Neural network-based encoding in free-viewing fMRI with precision models

Dora Gözükara[a,✉], **Nasir Ahmad**[a], **Katja Seeliger**[b], **Djamari Oetringer**[a], **and Linda Geerligs**[a]

[a]Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, the Netherlands
[b]Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

Representations learned by convolutional neural networks (CNNs) exhibit a remarkable resemblance to information processing patterns observed in the primate visual system on large neuroimaging datasets collected under diverse, naturalistic visual stimulation but with instruction for participants to maintain central fixation. However, this condition diverges significantly from ecologically valid visual behaviour, suppresses activity in visually active regions, and imposes substantial cognitive load on the viewing task. We present a modification of the encoding model framework, adapting it for use with naturalistic vision datasets acquired under fully natural viewing conditions – without fixation – by incorporating eye-tracking data and receptive field maps. Our *precision encoding models* were trained on the StudyForrest dataset, which features fully naturalistic movie viewing. By combining voxel-specific population receptive field estimates with eye-tracking data for each frame, we generate subject- and voxel-specific feature time series. These time series are constructed by sampling only the locally and temporally relevant elements of the CNN feature map for each voxel. Our results demonstrate that *precision encoding models* outperform conventional encoding models. This framework provides a strong foundation and justification for future large-scale data collection under fixation-free, fully naturalistic viewing conditions.

fMRI; CNN; encoding; naturalistic
Correspondence: ***dora.gozukara@donders.ru.nl***

Introduction

Cognitive neuroscience is increasingly adopting integrative research paradigms that combine naturalistic experimental designs with advanced analytical tools. For instance, the use of complex, multimodal and naturalistic stimuli in experiments, paired with the application of artificial neural networks (ANNs), has gained traction in recent years. Representations learned by convolutional neural networks (CNNs), in particular, have become a prominent model for understanding stages of information processing in the primate visual system ((1, 2)). These developments align with a broader effort to make experimental designs and analytical methods more ecologically valid. This paper builds on these advances by revisiting assumptions underlying brain-encoding models that rely on CNN features. Specifically, we propose that incorporating natural eye movements into the modelling process can significantly enhance the ecological validity of the underlying data, and predictive performance of these models. By integrating gaze dynamics and voxel-specific receptive field estimates, we demonstrate that CNN-based brain encoding can accommodate naturalistic stimuli without the need for restrictive fixation in the experimental designs. Our findings provide support for a more holistic approach to studying visual information processing, enabling future research to explore brain function under more ecologically valid conditions.

Naturalistic Experimental Design. Natural visual behaviour is inherently permanently active, with frequent gaze shifts to explore salient features in the environment (3). Despite this, almost all experiments underlying recent deep neural network (DNN)-based brain modelling studies imposed central fixation constraints during neuroimaging experiments to control attentional focus and minimize variability introduced by eye movements (4–9). While this approach ensures stimulus-response measurements that are simple to handle analytically, it diverges from natural viewing behaviour and imposes substantial cognitive load.

Fixation protocols also suppress activity in visually dynamic brain regions, potentially skewing insights into how the brain processes naturalistic stimuli. For example, (10) highlight that fixating on a static location during movie viewing disrupts the natural flow of perception. Nevertheless, some large neuroimaging data collections required participants to maintain fixation for extended periods, up to several dozens of hours (7, 11). Even in studies incorporating eye-tracking data, its use is typically limited to verifying fixation compliance or assessing participant alertness (12, 13).

Here we propose to use eye-tracking data to inform feature selection in DNN-based brain-encoding models. This approach allows for fixation-free experiments that better capture the dynamic and active nature of visual behaviour, offering a path toward more ecologically valid experimental designs and improved model accuracy. This research is particularly relevant since some of the largest ongoing neuroimaging data recording efforts have decided against fixation protocols for aforementioned reasons (14).

Holistic Analysis. Advances in artificial neural networks, particularly CNNs, have revolutionized our ability to model brain activity. Early brain-encoding models relied on manually engineered features or simple statistical representations (4, 5). By contrast, CNNs learn hierarchical visual representations directly from large-scale naturalistic data, making them a powerful tool for studying information processing in the visual system (1, 2, 15).

One key similarity between CNNs and the brain lies in their use of spatially selective receptive fields, where individual neurons or units respond to specific regions of the visual field.

DRAFT

For example, (16) showed that CNN units labelled as "face detectors" correspond to activity in multiple face-selective cortical areas, including the fusiform face area (FFA), occipital face area (OFA), and posterior superior temporal sulcus (pSTS-FA).

Despite these parallels, most fMRI studies using CNN-based encoding models ignore spatial selectivity by pooling features from entire CNN layers into single feature vectors to predict voxel activity (6, 16–20). This inflates model parameter spaces, requiring large datasets to achieve reliable fits and introducing ambiguities in feature selection, given that there are likely multiple combinations of features and locations that could explain the signal of a voxel. By sampling CNN features with voxels' eye-movement-adjusted receptive fields instead, we can significantly reduce parameter space complexity, improving both data efficiency and model interpretability.

Precision Brain Encoding. The primary goal of this paper is to harness the power of convolutional neural networks (CNNs) to model the brain's visual processing system in experimental conditions that are closer to naturalistic viewing scenarios, that is experimental neuroimaging scenarios that allow eye movements. Specifically, we introduce a method that modifies brain-encoding models to significantly reduce both the parameter space and thus the amount of data required while accommodating datasets where participants engage in free viewing. To demonstrate this, we use the publicly available StudyForrest dataset (21), which includes approximately two hours of fixation-free movie-watching runs, comprehensive eye-tracking data during the movies, and retinotopic mapping for each subject. This mapping enables us to estimate the receptive fields of individual voxels. Using a CNN pre-trained for natural image classification, we extracted features from all convolutional layers in response to the movie frames. This generated a time series of features for each spatial location and layer corresponding to each movie frame.

