UTRECHT UNIVERSITY

Department of Experimental Psychology

Artificial Intelligence Master Thesis

Modelling relationships between neural responses and cortical organisation

First examiner:

Dr. Ben Harvey

Candidate:

Maximilian J. Schwalenberg

Second examiner:

Dr. Leendert van Maanen

July 25, 2025

Abstract

Object recognition, a fundamental function of the brain's visual system, has been extensively studied, yet the precise neural mechanisms underlying it remain only partially understood. This study aims to deepen our understanding of object-selective responses in the human brain, focusing on faces as highly salient stimuli. Faces provide a uniquely powerful lens for investigating how complex object representations emerge. Using high-resolution fMRI data from the Natural Scenes Dataset (NSD), an analysis pipeline identified face-selective regions and examined their feature encoding.

Key findings showed that the face-selective middle temporal sulcus (+MTS/OFA), face-suppressive fusiform gyrus (–FFG), and face-selective fusiform gyrus (+FFG) exhibited strong trial-repeatability and encoded low-level features like face position and size. High-level features (gender, age) were not consistently encoded, possibly due to NSD's passive viewing and limited statistical power. Representational similarity analysis (RSA) on DNNs (GenderAge, RetinaFace) revealed hierarchical encoding, with low-level features dominating early layers and task-specific high-level features emerging deeper, reflecting training objectives. DNN representations were deterministic, contrasting with fMRI data's inherent noise.

Despite limitations (NSD stimulus constraints, passive task, manual ROI definition), this research offers insights into early face representation, emphasizing spatial configuration. By comparing human fMRI patterns with computational models, the study highlights fundamental differences in noise, task specificity, and feature progression, providing a versatile pipeline for future semantic face encoding investigations.

I extend my sincere gratitude to my supervisor, Dr. Ben Harvey, for his trust in me with this project, his guidance throughout, and for the many things he taught me along the way. I truly enjoyed our cooperation and can look back to a time of lots of learning.

Contents

1. Introduction

Object recognition is one of the primary functions of the brain's visual system and has been extensively studied over the years. However, despite significant progress, the exact neural mechanisms underlying object recognition remain only partially understood — particularly in terms of how different areas of the visual cortex interact to extract and represent object information. This study aims to provide a deeper understanding of object-selective responses in the human brain and how these responses emerge through transformations across visual regions. By inspecting these processes, it is expected to gain insights into the hierarchical organization of object recognition.

A substantial body of research has explored object recognition in macaques, particularly focusing on their responses to faces (Hesse & Tsao, 2020). These studies have provided critical insights into the anatomical organization of face-selective regions and the functional interactions between different face-processing areas. Additionally, research comparing macaque and human face-processing systems suggests a strong anatomical correspondence between the two species (Kriegeskorte, Mur, Ruff, et al., 2008; Tsao et al., 2008). This similarity justifies the extension of face-selective analyses to human participants, offering the potential to bridge findings across species and enhance our understanding of visual object recognition.

Faces serve as an ideal object category for studying recognition processes in the brain. As highly salient stimuli, they elicit strong and reliable neural responses, making it easier to identify face-selective areas (Kanwisher & Yovel, 2006). Moreover, using faces allows for direct comparability with previous macaque studies, where responses have been mapped in detail across multiple cortical regions.

A key difference between macaque and human studies, however, lies in the experimental approach. Invasive electrophysiological recordings in macaques enable precise measurements of neural activity in single neurons in response to a carefully curated set of face images. In contrast, human studies must rely on non-invasive techniques such as functional magnetic resonance imaging (fMRI) (and other neuroimaging methods that reflect the combined responses of large neural populations), which measures changes in blood flow as an indirect indicator of neural activity (Kwong

et al., 1992). This methodological difference necessitates alternative approaches for identifying face-selective regions in humans.

This study utilizes the publicly available Natural Scenes Dataset (NSD) (Allen et al., 2022), which includes fMRI recordings from eight participants, each of whom viewed 10.000 naturalistic images drawn from the Common Objects in Context (COCO) dataset (Lin et al., 2015). These images span a wide range of object categories, including people, animals, vehicles, and furniture.

Most human fMRI investigations of face perception begin by contrasting face stimuli with non-face controls to map face-selective regions. In their study, (Kanwisher et al., 1997) identified the fusiform face area (FFA) in the mid-fusiform-gyrus, which demonstrated a stronger response to faces than to objects or scenes, thus establishing the first reliable localization of face processing in the intact brain. That simple face-versus-non-face functional localizer quickly became the default for defining regions of interest across cognitive domains. Here, I follow the same modular, region-of-interest (ROI) based strategy by first using category contrasts to isolate object-selective regions, after which further analyses are performed in these ROIs.

A traditional method for mapping neural responses to continuous stimulus parameters (like visual position, numerosity, duration or auditory frequency) within a responsive region involves population receptive field (pRF) mapping (Dumoulin & Wandell, 2008). This characterizes the response functions of individual recording sites as a function of these stimulus parameters. However, faces and other objects lack such simple parameters. They differ in complex ways that do not allow easy quantification. Instead, Representational Similarity Analysis (RSA) is used, which allows for a model-free characterization of complex, high-dimensional stimuli by focusing on the relational structure of multivoxel activation patterns within a whole responsive region rather than requiring explicit, low-dimensional parameterizations.

The core idea behind RSA is to compare a responsive region's neural activity across different stimuli and summarize them in a Representational Dissimilarity Matrix (RDM), where each element of the RDM represents the dissimilarity between neural activation patterns for two different stimuli, with smaller values indicating higher similarity and larger values indicating greater dissimilarity. This matrix provides insight into how the responsive region represents different stimuli.

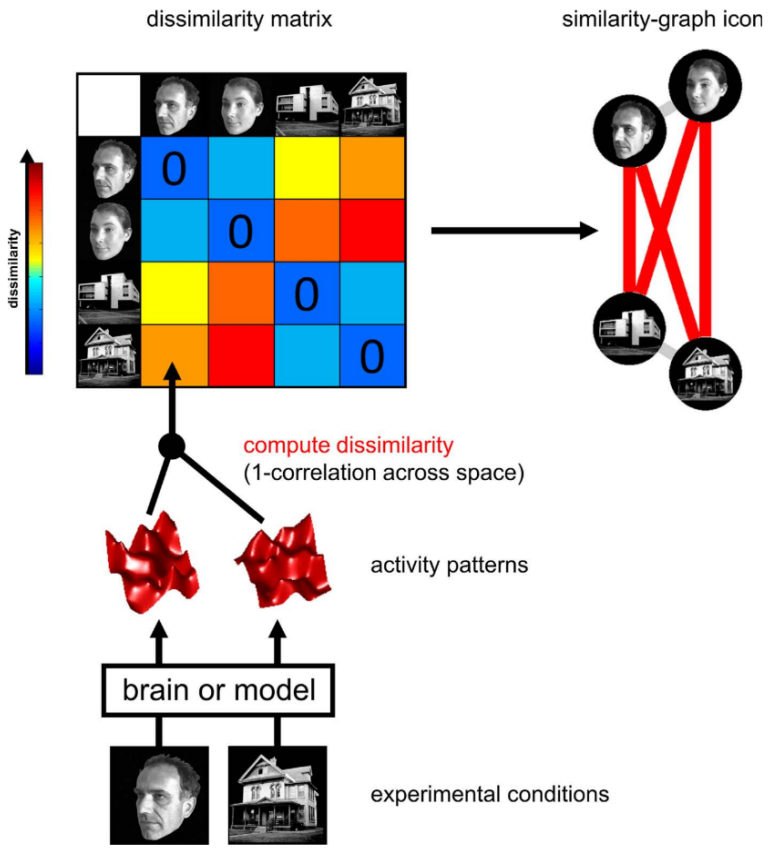


Figure 1.1: RSA framework for analyzing neural representations. The dissimilarity between neural response patterns is computed for each stimulus pair, forming a Representational Dissimilarity Matrix (RDM). Figure from (Kriegeskorte, Mur, & Bandettini, 2008).

Using RSA, previous research has shown that representations of faces and objects in the inferior temporal cortex (IT) exhibit a structured organization (Kriegeskorte, Mur, Ruff, et al., 2008). By comparing neural responses to different object categories (faces, animals, inanimate objects), a clear pattern emerges, distinguishing animate and inanimate stimuli, and (among animate stimuli) faces and non-faces (see figure 1.2). However, such analyses have typically used larger regions of interest, like the entire inferior temporal cortex, and demonstrated category-specific responses (inanimate objects, animate objects, and faces) that reflect large clusters of recording sites. Therefore, these results can be understood to reflect the presence of face-responsive regions within the larger inferior temporal cortex. Furthermore, RSA examines the pattern of responses across the whole area and is therefore unable to characterize responses at the finer scale of individual recording sites.

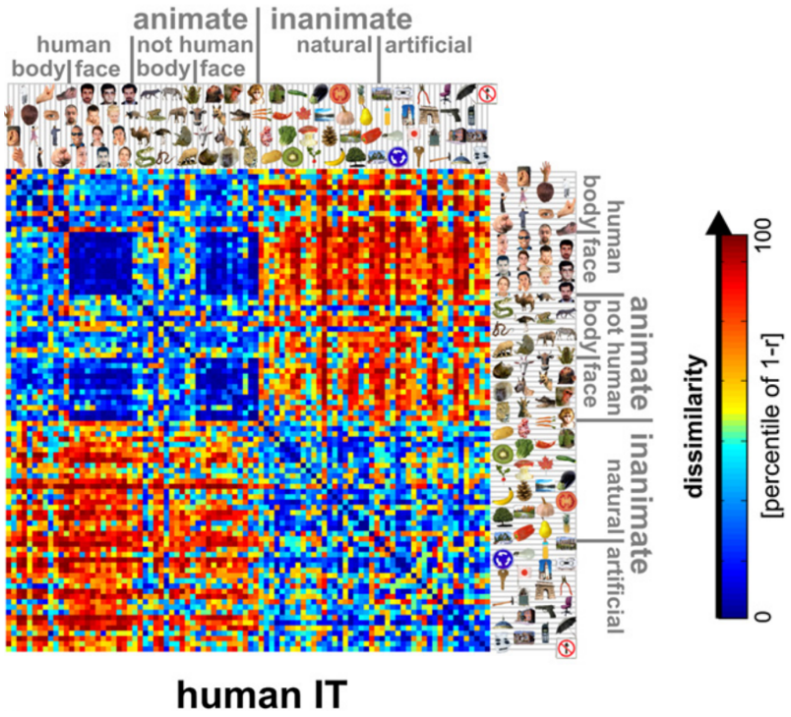


Figure 1.2: Representational Dissimilarity Matrix (RDM) showing distinct clustering of neural responses to animate and inanimate objects in the human inferior temporal cortex. Adapted from (Kriegeskorte, Mur, Ruff, et al., 2008).

Therefore, while prior RSA-work on large-scale inferotemporal clusters clearly distinguished animate versus inanimate and faces versus non-face animals, it left unanswered whether those broad clusters conceal finer internal structure and how individual sites respond to specific subsets of face exemplars. This study aims to address these limitations using finer-grained ROIs and analyzing their internal structure in further experiments.

Face-selective ROIs were identified by contrasting brain responses to human faces versus non-face objects, isolating face-specific processing from general animacy effects. After identifying face-responsive areas and constructing clusters/ROIs, RSA was applied to examine the structure of neural responses within these regions.

By applying RSA to face-responsive areas, representational spaces in form of RDMs are obtained, revealing organizational principles within face-selective cortical areas. Since RDMs are inherently high-dimensional, their direct interpretation can be challenging. To facilitate visualization and further analysis, Multidimensional Scaling (MDS) is employed to project the RDM into a lower-dimensional space while preserving the relative pairwise distances between neural representations as closely as possible. This transformation allows us to look at the (dis)similarity as a continuous

metric in the 2d space and allows for the voxelwise gaussian fitting at a later stage.

To further analyze the structure of face representations and assess if the ROIs have any internal structure going beyond just reflecting face-presence, I perform several downstream analyses. First, in order to analyze if any face-related-features such as the faces position or the persons gender are encoded in the single ROIs, a permutation test is performed, comparing the pairwise distances of face-stimuli in the MDS space to the corresponding pairwise-distances of features in the same MDS space. In the second analysis, I fit Gaussian functions to voxel activity in the MDS space. This step serves a crucial role in determining whether specific voxels encode subsets of face images. By fitting Gaussians over the MDS-projected space, it becomes possible to identify voxels that respond selectively to particular subsets of faces, for example faces with a particular identity, expression, or pose. Additionally, the resulting Gaussian parameters provide a quantitative measure of the spread and specificity of neural responses, offering insight into whether certain face-responsive-regions encode fine-grained distinctions between faces, while others encode more general face properties. Lastly, for each ROI and hemisphere, this study investigates the correlation between the pairwise geodesic distances of voxels on the cortical surface and the pairwise distances of the centers of their fitted responses in the MDS space. This analysis serves to determine the topographic organization of face representations, revealing whether physically close voxels on the cortex encode similar face features.

Finally, this study compares the RSA findings in the brain with those obtained from neural network models trained to identify, localize and classify faces. To allow for a fair comparison, the same methods that were applied to the fMRI data, were then applied to the neural network. The focus lies in the comparison of the quantifiable feature encodings across different layers of the neural network so it can be assessed whether computational models of face processing exhibit representational structure similar to those observed in the human brain.

By integrating RSA, quantifying feature encoding in representational spaces, Gaussian fitting, correlating representational geometry to cortical topography and neural network comparisons, this study aims to characterize the representational structure underlying face processing in the human brain. The findings will contribute to a more detailed understanding of how different regions encode facial information to support object recognition, and how biological and artificial systems compare in their representation of faces.

2. Methods

2.1 Dataset

All analyses in this study used the pre-computed single-trial beta maps provided by the Natural Scenes Dataset (NSD) (Allen et al., 2022), a publicly available 7 T whole-brain fMRI resource explicitly designed to meet the field's need for massive, high-quality neural data. The NSD comprises eight participants, each scanned in 30–40 one-hour sessions over the course of roughly one year. During these sessions, every subject viewed between 9,000 and 10,000 unique, richly annotated color images drawn from the COCO dataset (Lin et al., 2015), with a core subset of 1,000 images seen by all participants and each image presented up to three times. In total, NSD provides neural response amplitude (beta) estimates for over 70,000 distinct natural scenes and nearly 200,000 trials across all participants, making it more than an order of magnitude larger than prior fMRI image-sampling studies.

