# A self-supervised deep neural network for image completion resembles early visual cortex fMRI activity patterns for occluded scenes

**Michele Svanera** Centre for Cognitive Neuroimaging, Institute of Neuroscience and Psychology, University of Glasgow, UK

**Andrew T. Morgan** Centre for Cognitive Neuroimaging, Institute of Neuroscience and Psychology, University of Glasgow, UK

**Lucy S. Petro** Centre for Cognitive Neuroimaging, Institute of Neuroscience and Psychology, University of Glasgow, UK

**Lars Muckli** Centre for Cognitive Neuroimaging, Institute of Neuroscience and Psychology, University of Glasgow, UK

**The promise of artificial intelligence in understanding biological vision relies on the comparison of computational models with brain data with the goal of capturing functional principles of visual information processing. Convolutional neural networks (CNN) have successfully matched the transformations in hierarchical processing occurring along the brain's feedforward visual pathway, extending into ventral temporal cortex. However, we are still to learn if CNNs can successfully describe feedback processes in early visual cortex. Here, we investigated similarities between human early visual cortex and a CNN with encoder/decoder architecture, trained with self-supervised learning to fill occlusions and reconstruct an unseen image. Using representational similarity analysis (RSA), we compared 3T functional magnetic resonance imaging (fMRI) data from a nonstimulated patch of early visual cortex in human participants viewing partially occluded images, with the different CNN layer activations from the same images. Results show that our self-supervised image-completion network outperforms a classical object-recognition supervised network (VGG16) in terms of similarity to fMRI data. This work provides additional evidence that optimal models of the visual system might come from less feedforward architectures trained with less supervision. We also find that CNN decoder pathway activations are more similar to brain processing compared to encoder activations, suggesting an integration of mid- and low/middle-level features in early visual cortex. Challenging an artificial intelligence model to learn natural image representations via self-supervised learning and comparing them with brain data can help us to constrain our understanding of information processing, such as neuronal predictive coding.**

## Introduction

Investigating the functional dichotomy of forward and backward pathways in the visual system is necessary for describing how the brain performs vision. The conceptual divergence is broadly understood; forward pathways carry unlabelled sensory information into the brain while feedback pathways carry signals from higher visual and nonvisual areas back to earlier areas (Pennartz et al., 2019), in the opposite direction to sensory information. Feedback has an important role in the contextual modulation of the feedforward stream, carrying top-down signals such as attention and expectations (Van Essen & Anderson, 1995; Gilbert & Li, 2013; Roelfsema & de Lange, 2016; Takahashi et al., 2016; Angelucci et al., 2017; Klink et al., 2017; Morgan et al., 2019; Pennartz et al., 2019). Neuronal information processing in visual perception can be formalized as a statistical inference based on hierarchical internal models, which can theoretically be implemented by schemes such as predictive coding (Rao & Ballard, 1999; Friston, 2005). Predictive processing is a compelling framework for describing the neural phenomena observed in human primary visual cortex using brain imaging (Alink et al., 2010; Edwards et al., 2017). Predictive coding describes a neuronal coding process in which predicted information from top-down streams

Citation: Svanera, M., Morgan, A. T., Petro, L. S., & Muckli, L. (2021). A self-supervised deep neural network for image completion resembles early visual cortex fMRI activity patterns for occluded scenes. *Journal of Vision*, *21*(7):5, 1–17, https://doi.org/10.1167/jov.21.7.5.

https://doi.org/10.1167/jov.21.7.5 Received April 9, 2020; published July 14, 2021 ISSN 1534-7362 Copyright 2021 The Authors

is explained away from the feedforward stream (e.g., Rao & Ballard, 1999). Less narrow implementations of predictive coding recognize the necessity of the feedforward stream to not only compute prediction error, but also to communicate information that reinforces the internal models that were successful in predicting incoming information (Rao & Ballard, 1999; Shipp, 2016). We use the term predictive processing to allow for an even broader implementation of coding principles, including those that provide context (e.g., contour ownership, spatial information from auditory cues) without being limited to predicting a narrow set of precise features (e.g., a specific contour). For example, illusions of motion increase activity on the nonstimulated motion path in the primary visual cortex (Chong et al., 2016; Erlikhman & Caplovitz, 2017), and predictable stimuli presented on the motion path cause less activity than surprising stimuli (Alink et al., 2010; Edwards et al., 2017). Moreover, visual contours predicted by flanking stimuli increase activity along the illusory contour (Kok et al., 2016; Bergmann et al., 2019). These data reveal that some neuronal activity in early visual areas is not directly related to retinotopic sensory inputs, but rather to the brain's inference of the world, transmitted in cortical feedback pathways to earlier areas. A comprehensive biologically constrained artificial model of vision should account for this top-down brain processing.

In computational neuroscience, there is a tradition to exploit recent developments in statistical modeling and algorithms to develop more sophisticated models of visual processing (Simoncelli & Olshausen, 2001). In this respect, recent developments in artificial intelligence (AI), and in particular deep learning (DL), offered significant contributions in the last decade (Hassabis et al., 2017), showing promising results in the attempt to improve our understanding of visual counterstream computations (Kietzmann et al., 2019; Qiao et al., 2019). In the DL field, and more broadly in AI, a textbook categorization in the literature is the distinction between supervised and unsupervised learning; while the first learns statistical representations based on labelled datasets, including tasks such as classification, regression, and segmentation, the latter tries to extract features and inherent patterns from unlabelled data, with tasks such as clustering, anomaly detection, and dimensionality reduction (Hastie et al., 2009). Whereas supervised learning usually allows to obtain good performance in a predefined task, being able to make sense of the large amount of unlabelled data often available is a promising and exciting goal for the field. In this respect, the recent trend of self-supervised learning tries to use the data as a supervision strategy (Jing & Tian, 2020); the idea is to challenge the model to solve a specific unrelated task to learn representations of the data that can be later applied to solve supervised tasks, such as object classification or to automatically label the dataset. Examples include learning to predict some part of the image from other parts, predicting relative locations of two image patches, solving a jigsaw puzzle, and colorizing an image. Originally designed as unsupervised methods, in generative adversarial networks (Goodfellow et al., 2014), have proven successful in supervised and reinforcement learning tasks; in generative adversarial networks, a discriminator is trained by evaluating whether the data created by a generator is part of the training data (i.e., is a real image) or not (fake). Despite their great promise in understanding data, unsupervised learning methods are not often used for model comparisons with brain data.

Supervised learning models, under the flagship of convolutional neural networks (CNN), have advanced our understanding of visual information processing in the brain in a number of breakthrough studies (Hassabis et al., 2017). Using this approach to model spiking computations, CNN models can predict neural activations in the macaque visual ventral streams at early time points after stimulus presentation, suggesting they may capture important aspects of feedforward visual processing (Yamins & DiCarlo, 2016). Natural images have been found to cluster together in similar ways in the internal feature spaces of CNNs as in human inferior temporal cortex (Khaligh-Razavi & Kriegeskorte, 2014). Cichy et al. (2016) described how a CNN captured the stages of human visual processing in time and space from early visual areas towards the dorsal and ventral streams. Kay et al. (2008) and Güçlü and van Gerven (2015) compared two different encoding models to human functional magnetic resonance imaging (fMRI) data, one based on decomposing visual information into Gabor elements, and the other based on trained feedforward CNNs. Comparisons showed improved results of the supervised CNN with respect to the Gabor-based approach in explaining fMRI activity patterns. Using decoding modeling, different studies (Eickenberg et al., 2017; He et al., 2018) proposed to derive CNN representations from brain data, or they showed how to improve fMRI-based decoding performance mapping functional data and CNN features (Svanera et al., 2019). In a broader effort to validate CNNs as models of the visual system, networks supervised on image recognition were used to model differences in retinal and cortical networks, showing which architectural constraints help similar representations to emerge in CNNs as found in the brain (Lindsey et al., 2019). To better navigate between different object classification artificial networks, (Schrimpf et al., 2018) introduce multiple neural and behavioural benchmarks to score any artificial neural network on how similar it is to the brain's mechanisms for core object recognition. CNNs with recurrent connections better predict visual responses than feedforward models, and increase our ability to capture the cortical dynamics in MEG data and fMRI data (Kietzmann et al., 2019; Qiao et al., 2019), opening new opportunities on understanding visual counterstreams by modeling brain architectures and processing principles (Hong et al., 2016; Tang et al., 2018; Kar et al., 2019).