We sampled these features using subject-specific gaze data and voxel-specific receptive field estimates to construct personalized feature time series. This process involved sampling only the relevant portions of the visual scene and the corresponding time points for each voxel. The resulting feature time series were highly efficient in size, reducing feature data size by factors ranging from 112-fold for the smallest layer to approximately 30,000-fold for the largest layer.

Finally, we trained linear encoding models to learn optimal feature combinations from these reduced temporal feature series that predict voxel activity. By integrating gaze dynamics and voxel-specific receptive field properties, this pipeline not only addresses the variability across and within individuals but also enables robust predictions using substantially smaller datasets. This represents a meaningful step forward in making neural encoding models more efficient and practical for real-world applications, particularly in scenarios where traditional experimental constraints, such as fixation, are relaxed.

Methods

fMRI Dataset. We make use of the open 'Study Forrest' dataset which, alongside the taskless movie viewing sessions (of the movie Forrest Gump dubbed in German, presented across 8 runs of around 15 minutes), includes retinotopic mapping runs as well as eye-tracking for 15 subjects (aged 19 to 30). T2*-weighted echo-planar images (gradient-echo, 2s repetition time (TR), 30ms echo time, 90° flip angle, 1943 Hz/px bandwidth, parallel acquisition with sensitivity encoding (SENSE) reduction factor 2) were acquired during stimulation using a whole-body 3 Tesla Philips Achieva dStream MRI scanner equipped with a 32 channel head coil. 35 axial slices (thickness 3.0mm) with 80×80 voxels (3.0×3.0mm) of in-plane resolution, 240 mm field-of-view (FoV), (anterior-to-posterior phase encoding direction) with a 10 percent inter-slice gap were recorded in ascending order—practically covering the whole brain (21).

Movie Viewing Runs. The fMRI dataset is provided both preprocessed and denoised by (22). The pre-processing steps are described in detail in their paper and are briefly summarized here. Their pre-processing steps included motion correction, slice timing correction, brain extraction, and high-pass temporal filtering with a 200 s cut-off. They also used spatial ICA decomposition to remove head motion, physiology, and hardware related artefacts. We further preprocessed the data by carrying out co-registration with each subject's T1 anatomical image, followed by segmentation and normalisation to Montreal Neurological Institute (MNI) space. We used SPM 12 for all pre-processing steps above (www.fil.ion.ucl.ac.uk/spm). After anatomical normalisation, each voxel's timecourse within each run was z-scored independently. Finally, we concatenated the 8 runs resulting in a single functional timeseries. We limited our analyses to brain regions that were expected to encode visual information, based on the AAL atlas (23). This includes areas in the ventral visual stream, parts of the dorsal stream and the temporal lobe. We used the selected areas as a mask for both the movie viewing runs and the retinotopic runs.

Retinotopy Runs. To estimate the receptive fields of different voxels, we used the retinotopic mapping data that is included in the Study Forrest dataset. In these retinotopic mapping data, travelling wave stimuli were used to encode visual field representations in the brain via temporal activation patterns (24). Expanding/contracting rings and clockwise/counter-clockwise wedges consisting of flickering radial checkerboards (flickering frequency of 5 Hz) were displayed on a gray background to map eccentricity and polar angle.

To estimate receptive fields for each voxel we use the 'coarse to fine' method described by (25), as it is one of the most frequently used methods and is also computationally convenient. This method estimates the most likely horizontal (x) coordinate, vertical (y) coordinate, and size of a voxels pRF based upon an iterative process, starting with a coarse estimate and proceeding to a finer one. We used the x and y

coordinates of the fine estimate for each voxel. As our models also covered brain areas outside of the immediate visual stream, we did not exclude voxels using an explained variance threshold of each voxel's pRF model. The fine to coarse method estimates x and y coordinates in visual degrees in the range $[-10, 10]$. Based on the distance of the screen from the participants eyes (65 cm), we convert the visual degrees to pixel coordinates on the screen. The pRFs were estimated based on the native subject space images that were aligned with the movie-viewing-runs. After estimation, the x and y coordinates and the pRF size were transformed to MNI space.

Eye-Tracking. Eye tracking was performed using monocular corneal reflection and pupil tracking with an Eyelink 1000 (software version 4.594) equipped with an MR-compatible telephoto lens and illumination kit. More information about the details can be found in the original paper (21). To get a list of all fixation events per participant, and their coordinates on the screen, we used Remodnav (26). Remodnav is an algorithm for classifying eye-tracking events, which is founded on a pre-existing velocity-oriented approach, and can categorise saccades, post-saccadic oscillations, fixations, and smooth pursuit occurrences. For simplicity, we only used fixation events and ignored saccades or pursuits in our models. We used the default parameters of Remodnav to define our fixation events. We then matched the fixation event timestamps to the movie frames. In addition to determining the spatial location of our CNN features, we also used the fixation events to select which movie frames to include in our precision encoders. Instead of sampling the same equidistant frame set for all participants, we sampled the frames based on when each participant fixated. As fixation events span multiple frames, we only used the middle frames of each fixation event to extract the CCN features for the encoding model. Because of the low quality of the eye-tracking data of 2 participants, we excluded them from our study, leaving us with 13 participants.

CNN Features. Our proposed approach can in principle be used with any artificial neural network model which has accessible spatially variant features. In this work, we used a pre-trained CNN network but alternative architectures such as image transformers, generative adversarial networks, or diffusion networks could also be used.

Model: The CNN model we used to extract movie frame features is the Pytorch implementation of VGG-19, pre-trained to classify ImageNet images (27). We used the responses of 5 max pooling layers, each of which follows a block of convolutional layers, to build our encoding models. Fully connected layers were removed from the architecture as they lack spatial structure and have no spatial dimension. Furthermore, the fully connected layers of the model require images with a fixed 1:1 aspect ratio. Therefore, removing the fully connected layers made it possible to use the custom (cinematic) aspect ratio of 2.35:1 of the movie frames.