The NSD includes voxel-wise beta coefficients obtained via an optimized GLM pipeline, yielding denoised estimates of the hemodynamic response amplitudes on each trial. By drawing directly on these trial-level betas, this study focuses on exploiting the dataset's unprecedented scale within individual brains rather than on low-level preprocessing or averages responses within large populations of people.

From the full set of betas, I extracted the maps corresponding to our selected stimulus subsets and applied a per-voxel, within-participant standardization across all sessions. Specifically, for each voxel, I z-scored beta values across each session to correct for slow drifts in baseline activation and ensure consistency of neural representations over time, as shown in an earlier study (Roth & Merriam, 2023). This standardization mitigates session-to-session variability and yields unit-variance response profiles, providing a stable foundation for subsequent multivariate analyses.

2.2 Locating face-responsive areas

The first step towards investigating the structure of face representations in the brain is to isolate regions that are selectively engaged in face processing. This step is critical, as all downstream analyses in this study, such as representational similarity analysis and voxel-wise modeling, are constrained to these face-responsive areas to ensure interpretability and relevance to the domain of face perception.

To reliably identify face-responsive regions in the brain, I constructed the stimulus sets used in the statistical comparisons with great care. Previous studies indicate that faces exhibit a greater representational similarity to other animate objects than to inanimate ones (Kriegeskorte, Mur, Ruff, et al., 2008). Therefore, comparing face stimuli to a general inanimate baseline risks conflating face selectivity with broader animate–inanimate distinctions. To isolate neural responses that are specific to faces, I constructed two sets: the face set containing only (human) face stimuli and a non-face set consisting of animate but non-face stimuli (no human faces). This choice ensures that the statistical test, conducted voxel-wise using the beta responses from the NSD dataset, detects activity that is specific to face processing rather than general animacy. The resulting face-selective voxels form the basis for all subsequent analyses.

2.2.1 Constructing stimulus sets

The construction of the face and non-face stimulus sets is critical for ensuring the validity of the statistical tests and the reliability of identified face-responsive regions. Due to the large-scale nature of the dataset, a manual approach to creating the stimulus sets for each of the subjects would be infeasible. Instead, I utilized a semi-automated two-step approach that takes advantage of the image labels provided by the COCO dataset.

First, I extracted all images labeled as containing animate content using the 74 object categories provided by the COCO dataset. This yielded a candidate pool of images that included people as well as various animals.

Second, I divided this animate pool into two mutually exclusive subsets: the face set and the non-face set. Several criteria must be satisfied to avoid confounds. Images in the face set must depict a single, clearly visible human face, whereas images in the non-face set must contain one or more animate entities but no visible human faces.

To construct these sets, I used a combination of automated and manual procedures. I prefiltered the face set by applying a deep neural network, the RetinaFace detector (Deng et al., 2020), to all candidate images and selecting those with exactly one predicted (face) bounding box. I selected remaining ambiguities, as well as the entire non-face set, through manual annotation for each subject to ensure label accuracy and consistency. See Figure 2.1 for exemplary images of the face and non-face sets.

Figure 2.1: Exemplary images for the constructed face (top) and non-face (bottom) sets from the COCO dataset.

To ensure data quality for downstream analyses, I excluded trials with insufficient valid measurements. Specifically, the dataset occasionally contains faulty responses represented by "Not a Number" values (NaN). I retained an image only if at least two out of the three available beta responses were free of NaNs, allowing for a train/test split in later steps.

2.2.2 Statistical test

With the face and non-face stimulus sets defined, I performed a statistical test to identify voxels that respond selectively to faces. Given the two independent groups of face stimuli and animate non-face stimuli and the heterogeneity typical of fMRI signal variance, I employed Welch's t-test. This test does not assume equal variances nor class balance between the two groups and is therefore well-suited to the characteristics of the data (Welch, 1947).

For each subject, I computed a voxel-wise test statistic by comparing the distribution of beta response amplitudes between the two sets. Importantly, I used two of the available observations for each stimulus for the voxel-wise testing. Specifically, each beta value represents the average fMRI response for a given image across a small

temporal window around stimulus presentation, as derived from the NSD's native-surface beta estimates. The result was a t-value for every voxel, quantifying the degree to which its response differs between the face and non-face conditions. The code for the statistical testing can be found here *t_testing/t_testing.py*.

2.2.3 Defining ROIs

To localize face-responsive regions of interest (ROIs), I used a combination of statistical thresholding and manual delineation, implemented in MATLAB using the FreeSurfer suite (Fischl, 2012). After computing voxel-wise t-values (see previous subsection), I visualized these values by projecting them onto the cortical surface for each subject.

I applied an initial threshold (e.g., $|t| \geq 2.0$) to highlight voxels that show a statistically significant difference in activation between face and non-face conditions. Voxels with positive t-values above the threshold are interpretable as face-responsive, indicating stronger activation to face stimuli than to animate non-face stimuli. Conversely, regions with significantly negative t-values reflect face-suppressive responses, i.e., voxels that respond more strongly to non-face animate stimuli. While these negative regions are not face-selective, they show different responses between sets and contain information about faces, so I included them in subsequent analyses to explore broader representational distinctions.

Based on these thresholded maps, I manually drew ROIs around clusters of significant voxels. Importantly, I then computationally refined each initially drawn ROI to retain only those voxels that strictly exceed the statistical threshold.

To ensure the generalizability and interpretability of results, I selected ROIs based not only on their statistical and anatomical plausibility within individual subjects, but also on cross-hemisphere and cross-subject consistency. Ideally, I included an ROI in further analysis only if I could identify a corresponding region in all participants. This constraint guarantees that ROIs reflect a shared face-processing network rather than individual-specific activations, therefore improving the consistency and interpretability of downstream representational similarity.

2.3 Representational Similarity Analysis

Building upon the identification of face-responsive regions, I applied Representational Similarity Analysis (RSA) to examine the structure of neural responses within these regions. As outlined in the introduction, RSA provides a framework to quantify and visualize how different stimuli are represented across patterns of neural activity (Kriegeskorte, Mur, & Bandettini, 2008). Rather than focusing on mean activation levels across a group of stimuli within a set, RSA characterizes the geometry of representations, revealing whether and how distinct responses to specific stimuli differ within activation patterns.

Importantly, I conducted all RSA analyses on a region-of-interest (ROI) basis. The ROIs, defined in the previous section via voxel-wise contrasts between face and non-face stimuli, serve as the spaces within which representational structure is analyzed. This localized approach not only aligns conceptually with prior studies using larger-scale ROIs (e.g., inferior temporal cortex), but also leverages the higher spatial specificity provided by ultra-high-field fMRI to assess finer-grained representational differences.

I performed the analysis in several steps, which I will first outline briefly and then describe in detail. First, I constructed Representational Dissimilarity Matrices (RDMs) for each ROI and subject, capturing pairwise dissimilarities between response patterns. To facilitate interpretation and allow for further analyses, I then projected these high-dimensional matrices into two dimensions using Multidimensional Scaling (MDS).

Before interpreting the geometry of these spaces, I assessed the consistency of neural responses across repeated presentations of the same stimulus. This step is essential, as meaningful structure in the representational space can only be inferred if the underlying signal is stable across trials.

Next, I tested whether specific facial features (e.g., face position, age, or gender) systematically affect the responses to individual faces within the RDM/MDS space. I did this by correlating feature-based distances with distances in the representational spaces, enabling a quantification of how strongly different ROIs reflect particular stimulus attributes.

To further interpret the spatial layout of neural representations, I fitted Gaussian re-

sponse functions to the MDS embeddings of individual voxels, treating the beta value of each voxel as a third dimension. This fitting provides a compact summary of how individual voxels encode subsets of the representational space.

Finally, I examined whether these fitted voxel response functions are related to positions of voxels within the responsive areas. By correlating pairwise distances in MDS space with geodesic distances along the cortical surface, I assessed whether functional similarity is spatially organized, offering insight into the cortical mapping of representational content.

2.3.1 Constructing training/test set

For each stimulus, I obtained two or three trial-specific beta estimates. I defined the training set as either the voxel-wise average of the first two estimates (if three trials were available for the stimulus) or just the first estimate, while I reserved the third estimate as a leave-one-out test sample. This procedure ensured that test data remain completely independent during voxel fitting and that RDM spaces are computed exclusively from training data, thus preventing any indirect data leakage.

2.3.2 Constructing RDM and MDS spaces

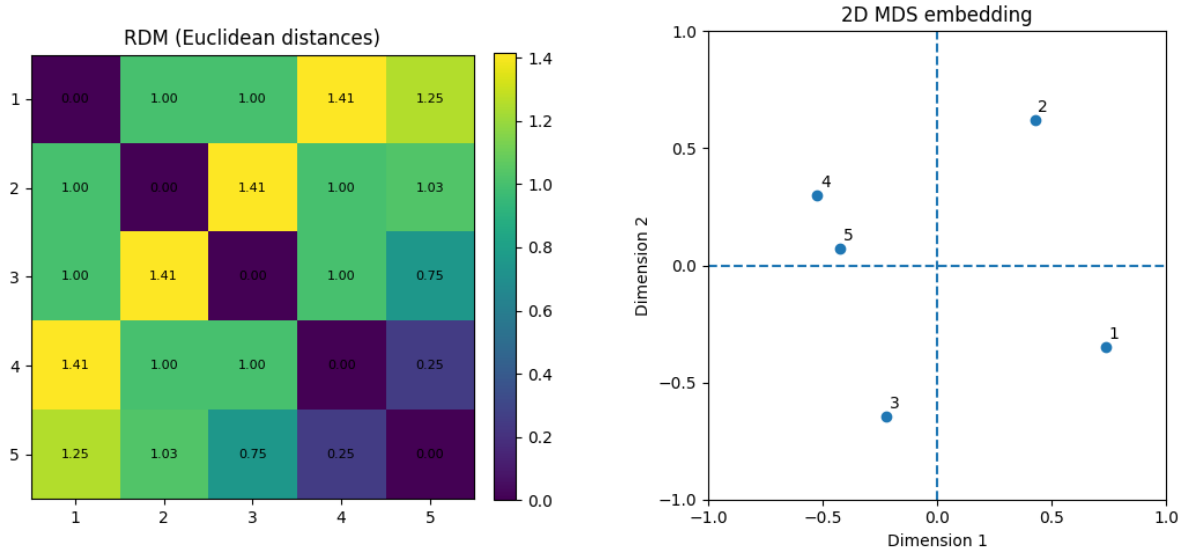
To investigate the internal structure of neural representations, I first constructed Representational Dissimilarity Matrices (RDMs) for each subject and each defined region of interest (ROI). These RDMs capture the pairwise dissimilarities between neural activation patterns elicited by different face stimuli.

For each subject, I selected all face images previously labeled for the stimulus set for statistical testing. I extracted corresponding beta values from the training set for these stimuli for each ROI. These beta patterns formed the basis for calculating the RDM. As suggested in previous studies, correlation is a useful measure to use as the similarity metric in construction of the RDM (Kriegeskorte, Mur, & Bandettini, 2008), so I summarized response dissimilarity between images as one minus the Pearson correlation of their response patterns.

The resulting RDM is a symmetric matrix where each entry quantifies the dissimilarity between the neural responses to two face stimuli. Because the number of pairwise comparisons grows quadratically with the number of stimuli, RDMs are inherently high-dimensional and difficult to interpret directly.

To facilitate visualization and further analysis, I applied Multidimensional Scaling (MDS) to each RDM. MDS transforms the high-dimensional dissimilarity space into a two-dimensional embedding, in which each stimulus is represented as a point. The Euclidean distances between these points approximate the dissimilarities captured in the original RDM, allowing us to visually assess the structure of the neural representation. Therefore, stimuli that evoke similar activation patterns will be located close together, while dissimilar stimuli will be positioned further apart (visualized in Figure 2.2).

This procedure yields a 2D MDS space for each ROI and subject, forming the foundation for all subsequent analyses, including assessments of representational stability, feature encoding, and topographic correspondence.



(a) Representational dissimilarity matrix (RDM) for five toy stimuli (1–5), where each entry shows the pairwise Euclidean distance between their feature vectors.

(b) Two-dimensional MDS embedding of those distances: each point is a stimulus (labeled 1–5), with dashed lines marking the origin. Stimuli pairs 2 & 3 and 4 & 1 are most dissimilar (d = 1.41), whereas stimuli 4 and 5 are most similar (d = 0.25).

Figure 2.2: (a) RDM matrix showing pairwise Euclidean distances among five toy stimuli. (b) Corresponding 2-D MDS embedding of those distances.

2.3.3 Assessing Trial Repeatability and Representational Stability

Before interpreting representational structures derived from RDM/MDS spaces, it is essential to establish whether and where neural responses are repeatable across multiple presentations of the same stimulus. If repeated trials of the same image evoke unrelated activation patterns, it would not be possible to gain any valuable insights

by performing further analyses. Importantly, I applied this stability analysis in both the original RDM space (where pairwise dissimilarities are given directly by the RDM entries) and in the low-dimensional MDS embedding (where pairwise distances are computed as Euclidean distances in 2D).

In this approach, I leveraged the fact that each face stimulus in our set was presented three times and yields at least two valid beta estimates per ROI and hemisphere. I therefore constructed two (or, when all trials passed quality control, three) independent RDMs by assigning each RDM one non-overlapping repetition of every stimulus. For the RDM-space analysis, I vectorized the upper triangle of each RDM (i.e., the full set of pairwise dissimilarities). For the MDS-space analysis, I first projected each RDM into two dimensions via classical multidimensional scaling, then vectorized the upper triangle of the resulting 2D distance matrix.

To quantify stability, I computed the Pearson and Spearman correlation across every pair of these vectors (e.g., Rep 1 vs 2, 1 vs 3, and 2 vs 3). A high mean correlation indicates that the representational geometry—whether measured in RDM or MDS space—is preserved across repetitions, reflecting robust encoding.

I then employed resampling procedures to assess the reliability and significance of these stability estimates. First, I bootstrapped the entire pipeline by randomly reassigning trials to repetitions for each stimulus, then recomputed RDMs (or their MDS embeddings) and pairwise correlations 1,000 times to derive a confidence distribution for the mean correlation. Next, I generated a null distribution by permuting the entries of one vector before each correlation, thereby breaking true stimulus alignment and capturing chance-level correlations. Comparing the observed mean correlation to this null distribution yields empirical p-values for representational stability in both spaces.

This analysis, executed in both RDM and MDS spaces, provides a robust assessment of representational stability. Only ROIs exhibiting strong between-trial repeatability were deemed reliable for downstream analyses. The corresponding code for this analysis can be found in the script *rsa/sample_repeatability.py*.