Although these achievements are noteworthy in terms of being the best available models of biological vision, the possible impact of unsupervised models remains comparatively underexplored. As in AI, unsupervised learning models are promising in terms of data exploitation and understanding. Furthermore, in modeling vision we could potentially narrow the gap created by the biological implausibility of supervised learning objectives and gain progress in understanding the role of cortical feedback processing. In addition, the setting (i.e., the task) in which the two models, biological and artificial, are compared could provide a better testbed and bring potential benefits, such as giving insights on the learning dynamics or identifying differences with respect to other tasks. We begin in this direction by comparing brain data capturing contextual feedback in vision (in amodal filling-in) to an artificial model trained to fill-in missing visual information. The human brain imaging data were recorded from early visual cortex (V1–V2), and subjects viewed partially occluded natural scene images (Smith & Muckli, 2010; Muckli et al., 2015; Revina et al., 2018; Morgan et al., 2019). By partially masking the visual stimulus and recording from the corresponding retinotopic region of V1, this paradigm allows us to study feedback information in the absence of stimulus-specific feedforward signals. We use these data as a testbed for investigating a vision model that includes feedback components. We selected a visual occlusion paradigm because we sought comparable functionality between the brain data and computer vision algorithms performing the same task. The predictive coding framework suggests that the brain is trying to reconstruct the image under the occlusion, using information from cortical feedback processing to infer the content of missing image sections. We showed recently that the brain's filling of the missing image section can be modelled as a behavioural line drawing (Morgan et al., 2019). This operation of reconstructing an unseen image is conceptually similar to the task of inpainting (also known as image completion) in computer vision. In this task, an artificial model predicts the missing part of the image (unseen or damaged; Pathak et al., 2016) relying on image statistics learned during training. This learning can be obtained using different techniques: with classical image processing techniques, such as the application of statistically similar patches (He & Sun, 2014), or with CNN approaches (Isola et al., 2017; Lempitsky et al., 2018; Yu et al., 2018). To solve the task of image reconstruction, we used a CNN with encoder/decoder architecture, trained in a self-supervised fashion as in Isola et al. (2017), to perform inpainting on the lower right quadrant that is, reconstructing the unseen (occluded) portion of the image (Figure 1A). We compared the encoder/decoder network trained to solve inpainting with brain data collected in a previous 3-Tesla fMRI experiment during the viewing of images with the lower right quadrant occluded (Figure 1B); see (Morgan et al., 2019). We investigated similarities between the network and brain data using representational similarity analyses (RSA, Kriegeskorte et al., 2008; Nili et al., 2014).

## Material and methods

We compared representations of two neural network models to brain imaging data acquired during an fMRI experiment. We first describe the fMRI experiment (see *fMRI experiment*), before specifying the two CNN models: the well-known supervised (feedforward) network VGG16 and our encoder/decoder model trained to fill the missing quadrant of the image (see *Artificial neural network models*). Last, we explain how we compared the brain data to the CNNs using RSA (see *Data analysis: RSA*). The experimental framework is shown in Figure 1.

### The fMRI experiment

Our fMRI dataset was collected previously and published in Morgan et al. (2019). Eighteen healthy volunteers with normal or corrected-to-normal vision participated in this study. Twenty-four real-world scenes from six categories (beaches, buildings, forests, highways, industry, and mountains), from the dataset in Walther et al. (2009) were shown to participants (Morgan et al., 2019; see Supplementary materials for stimulation images). Images were displayed in grayscale on a rear-projection screen using a projector system. Stimuli spanned 19.5° × 14.7° of visual angle and were presented with the lower right quadrant occluded by a white box (occluded region spanned ≈9° × 7°). A centralized fixation checkerboard marked the centre of the scene images. To ensure central fixation and minimize eye movements, we instructed participants to respond via a button press to a temporally random fixation color change (Morgan et al., 2019). Over the course of the experiment, each image was presented 16 times. Refer to Morgan et al. (2019) for further details.[1]

#### *Data acquisition*

We collected the fMRI data previously at the Centre for Cognitive Neuroimaging, at the University of Glasgow. Participants gave written informed consent

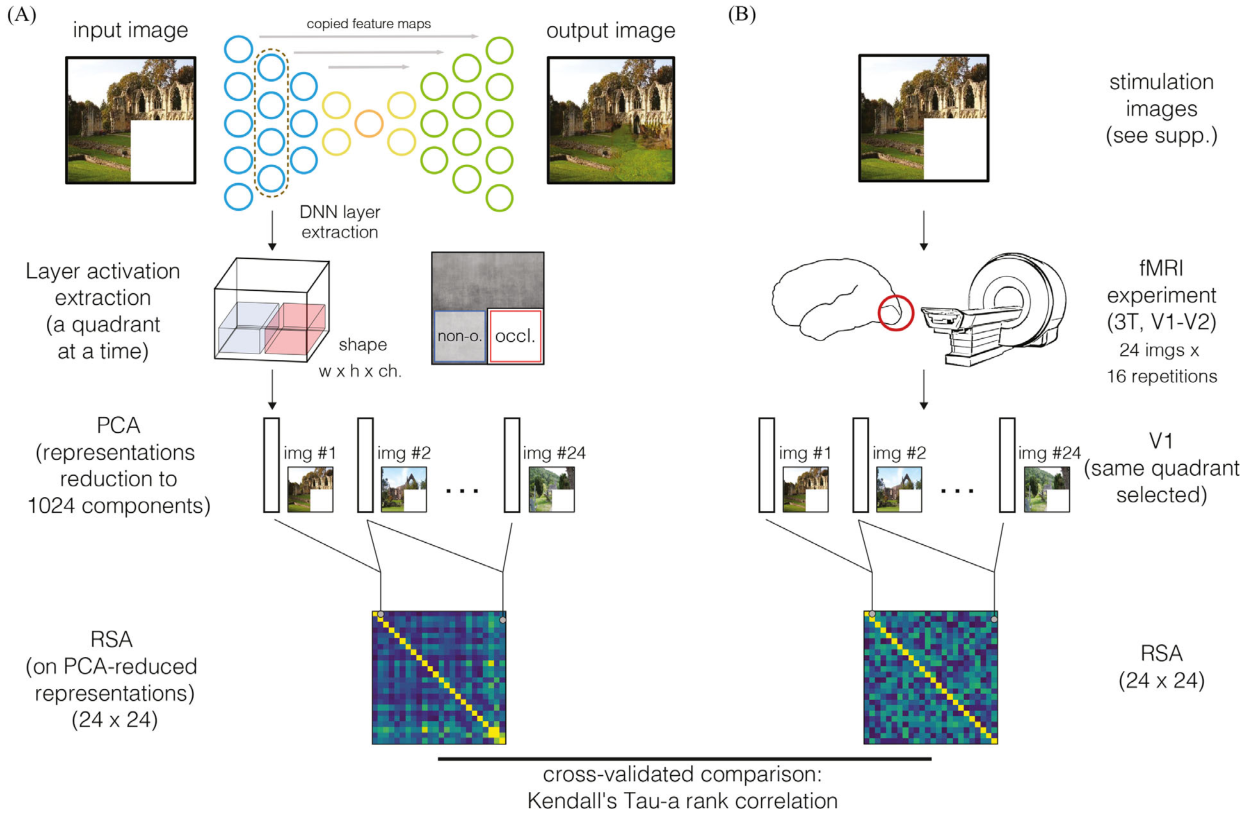


Figure 1. The analysis framework is composed of two parts: the encoder/decoder (artificial) neural network and brain imaging data collection. (A) The image passed through the network, we extracted activations for one layer, we selected a quadrant at the time, and we applied PCA transformation to reduce the dimension to 1024 components; we then obtained one 1024-d vector per layer (15), per quadrant (2), and per image (24). We used these vectors to compute representational dissimilarity matrices (RDMs). (B) fMRI data were collected from participants while viewing the same images that were fed into the network (in testing) and RDMs computed. We compared RDMs of the network and brain data (cross-validated; see Walther et al., 2016), for every CNN layer (15 layers analyzed), human visual area (V1 and V2), and image space quadrant (occluded and nonoccluded quadrants).

to participate, in accordance with the institutional guidelines of the local ethics committee of the College of Science & Engineering at the University of Glasgow (CSE01127). We used EPI sequences to acquire partial brain volumes aligned to maximise coverage of the visual pathway (18 slices; voxel size: 3 mm, isotropic; 0.3 mm interslice gap; TR = 1000 ms; TE = 30 ms; matrix size = 70 × 64; FOV = 210 × 192 mm). We used retinotopic mapping data (Sereno et al., 1995; Wandell et al., 2007) to locate the V1 and V2 (polar angle and eccentricity mapping). Additionally, we used three flashing checkerboard locations to map the cortical subregions of V1 and V2 corresponding with occluded and nonoccluded portions of the visual field. The first checkerboard was located in the bottom right visual quadrant (target mapping condition), the second was located only at the border (surround mapping condition; extending 2° visual angle into the occluded quadrant) and the third covered the remaining three quadrants (nonoccluded mapping condition). We declared cortical subregions that responded statistically more to the contrast of target versus surround conditions as occluded, that is, not stimulated with scene information. In our nonoccluded condition, the visual field was presented with scene information, activating the corresponding retinotopic region of V1 and V2.

### *Data preprocessing*

Functional data passed through different preprocessing steps; slice time, three-dimensional motion correction, and temporal filtering (high pass), before being normalized to Talairach space. We used data from retinotopic mapping runs to define early visual areas V1 and V2 using linear cross-correlation of eight polar angle conditions. To define the regions of interest corresponding with the nonoccluded and occluded quadrants, lower left and right image sections, respectively, we computed population receptive fields (pRF; Dumoulin & Wandell, 2008) and excluded voxels whose response profiles were not fully contained (within $2\sigma$ of their pRF center) by the respective visual regions of interest. The patterns of brain activities we obtained, which consisted of data from V1 and V2 for each of the two image quadrants analysed, were not preprocessed with any dimensionality reduction technique (contrary to CNN activations that went through a principal component analysis (PCA) analysis). We discarded the upper quadrants owing to the visual content of stimulation images (shown in supplementary). Most of the images included depict landscapes, and the upper parts of the images therefore included large portions of sky (i.e., uniform values).