Input Images: To convert the movie stimulus into images that could be used as input to VGG-19, we cropped the grey bars above and below the movie frames, and downsampled them while respecting the aspect ratio (original (1280, 544), resized (527, 224)). We then applied the default VGG input image normalisation to the pixel values for each RGB channel with means of 0.485, 0.456, 0.406, and standard deviations of 0.229, 0.224, 0.225 respectively.

Feature Maps: After getting a list of all fixation events as described above, we obtained feature maps from our CNN model for each frame corresponding to the timepoint of fixation. The five feature maps are extracted from different layers of the CNN model, each with differing spatial and channel dimensions. We temporally binned these features by averaging all frame features that fall within a TR of the fMRI measurements. For computational feasibility, and to be able to combine features across different layers in our models, we fixed the spatial sizes of feature maps of different layers to the same size (7,16) by spatially averaging the features, and concatenated features from all layers to form a 'hyperlayer' feature map. However, precision encoding models could also be built for separate layers.

Precision Encoding Model. To create our proposed precision encoding models, we combine the population receptive field estimate of a voxel with gaze locations at each frame for every participant and each voxel. This allows us to extract a voxel and subject specific feature timeseries by sampling only the parts of the visual scene that fall in a voxels receptive field given the current subject gaze and timepoint. This feature timeseries is then used to train a linear encoder that estimates the mapping from features in the timeseries to voxel activity.

Herein we describe the pipeline by which this data is extracted and trained, see Figure 1 for a visualization.

Feature map extraction. Given N fixation timepoints, corresponding movie frames are extracted to create a dataset of images, $I \in \mathbb{R}^{N \times H \times W \times 3}$ where N is the number of frames, H is the frame height, W the frame width, and each pixel has three values (RGB). Each frame is then preprocessed (normalized and downsampled) and then passed through the CNN model (VGG-19) while collecting all activations at max-pooling layers. This results in a set of feature maps, A_0, A_1, A_2, A_3, A_4, each of which has a dimension of the number of frames, as well as independent dimension of height, width, and channel. In particular, $A_n \in \mathbb{R}^{N \times H_i \times W_i \times C_i}$ where subscripts of $i \in [0, 1, 2, 3, 4]$ indicate specific layer feature map heights (H_i), widths (W_i), and channels (C_i).

Finally, all feature maps are combined to create a "hyperlayer" feature map via a spatial down/up-scaling to a height and width of $\bar{H} = 7$, and $\bar{W} = 16$, respectively. All channel dimensions are concatenated, $\bar{C} = (C_0 + C_1 + C_2 + C_3 + C_4)$ This hyperlayer feature map can be denoted $M \in \mathbb{R}^{N \times \bar{H} \times \bar{W} \times \bar{C}}$, note that channel-wise concatenation simply results in a summed channel dimension.

Voxel pRF and Gaze Adjusted Features. Given a unified hyperlayer feature map, M, we make use of voxel pRF and gaze

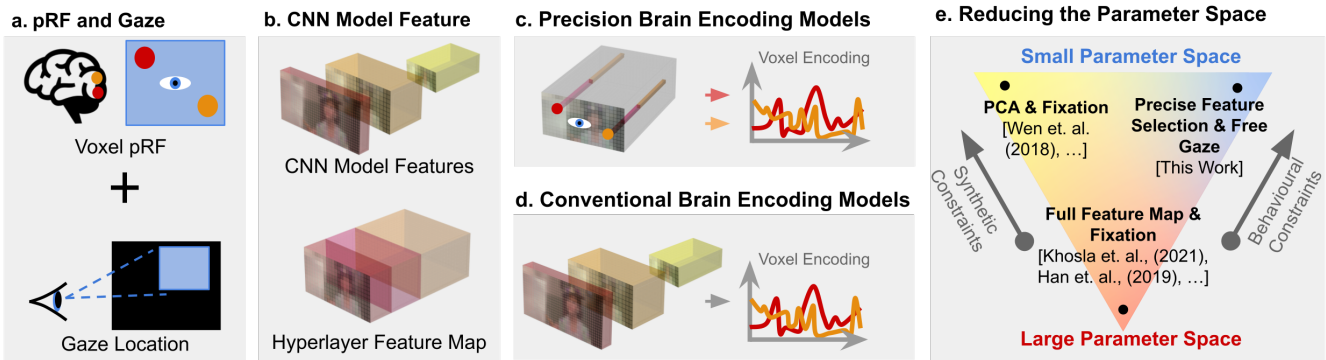

Fig. 1. Schematic figure that outlines the most important aspects of precision encoding models. (**a**) Each visual voxel has a receptive field (shown as red and orange dots). This information is combined with the unique gaze pattern for each individual. (**b**) For computational efficiency, and to be able to build one model across all layers, we combined CNN model features into a hyperlayer feature map. However, precision encoding models could also be built for separate layers. (**c**) This way, the precision encoding model in each voxel is based specifically on the CNN features that are relevant for information processed in that brain area. (**d**) Traditional CNN based encoding models use the whole feature sets of layers. (**e**) Diagram which shows how our approach compares to other approaches in the literature in terms of the number of parameters in the model and the type of constraints that are used to reduce the number of parameters (synthetic or behavioural constraints).

estimations in order to further refine the data used for the fitting of the encoding model.

Consider the voxel's pRF height and width coordinate as x_v^{pRF} and y_v^{pRF}, respectively where v indexes the particular voxel. Additionally, let the gaze coordinates for a given frame, n be represented by x_n^{gaze} and y_n^{gaze}. Both of these coordinates are assumed to be computed in pixel-space. The specific spatial location for a specific frame n and voxel v $(x_{n,v}, y_{n,v})$ is thus simply determined by:

$$x_{n,v} = x_v^{\text{pRF}} + x_n^{\text{gaze}} \quad \textbf{(1)}$$

$$y_{n,v} = y_v^{\text{pRF}} + y_n^{\text{gaze}} \quad \textbf{(2)}$$

Using these as (normalized) coordinates, one can extract a voxel and gaze specific feature set from the hyperlayer feature map, such that

$$X_{n,v} = M_{n,x_{n,v},y_{n,v},:} \quad \textbf{(3)}$$

where $X_{n,v}$ is a frame, voxel, and gaze specific feature vector. From this, the full feature vector for a given voxel across all frames, each with their own specific gaze, can be written $X_n = [X_{n,0}, X_{n,1}, X_{n,2} ... X_{n,V}] \in \mathbb{R}^{V \times \bar{C}}$, where V is the total number of voxels. From this, the full dataset for all frames and voxels can be written $X \in \mathbb{R}^{N \times V \times \bar{C}}$.