2.3.4 Quantifying Feature Encoding in Representational Space

Having established the representational stability of the individual ROIs, I next assessed whether specific stimulus features are systematically reflected in these repre-

sentational spaces. In particular, I examined whether distances in feature spaces correlate with the neural dissimilarities captured by the RDM/MDS embeddings. For this, I considered several features extracted from the face stimuli using two different neural networks provided by the InsightFace library. This includes the RetinaFace neural network (Deng et al., 2020) for detecting the position of faces by estimating their bounding boxes as well as a regression model that takes the cropped images of the predicted faces and estimates their gender and age. Therefore, the available features that are tested in this analysis are gender and age as well as the position of the face (center of the predicted bounding box) and its size (bounding box area).

Previous work (Henriksson et al., 2015) has shown that spatial location of faces within images can significantly impact neural face-selective responses, especially in high-level visual areas. Therefore, assessing the influence of such features is not only of methodological interest but also theoretically motivated.

For each ROI and subject, I applied a permutation-based Mantel test (Mantel, 1967) to evaluate whether the spatial arrangement of stimuli in RDM/MDS space reflects similarity in a given feature. The Mantel test quantifies the correlation between two distance matrices. In this case, it is the pairwise Euclidean distances between stimuli in the 2D MDS embedding or the dissimilarities given by the RDM and the pairwise distances between the same stimuli in a given feature space (e.g., Euclidean distance between face centers).

To determine the significance of the observed correlation, I performed a large number of permutations ($n = 2000$), randomly shuffling the rows of the feature matrix before recalculating the correlation. The resulting p-value reflects how likely the observed alignment between RDM/MDS space and feature space would arise by chance while the effect size r describes the magnitude of the correlation between neural representational dissimilarities and feature-space distances.

A significant correlation implies that the neural representational geometry within the ROI is not arbitrary, but systematically encodes information about that particular stimulus attribute. Conversely, a lack of significant correlation suggests that the feature is either not encoded or not captured by the RDM/MDS spaces.

For each ROI, I then combined its per-subject Mantel results into one summary

effect and one overall p-value. Effect sizes r_s were Fisher–z transformed,

$$\bar{z} \;=\; \frac{1}{S}\sum_{s=1}^{S} \operatorname{arctanh}(r_s),$$

and back-transformed via $r_{\mathrm{ROI}} = \tanh(\bar{z})$. P-values p_s were aggregated using Fisher's method,

$$X^2 = -2\sum_{s=1}^{S} \ln(p_s) \sim \chi^2_{2S},$$

from which the combined p_{ROI} is obtained.

By systematically probing different facial features, this analysis identifies which aspects of the stimuli are preserved in the representational structure of different brain regions. The results provide insight into the functional specialization of ROIs. The corresponding code for this analysis can be found in the script *rsa/permutation_analysis.py*.

2.4 Voxel-wise Gaussian fitting in MDS space

Having established the geometry of the face-space via MDS embeddings of our ROI responses, I further characterized the tuning of individual voxels within that space by treating each voxel as a circular-symmetrical Gaussian detector whose response varies as a function of position in the 2D MDS coordinates, mirroring classical models of visual-spatial receptive fields in early visual neurons (Hubel & Wiesel, 1962) and population-level tuning in later visual areas modeled as the integrated response over Gaussian fields on the cortical surface (Haak et al., 2013). Specifically, for each voxel v I parameterized its location (x_v, y_v) in MDS-space in polar coordinates (θ_v, r_v) within the unit circle and defined its response function as

$$\hat{y}_v(x) = A_v \exp\Big(-\frac{\|x-\mu_v\|^2}{2\sigma_v^2}\Big) + B_v \tag{2.1}$$

Parameter	Initialization range	Hard bounds
θ_v	$[0, 2\pi)$	$[0, 2\pi)$
r_v	$[0, 0.5]$	$[0, 1]$
A_v	$[0.1, 10]$ or $[-10, -0.1]$	$A_v \geq 0.1$ or $A_v \leq -0.1$
B_v	$[-2, 2]$	unconstrained
σ_v	$[0.01, 2]$	$0.01 \leq \sigma_v \leq 20$

Table 2.1: Initialization ranges and hard bounds for the Gaussian model parameters.

where $\mu_v = \left(r_v \cos\theta_v,\ r_v \sin\theta_v\right)$. The amplitude is represented by A_v, the intercept by B_v and the tuning width by σ_v. I bounded parameters to plausible ranges and enforced the center of the fit to be located inside of the unit circle. Initial guesses are uniformly sampled within restricted limits near the MDS center while assuring that the initial receptive field is not greater than the sampling space itself (limiting σ). Table 2.1 summarizes these initialization ranges and parameter limits. It should be noted that the amplitude is fitted either bounded to only positive values or only negative values. This is because, I wanted to establish a clear comparison between the Gaussians fitted in the face-responsive-regions (positively significant t-value) and the face-suppressive-regions (negatively significant t-value) where negative amplitudes as optimal values are to be expected resulting in a very different encoding in the representational space.

I fitted each voxel's parameters by running four separate bounded non linear least-squares optimizations (Trust-Region Reflective, using SciPy's default tolerances), each from a different random start. To guard against failures, each fit is retried up to three times on error, and if all retries fail the solution defaults to predicting the mean response (amplitude = mean of the training values), centered at the origin, with $\sigma = 20$ and zero intercept. I allowed early stopping as soon as a run's cross-validated R^2 on the test set came within 0.04 of the current best, then kept the solution with the highest cross-validated R^2. This strikes a balance between computational efficiency and fit quality.

Held-out predictions $\hat{y}_v(x_{\text{test}})$ are compared to observed responses via cross-validated variance explained (CVVE)

$$\text{CVVE} \;=\; 1 \;-\; \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2},$$

where y_i and $\hat{y}_i$ are the observed and predicted responses for condition i, and $\bar{y}$ is the mean observed response.

For each voxel, I computed the proportion of variance explained (R^2) on the test set, yielding a quantitative metric of how well the voxel's response can be captured by a smooth, spatially localized function in representational space. Additionally, to account for potential differences in overall gain and baseline between training and test sets, I also computed a rescaled variance explained metric ($\text{CVVE}_{\text{rescaled}}$) by fitting voxel-specific scaling parameters a_v and b_v on the test set such that

$$y_i \;=\; a_v\, \hat{y}_v(x_i) \;+\; b_v,$$

and then calculating

$$\text{CVVE}_{\text{rescaled}} \;=\; 1 \;-\; \frac{\sum_i \big(y_i - (a_v \hat{y}_i + b_v)\big)^2}{\sum_i (y_i - \bar{y})^2}.$$

This rescaled CVVE reflects the model's ability to capture response patterns independent of absolute amplitude and offset discrepancies between the two sets.

Voxels with high R^2 and low σ values are interpreted as exhibiting localized and reliable tuning within MDS space, implying that their activity reflects specific dimensions of representational geometry. In contrast, voxels with broad Gaussians (high σ) or low explained variance either encode diffuse, unspecific features or exhibit noisy activation patterns. The impact of the tuning width is visualized in Figure 2.3, where each panel shows a normalized 2D Gaussian at the MDS-space origin for $\sigma = 0.05$ (highly localized), $\sigma = 0.5$ (moderate), and $\sigma = 3$ (broad), illustrating how σ governs spatial selectivity.

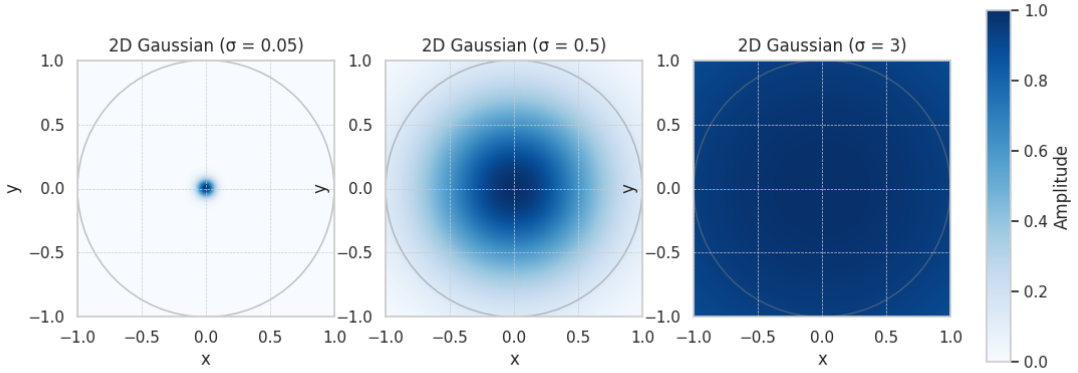


Figure 2.3: Normalized 2D Gaussians for $\sigma = 0.05$ (left), $\sigma = 0.5$ (center), and $\sigma = 3$ (right) within the unit sampling circle. The 2D Gaussian function is defined as $\hat{y}_v(x) = A_v \exp\left(-\frac{\|x-\mu_v\|^2}{2\sigma_v^2}\right) + B_v$. For these plots, the function is centered at the origin, so $\mu_v = 0$. The amplitude is set to $A_v = 1$ and the baseline offset is $B_v = 0$.

This analysis enables a voxel-wise quantification of spatial selectivity in representational space, and further identifies functional substructure within an ROI. The corresponding code for this analysis can be found in the script *gaussian/fit_gaussian.py*.

2.5 Relating Representational Geometry to Cortical Topography

Having modeled each voxel's tuning as a Gaussian with center (x_0, y_0) in the 2D MDS embedding, I then analyzed if these functional preferences align with the physical layout of the cortex by checking if voxels that are close together in representational space are also close together on the cortical sheet.

To address this, I first computed the pairwise geodesic distance $d_{\text{geo}}(v, v')$ between every pair of voxels within each ROI by running Dijkstra's algorithm on the subject-specific cortical surface mesh, which is given by the white matter surface neuroimaging file. I then formed the corresponding matrix of representational distances

$$d_{\text{repr}}(v, v') \;=\; \left\|(x_{0,v}, y_{0,v}) - (x_{0,v'}, y_{0,v'})\right\|$$

between the Gaussian centers in MDS space. Finally, for each ROI and each hemi-

sphere separately, I quantified the relationship between anatomical and functional distances using Spearman's rank correlation between $\{d_{\text{geo}}(v, v')\}$ and $\{d_{\text{repr}}(v, v')\}$.

To account for the effects of the downscaling of the original cubic voxel size of 1.8mm to 1.0mm (Allen et al., 2022), I performed the spearman correlation with a randomly sampled subset of the available voxels for each hemispheres ROIs by the size of

$$sample\_size = n\_voxels\_roi \cdot \frac{1}{1.8^3}$$

To obtain a robust, statistically conclusive estimate, I repeated this subsampling procedure 1,000 times with different random subsets. I also generated a null distribution by randomly shuffling the entries of the representational-distance matrix (keeping the anatomical distances fixed) and recomputing the Spearman ρ, which allowed me to assess the significance of the observed correlations against chance.

A strong positive correlation would suggest that the representational space is not arbitrarily organized but instead reflects the spatial layout of cortical tissue. This supports the hypothesis that neighboring cortical regions tend to encode similar images, consistent with a topographic organization of functional tuning. The corresponding code for this analysis can be found in the script *rsa/cortical_correlation.py*.

2.6 RSA for Neural Networks

To answer the second research question and allow for comparability between neural networks and the brain, I extended the same analytical framework to a deep neural network trained for face detection. This comparison served two key purposes. First, it allowed me to evaluate whether representational structures observed in human brain areas are mirrored in artificial models trained on similar tasks. Second, it provided a tractable way to probe encoding mechanisms at different stages of visual processing within a network, offering a conceptual bridge between biological and artificial vision systems.

Specifically, I analyzed the *RetinaFace* model, a neural network trained to detect human faces in natural images. This is the same model used earlier in subsection 2.2.1

to construct the face set. Additionally, the *GenderAge* model, which was used to obtain the ages and genders was analyzed in a similar way. For our analysis, the original model weights (provided in the Open Neural Network Exchange (ONNX) format) were converted into an equivalent PyTorch model, allowing full access to internal activations and facilitating representational analysis layer by layer. Figure 2.4 shows the pipeline for obtaining predictions from both models. Note that they do not share any weights and operate independently, except that the *GenderAge* model uses the bounding boxes predicted by the *RetinaFace* model to obtain face crops.

The analysis consists of five key steps. First, for each subject, I extracted the exact set of face images that were shown during the experiment and passed them through the RetinaFace network in evaluation mode. Then, activations were captured at each intermediate layer following the application of the nonlinearity (e.g., ReLU), as these layers correspond to functional stages of processing similar to cortical regions in the brain. After this, each activation layer was treated as a functional analogue to a brains ROI, with the resulting activation pattern across units taken as the "layer-ROI" response to a given stimulus. Subsequently, I constructed a Representational Dissimilarity Matrix (RDM) for each layer by computing pairwise dissimilarities (using $1 - \textit{Pearson\_correlation}$ as a dissimilarity metric) between activation vectors across all stimuli, and projected these RDMs into a two-dimensional MDS space. Finally, feature-based analyses analogous to the ones performed in subsection 2.3.4 were carried out within these spaces to assess whether specific visual features such as face position or gender are encoded across layers.

This approach enables a direct comparison between the representational structure in brain regions and that in artificial neural networks. Applying the same RSA methodology across domains allows to examine where and how the human visual system and computational models optimized for tasks align, both in representational geometry and in feature selectivity.

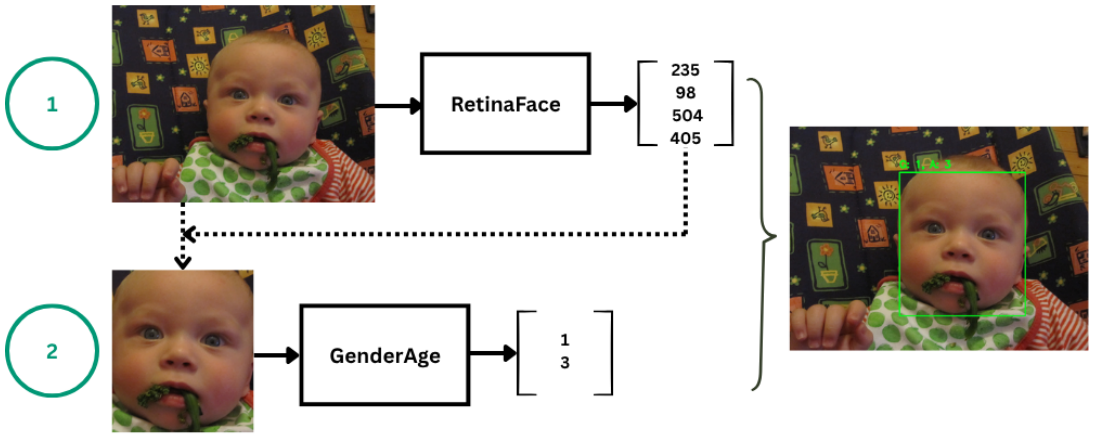


Figure 2.4: Visualizing the flow of information between the two utilized neural networks RetinaFace & GenderAge. In (1), faces bounding boxes are regressed, while (2) uses this predicted bounding box to pass the cropped face for regressing age and gender.