## Artificial neural network models

### *VGG16: A supervised image classification network*

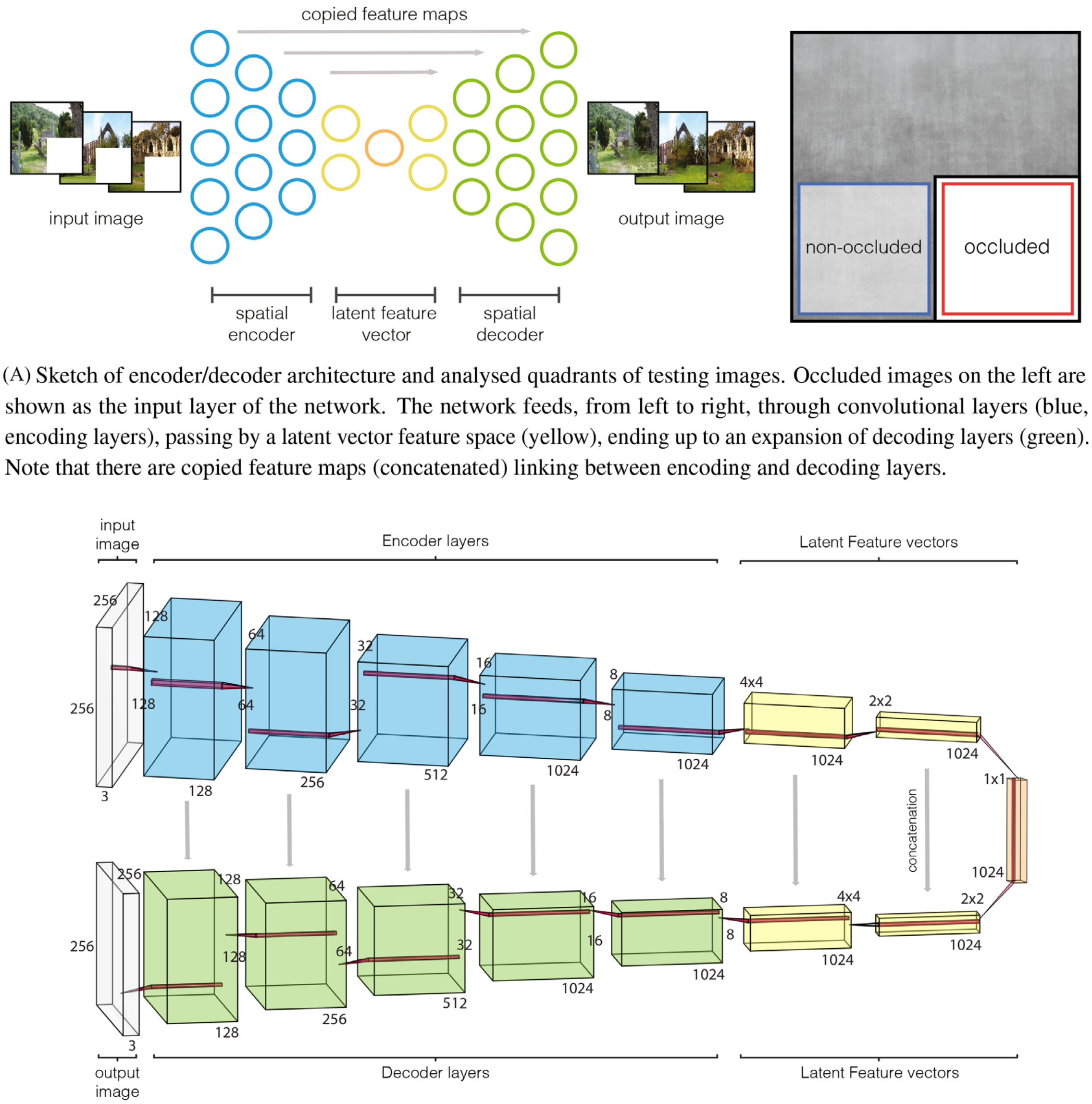
To compare brain data with a supervised network, we used the VGG16 network (Simonyan & Zisserman, 2014), a well-known model used for comparing visual pathway activity with CNNs (Güçlü & van Gerven, 2015; Cichy et al., 2016). The network is a supervised model (Simonyan & Zisserman, 2014) pre-trained to solve image classification – the task to attribute an object class label to an image – on the ImageNet database with 1000 classes (Russakovsky et al., 2015). The architecture is strictly hierarchical – image-to-labels — with 16 convolutional layers, and it uses repetitive and simple base structure made by 3 × 3 convolutional layers, for a total of ≈138 M parameters. In our analyses, we used only convolutional layers before pooling, leading to a total of 5 layers analysed (layers: `conv1_2, conv2_2, conv3_3, conv4_3, conv5_3`).

### *Encoder/decoder model: A self-supervised image completion network*

Our self-supervised image-to-image model is trained to reconstruct the occluded image section (always the lower-right quadrant); it is a fully-CNN, with a encoder/decoder architecture and with skip connections (known as U-Net and described in Ronneberger et al. (2015). Details on layer activation dimensions are shown in Figure 2B. Every layer implements the same components: a two-dimensional convolution with a 3 × 3 filter and stride= 2 (that downsamples the activations, no pooling used), batch normalisation, and a rectified linear unit (ReLu) function for encoder and a leaky ReLu[2] for decoder. We added padding to keep the activation dimensions constant after convolution and adopted dropout during the training to decrease overfitting.

We trained the network similarly to what has been proposed in Isola et al. (2017), in which not only the mapping from input image to output image (i.e., network filters) is learned, but also the best loss function to evaluate (or train) this mapping. The selection of the learning strategy, and in particular the loss, is crucial in unsupervised learning, where labels or categories are not available. Early unsupervised models adopted custom loss functions hand crafted by the experimenter; for example, autoencoders use the mean squared error (MSE) between the original image and the generated to compute the loss and the relative gradient. Variational autoencoders (Kingma & Welling, 2013), a combination of autoencoders with statistical inference, add a term to the MSE loss measuring how closely the latent variables match specific distributions (for example, unit gaussian). However, MSE produces blurry images and VAEs fail to generate good-looking images because they are not able to parametrise precisely the complex distributions of images. A successful approach to overcome the hand-crafted selection of a loss function is to let the network learn it: this can be accomplished using conditional generative adversarial neural networks (Goodfellow et al., 2014). The training was carried out using images from the SUN database (Xiao et al., 2010, the same database where the testing images came from), with occluded images as input to the network and original images (without occlusion) as output. We discarded some images from the database because they were too low resolution ($< 200 \times 200$ pixels), leading to a total of ≈107, 000 images for the training.[3] We trained the model and its 228M parameters with epochs = 5 and batch_size = 5 and the training process lasted for ≈49 hours on a GeForce 1080Ti. All the code is implemented in tensorflow. An overview of the training and testing procedures is shown in Figure 3 of the Supplementary material.

We extracted layer activations from both encoder and decoder streams, as shown in Figure 1 for two quadrants of the image: lower left (nonoccluded), and lower-right (occluded). The contractions in the architecture produce an increase of the CNN RFs as the layers increase, causing overlaps between quadrants. To avoid confounding results, we selected only an area within the quadrants, removing borders between them. Calculating the RF of the network (see the Supplementary material for RF sizes), we noted how after the fifth encoder layer, the RF size (i.e., the region of the input image which contribute to filter activations) spanned the entire image, and layers representations no longer contained spatiotopic activity maps (this factor was verified displaying activation maps). It was then not possible to distinguish quadrants by spatial location, so we decomposed the analysis into three network sections: the *spatial encoder*, in which activations could be split into quadrants; *latent feature vectors*, which included activations about the context only and without the spatial information; and *spatial decoder*, with again the concept of spatial separation in reconstruction.



(A) Sketch of encoder/decoder architecture and analysed quadrants of testing images. Occluded images on the left are shown as the input layer of the network. The network feeds, from left to right, through convolutional layers (blue, encoding layers), passing by a latent vector feature space (yellow), ending up to an expansion of decoding layers (green). Note that there are copied feature maps (concatenated) linking between encoding and decoding layers.

(B) Layer activation dimensions for encoder, decoder, and latent feature layers in blue, green, and yellow respectively.

Figure 2. Model architecture with layer dimensions. Every layer implements: convolution, batch normalization, and activation function (leaky relu for encoder and relu for decoder).

### *Dimensionality reduction of layer activations: PCA*

As described next (see *Data analysis: RSA*, we performed an RSA) through predictions of similarities between CNN and early visual cortex representations. Because layer activations from the CNN have large dimensions (see sizes in Figure 2B), we applied a dimension reduction technique to obtain stable and feasible RSA predictions in terms of variance explained (avoiding overfitting) and computation time. We therefore applied PCA to every layer activation of the set.

To obtain comparable comparisons in the RSA analysis, we reduced every layer activation to the same dimension, for every layer analyzed, as done in Cichy et al. (2016). The dimensionality reduction is learned randomly by selecting 10,000 images from the training set, extracting the corresponding activations, learning the transformation through an incremental PCA (Ross et al., 2008; Pedregosa et al., 2011), and eventually applying the transformation on activations extracted from the testing set. We tested different values for the number of principal components (within $2^n$, with

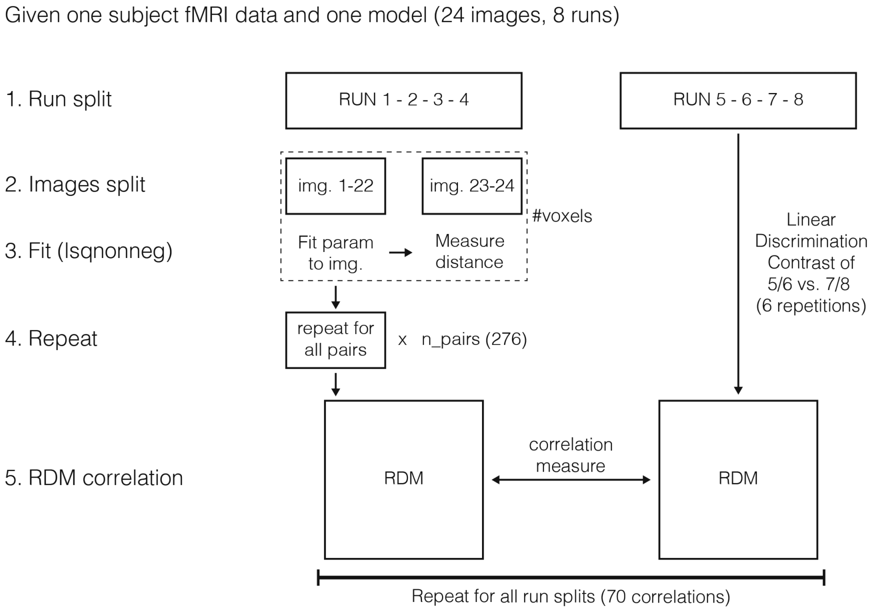


Figure 3. A schematic overview of the RSA computation.

