbers T, Klatzky RL, Loomis JM, Wutte MG, Giudice NA (2011a) Modality-
854 Independent Coding of Spatial Layout in the Human Brain. *Curr Biol* 21:984–
855 989.
- 856 Yamins DLK, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ (2014)
857 Performance-optimized hierarchical models predict neural responses in higher
858 visual cortex. *Proc Natl Acad Sci* 111:8619–8624.
- 859 Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2015) Object Detectors Emerge in
860 Deep Scene CNNs. *Int Conf Learning Rep (ICLR 2015)*.
- 861 Zhou B, Lapedriza A, Xiao J, Torralba A, Oliva A (2014) Learning Deep Features for
862 Scene Recognition using Places Database. *NIPS* 27.
- 863

864 **7 TABLES**

| A | | |
|-----------------------------|----------------------|---------------------|
| | Onset latency | Peak latency |
| Clutter level | 56 (42 – 71) | 107 (103 – 191) |
| Luminance level | 644 (68 – 709) | 625 (146 – 725) |
| Contrast level | 53 (42 – 128) | 74 (68 – 87) |
| B | | |
| Size across clutter level | 226 (134 – 491) | 283 (191 – 529) |
| Size across luminance level | 183 (138 – 244) | 217 (148 – 277) |
| Size across contrast level | 138 (129 – 179) | 238 (184 – 252) |

865

866 **Table 1. Onset and peak latencies for MEG classification analyses.** Onset and peak latency (n
867 = 15, $P < 0.05$, cluster-level corrected, cluster-definition threshold $P < 0.05$) with 95% confidence
868 intervals. **A)** Clutter, luminance and contrast level representation time course information. **B)**
869 Time course of cross-classification for scene size. 95% confidence intervals are reported in
870 brackets.

871

872

873

874

875

876

877

| A | | |
|--|----------------------|---------------------|
| | Onset latency | Peak latency |
| GIST | 47 (45 - 149) | 80 (76 - 159) |
| HMAX | 48 (25 - 121) | 74 (61 - 80) |
| Deep object network | 55 (20 - 61) | 97 (83 - 117) |
| Deep scene network | 47 (23 - 59) | 83 (79 - 112) |
| B | | |
| Deep scene network minus GIST | 58 (50 - 78) | 108 (81 - 213) |
| Deep scene network minus HMAX | 75 (62 - 86) | 108 (97 - 122) |
| Deep scene network minus deep object network | - | - |

878

879 **Table 2. Onset and peak latencies for model-MEG representational similarity analysis.**
880 Onset and peak latency ($n = 15$, $P < 0.05$, cluster-level corrected, cluster-definition threshold $P <$
881 0.05) with 95% confidence intervals. **A)** Correlation of models to MEG data. **B)** Comparison of
882 MEG-model correlation for the deep scene network and all other models. 95% confidence
883 intervals are reported in brackets.

884

| Layer | Conv1 | Pool/ Norm1 | Conv2 | Pool/ Norm2 | Conv3 | Conv4 | Conv5 | Pool 5 | FC1 | FC2 | FC3 |
|---------|-------|----------------|-------|----------------|-------|-------|-------|-----------|------|------|-------------|
| Units | 96 | 96 | 256 | 256 | 384 | 384 | 256 | 256 | 4096 | 4096 | 683/ 216 |
| Feature | 55×55 | 27×27 | 27×27 | 13×13 | 13×13 | 13×13 | 13×13 | 6×6 | 1 | 1 | 1 |

885

886 **Table 3: Number of units and features for each CNN layer.** Units and features of the deep
887 neural network architecture were similar as proposed in (Krizhevsky et al., 2012). All deep neural
888 networks were identical with the exception of the number of nodes in the last layer (output layer)
889 as dictated by the number of training categories, i.e. 683 for the deep object network, 216 for
890 deep scene network. Abbreviations: Conv = Convolutional layer, Pool = Pooling layer; Norm =
891 Normalization layer; FC1-3 = fully connected layers. The 8 layers referred to in the manuscript
892 correspond to the convolution stage for layers 1-5, and the FC103 stage for layers 6-8
893 respectively.