3. Results

3.1 Characterization of constructed stimulus sets

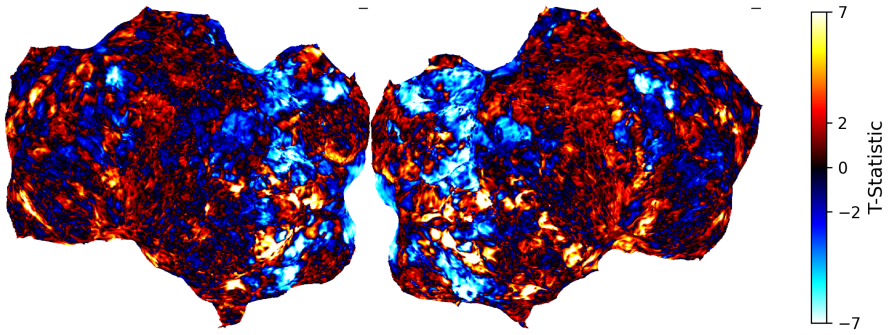
To identify face-responsive regions using statistical testing, I first created the two stimulus sets (face and non-face set) for each subject. As shown in Table 3.1, each subject's stimulus set comprised a variable number of face and non-face images, with the numbers in parentheses marking the subset of images that were seen by all participants. For some participants not all of those images could be utilized due to faulty values. In total, each subject's combined sets consisted of 190 to 583 images.

Subject	Faces	Non-Faces	Total Images
1	251 (74)	332	583
2	213 (74)	212	425
3	186 (68)	178	364
4	163 (65)	169	332
5	195 (69)	190	385
6	95 (39)	95	190
7	215 (74)	215	430
8	166 (62)	165	331

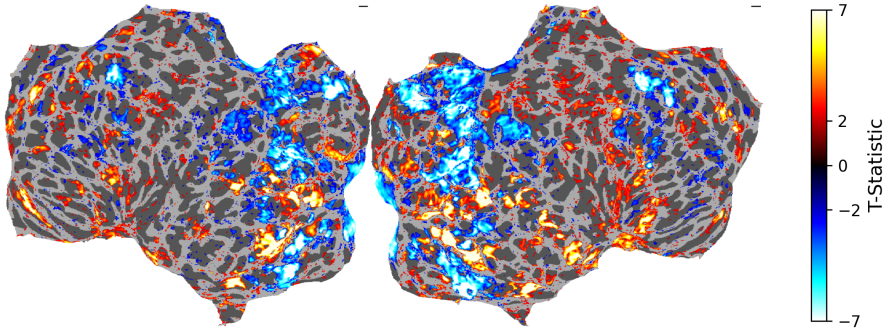
Table 3.1: Distribution of images in face vs. non-face set per subject. The numbers in parentheses mark the valid images seen by all participants.

3.2 Identification of Face Responsive Regions

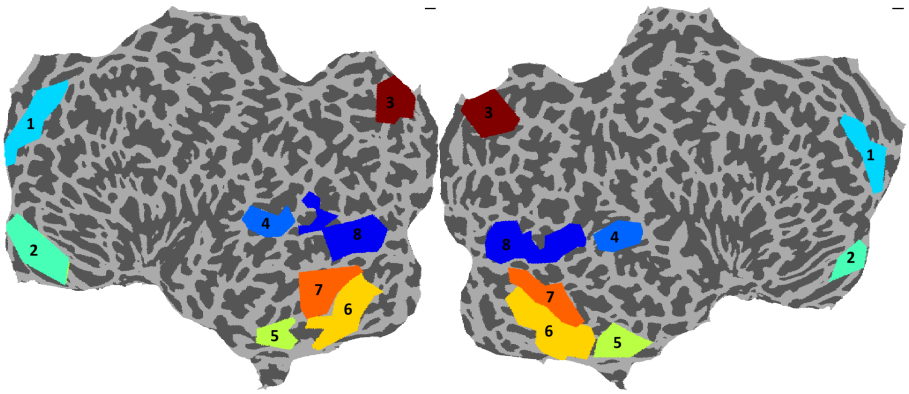
Next, for each subject, I performed a voxel-wise t-test using the constructed stimulus sets to locate face-responsive regions. Figure 3.1 shows the resulting t-statistic mapped to the cortical surface for subject 2. Figure 3.1a displays the unthresholded t-values on the cortical surface, and Figure 3.1b shows those values thresholded at $|t| \geq 2$. From these thresholded maps, I identified eight ROIs that were largely consistent across participants. The anatomically labeled ROIs appear in Figure 3.1c, while Figure 3.1d illustrates the ROIs that survived the $|t| \geq 2$ threshold.



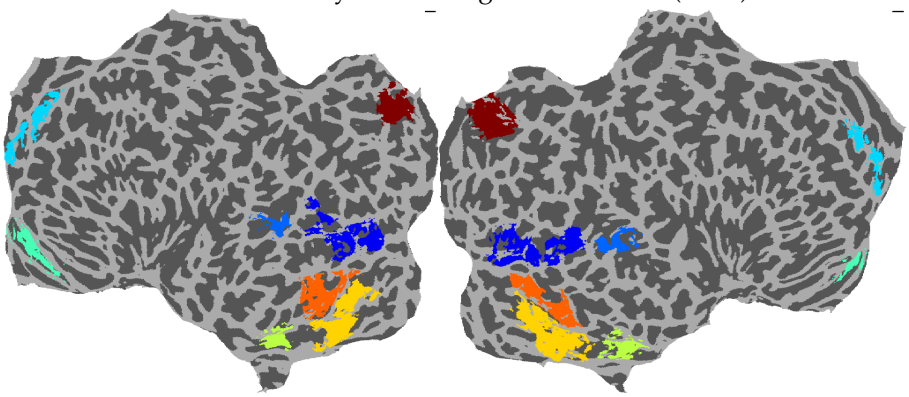
(a) Voxel-wise t-statistic on cortical surface.



(b) Thresholded voxel-wise t-statistic ($|t| \geq 2$).



(c) Anatomically labelled regions of interest (ROIs).



(d) ROIs containing voxels with $|t| \geq 2$.

Figure 3.1: Voxel-wise t-statistic for subject 2 mapped on the cortical surface: (a) the unthresholded t-statistic; (b) the t-statistic thresholded at $|t| \geq 2$, showing only supra-threshold voxels; (c) the anatomically labelled ROIs used for analysis; (d) those ROIs containing voxels for which $|t| \geq 2$.

I mapped the determined ROIs to their anatomical labels that are used throughout this work to refer to each of the ROIs.

ROI nomenclature:

1. the face-selective lateral occipital complex (+LOC)
2. the face-selective extrastriate body area (+EBA)
3. the face-suppressive precuneus (-PCu)
4. the face-selective superior temporal sulcus (+STS)
5. the anterior temporal lobe (+ATL)
6. the face-suppressive fusiform gyrus (–FFG)
7. the face-selective fusiform gyrus (+FFG)
8. the face-selective middle temporal sulcus (+MTS/OFA)

Finally, Table 3.2 summarizes which of these eight ROIs were detected (in both hemispheres) in each of our eight participants.

Subject	+LOC	+EBA	–PCu	+STS	ATL	–FFG	+FFG	+MTS
1	✓	✓	✓	✓	✓	✓	✓	✓
2	✓	✓	✓	✓	✓	✓	✓	✓
3		✓	✓	✓	✓	✓	✓	✓
4	✓	✓	✓	✓	✓	✓	✓	✓
5	✓	✓	✓	✓	✓	✓	✓	✓
6	✓			✓	✓	✓	✓	✓
7	✓		✓		✓		✓	✓
8					✓		✓	✓

Table 3.2: ROI presence across subjects

3.3 Representational Similarity Analysis

After having defined the ROIs for all subjects, I then applied RSA to each of those ROIs and performed the subsequent downstream analyses.

3.3.1 Assessing Trial Repeatability

To evaluate the stability of representational geometries across repeated presentations, I applied the “independent-replicate” analysis in RDM space and in the low dimensional MDS embeddings.

First, for each ROI and hemisphere, I computed Pearson/Spearman correlations between the vectorized upper-triangles of independent RDM replicates (Rep 1 vs 2, 1 vs 3, and 2 vs 3), and analogously for the 2D MDS-derived distance matrices.

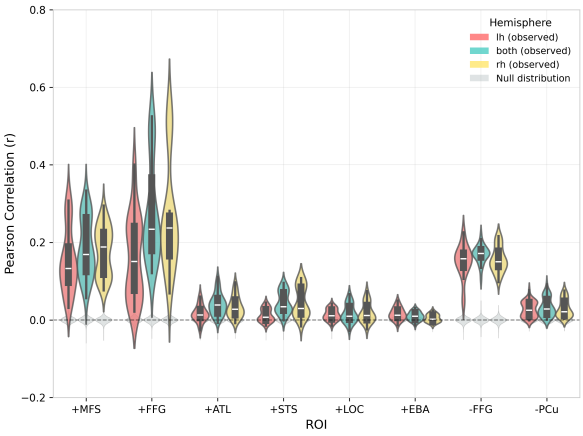
In both Pearson (Fig. 3.2a) and Spearman (Fig. 3.2b) analyses, the shuffled null distributions (gray) are tightly centered on zero, demonstrating that all observed correlations exceed chance. Critically, only ROIs +MTS, –FFG, and +FFG exhibit reliable trial-to-trial stability (mean Pearson $r = 0.19, 0.17, 0.28$; mean Spearman $\rho = 0.18, 0.16, 0.27$), whereas all other ROIs remain at or below $r, \rho \leq 0.05$. Pearson and Spearman yield virtually identical ROI rankings and effect sizes. While they still show above chance repeatability, the effect is rather low and might not be strong enough for any further downstream analyses. In Figure 3.3 the ratio of significant ($p \leq 0.05$) to insignificant p-values, that were obtained by the Pearson/Spearman correlation testing, is shown for each ROI. The ROIs showing the strongest effect sizes (+MFS, +FFG, –FFG) only contain significant p-values, whereas the ROI +EBA contains around $\approx$ 60% of p-values greater than the signficance threshold of 0.05.

To quantify whether even the worst-performing mask exceeds chance, we treat each of its N tests (e.g., 72 tests for +EBA) as a Bernoulli trial under the null hypothesis ($p_0 = 0.05$ for significance) and ask for the smallest integer k such that

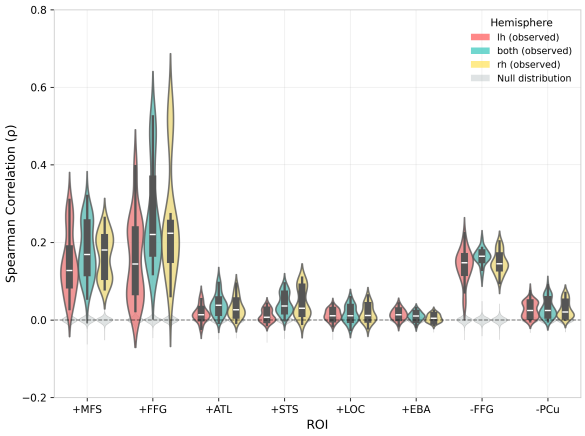
$$P(X \geq k \mid N, p_0) = \sum_{i=k}^{N} \binom{N}{i} p_0^i (1 - p_0)^{N-i} \leq 0.05$$

For +EBA, with N=72, solving this gives k=8. This means that at least 8 significant tests (a threshold ratio $k/N \approx 0.111$) are required to reject the null hypothesis that the observed significance rate is due to chance. +EBA actually yields 28 significant tests out of 72 ($\frac{28}{72} \approx 0.389$), comfortably above the 11.1% chance level and the required 8 significant tests, but still markedly lower than the other ROIs, reflecting its relatively weak trial-to-trial stability.

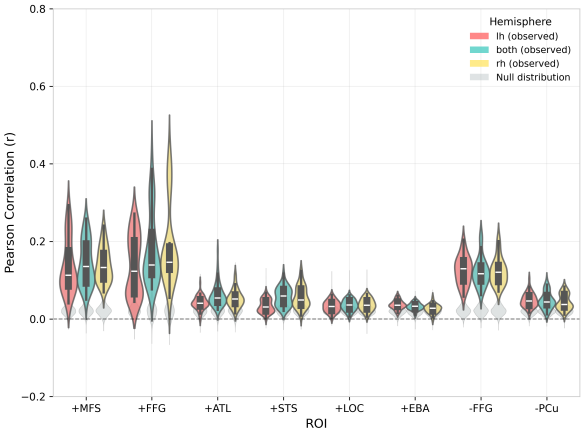
As they are used in later downstream analyses, I applied the same procedure to the 2D MDS embeddings (Fig. 3.2c & Fig. 3.2d), producing lower absolute correlations—as expected from dimensionality reduction—yet preserving the ordinal stability across masks.



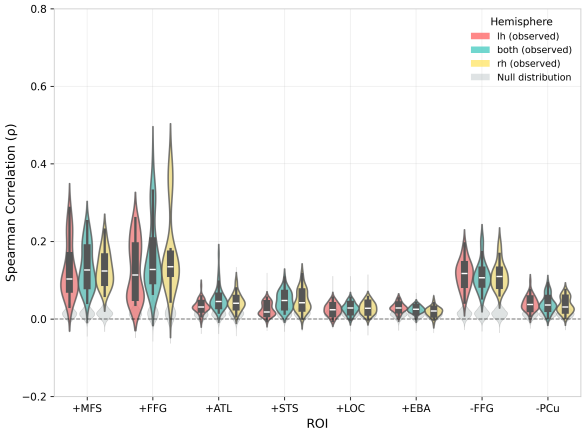
(a) RDM stability (Pearson correlation)



(b) RDM stability (Spearman correlation)



(c) Stability of 2D MDS embeddings across repeats (Pearson correlation)



(d) Stability of 2D MDS embeddings across repeats (Spearman correlation)

Figure 3.2: Trial-to-trial representational stability across ROIs: (a) RDM Pearson correlations; (b) RDM Spearman correlations; (c) Pearson correlations of 2D MDS embeddings; (d) Spearman correlations of 2D MDS embeddings.