$n = 3, 4, \ldots, 10$, i.e., 8, 16, ..., 1024 components), as reported in *Dimensionality reduction: PCA*.

## Data analysis: RSA

We computed representational dissimilarity matrices (RDMs) from fMRI data and PCA-reduced CNN activations using the linear discriminant contrast method (Walther et al., 2016). For each scene comparison, we fit models to the RDM of the 22 other scenes in the first set (e.g., runs 1/2 vs. 3/4) using non-negative least squares (Khaligh-Razavi & Kriegeskorte, 2014). We repeated this analysis for all scene comparisons, producing a predicted RDM based on model parameters, which was then compared with an RDM produced from the other half of the dataset (e.g., runs 5/6 vs. 7/8) using Kendall's Tau, a rank correlation (Nili et al., 2014). We repeated this procedure for all 70 possible split-quarter combinations, and averaged values over splits to produce one correlation value per subject per region of interest and model combinations (Morgan et al., 2019). A brief outline of the computation is shown in Figure 3.

# Results

First, we show the visual results of the network output and a description of what layer activations represent (sec *DNN model training*). Then we show similarity performance which drove the selection of the number of PCA components (sec *Dimensionality reduction: PCA*). Last, we compare VGG and our encoder/decoder in terms of RSA and encoder/decoder layers (see *Comparison between VGG16 and encoder/decoder*).

## DNN model training

A graphical result to demonstrate the quality of the output of the network is shown in Figure 4 for color images. However, to create a model that might capture fMRI brain activation to occluded regions of images, we trained a network to reconstruct grayscale images with occlusion. Images used in the experiment were not part of the training set.

To create a model that might capture fMRI brain activation to occluded regions of images, we trained a network to reconstruct grayscale images with occlusion. Images used in the experiment were not part of the training set.

To better understand the processing performed by the network, we show a selection of stimuli eliciting maximal activations for every layer of the network encoder in Figure 5. Patches were obtained as follows: given activations for a specific channel[4] from 10,000 images of the training set, we found the top five activations (maximum responses for that filter). Starting from the location of the maximum, we found the corresponding patch (the RF) in the image space that caused that activation. In the classical hierarchical structure of different feedforward networks, such as VGG16 (Simonyan & Zisserman, 2014), it is possible to find features from low to high level of complexity. In this

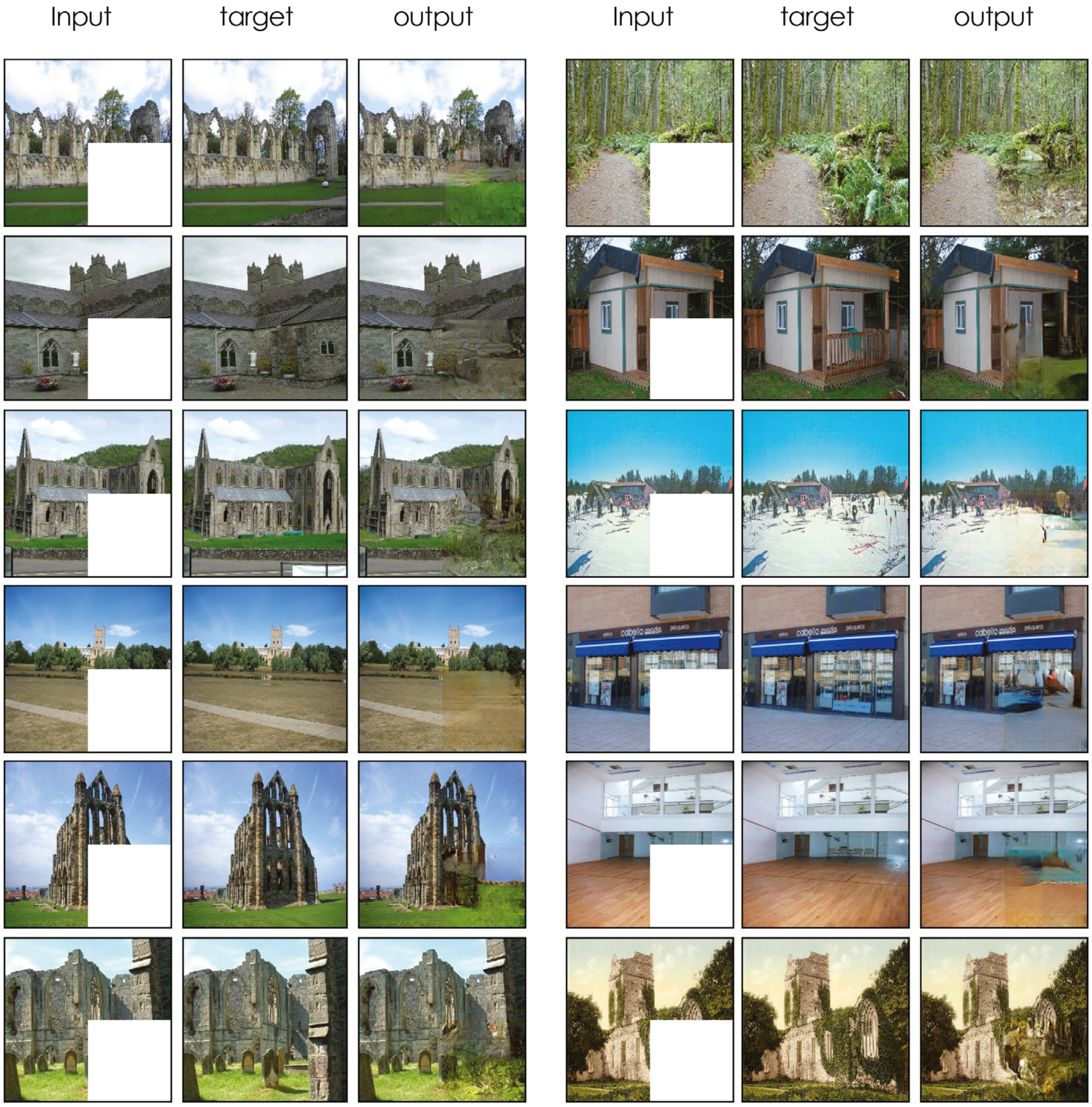
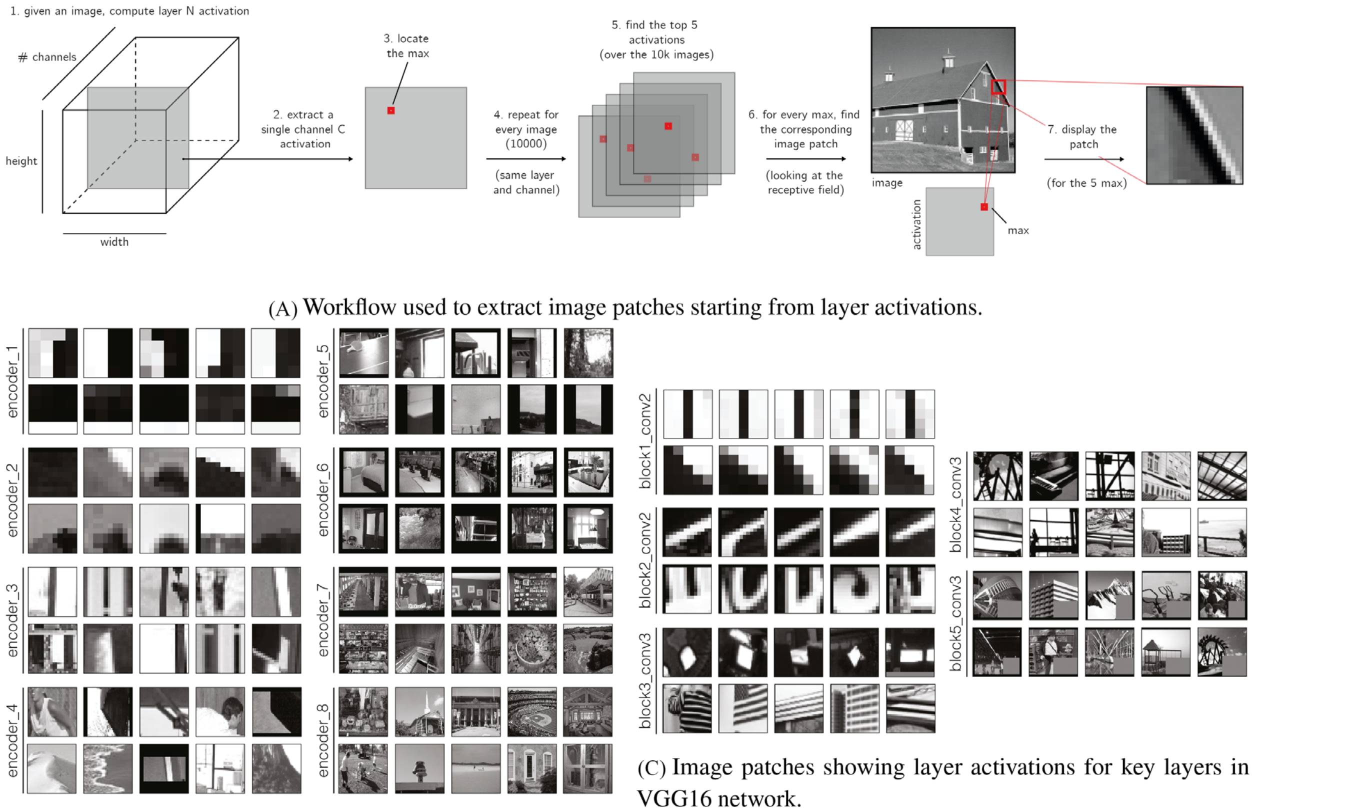