Linear encoding model. Using the voxel, and gaze specific set of feature vectors, X, one may build a linear encoding model:

$$Y = XW \quad \textbf{(4)}$$

where $W \in \mathbb{R}^{\bar{C} \times 1}$ is the weight matrix, and the encoding $Y \in \mathbb{R}^{N,V}$ represents the voxel-wise timeseries signal. Our choice of concatenating all feature maps into a hyperlayer allows our encoding model to train a single matrix of parameters, computationally simplifying the setup. Furthermore, alternative methods often train an encoding model for each individual layer (feature map) of a CNN where this method trains only one based upon the hyperlayer.

Finally, the optimal weights W that best describe the relationship between the stimulus X and the neural response Y can be determined via ridge regression using the Pearson correlation as a loss function (1-Pearson r):

$$\mathcal{L}(W) = (1 - \frac{\text{cov}(XW, Y)}{\sigma_{XW}\sigma_Y}) + \lambda ||W||_2^2 \quad \textbf{(5)}$$

Experiments and Baselines. Precision model: The number of fMRI samples to be predicted were 3,599 per voxel, each of which potentially included multiple fixation events. To combine the information across all fixation events within an fMRI sample duration, corresponding hyperlayer features for fixation events within this duration were averaged as described below. Note, the hyperlayer feature maps were constructed from successive VGG-19 maxpool layers with 64, 128, 256, 512, and 512 channels respectively (totalling to 1,472). Thus, each hyperlayer feature map had spatial dimension $\bar{H}, \bar{W} = (7, 16)$ and channel dimension $\bar{C} = 1472$. To construct the dataset for voxel activation prediction, averaging of features over multiple fixations was required. For all fixation events within an fMRI sample duration, the corresponding frames from the movie were processed (via the CNN model) and used to form a set of hyperlayer feature maps. Next, for each fixation the specific gaze and pRF locations were used to extract a feature vector (size $[1, 1, 1472]$) from each hyperlayer feature map at the corresponding location. These feature vectors were finally averaged over all fixations within a single fMRI sample duration for every voxel to form a full features stack of size $[\#\text{Voxels}, 1472]$. This feature stack, shifted 4.5 forward in time to account for the hemodynamic response function (HRF), was the final data used to predict voxel activity.

Baseline model: As a baseline model, we used the entire hyperlayer feature maps, without taking pRF or gaze location into account. The feature timeseries for the baseline model are still subject specific because they were sampled

from each participants' fixation events (just like in the precision model). We then trained a linear encoder that learns the best combination of features in this timeseries that predicts the activity of the voxel. The resulting feature timeseries had the dimensions of 3,599 timepoints by 164,864 features ($7 \times 16 \times 1472$).

Reverse Gaze Precision model: One important difference between the baseline model and the precision model is the size of the parameter space. To make sure this did not bias our results and to confirm that the combination of the voxel pRF and subject gaze provided a meaningful constraint on the features, we also provided a control model with the same number of parameters. Specifically, we constrained our parameter space by combining the pRFs with the eye-tracking data played in reverse (the gaze location of the last frame gets assigned to the first frame, and vice-versa). We then trained a linear encoder that tries to learn the best combination of features in this timeseries that predict the activity of the voxel. The resulting feature timeseries had the same dimensions as the precision model.

Determining the Ridge regularization space for each subject: Initially, we aimed to estimate 5 candidate regularization terms for each unique model (and therefore for each voxel) using singular value decomposition. Due to the computational load this introduced, we decided to estimate candidate regularization terms per participant. We created a "maximally active" feature timeseries by sampling the greatest activation in a feature map, instead of sampling the gaze adjusted pRF location like in precision models. We used this feature set to estimate the best regularization term for each subject for precision and reverse gaze models. Subject specific regularization terms for the baseline models were estimated using the whole feature set.

Analysis

Training and testing data. The fMRI dataset has 3599 total volumes (with 2 seconds TRs) divided between 8 runs. We separated the middle 20 percent of all timepoints (655 volumes) from each run to use as our test set. We further divided the training set into a validation set (20 percent, 574 volumes) and a training set (80 percent, 2306 volumes) to search for the best regularization value for ridge regression. We also removed 4 TRs from beginning and end of each section to remove the remaining autocorrelation between our test and training sets (64 volumes). Using the validation and training sets, we trained 5 different encoders using the same set of features but different regularization values and determined the best regularization parameter. We then trained the encoder with the entire training set using the best regularization value before measuring its accuracy (Pearson correlation) on the test set.

Statistical testing and visualization. As each of our datasets was projected to MNI space, we were able to test the significance of model performance across all 13 participants. For each model and voxel, we performed a one-sided (alternative hypothesis being the performance is greater than zero) paired t-test across participants to determine if a particular model can statistically significantly predict that voxels activity. We then applied False Discovery Rate (28) correction with the total number of voxels (19629) to correct for multiple testing. To test for significant differences across models, we first selected the voxels that were statistically significantly predicted by either of the three models. Within this smaller subspace of voxels, we performed a one-sided (alternative hypothesis being the performance difference is greater than zero) paired t-test across participants to determine if precision models are statistically significantly better than the baseline and reverse gaze models. In order to identify which areas are predicted better than the others, we used the Julich atlas version 2.9 (29) and computed the average performances of voxels within a region of interest (ROI) for each participant and model. We used the following six ROIs: V1 (human Occipital cortex 1 (hOc1)), V2 (hOc2), V3 (hOc3), lateral occipital (LO) areas (made up of hOc4p, hOc4la, and hOc5), fusiform gyri (FG; made up of FG1, FG2, FG3, FG4), and superior temporal sulcus (made up of STS1 and STS2). For generating the surface maps, we used the volume to surface projection tools of the Python package Nilearn (30), by projecting the fMRI data in MNI space to a common mesh.