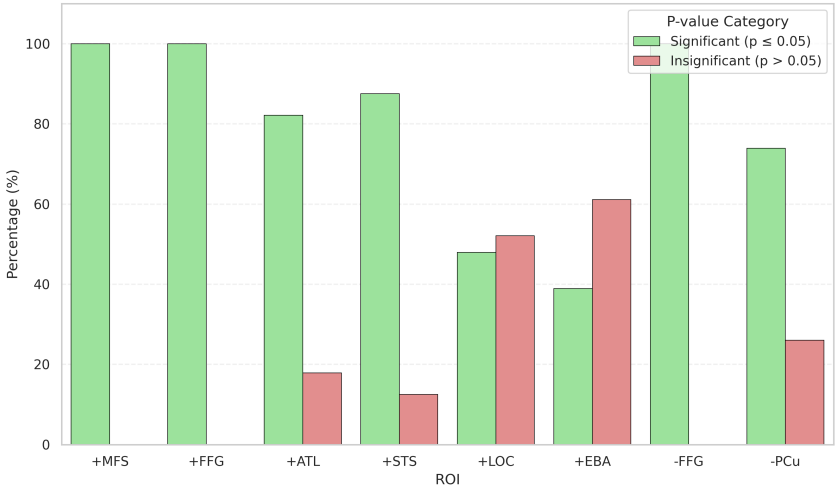


Figure 3.3: Ratio of significant ($p \leq 0.05$) to insignificant p-values obtained from trial repeatability correlation tests for each ROI.

Together, these RDM- and MDS-based analyses converge on the conclusion that only the face-selective and face-suppressive ROIs +MTS, –FFG, & +FFG support robust, repeatable representational geometry, and that this hierarchy endures even after embedding in a lower-dimensional space.

3.3.2 Quantifying encoded features

Having analyzed the ROIs representational stability, I then went on to examine if features are encoded in the ROIs using the Mantel-test. As seen in the previous section, downscaling from RDM spaces to MDS spaces resulted in a loss of information, which is why I only present the results of the mantel-test applied to the RDM space in the following. While Figure 3.4 presents one combined Fisher-z effect size and one combined p-value per ROI, Figure 3.5 displays, for each ROI and feature, the Fisher-z effect sizes (left panels) and combined $-\log_{10}(p)$ values (right panels).

It is shown that age and gender are effectively absent from the representational geometry: all Fisher-z values lie between -0.04 and $+0.01$, with the lowest combined p-value being $p = 0.005$ for gender in ROI +ATL. In contrast, face size and location yield more robust signals in a subset of ROIs. For face size, ROI +MTS ($z = 0.12$, $p < 1^{-12}$) and ROI +FFG ($z = 0.12$, $p < 1^{-12}$) show the largest values. For face location, ROIs +MTS ($z = 0.07$, $p < 1^{-6}$), -FFG ($z = 0.07$, $p < 1^{-6}$), and +FFG ($z = 0.10$, $p < 1^{-6}$) stand out. Thus, low-level geometric features (size and position) are encoded much more strongly than demographic attributes in these regions.

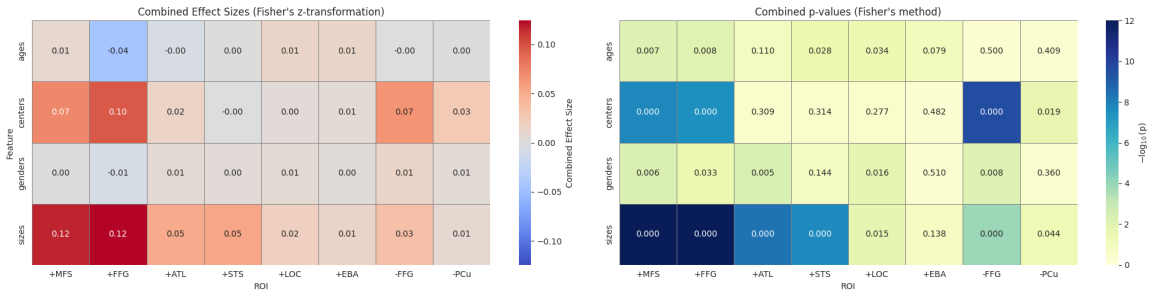


Figure 3.4: Summary per ROI of combined effect sizes (*left*) and combined p-values (*right*).

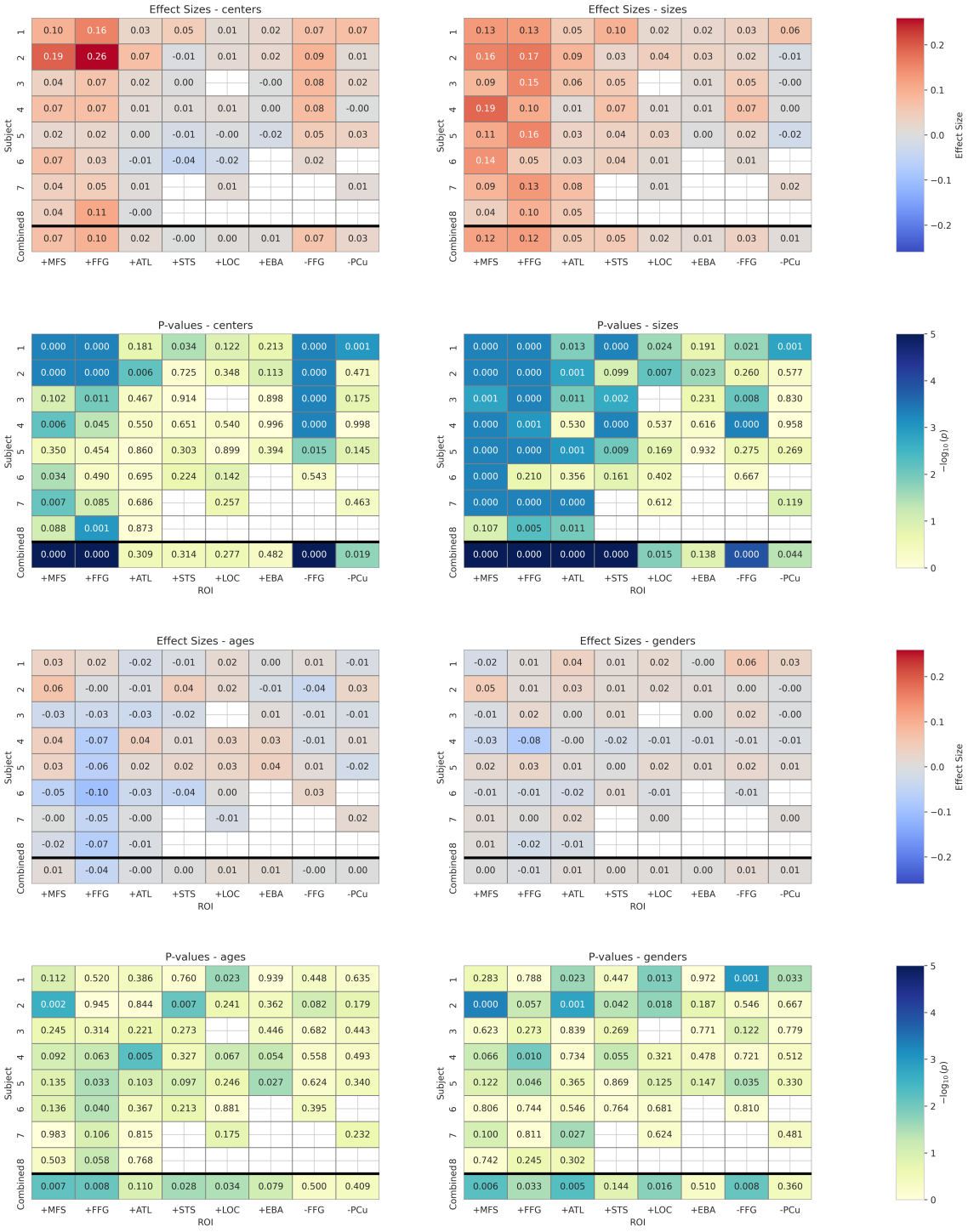
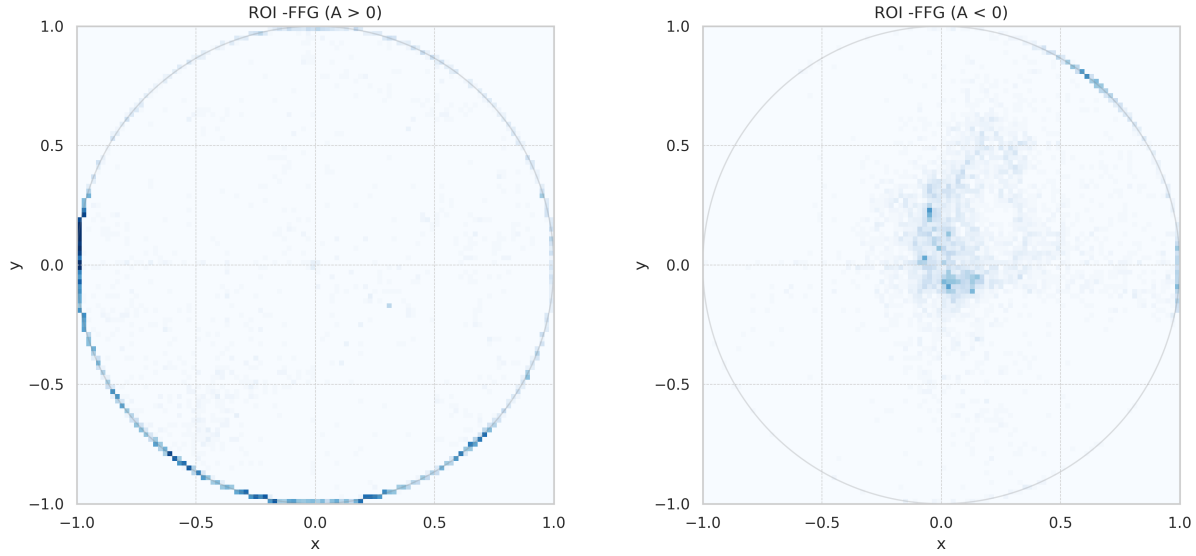


Figure 3.5: Complete Mantel-test (effect sizes and p-values in $-\log_{10}(p)$) results for each feature and per subject and ROI in RDM space. The bottom row shows the combined effect sizes or p-values across subjects for each ROI.

3.4 Voxel-wise Gaussian fitting

Having obtained the voxel-wise gaussian fits, I analyzed their characteristics and metrics with respect to the research question, which included analyzing their Gaussian

center's MDS position (x_0, y_0), their extent (sigma, σ, the standard deviation of the Gaussian) and the explained variance. Importantly, each face-selective ROI was optimized with positive amplitude bounds, while the face-suppressive ROIs were optimized with strictly negatively bounded amplitudes. If these restrictions were not explicitly imposed, the Gaussian center of the majority of the fits was optimized to be on the edge of the MDS space, as shown in Figure 3.6. Figure 3.6a shows that, when a positive amplitude bound is imposed on face-suppressive ROI -FFG, most Gaussian centers cluster at the perimeter of the MDS map. By contrast, with a negative amplitude bound (Figure 3.6b), the centers distribute more centrally.



(a) Fits with a positive amplitude bound cluster at the perimeter of the MDS

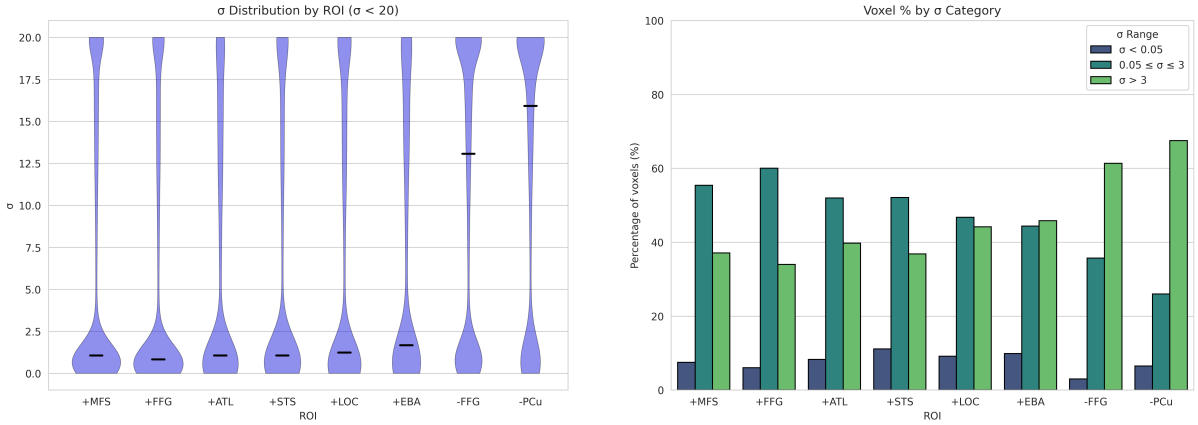
(b) Fits with a negative amplitude bound exhibit a more distributed pattern of Gaussian centers

Figure 3.6: Gaussian center locations of voxel-wise Gaussian fits in face-suppressive ROI – FFG under different amplitude constraints.

First, I analyzed the distribution of the sigmas of the fittings, as they serve as a proxy of how localized a voxel's tuning is in the representational space. Our stimuli are embedded by MDS and lie within the unit circle (so that the maximum pairwise distance is ≈ 2). I therefore defined three σ-ranges:

$$\begin{aligned} \sigma < 0.05 \quad & \big(0.025 \times \text{diameter},\ \text{single-stimulus tuning}\big), \\ 0.05 \leq \sigma \leq 3 \quad & \big(0.025 \text{ to } 1.5 \times \text{diameter},\ \text{group tuning}\big), \\ \sigma > 3 \quad & \big(\text{flat over the space},\ \text{no specific tuning}\big). \end{aligned}$$

Applying these definitions, Figure 3.7 shows that the face-selective regions' voxels mostly fit Gaussians with a meaningful spread ($0.05 \leq \sigma \leq 3$), with ROI +FFG having around 60% of the fits in that range. For face-suppressive regions, however, the opposite is true, as the majority of fits have very large sigmas (e.g. ROI –PCu with 60% in the $\sigma > 3$ range). However, across all ROIs, few voxel fits correspond to a single stimulus, as indicated by the small proportion ($\leq 10\%$) of sigma values in that range.