Figure 4. Model results with the input to the network (occluded images), the original image (i.e., shown here only for graphical comparisons of input and output with the original image), and the output of the network. Images used in this figure were not part of the training set. Please notice that images are here shown in colour for displaying purposes only; the comparisons below are made with activations from a model trained on grayscale images, consistent with the brain imaging experiment.

network, instead, features seem to be more dedicated to the detection of edges even in the middle and upper layers. Only in layer `encoder_6` to `encoder_8` do features start to become more complex, incorporating larger areas of the scene.

The functional characterization of these blocks can be challenging, but a possible interpretation follows. The model architecture is composed of two branches: encoder and decoder, also known as contraction and expansion paths. The encoder takes in an original image and produces a compressed vector, or latent variable vector; the decoder takes this vector and tries to reproduce the original image. The encoder branch increases, layer by layer, the representation of "what," that is, the feature complexity, and decreases the "where," that is, the localization of a specific feature in the image space. In terms of early visual processing, this is equivalent to compressing retinotopic features into high-level representations. In contrast, the expansion path creates a high-resolution image on the output through a sequence of up-convolutions (that is, transposed convolution) and concatenation with high-resolution features from the contracting path, that is, reconstructs retinotopic features from high-level representations. We chose a contraction (similar to what happens with CNNs for object recognition (Krizhevsky et al., 2012) and an expansion path, to accomplish increased computational efficiency, rather than learning a full-size convolutional net, forcing the network to learn features at different scales (Noh et al., 2015).


1. given an image, compute layer N activation
# channels
height
width
2. extract a single channel C activation
3. locate the max
4. repeat for every image (10000)
(same layer and channel)
5. find the top 5 activations (over the 10k images)
6. for every max, find the corresponding image patch
(looking at the receptive field)
image
activation
max
7. display the patch
(for the 5 max)
encoder_1
encoder_2
encoder_3
encoder_4
encoder_5
encoder_6
encoder_7
encoder_8
block1_conv2
block2_conv2
block3_conv3
block4_conv3
block5_conv3

(A) Workflow used to extract image patches starting from layer activations.

(B) Image patches showing layer activations for every layer of the encoder stream of our network.

(C) Image patches showing layer activations for key layers in VGG16 network.

Figure 5. (A) Schematic overview of the method used to visualize layer features: the method identifies, for a given activation layer, the five maximum unit activations, and visualizes the image region corresponding with the RF of these maximum points. (B) Image patches obtained for the encoder stream of our network: for every layer of the encoder, two channels are displayed. (C) The same method is applied to visualize the features within key layers of VGG16 network. These images are randomly selected from a bigger pool of analysed layers; the remaining are displayed in the Supplementary material.

### *Dimensionality reduction: PCA*

Considering the type of RSA conducted (see *Data analysis: RSA*) to avoid overfitting and decrease the amount of regularization needed, we decreased the layer activation dimensions to a lower dimension. Therefore, we applied different numbers of principal components (8, 16, . . . , 1024), and tested the performance achieved (in terms of similarity, Figure 6). The results describe the mean and the standard deviation of similarity across every layer analyzed – between V1 and CNN layer RDMs – for every number of components, in terms of Kendall’s tau-a. Results are averaged across all subjects. The two sections (i.e., subplots) are nonoccluded and occluded: the difference is the relative portion of the cortex analysed through the selection of the correspondent RFs (pRF analysis). In the nonoccluded section, we analyzed the corresponding region of retinotopic cortex that received feedforward stimulation from the image; in the occluded, the region of cortex that processed only the occluder.

As expected, with increasing the number of components, correlations become higher and more stable for both V1 and V2. Based on this finding, we kept the number of components equal to 1024. This strategy allowed us to explain the most possible variance in CNN layer activations. Note that 1024 is the maximum reachable number of components, as it is equal to dimensionality of the smallest layer (see Figure 2B). All results are reported using this number of components; results for VGG are here omitted for brevity, but the same logic is applied.

## Comparison between VGG16 and encoder/decoder

We assessed the ability of the two models to describe brain data using RSA. The first was VGG16 – a supervised network designed for image classification. The second was our encoder/decoder scheme – a model trained in a self-supervised fashion to fill in an occluded

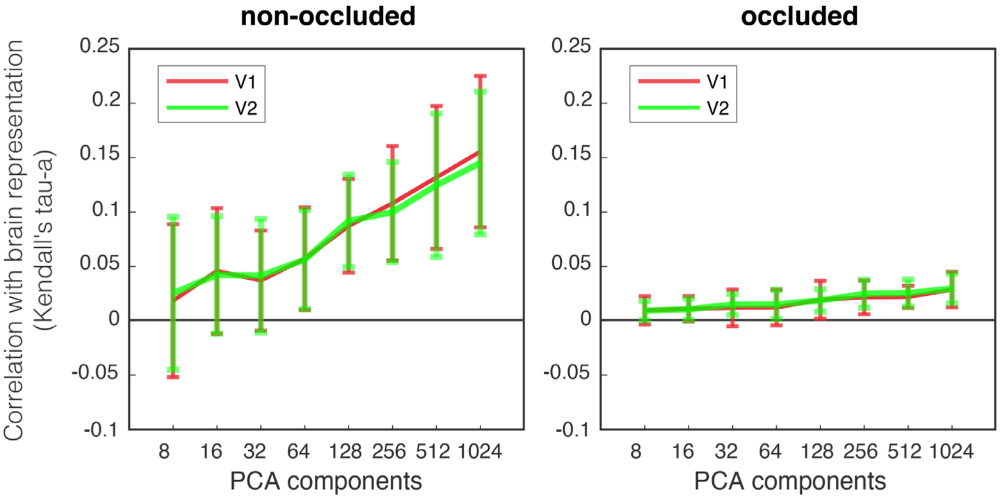
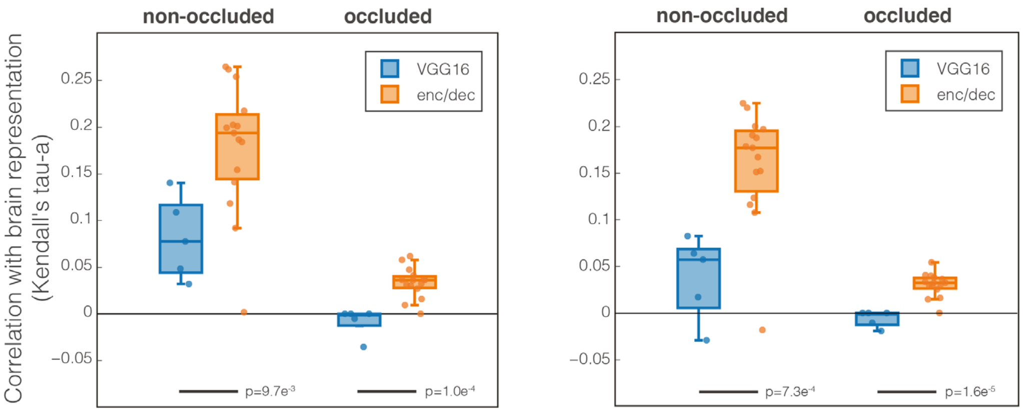


Figure 6. Similarity results (between V1 and CNN layer RDMs) for the nonoccluded and occluded quadrants in terms of Kendall’s tau-a. For every component in the range $2^n$, with $n = 3, 4, \ldots, 10$, we report mean similarity and standard deviation averaged across layers and subjects.



(A) V1

(B) V2

Figure 7. Kendall’s Tau-a rank correlation between VGG16 (blue) and our model (orange) with visual areas (A) V1 and (B) V2. Every dot is a CNN layer (5 conv for VGG16 and 15 for encoder/decoder); results are averaged across subjects. We performed a *t*-test to determine statistical significance of the difference between the models; results are reported below the bars.

portion of images. In Figure 7, we show VGG16 and encoder/decoder response similarities with V1 and V2 in terms of Kendall’s Tau-a rank correlation.

Focusing on the occluded quadrant, we observed a significant difference between VGG16 and our model [$p <= 0.0097$] (differences within our model between encoder and decoder are reported below). This result shows how our model, trained in a self-supervised way to perform inpainting, has representations more similar to the brain data compared with VGG16s, a trained supervised model for object recognition. In the nonoccluded area, a cortical region combining feedforward and feedback processing, we notice a remarkable increase of similarity between the brain data and our model. We, therefore, have greater similarity to brain data with our network representations rather than VGG16, for both occluded and nonoccluded areas.