Results

Model performances. We assessed the performances of different encoding models for each voxel by using a paired-sample t-test across all 13 participants. We found that precision encoding models predicted 42 percent of all voxels in our mask statistically significantly after FDR correction, whereas reverse gaze models predicted 18 percent, and baseline models predicted none. Figure 2a shows all voxels that are statistically significantly predicted by the precision and reverse gaze models on the cortex. Figure 2b shows the histograms of group average voxel performances for each model. Baseline models do not statistically significantly predict any voxels and therefore are not visualised. These figures show that precision models significantly encode voxels in a range of visual stream areas from as early as V1 to as late as LO, FG, and STS, while the reverse gaze models show significant encoding only in later parts of the ventral hierarchy. These results show that the precision encoding model massively outperforms the traditional approach to building encoding models that includes information from all CNN features.

The results show that precision encoding models results in a larger number of voxels being significantly modelled as well as an increase in the average model performances. Next, we also performed a direct comparison between the encoding model performances across the three models (see Figure 2c). We found that the precision models especially outperformed the baseline model the and reverse gaze models in areas with smaller pRFs (from V1 to V3). Reverse gaze models and baseline models did not have any statistically significantly different voxels.

In order to identify which areas are predicted better than the others, we computed the average performances of vox-

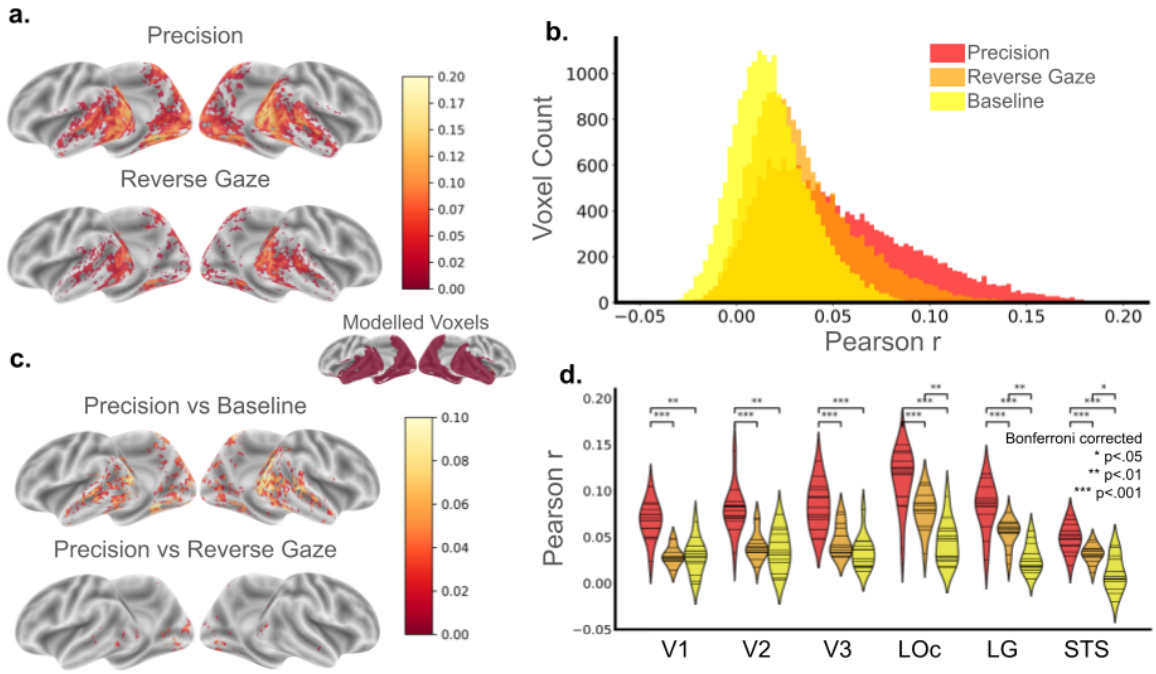

Fig. 2. Precision models outperform baseline and reverse gaze models. (**a**) Cortical maps show group average model performances for voxels that are statistically significantly predicted by each model. Precision models statistically significantly predict twice as many voxels as the reverse gaze models. Baseline models do not reach statistical significance. (**b**) Histograms showing group average model performances for all modelled voxels. (**c**) Cortical maps show group average precision model performances for voxels that are statistically significantly better than Top: baseline, Bottom: reverse gaze models. Reverse gaze models are not statistically significantly different than baseline models. (**d**) Violin plots show model performances in bilateral V1, V2, V3, LOc, and FG. Each violin plot shows the distribution of average model performance within an ROI across all participants. Each horizontal line shows one participant.

els within regions of interest (ROIs) for each participant and model (see Figure 2d). We found that the precision models performed best at LO, followed by FG and V3. Performance was lower for V1 and V2 and lowest for STS. Compared to the reverse gaze models, we found a clear performance improvement in all areas, but most strongly in V1, V2 and V3 where population receptive fields are smaller. Mirroring this, the reverse gaze models were not significantly different from baseline models especially at the early visual stream areas. This is further confirmation that gaze adjusted pRF sampling is an effective dimensionality reduction tool.

Single-subject model performances. To illustrate the actual model performances instead of a group averaged approximation, we plotted the precision model performances (see Figure 3a) and precision model and reverse gaze model differences (see Figure 3b) of example individual participants separately on the cortex. These figures show that performance is consistently high for most participants while one participant shows lower performance. This participant also had noticeably lower number of detected fixations, with the detected ones falling outside of the movie frame more often than the other participants. At the same time, we observe that the best performing voxels vary considerably between individuals. This suggests that the group-level statistical testing we performed may be conservative in describing the actual benefit of the precision encoding models.