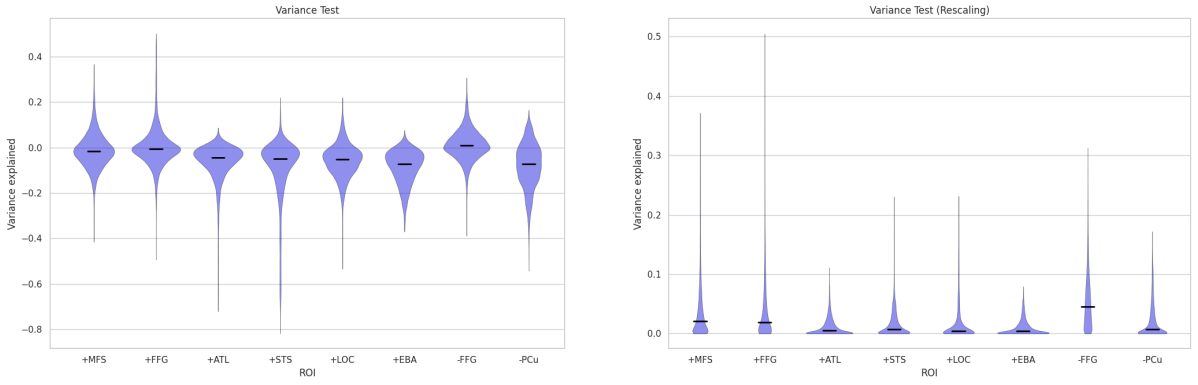


(a) Violinplot showing distribution of sigmas.

(b) Barplot of sigmas categorized by three sigma-ranges.

Figure 3.7: Distributions of sigma values across voxels: (a) violinplot of the raw sigma distribution and (b) barplot of sigmas grouped into low, medium, and high ranges.

Next, I analyzed the variances of the fits for each ROI and across subjects. In Figure 3.8a you can see the explained variance of the fits on the test set, and in Figure 3.8b the explained variance after optimal amplitude rescaling on the test set.



(a) Explained variance of the fits on held-out test data.

(b) Explained variance after optimal amplitude rescaling on the test data.

Figure 3.8: Violin plots of explained variance for each ROI across subjects: (a) on held-out test data and (b) after optimal amplitude rescaling. The width of each violin reflects the density of voxel-level variance values.

Figure 3.8a shows that all eight ROIs yield very modest explained variance on unseen data, with ROIs +MTS, -FFG and +FFG centered near zero and the remaining ROIs consistently negative. However, this seems likely to result from global differences in response amplitude between the training and test data's scan sessions that were larger than differences between images.

We compensate for this by applying an optimal rescaling of each Gaussian's amplitude & offset (see Figure 3.8b) between train and test data. This increases variance explained across the board but the rank order remains unchanged—with ROIs +MTS, -FFG and +FFG still achieving the highest (or least negative) values.

Finally, to emphasize voxels with spatially specific tuning, I retained only those with fitted $\sigma < 3$ due to the broad Gaussian. Figure 3.9 shows the Gaussian center (x_0, y_0) coordinates in the MDS embedding for each ROI, pooled across subjects. Clear, ROI-specific clustering patterns emerge. Some regions concentrate near the center of the face-space, while others form arcs along its periphery.

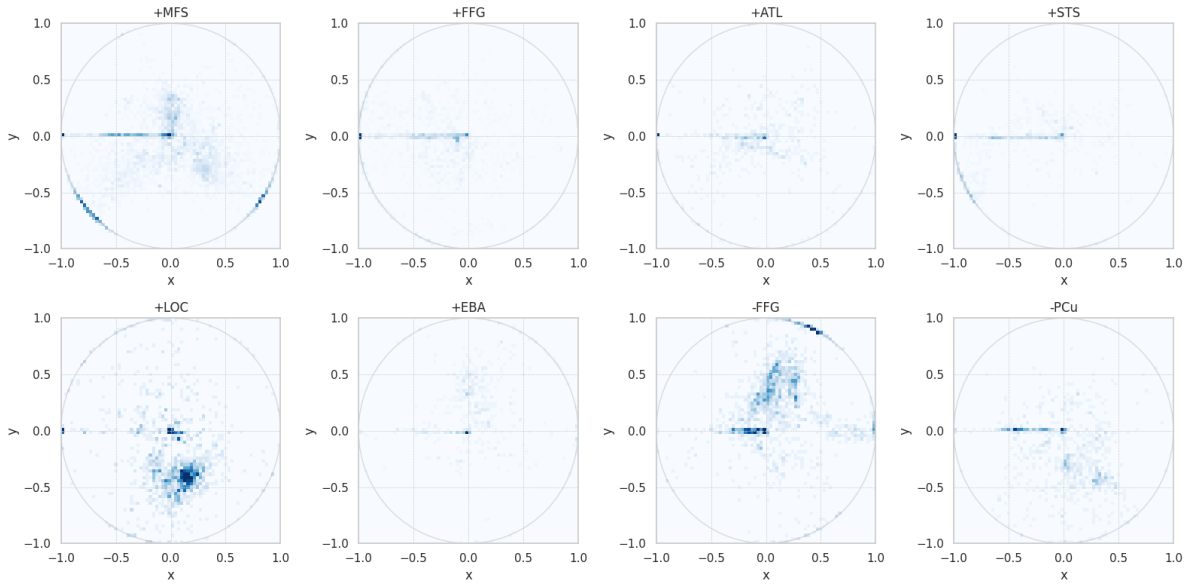


Figure 3.9: Density plots of Gaussian centers (x_0, y_0) for all voxels with $\sigma < 3$, by ROI. Each subplot shows the pooled distribution across subjects; darker shading indicates higher voxel counts.

3.5 Relating Representational Geometry to Cortical Topography

To test the hypothesis that the structure of voxel-wise tuning reflected in the 2D MDS embedding is not arbitrary but instead mirrors the physical layout of the cortex I next asked whether voxels that lie close together in representational space also sit close together on the cortical sheet. Concretely, for each ROI and hemisphere I correlated pairwise geodesic distances along the white-matter surface with distances between Gaussian-fit centers in MDS space.

In the following, I show the result of this analysis. The distribution of median Spearman's ρ across subjects for each ROI is shown in Figure 3.10. Medians ranged from 0.20 in +EBA to 0.41 in +MFS. Regions +MFS and –FFG showed the highest consistency across subjects (both 0.41), whereas +EBA and –PCu were lowest (0.20 & 0.21). An example mapping of Gaussian-fit centers onto the cortical surface for the +MFS ROI in Subject 2, represented in polar coordinates, is shown in Figure 3.11.

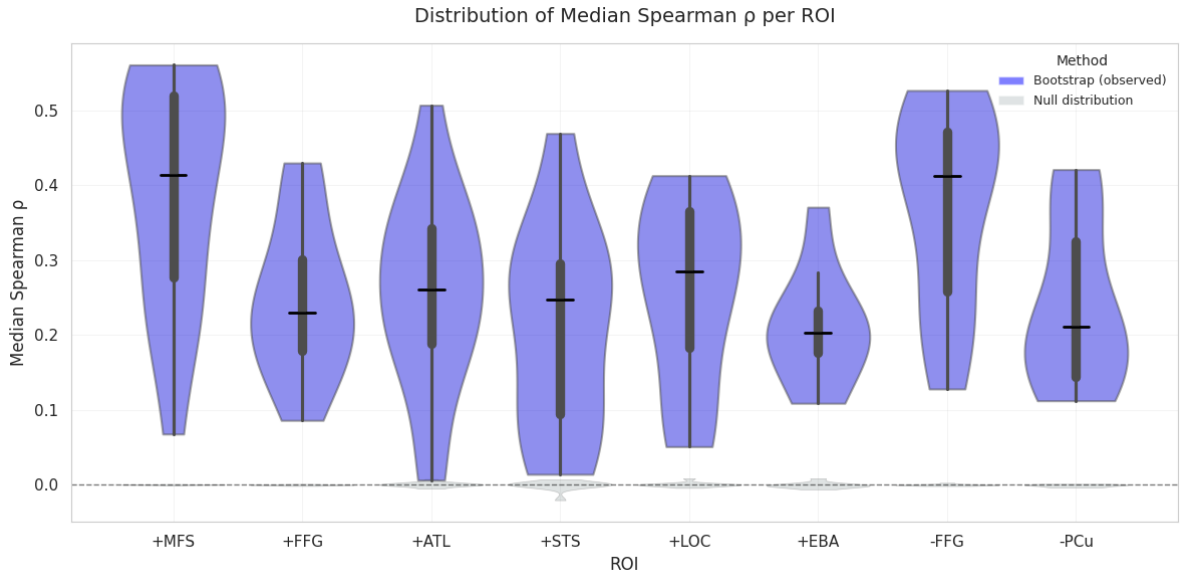
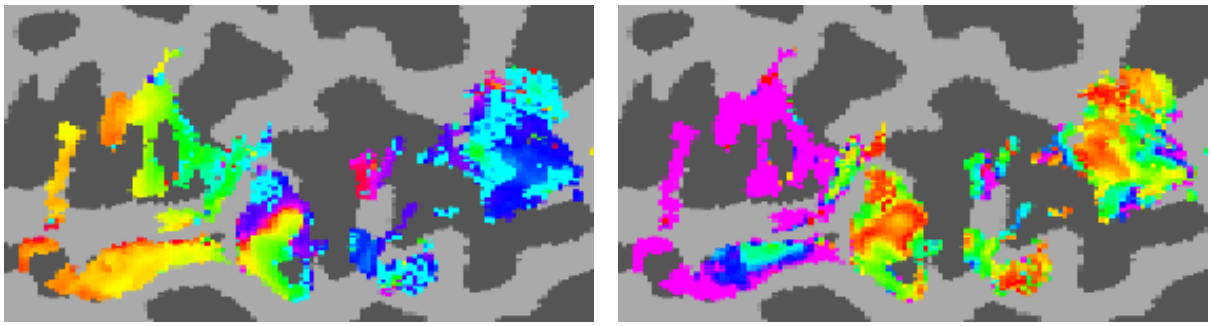


Figure 3.10: Distribution of median Spearman's ρ across subjects for each ROI. Black bars mark the median for each violin.



(a) Polar angle of Gaussian centers (0 to 2π) in +MFS.

(b) Radial distance of Gaussian centers (0 to 1) in +MFS.

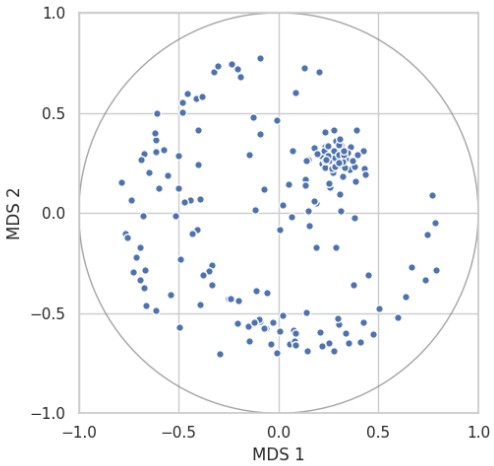
Figure 3.11: Polar coordinate representation of Gaussian center locations for Subject 2 in the +MFS ROI.

3.6 Representational Similarity Analysis of Neural Network Activations

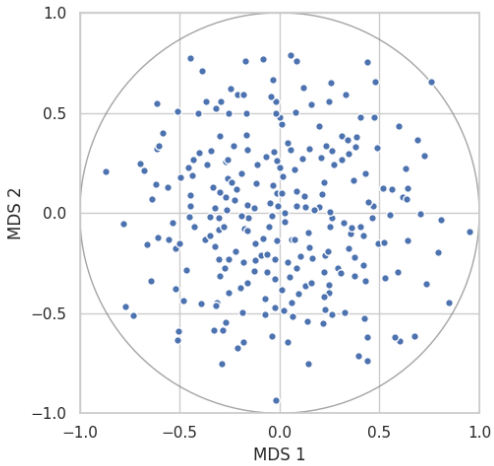
In order to directly compare cortical representational geometries with those of deep neural networks, I extracted activation patterns from every layer of the RetinaFace and GenderAge models by presenting them with the same face-set stimuli. Using these network activations alongside each subject's ROI responses, I then computed representational dissimilarity matrices (RDMs) and visualized them via multidimensional scaling (MDS).

Figure 3.12 compares the 2D MDS embeddings for the final layer of the RetinaFace network (3.12a) and ROI +FFG of subject 1 (3.12b). The network's embedding reveals several tight clusters indicating a highly structured face-space, whereas the +FFG embedding is much more diffuse, with no obvious cluster separation.

To test whether basic image features are encoded in that network embedding, I overlaid two stimulus properties onto the RetinaFace's layers MDS coordinates (Figure 3.13). In Figure 3.13a, each point is linked by an arrow from the image center to the predicted bounding-box center, with arrow direction and length encoding spatial offsets. In Figure 3.13b, points are colored by predicted bounding-box area (pixel^2), as shown by the adjacent colorbar. Both plots exhibit systematic gradients. Points with similar box centers and sizes cluster together indicating that spatial location and scale information are indeed reflected in the representational geometry of the final layer of the RetinaFace network.

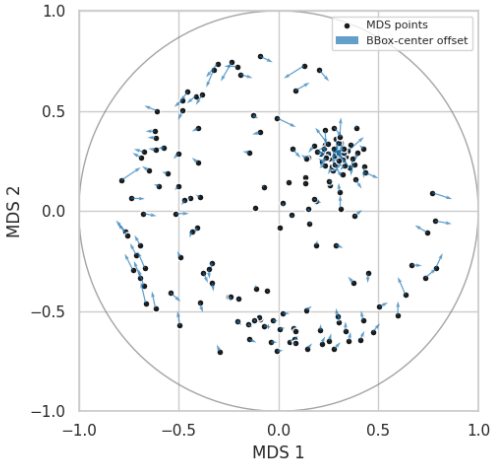


(a) RetinaFace final-layer activations showing clearly separated clusters of stimuli.

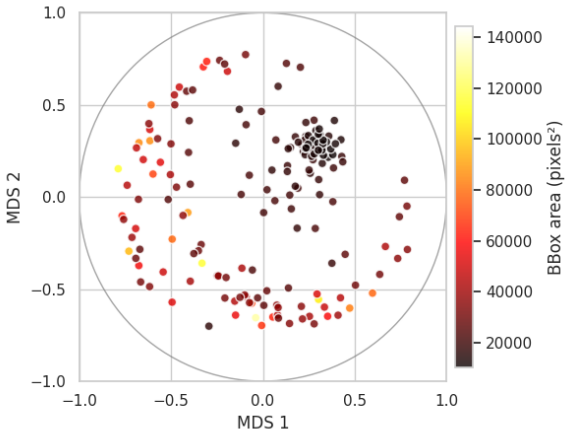


(b) +FFG responses (Subject 1) with a relatively homogeneous embedding, suggesting less differentiated representational geometry.