### Encoder/decoder layers detail

Focusing on our model, we report results for every layer in Figure 8, analyzing the differences between encoder (or contraction path) and decoder (or expansion path) parts. We decompose results in the three sections, *spatial encoder* (blue), *latent feature vectors* (yellow), and *spatial decoder* (green), as described in *Encoder/decoder model: A self-supervised image completion network*.

Four plots are shown: the first row displays V1 results, and the second row displays V2; the two

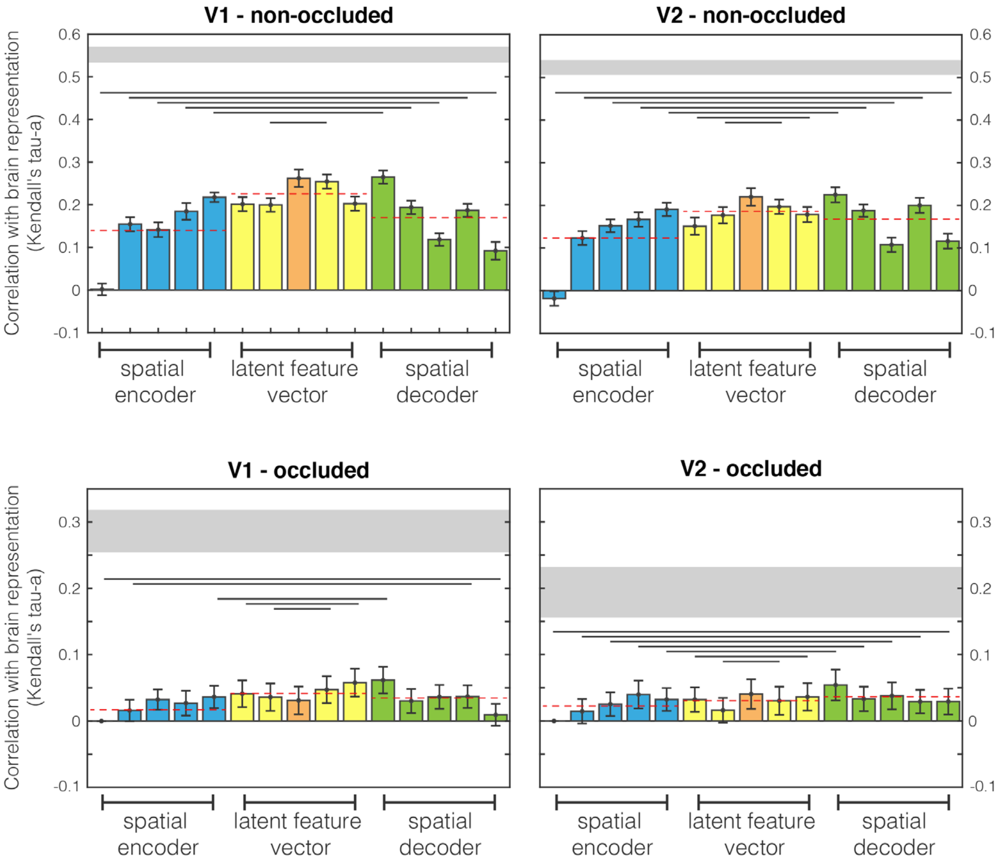
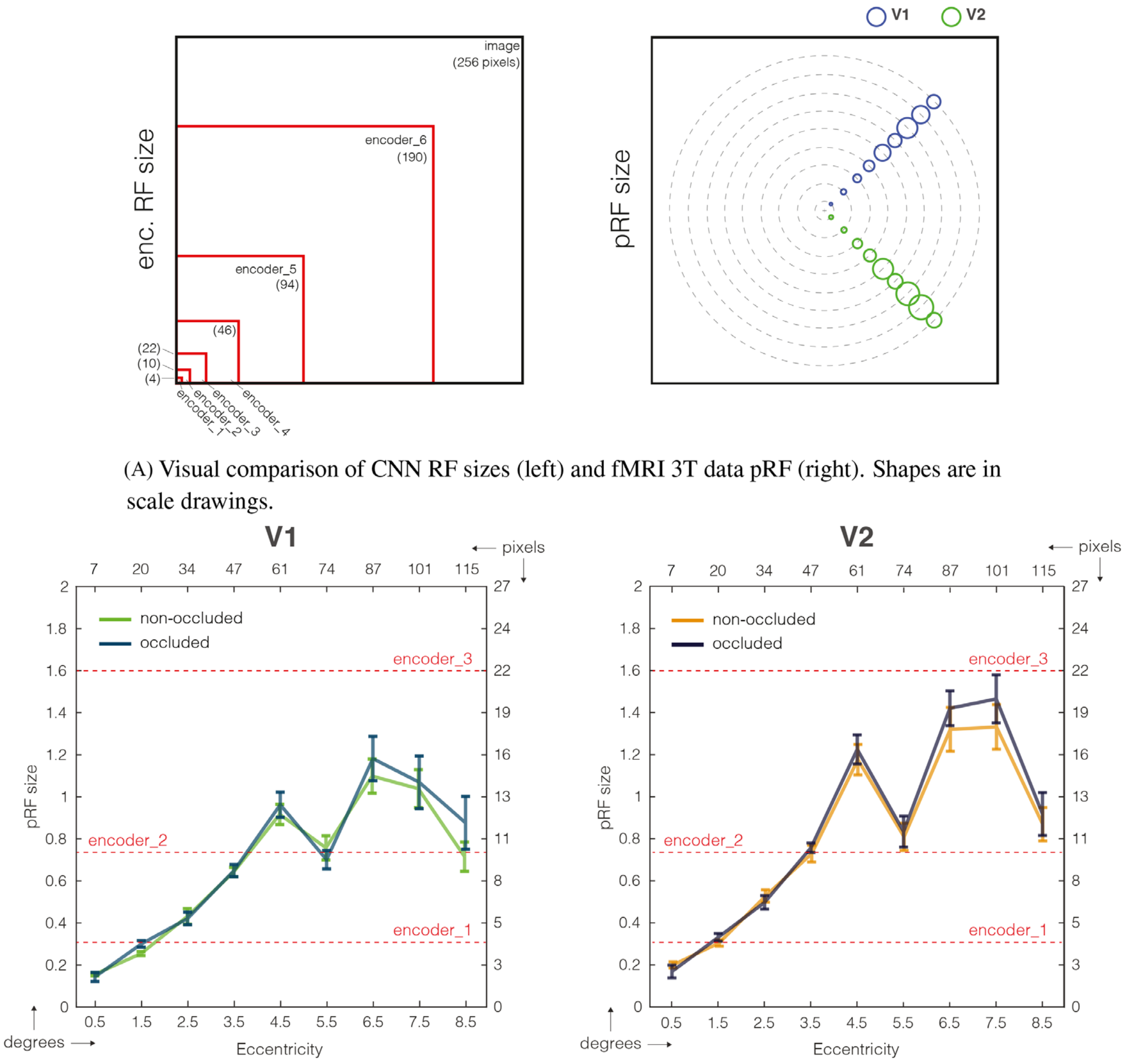


Figure 8. Comparison between brain and encoder/decoder network RDMs. The first and second columns are different human visual areas, V1 and V2, respectively. The first and second rows relate to nonoccluded and occluded quadrants. Results are highlighted with different colours for spatial encoder, latent features, and spatial decoder network sections. Dashed red lines are mean values for the three sections. Lines below bars show when the encoder and decoder populations difference is significant (Wilcoxon signed rank test). Grey horizontal shaded region indicates the upper and lower bounds of the noise ceiling, that is, the maximum correlation possible within this dataset.

columns correspond with nonoccluded and occluded quadrants. Every bar is an analyzed network layer and colours highlight the three different network sections: encoder (blue), latent feature vector (yellow, with the middle layer in orange), and decoder (green). The bar indicates the mean value across subjects, and the error bar the standard error (standard deviation divided by the square root of the number of subjects; i.e., 18).

Grouping layers together we can highlight large differences between sections. As depicted by dashed lines, decoder layers (which integrate information from the latent vector space – middle level features – and the correspondent encoders – lower level features) have greater and statistically significant similarity to brain data with respect to encoders. Surprisingly, the latent feature vector is even higher than decoder, showing the highest similarity with brain data (see Supplementary material for statistical test). Analyzing layer by layer, we notice that progressive encoding layers (blue bars) have better and statistically significant similarities to brain activity patterns (see Supplementary materials for statistical test). Once the encoder is devoid of retinotopic organisation owing to the absence of spatial information, there are the latent feature vectors (yellow bars), which only have context, or higher level features. With these layers, we have an immediate decrease in performance with the first layer after the retinotopic layers, which often follows an increase of similarity. In particular, it seems that in the occlusion, for V1, there is a decrease and increase of performance. The orange bar represents the inner layer activation (higher level) that in some cases represents the higher activation among the context layers. This is until the decoder layers with spatial information (green bars), in which the network reobtains retinotopic organisation. Here, there is a statistically significant overall decrease of similarity (see Supplementary materials for statistical test), in particular, after an initial increase, there is a stabilisation of similarity with brain activity patterns.

Interestingly, the first decoder layer is always the most similar to brain activity patterns. This is the first retinotopically organized layer after integrating the context information from latent feature layers. This



(A) Visual comparison of CNN RF sizes (left) and fMRI 3T data pRF (right). Shapes are in scale drawings.

(B) pRF estimation for fMRI data with overlaid the CNN receptive fields.