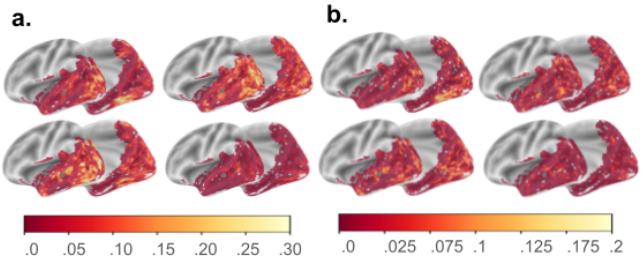

Fig. 3. Single subject model performances. Cortical maps show precision model performances (a) and precision model performance gains against the reverse gaze model (b) for four example participants. The best performing voxels vary considerably between individuals. Notice the bad performance for the participant on the bottom right both (a) and (b). Note that this participant had bad eye-tracking data.

Discussion

We have demonstrated that precision brain encoding models, which leverage CNN features accounting for spatial characteristics derived from eye movements and voxel receptive fields, outperform traditional spatially agnostic models across all participants in our dataset. This improvement is most evident in mid-level visual areas, such as the Lateral Occipital (LO) cortex, but also extends to both higher (e.g. LG) and lower-order (e.g. V1) areas. Our results also show that reducing the parameter space is an important step toward reaching high levels of encoding model performance with limited data. Even when we reduced the parameter space by including features from irrelevant spatial locations (in our reverse gaze model) we found that this improved model performance, compared to the spatially agnostic baseline approach. How-

ever, our precision encoding models still significantly outperformed the reverse gaze models, indicating that the reduction in parameter space alone cannot fully explain the improvements we observed. By reducing the parameter space in a spatially informed way, precision models are able to achieve meaningful performance with less data. Moreover, incorporating naturalistic behaviour, such as eye-tracking, into the model opens new avenues for studying the brain under ecologically valid scenarios, providing deeper insights into how visual processing works in real-world contexts.

Precision models compared to the literature. A direct comparison of our model with conventional models is not straightforward due to variations in data collection (fMRI recording parameters, static vs. video stimuli, fixation vs. naturalistic) and analysis methods (preprocessing, linear response modelling). Nonetheless, the statistically significantly predicted areas across the visual stream by precision models are very similar to studies which either use large (100+) group-averaged brain responses, or at least five times more scanning hours; most of which requiring fixation (16, 19, 20, 31). The first papers in the field which used DNN layers to predict BOLD activity have done so using fixation combined with static images that were presented multiple times (19). This allowed them to predict a wide range of the ventral stream, from V1 until LO. Later work has shown that encoding models can also be trained with naturalistic datasets, like different short video clips that participants watched while fixating (16, 20). Due to the dynamic nature of their stimuli, their models show significant prediction extending beyond earlier ventral stream areas and into temporal areas, similar to our data. More recently, a multimodal approach to brain encoding from naturalistic video has been proposed, which used longer movies without fixation (31). This study showed high encoding performance along both the ventral stream and into temporal areas, though performance was lower in early visual areas, potentially due to the eye movements. It should be noted that this study focused on group-average responses across 158 participants to reduce the large amount of variability caused by inter-individual variations such as eye-movements. Our own baseline models perform very poorly with the limited (2 hour) single subject data we used, with no voxels statistically significantly predicted. With this work we showed that across all subjects, precision models achieve statistically significant encoding accuracies with only two hours of fMRI data, and explain a similar breadth of the ventral stream areas as the literature. This is especially impressive given the limited number of scanning hours and no repeated stimuli, the behavioural freedom granted to subjects, and the big analysis mask we used which resulted in a substantial degree of multiple testing correction.

Task-free naturalistic vision. While experimental control is important in cognitive neuroscience research, it is also important to recognize that the brain operates in a complex and dynamic environment. Complete control over the data often comes with simplifications on our conceptual as well as technical models. Brain encoding models that capture behaviour rather than require its constraint offer a powerful tool for studying the brain's visual processing under naturalistic conditions. This is important because the way experimental designs are set up can strongly affect the conclusions that can be drawn from them. An example is the way researchers have often investigated category selectivity in ventral stream areas. Due to the nature of the experimental designs that were used, regions in occipitotemporal cortex have been thought to be uniquely responsive to clearly defined category features driven by factors like animacy or object size (32–34). However, subsequent studies have shown that this same selectivity could be the result of a set of mid-level textural visual features (35) or the functional requirements of a task (i.e. navigation, recognition, tool use etc.) (36). Instead of discussing/searching for the most appropriate category label for a brain areas selectivity, naturalistic approaches have the potential to allow us to detect and discuss the naturally occurring cross correlated features in behaviour as well as in the environment, and perhaps investigate our ideas of brain area function in a more data driven manner (37, 38).
One demonstration of how precision models can help us move towards a more integrative account of ventral stream representations is the high performance of our models in the LOc, and fusiform gyrus. LOc is often associated with object recognition (33, 34). Interestingly, LOc was predicted worse than the earlier V1, V2 and V3 in the paper of (19) who used static images. In contrast, similar to (20) and (31), who also used videos, we see that LOc is better predicted than earlier ventral stream areas. This might partially be because of the rapidly changing visual input in the earlier visual areas that are selective for finer features combined with poor temporal resolution of fMRI, but could also reflect a general difference in the neural responses elicited by static vs. naturalistic stimuli.