Figure 3.12: Comparison between MDS embeddings of representational dissimilarity matrices.



(a) Arrows show each stimulus's bounding-box center offset from the image center (direction=angle, length=distance)



(b) Stimuli colored by bounding-box area (pixel²)

Figure 3.13: Feature encoding in MDS space. Systematic clustering along these arrows and color gradients demonstrates that the network's final-layer representation embeds both spatial and scale features.

Finally, to move beyond visual inspection, I ran Mantel tests between each feature RDMs (age, gender, bounding-box-center, bounding-box-size) and each model/ROI RDMs (all layers of both networks plus each functional ROI). The full matrix of correlation coefficients and p-values is presented in Figure 3.14.

Metric	p-value				Effect size			
Feature	ages	centers	genders	sizes	ages	centers	genders	sizes
layer								
Relu_2	0.577	0.301	0.066	0.596	-0.018	0.036	-0.045	0.018
Relu_5	0.881	0.907	0.087	0.267	0.005	0.005	-0.046	0.043
Relu_8	0.649	0.801	0.210	0.227	0.017	0.009	-0.034	0.045
Relu_12	0.659	0.063	0.969	0.295	0.014	0.061	0.001	0.034
Relu_16	0.956	0.481	0.740	0.784	0.002	0.025	0.008	-0.010
Relu_19	0.635	0.032	0.838	0.720	0.014	0.069	0.004	-0.011
Relu_23	0.919	0.306	0.434	0.621	0.003	0.037	0.019	-0.017
Relu_26	0.659	0.036	0.599	0.416	0.014	0.071	0.012	0.026
Relu_30	0.923	0.319	0.142	0.351	0.003	0.036	0.038	-0.033
Relu_33	0.667	0.010	0.090	0.702	0.014	0.087	0.040	-0.012
Relu_40	0.688	0.113	0.043	0.186	0.013	0.056	0.049	-0.044
Relu_43	0.990	0.000	0.150	0.633	0.000	0.159	0.028	-0.012
Relu_47	0.504	0.029	0.020	0.239	0.022	0.076	0.055	-0.038
Relu_50	0.698	0.000	0.137	0.600	0.012	0.135	0.031	-0.014
Relu_54	0.316	0.032	0.028	0.198	0.032	0.075	0.051	-0.042
Relu_57	0.590	0.000	0.200	0.854	0.017	0.113	0.028	0.006
Relu_61	0.301	0.027	0.066	0.254	0.034	0.077	0.043	-0.037
Relu_64	0.450	0.000	0.158	0.851	0.022	0.135	0.029	0.006
Relu_71	0.192	0.000	0.050	0.780	0.040	0.137	0.043	-0.008
Relu_74	0.347	0.000	0.115	0.804	0.028	0.175	0.033	0.008
Relu_78	0.209	0.000	0.074	0.681	0.037	0.145	0.038	0.012
Relu_81	0.445	0.000	0.076	0.753	0.021	0.192	0.035	0.009
Relu_88	0.375	0.000	0.031	0.010	0.021	0.282	0.037	0.058
Relu_91	0.681	0.000	0.169	0.000	0.009	0.314	0.022	0.097
Relu_95	0.709	0.000	0.103	0.000	0.009	0.307	0.029	0.089
Relu_98	0.695	0.000	0.265	0.000	0.011	0.328	0.021	0.088
Relu_102	0.715	0.000	0.231	0.000	0.009	0.339	0.021	0.121
Relu_157	0.362	0.000	0.501	0.665	0.028	0.159	0.015	0.014
Relu_160	0.570	0.000	0.609	0.177	0.018	0.136	0.012	-0.042
Relu_163	0.222	0.000	0.960	0.514	0.037	0.140	0.001	-0.020
Sigmoid_171	0.867	0.094	0.271	0.000	0.005	0.053	-0.025	0.231
Relu_180	0.977	0.000	0.002	0.043	0.001	0.183	0.077	-0.069
Relu_183	0.774	0.000	0.010	0.985	0.010	0.171	0.064	0.001
Relu_186	0.751	0.000	0.008	0.091	0.011	0.279	0.069	-0.057
Sigmoid_194	0.087	0.000	0.270	0.000	-0.025	0.220	0.012	0.056
Relu_203	0.277	0.001	0.019	0.000	0.041	0.129	0.065	0.318
Relu_206	0.278	0.000	0.017	0.000	0.041	0.188	0.065	0.278
Relu_209	0.171	0.000	0.047	0.000	0.050	0.224	0.054	0.350
Sigmoid_217	0.446	0.000	0.010	0.000	0.024	0.332	0.061	0.336

(a) Result of face-detection network RetinaFace

Metric	p-value				Effect size			
Feature	ages	centers	genders	sizes	ages	centers	genders	sizes
layer								
conv_1_relu	0.023	0.062	0.938	0.007	-0.052	0.045	-0.001	0.064
conv_2_dw_relu	0.186	0.025	0.908	0.003	-0.032	0.056	0.002	0.070
conv_2_relu	0.572	0.404	0.491	0.003	-0.012	0.019	0.011	0.061
conv_3_dw_relu	0.204	0.002	0.243	0.000	-0.033	0.100	0.024	0.161
conv_3_relu	0.316	0.001	0.438	0.000	-0.023	0.088	0.013	0.170
conv_4_relu	0.446	0.002	0.354	0.000	-0.018	0.087	0.016	0.161
conv_4_dw_relu	0.331	0.001	0.517	0.000	-0.023	0.094	0.012	0.168
conv_5_relu	0.454	0.027	0.084	0.000	-0.020	0.061	0.034	0.102
conv_5_dw_relu	0.967	0.002	0.154	0.000	-0.001	0.080	0.026	0.132
conv_6_relu	0.553	0.046	0.067	0.000	-0.016	0.055	0.037	0.107
conv_6_dw_relu	0.508	0.022	0.132	0.000	-0.017	0.064	0.029	0.111
conv_7_dw_relu	0.912	0.106	0.003	0.001	-0.003	0.047	0.064	0.104
conv_7_relu	0.852	0.486	0.000	0.002	0.005	0.019	0.100	0.082
conv_8_relu	0.049	0.200	0.000	0.001	0.048	0.032	0.151	0.080
conv_8_dw_relu	0.220	0.342	0.000	0.001	0.031	0.025	0.115	0.082
conv_9_dw_relu	0.003	0.262	0.000	0.001	0.062	0.025	0.192	0.069
conv_9_relu	0.000	0.117	0.000	0.001	0.086	0.033	0.255	0.071
conv_10_dw_relu	0.000	0.169	0.000	0.002	0.113	0.028	0.307	0.063
conv_10_relu	0.000	0.090	0.000	0.004	0.145	0.036	0.359	0.061
conv_11_dw_relu	0.000	0.214	0.000	0.013	0.165	0.025	0.395	0.048
conv_11_relu	0.000	0.191	0.000	0.031	0.195	0.028	0.444	0.045
conv_12_relu	0.000	0.976	0.000	0.409	0.241	-0.001	0.443	0.017
conv_12_dw_relu	0.000	0.404	0.000	0.112	0.217	0.018	0.467	0.032
conv_13_dw_t0_relu	0.000	0.475	0.000	0.686	0.183	0.013	0.621	0.006
conv_13_t0_relu	0.000	0.170	0.000	0.376	0.068	0.023	0.769	0.014
conv_13_t1_relu	0.000	0.047	0.000	0.914	0.297	-0.033	0.185	-0.002
conv_13_dw_t1_relu	0.000	0.101	0.000	0.709	0.344	-0.033	0.357	0.007
conv_14_dw_t1_relu	0.000	0.106	0.000	0.978	0.354	-0.026	0.138	0.000
conv_14_t0_relu	0.204	0.237	0.000	0.625	0.021	0.022	0.843	0.008
conv_14_t1_relu	0.000	0.004	0.000	0.986	0.298	-0.042	0.139	0.000
conv_14_dw_t0_relu	0.000	0.582	0.000	0.719	0.085	0.009	0.737	-0.005

(b) Result of gender & age prediction network

Figure 3.14: Mantel test results showing correlation coefficients and p-values between stimulus-feature RDMs (gender, age, bounding-box center, bounding-box area) and model/ROI RDMs for the face-detection network (a) and the gender & age prediction network (b). P-values are colorcoded in the $-log_{10}(p)$ scale with 5 being the maximum value marked by dark blue.

The tables show that in the face-detection network (3.14a), correlations with age and gender remain essentially zero ($|r_{\text{obs}}| < 0.07$) across all ReLU layers, even where a few p-values dip below the significance threshold 0.05, indicating no meaningful en-

coding of demographic attributes. By contrast, correlations with bounding-box center and size are minimal in early layers ($r_{\text{obs}} \approx 0.005$–$0.06$, $p > 0.2$) but gradually rise in the deeper ReLU blocks (reaching up to $r_{\text{obs}} \approx 0.34$ with $p < 1^{-5}$), suggesting that spatial localization features are refined only toward the network's end. In the gender-&-age network, box size shows a small but significant correlation already after the first few layers peaking at ($r_{\text{obs}} \approx 0.17$, $p < 1^{-5}$) before tapering off. In contrast, age and gender correlations remain near zero in early layers and then steadily increase in mid-to-late layers, ultimately peaking in late layers (gender $r_{\text{obs}} \approx 0.843$; age $r_{\text{obs}} \approx 0.354$ with $p < 1^{-5}$).

4. Discussion

The aim of this study was to understand how the brain represents individual faces and how that representation compares to deep neural networks trained on face tasks. To do this, I built a multistep analysis pipeline that (1) identifies face-selective regions in the high-resolution fMRI data from the Natural Scenes Dataset (NSD) and (2) examines what features those regions encode. Across subjects, I defined eight regions of interest (ROIs)—six face-selective patches and two non-face controls—and measured how consistent their responses were across trials and which face-related features they tracked. I found that the face-selective middle temporal sulcus (+MTS/OFA), the face-suppressive fusiform gyrus (–FFG), and the face-selective fusiform gyrus (+FFG) showed strong repeatability and encoded low-level features like face position and size.

4.1 Interpreting Results

4.1.1 Face-Selectivity vs. Trial-Repeatability

I began by comparing classical face-selective regions in human ventral occipitotemporal cortex to the well-characterized face patches in macaques, noting that, as in non-human primates, the precise number and strength of face-responsive locations varied notably between subjects (Tsao et al., 2008). Some participants exhibited only a handful of weakly face-responsive clusters (low-t statistics), whereas others showed robust, spatially clustered activations surrounding the fusiform gyrus and adjacent areas.

Despite this variability, a clear dissociation emerged when I examined trial-to-trial response consistency. The face-selective middle temporal sulcus (+MTS/OFA), the face-suppressive fusiform gyrus (–FFG), and the face-selective fusiform gyrus (+FFG) displayed significant trial-repeatability, whereas the remaining areas showed significantly less to minimal consistency. Although having only two to three presentations per image undoubtedly reduces statistical power, the persistence of these differences suggests genuine functional distinctions rather than sampling noise.

One explanation is that certain ROIs, while face-responsive, lack a reproducible

internal structure for my stimulus set. In other words, fMRI can detect whether a region responds more to faces than to other objects, but if that response pattern is not reliably evoked by the same images, further structure (for example, feature tuning) cannot be uncovered. Conversely, a region doesn't need to be face-responsive but can also be clearly face-suppressive to exhibit repeatable patterns if it contains stable subpopulations of voxels that consistently coactivate.

One possible explanation for the robust repeatability of the face-suppressive fusiform gyrus (–FFG) is that its immediate adjacency to the strongly face-selective, highly repeatable fusiform gyrus (+FFG) produces vascular or hemodynamic spillover, causing correlated activations in –FFG despite its lack of true face selectivity.

Another possibility is that the face-suppressive fusiform gyrus (–FFG), rather than merely reflecting passive spillover from its face-selective neighbor, may carry its own internally consistent pattern of deactivation that varies systematically with face stimulus parameters. In this view, –FFG suppresses its overall activity in response to faces, but the relative pattern of that suppression across voxels is modulated by factors such as face position and size, yielding high trial-repeatability even in a region that shows net negative activation to faces. This idea finds a close parallel in work on the Default Network (DN), where distinct spatial deactivation patterns encode purely visual stimulus attributes. The DN deactivates in a position-dependent manner, implying that high-level "task-negative" regions can nonetheless carry precise spatial information via deactivation topographies (Szinte & Knapen, 2019). By analogy, –FFG could use voxel-wise suppression profiles to represent where (and how large) a face appears, producing the stable, repeatable patterns we observed.

Supporting the hypothesis that only representationally stable ROIs potentially show some internal structure, my permutation analyses revealed that only the trial repeatable ROIs encoded any face-related features at above chance levels.
Face-selective but non-repeatable ROIs failed to show significant structure in their representational dissimilarity matrices, underscoring the tight link between repeatability and feature encoding.

Finally, voxel-wise Gaussian fitting mirrored these findings: mean test variances for the face-selective middle temporal sulcus (+MTS/OFA), the face-suppressive fusiform gyrus (–FFG), and the face-selective fusiform gyrus (+FFG) clustered around zero, whereas non-repeatable areas exhibited substantially negative variances. Even

after allowing a linear rescaling of the Gaussian models, the same rank ordering persisted, with the face-selective middle temporal sulcus (+MTS/OFA), the face-suppressive fusiform gyrus (–FFG), and the face-selective fusiform gyrus (+FFG) showing the highest retained variance. This pattern further demonstrates that repeatability of stimulus-evoked patterns underlies both my RSA results and spatial tuning estimates.

4.1.2 Low- vs. High-Level Feature Encoding

To assess the granularity of face-related information represented in each ROI, I performed a permutation analysis utilizing a mantel test comparing the representational spaces obtained through RSA and low-level (face position and size) and high-level (gender and age) features. The results (see Figure 3.4) revealed a clear dissociation. Low-level spatial attributes were significantly encoded in the trial-repeatable ROIs, as evidenced by greater similarity among responses to stimuli sharing the same position or size, whereas high-level attributes such as gender and age failed to produce reliable representational structure above-chance.

This pattern suggests that the face-selective middle temporal sulcus (+MTS/OFA), the face-suppressive fusiform gyrus (–FFG), and the face-selective fusiform gyrus (+FFG) are tuned to basic visual parameters and do not show any signs of encoding higher-level features. However, this result directly contradicts previous work that found signs of high-level features, such as sex and gender, being encoded in the FFA (Contreras et al., 2013) as well as a study suggesting that the FFA encodes "social face information" as opposed to "low-level image properties" (Tsantani et al., 2021). A possible explanation for this is that the NSD was collected while performing a very unrelated task of remembering if the currently shown image has been seen before. In the context of this study, this could be regarded as passive viewing conditions, while the participants in the other study had to perform the specific task of recognizing gender and sex. However, the absence of detectable age and gender coding could also stem from limited sample variability or insufficient statistical power for these higher-order features.