Figure 9. Both subplots show the same RF analysis in two forms: visually and numerically. (A) Visual comparisons of both RFs, CNN and V1/V2. The dimensions are in scale drawings so that it is possible to visually compare them. The indication of CNN RF sizes are displayed in parentheses. (B) Variation of fMRI pRF sizes in relation to the distance of the centre of fixation, in both degree and pixel units, for V1 and V2. Bars indicate the standard error across different subjects. Red lines indicate the first three CNN layer RFs, constant along eccentricity.

surprising result opens a series of possibilities for future modeling of brain data from early visual cortex using deep neural networks, such as using a multiscale approach or adding more complexity than simple edge detection to model V1 processing.

### *RF analysis*

We investigated whether there was a relationship between the size of V1-V2 RFs and encoder/decoder network RFs. In Figure 9, we show the RF analysis for (a) the CNN encoder layers and (b) the fMRI 3T data. RF has here two different meanings, based on the data on which it refers to. On fMRI data, it specifies the size of the visual field that each voxel captures; this is why voxels processing the fovea region have smaller RF size than the peripheral ones. Instead, in CNNs, every unit in a convolutional layer only depends – and processes – a specific region of the input image, called the RF. Note that every network layer processes the output of the previous layer, not the original image. In the encoder branch, because “arriving” skip connections are not present (only “departing”), every layer elaborates already processed information because this branch is a cascade of consecutive layers.

In the Figure 9, we can observe how the RF size of the fMRI data is increasing from the fovea to the periphery (we instructed participants in a central fixation task during the fMRI experiment), for both V1 and V2 (range = {3, 16} pixels for V1 (blue) and {3, 21} for V2 (green)). These data show high similarity with the RF sizes of the first three network layers (`encoder_1` to `encoder_3`, respectively 4 to 22 pixels, see Supplementary material for a full list). This may seem to be in opposition with the results shown in Figure 8, where the similarity increases going from `encoder_1` to `encoder_6`. This finding means that the increase of similarity further in the network is not due to the matching of the RF sizes, but to something else. We can read this result as evidence of the integration of mid- and low/middle-level features in early visual cortex, because in the decoder layers there is an integration of two different streams of processing: bottleneck and skip connections information.

## Discussion

We investigated the similarity in representations between an artificial neural network trained to fill occlusions and early visual cortical fMRI activity acquired from humans viewing occluded images. Specifically, we wanted to investigate whether an artificial neural network with encoder/decoder architecture would better approximate brain data acquired during a task involving cortical feedback signals, than a purely feedforward network. We have shown two main findings. First, forcing an artificial neural network to learn representations of natural image structure via self-supervised learning revealed increased similarity to brain data compared to a generic supervised network for classification. Furthermore, the CNN decoder pathway was more similar to brain processing than the encoder pathway. In fact, there was an increase in similarity going further along the network, in line with the integration of multiple feature levels in early visual cortex.

### Comparison between VGG16 and encoder/decoder network

As already speculated (Hassabis et al., 2017), a possible common ground for the AI and neuroscience communities is to challenge AI to replicate brain processing while performing common tasks. Having an AI solution could thus lead to better models of the brain for those tasks. With this idea in mind, we trained a network to solve inpainting and compared it with human brain activity while watching the same occluded images. We first compared a supervised network for image classification (VGG16) and a self-supervised network trained on the same task (our model) in terms of brain similarity. In both occluded and nonoccluded image locations, we obtained a higher similarity between our model activations and fMRI data; an expected result, considering the features learned by the CNN. We show in Figure 7 how our model outperforms VGG16, meaning that our model represents images more similarly to the brain in both nonoccluded and occluded quadrants. One of the possible reasons why our model has better similarities with the brain representations with respect to VGG16 could be linked to the type of task; speculatively, the brain performs the same task of the model, trying to reconstruct the image behind the occlusion.

### Encoder/decoder layers detail

What can our model tell us about brain processing? Our model attempted to reconstruct the occluded image section (lower right quadrant) using contextual information from the image surround (i.e., the remaining three quadrants) through several layers of processing. The model is composed of two branches, encoder and decoder, which analyzes images and reconstructs the missing part in the decoder. The network mainly learns low-level features, including edge detection and texture pattern recognition, at different resolutions (RFs), because these are the features considered useful to solve the task (i.e., reconstructing the image).

Going further with the analysis of our network, we tested whether the encoder (i.e., the contraction path) or the decoder (i.e., the expansion path) was more similar to the brain's representations. We found that the branch of the network more similar to the early visual cortex was not the portion compressing retinotopic features into higher level representations (i.e., the forward encoder pathway); but instead the portion reconstructing retinotopic features from higher level representations (i.e., the decoder pathway; see Supplementary material for statistical test). Predictive coding theories (Friston, 2008) suggest that the brain is trying to reconstruct the image under occlusion, using information in cortical feedback to probabilistically infer the content of missing image sections. Partial or complete occlusion, of objects for example, is commonplace in our visual environments. Predictive coding offers a neuronal mechanism by which the brain can minimize surprise by facilitating the negotiation between top-down predictions and feedforward input once the occlusion is removed. A small warning signal in a cluttered environment can become better detected if predicted information is filtered out. Consistent with such a neuronal framework, our data confirms some level of filling-in with information related to inpainting in computer vision.

### Eye movements and attention

In addition to the provided interpretation, there may be two external hypotheses for finding contextual information in nonstimulated regions of V1: eye movements and attention. Although we have confirmed over many studies that neither can account for our data (including Smith & Muckli, 2010; Muckli et al., 2015; Revina et al., 2018; Morgan et al., 2019), we provide possible disproofs of these hypotheses. In this line of research, it is essential that subjects do not make eye movements, and they categorically do not explore the entire image. To ensure that subjects fixate throughout the duration of the stimulation runs we used a central fixation point task while recording eye movements so that trials with saccades can be excluded from the analysis (see fMRI experiment). Aside from this fact, for eye movements to account for our data would require specific eye movements to each image exemplar and that these specific patterns were consistent across trials of those images (16 repetitions of 24 images shown to 18 subjects in the current case). Such a systematic pattern would be necessary to obtain consistency in brain-DNN similarities to still work. We have previously shown no significant differences in mean eye position across our occluded scenes (Smith & Muckli, 2010). In previous occlusion studies conducted in our laboratory (Smith & Muckli, 2010; Muckli et al., 2015; Revina et al., 2018; Morgan et al., 2019), the subjects overt attention is always on the central fixation point. Here, as in Morgan et al. (2019), subjects had to pay attention to a temporally random fixation color change, and to report the category of the scene being presented during the fixation color change using randomised response buttons. We assume a top-down attentional process monitoring the color of the fixation. There may also be (covert) attentional shifts toward the occluded region of the image, perhaps localized to regions of expected features behind the occluder. However, gain models of attention (that predict electrophysiologic data well) might result in increased signal magnitude or increased performance, neither of which we observe, or which accounts well for classification data. Further, we have shown in other studies, that when subjects direct attention away from a nonstimulated region of V1, there is still a filling in of expected information, likely owing to contextual feedback in Muckli et al. (2005) the case of apparent motion), demonstrating that contextual feedback processing is independent of attention, as expected.

### Future directions

In the future, it will be important to continue to improve network descriptions of the human visual system by employing more descriptive and specialized networks. Relatedly, we recently showed how line drawings provide a good description of internal model structure representing scene-specific features in human early visual cortex (Morgan et al., 2019). Being able to build a network able to predict sketches from an image would therefore provide additional value when comparing CNNs with brain data. In computer vision, hand sketches have been extensively studied, for example in sketch recognition, generation, and sketch-based image retrieval (Riaz Muhammad et al., 2018), and reveal that computers can classify line drawings in addition to digital images.

In terms of improving the biological plausibility of neural networks, recent work has also shown that CNNs with recurrent connections exhibit superior performance when recognizing occluded and nonoccluded objects (Spoerer et al., 2017) and that recurrent connections can help to describe cortical dynamics in early visual cortex (Kietzmann et al., 2019). Our plan for the future is to advance upon these important findings by training recurrent connections to predict occluded portions of images as well as behavioural sketches.

Along the same lines, a recently developed network has accurately predicted visual saliency using a similar encoder-decoder network (Kroner et al., 2020). Because we know the human visual system's coding of saliency must be robust to occlusion and clutter, it would be interesting to compare network objectives of predicting occluded visual features and predicting saliency. Such comparisons would allow us to understand whether aspects of these tasks could be shared by the same neuronal pathways.

## Conclusions

We investigated the representational similarity between a specialised artificial network built to reconstruct occluded images and fMRI data obtained while human subjects viewed the same stimuli. Results suggest that low- and mid-level features are present in early visual cortex (V1 and V2). Optimizing models to characterise feedback signals in human cortex will improve our understanding about the computations in early visual cortex where we know there is rich top-down information predicting feedforward input, that is currently not captured in feedforward networks. This work points to new experiments in which we challenge AI to replicate as closely as possible brain processing while performing cognitive tasks, testing models explaining memory, visual imagery, and auditory responses in early visual cortex.

*Keywords: self-supervised deep neural networks, encoder/decoder architecture, 3T fMRI, early visual cortex, representational similarity analysis (RSA)*

## Acknowledgments

Funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 720270 and 785907 (Human Brain Project SGA1 and SGA2).

**Author contributions**: All authors conceptualized the study. M.S. trained the network. A.T.M. collected and analyzed fMRI data. M.S. and A.T.M. computed the RSA analysis. M.S. and L.S.P. wrote the article, and all authors edited the article.