Limitations & Future Directions

In this paper, we did not aim to get the highest possible encoding performances, rather we wanted to draw a simple comparison between the standard approach in the field and our precision models. There are a few factors that might have limited the performance of our encoding models. Firstly, we reduced the spatial resolution of the features from early layers, which might have led to a poorer encoder performance in early visual areas where spatial selectivity is highest. Secondly, the conceptual simplicity and computational ease brought to the models by only using fixation events in our eye-tracking pre-processing might have hindered our models performances, as studies show that saccades and slow pursuits also elicit reliable responses in the early visual cortex (39). Thirdly, the limited dataset used in this study (two hours of fMRI data) and the absence of repeated stimuli might have constrained the model's ability to achieve higher accuracy. Last but not least, the encoding performance of our models is limited by features extracted from the CNN model we selected. The tasks and data CNN models are trained on have the ultimate influence on the capabilities of our approach. The CNN model used here was pre-trained on the ImageNet

competition dataset (40), which lacks ecological validity and does not optimally align with the statistical characteristics of the brain's visual processing in naturalistic settings (41). Although other analysis choices could have further improved the performance of our model, this does not undermine the strong evidence we found for the increase in model performance due to precision modelling when limited data is available.

In the future, we envision a general shift in more precise and personalized encoding models that take inter- and intra-individual variances into account. Mainstreaming this approach could not only reduce time and money spent on data collection but combined with promising new advancements like gaze prediction from fMRI recordings of the eyes (42), could allow us to move towards paradigms in which more naturalistic behaviour is possible. Precision models might be particularly beneficial for applications where training models efficiently with little data is imperative and individual behaviour is important to consider. Examples include online perceptual reconstruction or visually driven neurofeedback setups, especially combined with the fMRI data driven eye-tracking. Additionally, the targeted reduction in the dimensionality of CNN features, combined with the naturalistic and continuous brain data from movie watching could be further leveraged towards recurrent neural networks which can capture time dependent influences of visual stimuli on brain activity.

Conclusion

We demonstrated that by building precision brain encoding models, we can use less data while improving encoding performance and ecological validity. The dramatic reduction in the data needed to be collected to train encoding models can facilitate faster and cheaper data collection. Moreover, this reduction in data requirements comes with more clearly defined models and assumptions. One of our pillars in using CNNs to model visual perception in the brain is the mechanistic similarities in their architectures, like the receptive field and successive non-linear transformations of input. Conventional modelling practices use this pillar in defining their models, yet by ignoring the spatial selectivity in the modelling process, they not only perform worse, but also introduce the ambiguity about whether their models respect the correspondences in said architectures. Additionally, we showcased how trading imposed experimental control with captured individual variation can bring us closer to a more ecologically valid investigation of the brain.

Author Contributions

Dora Gözükara: Conceptualization, Methodology, Software, Formal analysis, Visualization, Writing - Original Draft. **Nasir Ahmad**: Writing - Mathematical formulation & Review & Editing. **Katja Seeliger**: Writing - Review & Editing. **Djamari Oetringer**: Methodology, Software, Writing - Review & Editing. **Linda Geerligs**: Conceptualization, Methodology, Resources, Writing - Review & Editing, Supervision.

Funding

Linda Geerligs was supported by a Vidi grant (VI.Vidi.201.150) from the Netherlands Organization for Scientific Research. Katja Seeliger was supported by a Max Planck Research Group grant to Martin N. Hebart.

ACKNOWLEDGEMENTS

We thank Umut Güçlü for the initial conceptualisation and guidance, and Jordy Thielen for his help with pRF mapping.

Bibliography

1. A. Doerig, R. Sommers*, **Seeliger, K.**, B. Richards, J. Ismael, G. Lindsay, K. Kording, T. Konkle, M. A. J. van Gerven, N. Kriegeskorte, and T. C. Kietzmann. The neuroconnectionist research programme. *Nature Reviews Neuroscience*, 24:431–450, 2023.
2. Grace W Lindsay. Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of cognitive neuroscience*, 33(10):2017–2031, 2021.
3. Mary M Hayhoe and Constantin A Rothkopf. Vision in the natural world. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(2):158–166, 2011.
4. Kendrick N Kay, Thomas Naselaris, Ryan J Prenger, and Jack L Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, 2008.
5. Shinji Nishimoto, An T Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current biology*, 21(19):1641–1646, 2011.
6. Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature communications*, 8(1):15037, 2017.
7. Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25 (1):116–126, 2022.
8. Nadine Chang, John A Pyles, Austin Marcus, Abhinav Gupta, Michael J Tarr, and Elissa M Aminoff. Bold5000, a public fmri dataset while viewing 5000 visual images. *Scientific data*, 6(1):49, 2019.
9. Martin N Hebart, Oliver Contier, Lina Teichmann, Adam H Rockter, Charles Y Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I Baker. Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *Elife*, 12:e82580, 2023.
10. M. Dorr, T. Martinetz, K. R. Gegenfurtner, and E. Barth. Variability of eye movements when viewing dynamic natural scenes. *Journal of Vision*, 10(10):28–28, August 2010. ISSN 1534-7362. doi: 10.1167/10.10.28.
11. K. Seeliger, R. P. Sommers, U. Güçlü, S. E. Bosch, and M. A. J. van Gerven. A large single-participant fMRI dataset for probing brain responses to naturalistic stimuli in space and time, 2021.
12. Janne Kauttonen, Yevhen Hlushchuk, Iiro P. Jääskeläinen, and Pia Tikka. Brain mechanisms underlying cue-based memorizing during free viewing of movie memento. *NeuroImage*, 172:313–325, May 2018. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2018.01.068.
13. A. Zadbood, J. Chen, Y.C. Leong, K.A. Norman, and U. Hasson. How we transmit memories to other brains: Constructing shared neural representations via communication. *Cerebral Cortex*, 27(10):4988–5000, August 2017. ISSN 1460-2199. doi: 10.1093/cercor/bhx202.
14. Julie A Boyle, Basile Pinsard, Amal Boukhdhir, Sylvie Belleville, Simona Brambatti, Jeni Chen, Julien Cohen-Adad, André Cyr, P Fuente Rainville, and Pierre Bellec. The courtois project on neuronal modelling-first data release. In *26th annual meeting of the organization for human brain mapping*, 2020.
15. Nikolaus Kriegeskorte. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, 1(1):417–446, 2015.
16. Haiguang Wen, Junxing Shi, Yizhen Zhang, Kun-Han Lu, Jiayue Cao, and Zhongming Liu. Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral Cortex*, 28(12):4136–4160, October 2017. ISSN 1460-2199. doi: 10.1093/cercor/bhx268.
17. Pulkit Agrawal, Dustin Stansbury, Jitendra Malik, and Jack Gallant. Convolutional neural networks mimic the hierarchy of visual representations in the human brain. *Journal of Cognitive Neuroscience*, 2020.
18. Pulkit Agrawal, Dustin Stansbury, Jitendra Malik, and Jack L Gallant. Pixels to voxels: modeling visual representation in the human brain. *arXiv preprint arXiv:1407.5104*, 2014.
19. U. Guclu and M. A. J. van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27): 10005–10014, July 2015. ISSN 1529-2401. doi: 10.1523/jneurosci.5023-14.2015.
20. Kuan Han, Haiguang Wen, Junxing Shi, Kun-Han Lu, Yizhen Zhang, Di Fu, and Zhongming Liu. Variational autoencoder: An unsupervised model for encoding and decoding fmri activity in visual cortex. *NeuroImage*, 198:125–136, September 2019. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2019.05.039.
21. Michael Hanke, Nico Adelhöfer, Daniel Kottke, Vittorio Iacovella, Ayan Sengupta, Falko R. Kaule, Roland Nigbur, Alexander Q. Waite, Florian Baumgartner, and Jörg Stadler. A study-forrest extension, simultaneous fmri and eye gaze recordings during prolonged natural stimulation. *Scientific Data*, 3(1), October 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.92.

22. Xingyu Liu, Zonglei Zhen, Anmin Yang, Haohao Bai, and Jia Liu. A manually denoised audio-visual movie watching fMRI dataset for the studyforrest project. *Scientific Data*, 6(1): 1–9, 12 2019. ISSN 20524463. doi: 10.1038/s41597-019-0303-3.
23. N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, 15 (1):273–289, 2002. ISSN 1053-8119. doi: 10.1006/nimg.2001.0978.
24. Ayan Sengupta, Falko R. Kaule, J. Swaroop Guntupalli, Michael B. Hoffmann, Christian Häusler, Jörg Stadler, and Michael Hanke. A studyforrest extension, retinotopic mapping and localization of higher visual areas. *Scientific Data*, 3(1), October 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.93.
25. Serge O. Dumoulin and Brian A. Wandell. Population receptive field estimates in human visual cortex. *NeuroImage*, 39(2):647–660, January 2008. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2007.09.034.
26. Asim H. Dar, Adina S. Wagner, and Michael Hanke. Remodnav: Robust eye-movement classification for dynamic stimulation. *bioRxiv*, 2020. doi: 10.1101/619254.
27. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.
28. Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 57(1):289–300, January 1995. ISSN 1467-9868. doi: 10.1111/j.2517-6161.1995.tb02031.x.
29. K. Amunts, H. Mohlberg, S. Bludau, S. Caspers, S. B. Eickhoff, and P. Pieperhoff. Whole-brain parcellation of the julich-brain cytoarchitectonic atlas (v2.9), 2021.
30. Nilearn contributors. nilearn.
31. Meenakshi Khosla, Gia H. Ngo, Keith Jamison, Amy Kuceyeski, and Mert R. Sabuncu. Cortical response to naturalistic stimuli is largely predictable with deep neural networks. *Science Advances*, 7(22), May 2021. ISSN 2375-2548. doi: 10.1126/sciadv.abe7547.
32. Nancy Kanwisher, Josh McDermott, and Marvin M. Chun. The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17 (11):4302–4311, 1997. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.17-11-04302.1997.
33. Paul E. Downing, Yuhong Jiang, Miles Shuman, and Nancy Kanwisher. A cortical area selective for visual processing of the human body. *Science*, 293(5539):2470–2473, 2001. doi: 10.1126/science.1063414.
34. Joshua B. Julian, Jack Ryan, and Russell A. Epstein. Coding of object size and object category in human visual cortex. *Cerebral Cortex*, 27(6):3095–3109, 06 2016. ISSN 1047-3211. doi: 10.1093/cercor/bhw150.
35. Bria Long, Chen-Ping Yu, and Talia Konkle. Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences*, 115(38), August 2018. ISSN 1091-6490. doi: 10.1073/pnas.1719616115.
36. Marius V. Peelen and Paul E. Downing. Category selectivity in human visual cortex: Beyond visual object recognition. *Neuropsychologia*, 105:177–183, October 2017. ISSN 0028-3932. doi: 10.1016/j.neuropsychologia.2017.03.033.
37. Saurabh Sonkusare, Michael Breakspear, and Christine Guo. Naturalistic stimuli in neuroscience: Critically acclaimed. *Trends in Cognitive Sciences*, 23(8):699–714, 2019. ISSN 1364-6613. doi: https://doi.org/10.1016/j.tics.2019.05.004.
38. Iiro P. Jääskeläinen, Mikko Sams, Enrico Glerean, and Jyrki Ahveninen. Movies and narratives as naturalistic stimuli in neuroimaging. *NeuroImage*, 224:117445, 2021. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2020.117445.
39. Ioannis Agtzidis, Inga Meyhöfer, Michael Dorr, and Rebekka Lencer. Following forrest gump: Smooth pursuit related brain activation during free movie viewing. *NeuroImage*, 216:116491, 2020. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2019.116491.
40. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.
41. Johannes Mehrer, Courtney J Spoerer, Emer C Jones, Nikolaus Kriegeskorte, and Tim C Kietzmann. An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, 118(8):e2011417118, 2021.
42. Markus Frey, Matthias Nau, and Christian F. Doeller. Magnetic resonance-based eye tracking using deep neural networks. *Nature Neuroscience*, 24(12):1772–1779, December 2021. ISSN 1546-1726. doi: 10.1038/s41593-021-00947-w. Publisher: Nature Publishing Group.