Commercial relationships: none.
Corresponding author: Michele Svanera.
Email: michele.svanera@glasgow.ac.uk.
Address: Centre for Cognitive Neuroimaging, Institute of Neuroscience and Psychology, University of Glasgow, UK.

## Footnotes

[1] Data available under EBRAINS knowledge graph at https://search.kg.ebrains.eu/instances/Dataset/c2df7995-f156-4953-bed1-fb4a003b364e.
[2] Leaky Relu: $y = f(x) = 0.6 * x + 0.4 * |x|$ (see figure in the Supplementary materials).
[3] The trained model and the code are available at https://github.com/rockNroll87q/self-supervised_inpainting.
[4] In this context, a channel – or feature map – is the resulting operation of a single convolution (input image with a single kernel). The number of channels of a layer activation is usually the third dimension after height and width (with two-dimensional input).

## References

Alink, A., Schwiedrzik, C. M., Kohler, A., Singer, W., & Muckli, L. (2010). Stimulus predictability reduces responses in primary visual cortex. *Journal of Neuroscience, 30*(8), 2960–2966.

Angelucci, A., Bijanzadeh, M., Nurminen, L., Federer, F., Merlin, S., & Bressloff, P. C. (2017). Circuits and mechanisms for surround modulation in visual cortex. *Annual Review of Neuroscience, 40*, 425–451.

Bergmann, J., Morgan, A. T., & Muckli, L. (2019). Two distinct feedback codes in v1 for 'real' and 'imaginary' internal experiences. *bioRxiv*.

Chong, E., Familiar, A. M., & Shim, W. M. (2016). Reconstructing representations of dynamic visual objects in early visual cortex. *Proceedings of the National Academy of Sciences, 113*(5), 1453–1458.

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.

Dumoulin, S. O., & Wandell, B. A. (2008). Population receptive field estimates in human visual cortex. *NeuroImage, 39*(2), 647–660.

Edwards, G., Vetter, P., McGruer, F., Petro, L. S., & Muckli, L. (2017). Predictive feedback to V1 dynamically updates with sensory input. *bioRxiv*.

Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017). Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage, 152*, 184–194.

Erlikhman, G., & Caplovitz, G. P. (2017). Decoding information about dynamically occluded objects in visual cortex. *NeuroImage, 146*(2016), 778–788.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences, 360*(1456), 815–836.

Friston, K. (2008). Hierarchical models in the brain. *PLoS Computational Biology, 4*(11), 1–24.

Gilbert, C. D., & Li, W. (2013). Top-down influences on visual processing. *Nature Reviews Neuroscience, 14*(5), 350–363.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., . . . Bengio, Y. (2014). Generative adversarial nets. In: Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (eds.), *Advances in neural information processing systems 27*, pp. 2672–2680. Cambridge, MA, USA: Curran Associates, Inc.

Güçlü, U., & van Gerven, M. A. J. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience, 35*(27), 10005–10014.

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron, 95*(2), 245–258.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer Science & Business Media.

He, K., & Sun, J. (2014). The statistics of similar patches. *Pattern Analysis and Machine Intelligence, IEEE Transactions on (2014), 36*(10), 1–13.

He, T., Kong, R., Holmes, A., Nguyen, M., Sabuncu, M., Eickhoff, S. B., Bzdok, D., Feng, J., . . . Yeo, B. T. (2018). Do deep neural networks outperform kernel regression for functional connectivity prediction of behavior? *bioRxiv*.

Hong, H., Yamins, D. L., Majaj, N. J., & DiCarlo, J. J. (2016). Explicit information for categoryorthogonal object properties increases along the ventral stream. *Nature Neuroscience, 19*(4), 613.

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. *CVPR*.

Jing, L., & Tian, Y. (2020). Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* pp. 1–1.

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience, 22*(6), 974–983.

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature, 452*(7185), 352–355.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Computational Biology, 10*(11), e1003915.

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K. A., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences, 116*(43), 21854–21863.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Klink, P. C., Dagnino, B., Gariel-Mathis, M.-A., & Roelfsema, P. R. (2017). Distinct feedforward and feedback effects of microstimulation in visual cortex reveal neural mechanisms of texture segregation. *Neuron, 95*(1), 209–220.

Kok, P., Bains, L., vanMourik, T., Norris, D., & de Lange, F. (2016). Selective activation of the deep layers of the human primary visual cortex by top-down feedback. *Current Biology, 26*(3), 371–376.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (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. In *Advances in Neural Information Processing Systems 25,* pp. 1097–1105. USA: Curran Associates, Inc.

Kroner, A., Senden, M., Driessens, K., & Goebel, R. (2020). Contextual encoder-decoder network for visual saliency prediction. *Neural Networks, 129*, 261–270.

Lempitsky, V., Vedaldi, A., & Ulyanov, D. (2018). Deep image prior. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* pp. 9446–9454.

Lindsey, J., Ocko, S. A., Ganguli, S., & Deny, S. (2019). A unified theory of early visual representations from retina to cortex through anatomically constrained deep CNNs. *arXiv preprint arXiv:1901.00945*.

Morgan, A. T., Petro, L. S., & Muckli, L. (2019). Scene representations conveyed by cortical feedback to early visual cortex can be described by line drawings. *Journal of Neuroscience, 39*(47), 9410–9423.

Muckli, L., De Martino, F., Vizioli, L., Petro, L. S., Smith, F. W., Ugurbil, K., Goebel, R., . . . Yacoub, E. (2015). Contextual feedback to superficial layers of v1. *Current Biology, 25*(20), 2690–2695.

Muckli, L., Kohler, A., Kriegeskorte, N., & Singer, W. (2005). Primary visual cortex activity along the apparent-motion trace reflects illusory perception. *PLoS Biol, 3*(8), e265.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS Computational Biology, 10*(4), e1003553.

Noh, H., Hong, S., & Han, B. (2015). Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision,* pp. 1520–1528.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., & Efros, A. A. (2016). Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 29*(34), 2536–2544.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Pennartz, C. M., Dora, S., Muckli, L., & Lorteije, J. A. (2019). Towards a unified view on pathways and functions of neural recurrent processing. *Trends in Neurosciences, 42*(9), 589–603.

Qiao, K., Chen, J., Wang, L., Zhang, C., Zeng, L., Tong, L., . . . Yan, B. (2019). Category decoding of visual stimuli from human brain activity using a bidirectional recurrent neural network to simulate bidirectional information flows in human visual cortices. *Frontiers in Neuroscience, 13*, 692.

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2*(1), 79–87.

Revina, Y., Petro, L. S., & Muckli, L. (2018). Cortical feedback signals generalise across different spatial frequencies of feedforward inputs. *NeuroImage, 180*(March 2017), 280–290.

Riaz Muhammad, U., Yang, Y., Song, Y.-Z., Xiang, T., & Hospedales, T. M. (2018). Learning deep sketch abstraction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 45*(6), 2064–2074.

Roelfsema, P. R., & de Lange, F. P. (2016). Early visual cortex as a multiscale cognitive blackboard. *Annual Review of Vision Science, 2*, 131–151.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention,* pp. 234–241. Springer.

Ross, D. A., Lim, J., Lin, R.-S., & Yang, M.-H. (2008). Incremental learning for robust visual tracking. *International Journal of Computer Vision, 77*(1-3), 125–141.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision, 115*(3), 211–252.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., . . . Schmidt, K. et al. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv, 22*(6), 407007.

Sereno, M. I., Dale, A., Reppas, J., Kwong, K., Belliveau, J., Brady, T., . . . Tootell, R. (1995). Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science, 268*(5212), 889–893.

Shipp, S. (2016). Neural elements for predictive coding. *Frontiers in Psychology, 7*, 1792.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience, 24*(1), 1193–1216.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Smith, F. W., & Muckli, L. (2010). Nonstimulated early visual areas carry information about surrounding context. *Proceedings of the National Academy of Sciences, 107*(46), 20099–20103.

Spoerer, C. J., McClure, P., & Kriegeskorte, N. (2017). Recurrent convolutional neural networks: A better model of biological object recognition. *Frontiers in Psychology, 8*, 1551.

Svanera, M., Savardi, M., Benini, S., Signoroni, A., Raz, G., Hendler, T., . . . Valente, G. (2019). Transfer learning of deep neural network representations for fMRI decoding. *Journal of Neuroscience Methods, 328*, 108319.

Takahashi, N., Oertner, T. G., Hegemann, P., & Larkum, M. E. (2016). Active cortical dendrites modulate perception. *Science, 354*(6319), 1587–1590.

Tang, H., Schrimpf, M., Lotter, W., Moerman, C., Paredes, A., Ortega Caro, J., . . . Kreiman, G. (2018). Recurrent computations for visual pattern completion. *Proceedings of the National Academy of Sciences, 115*(35), 8835–8840.

Van Essen, D. C., & Anderson, C. H. (1995). Information processing strategies and pathways in the primate visual system. *An Introduction to Neural and Electronic Networks, 2*, 45–76.

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., & Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. *Neuroimage, 137*, 188–200.

Walther, D. B., Caddigan, E., Fei-Fei, L., & Beck, D. M. (2009). Natural scene categories revealed in distributed patterns of activity in the human brain. *Journal of Neuroscience, 29*(34), 10573–10581.

Wandell, B. A., Dumoulin, S. O., & Brewer, A. A. (2007). Visual field maps in human cortex. *Neuron, 56*(2), 366–383.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). Sun database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on,* pp. 3485–3492. IEEE.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience, 19*(3), 356–365.

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., & Huang, T. S. (2018). Generative Image Inpainting with Contextual Attention. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* pp. 5505–5514.