Nonetheless, the robust low-level feature effects underscore the primacy of spatial configuration in early stages of face representation, in accordance with hierarchical models positing a progression from elemental to semantic encoding along the ventral stream (Hubel & Wiesel, 1962; Riesenhuber & Poggio, 1999).

4.1.3 Spatial Tuning & Cortical Topography

To probe the spatial tuning characteristics of each ROI, I fitted voxel-wise Gaussian models within the 2D MDS space and analyzed the distribution of fitted widths (σ). Strikingly, face-selective regions exhibited a substantially lower proportion of broad fits ($\sigma > 3$) compared to face-suppressive areas — approximately 40% versus 60% (see Figure 3.7). This suggests that even when variance explained is limited due to poor trial-repeatability, face-selective ROIs tend to exhibit more localized tuning within the representational space, indicating selective responses to restricted stimulus subsets.

These narrower distributions do not reflect anatomical receptive field sizes but instead point to functional specialization. A greater proportion of voxels in face-selective regions appear tuned to specific areas within the representational geometry. This observation is consistent with the notion of modular processing in the ventral visual stream, where face-selective areas may support fine-grained discriminations between similar stimuli. It aligns qualitatively with prior anatomical findings that face-processing regions reside in areas associated with high feature specificity (Wandell et al., 2007).

Thus, while low repeatability in some ROIs constrained overall model fits, the narrower σ values in face-selective regions nevertheless reveal robust and localized representational tuning.

4.1.4 RSA in Neural Networks

To bridge my fMRI findings with computational models, I applied representational similarity analysis (RSA) to two convolutional neural networks trained respectively for gender and age estimation (the "GenderAge" network) and for generic face recognition (the "RetinaFace" network). For each network, I extracted layerwise activations in response to the same face stimuli used in the NSD experiment, computed representational dissimilarity matrices (RDMs), and then correlated these RDMs with low-level (bounding-box-size, bounding-box-position) and high-level (age, gender) feature matrices.

Generally, I observed a consistent trend where effect sizes for encoded features increased with network depth (see Figure 3.14). This was particularly evident for features explicitly targeted by the networks' training objectives: age and gender in the GenderAge model, and implied face localization features (such as bounding box infor-

mation) in the RetinaFace model. This pattern highlights that the networks primarily learn and attend to features that are explicitly relevant to their defined loss functions. Moreover, it confirms a hierarchical processing scheme where early layers predominantly encode low-level visual features, while higher layers develop more abstract, high-level feature representations of faces.

However, the encoding of bounding-box size in the GenderAge network presented a more complex pattern. Initially, I observed strong, significant encoding of bounding-box size in the earliest convolutional layers of the GenderAge network, reinforced by a consistently low p-value. Figure 3.14b shows that, in the earliest convolutional layers (conv1_relu–conv5_relu), activations encode bounding-box size with a consistently strong, positive correlation (r_{obs}) and highly significant p-values. See columns `p_values` and `r_obs`. This initial effect likely reflects that different crop sizes induce systematic changes in input resolution. A 128x128 face patch versus a 32x32 patch, for instance, will elicit distinct early-layer activations sensitive to spatial frequency and edge content. In other words, size variations translate directly into low-level feature differences that early filters readily capture. Interestingly, after the first few layers, the effect size for bounding-box size in the GenderAge model gradually faded (e.g., conv_7_relu). I hypothesize that although all images are ultimately resized to the same input dimensions before entering the network, this rescaling operation might still leave detectable artifacts or patterns that are quantifiable by lower layers. These low-level artifacts, however, likely become irrelevant as the network progresses to extract more high-level, semantic features for gender and age estimation, leading to their diminishing representation in intermediate layers.

Moving deeper into the shared backbone of the GenderAge network, both age and gender effects gradually increased, suggesting these mid-level layers develop feature representations useful for both tasks. However, upon entering the task-specific regression heads, I saw a striking divergence. Gender encoding surged in the gender head while age encoding dropped nearly to zero (and vice versa in the age head). This alternation of high effect sizes between the two features mirrors the networks' explicit optimization objectives. Each head "focuses" on its assigned label at the expense of the other. Such sharp task-driven selectivity contrasts with the graded, overlapping feature representations I observed in the brain's face patches.

Moreover, unlike the variable trial-repeatability in fMRI data, the networks yielded perfectly deterministic RDMs across repeated passes. This consistency underscores a

fundamental difference: neural networks, absent biological noise and hemodynamic variability, will always reproduce the same representational geometry for a given stimulus set. While this determinism aids in theoretical interpretability, it also highlights that engineered models may lack the flexible encoding and context-dependent dynamics present in human cortex.

Overall, my RSA results in neural networks reinforce two key points. First, task specificity strongly shapes representational formats: models trained for a single objective show abrupt transitions in feature encoding aligned with their loss functions. Second, low-level spatial attributes dominate early processing stages, while high-level semantic features arise only where networks are directly supervised to extract them. By contrasting these findings with the brain's more nuanced, task-dependent yet noisy patterns, I gained insight into both the power and the limitations of current deep-learning analogues for human face recognition.

4.2 Limitations

This study has several clear limitations. First, the NSD dataset presented specific challenges for this study concerning the face stimuli. Out of approximately 10,000 total images seen by participants, only about 2% met the requirements for our face stimulus set, resulting in a low sample size of usable face images. Additionally, each of these selected face stimuli was presented only two or three times across the entire experiment. A higher number of repetitions would undoubtedly increase statistical power for detecting subtle neural patterns. Moreover, the NSD data collection spanned many sessions over a year, which introduces potential problems such as representational drift in neural responses and varying subject conditions across different sessions.

Second, participants were performing a memory task that involved seeing many different objects with only occasional face images, meaning they were not explicitly judging faces. Consequently, they might not have engaged the brain's full face-processing machinery. Third, I drew my ROIs by hand on each subject's brain scans. This introduces potential bias and makes it tougher to compare across individuals. An automated clustering approach would be more objective. Fourth, the face-suppressive fusiform gyrus (–FFG) sits immediately next to the strongly face-selective fusiform gyrus (+FFG). This proximity means that blood flow from one could spill over into the other, potentially making the face-suppressive fusiform gyrus (–FFG) appear more re-

peatable than it genuinely is. Fifth, I only looked at a handful of features (position, size, age, gender) and used simple, round Gaussian models for each voxel's tuning. There are likely other important dimensions (expressions, viewpoints) and tuning shapes I missed. Finally, the NSD dataset covers just a few well-sampled individuals, so it is unknown how well these results hold up in larger, more diverse groups.

4.3 Future Work

There are a number of straightforward ways to build on this work. Future research should also thoroughly analyze individual subject patches, rather than focusing predominantly on common or shared locations across subjects as was done in the current analysis. A new fMRI study could be conducted where participants explicitly judge identity, gender, age, and emotion, and more trials per condition could be included to boost power for detecting high-level features. Switching to an automated ROI definition, such as by clustering activation maps, would remove hand-drawing bias and improve consistency across subjects. The feature set for RSA could also be expanded to cover expressions, head angle, identity strength, and so on, using deep-network embeddings to capture richer dissimilarities. For receptive-field mapping, moving from simple isotropic Gaussians to anisotropic or multi-peak models could better match real neural tuning. Adding MEG or EEG (or intracranial recordings) alongside fMRI would allow for tracking when low- versus high-level features emerge in time (Cichy & Oliva, 2020). A complementary direction is to investigate alternative (dis)similarity metrics for constructing RDMs such as Mahalanobis distance, crossnobis estimators, or information-theoretic divergences (Walther et al., 2016) and to compare a range of embedding algorithms (e.g., non-metric MDS, t-SNE, UMAP) that do not rely on the parametric assumptions of classical MDS. This would test the robustness of the representational geometries observed and might reveal structures obscured by correlation-based RDMs and metric MDS. Finally, applying this pipeline to other category-selective areas, like the PPA for places, EBA for bodies, or VWFA for words, would reveal whether the link between repeatability and feature encoding holds across the whole ventral stream.

4.4 Conclusion

This study demonstrates a versatile analysis pipeline capable of uncovering both repeatable and spatially-tuned face-selective and face-suppressive regions in human visual cortex. Despite the NSD's passive-viewing design and limited repetitions, I successfully identified multiple face patches that reliably encode low-level spatial attributes such as position and size. However, my inability to detect gender or age coding underscores how task demands, sampling power, and ROI definition critically shape representational inferences. By contrasting human fMRI patterns with deterministic neural-network models, I highlight fundamental differences in noise, task specificity, and feature progression along a processing hierarchy. Overall, these results confirm that early face representations are grounded in spatial configuration, provide a roadmap for refining region definitions and feature sets, and pave the way for more targeted investigations of semantic face encoding in the brain.

A. Appendix

A.1 Code availability

All code used to carry out the data processing, statistical analyses, and figure generation presented in this study is publicly accessible at:

https://github.com/maxschwalenberg/master_thesis

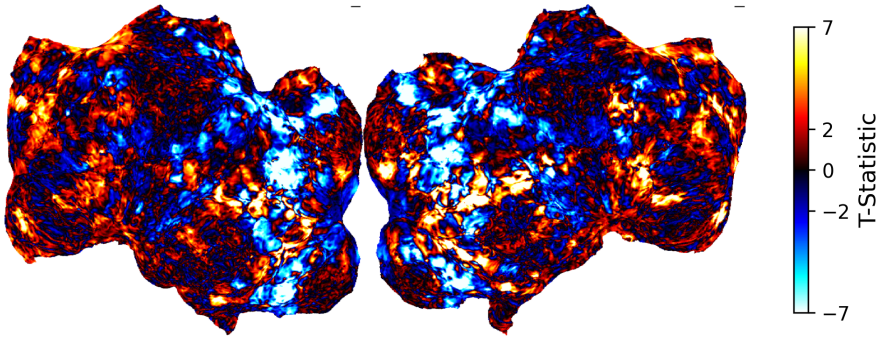
The repository is organized into subdirectories (*/t_testing*, */rsa* & */gaussian*) corresponding to each major analysis step and also contains:

- **README.md**: an overview of the repository structure and instructions for reproducing every result.
- **requirements.txt**: a complete list of software dependencies and version information.
- **notebooks/**: annotated Jupyter notebooks illustrating the analyses and visualizations.
- **data/**: contains all outputs from the analyses and inputs needed to run the analyses.

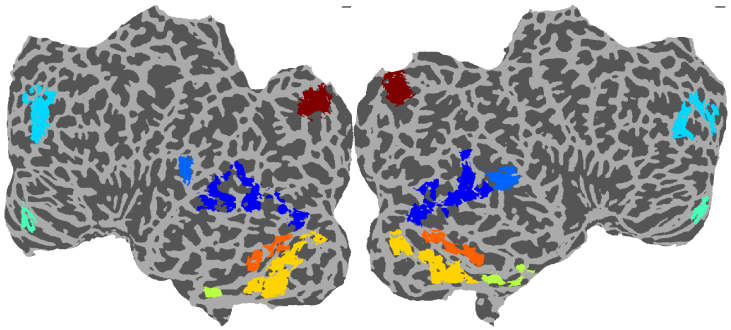
Step-by-step instructions are provided to recreate the computational environment.

I would like to thank Stan Bergey for providing the code from his thesis (Bergey, 2024), which I used and modified to successfully complete my project.

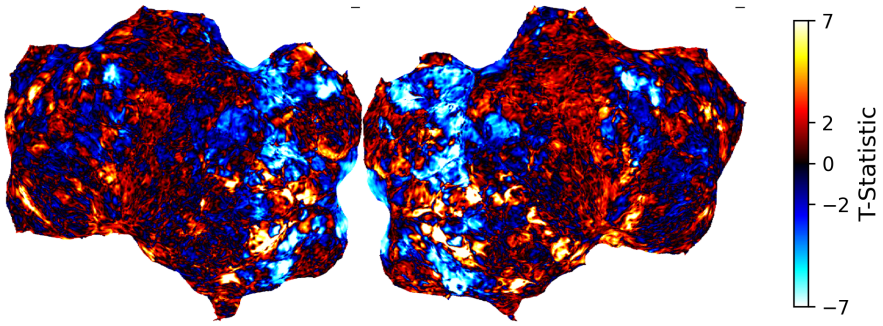
A.2 T-Statistics Outputs



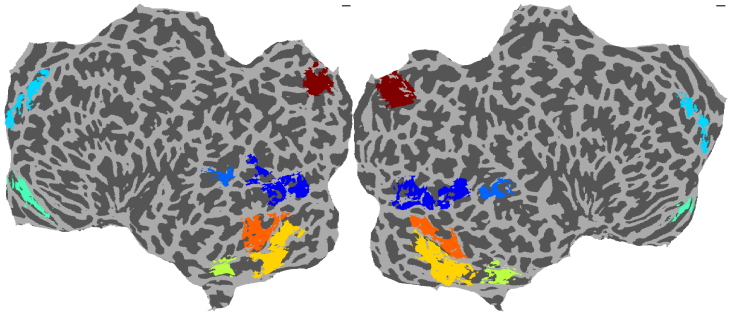
(a) Subject 1: voxel-wise t-statistic.



(b) Subject 1: thresholded ROIs ($|t| \geq 2$).

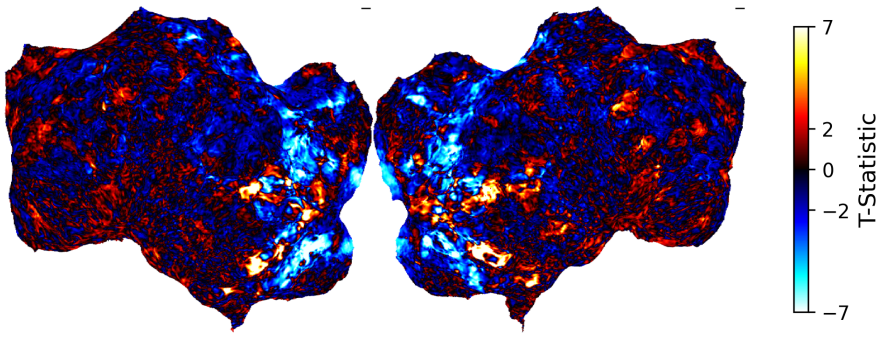


(c) Subject 2: voxel-wise t-statistic.

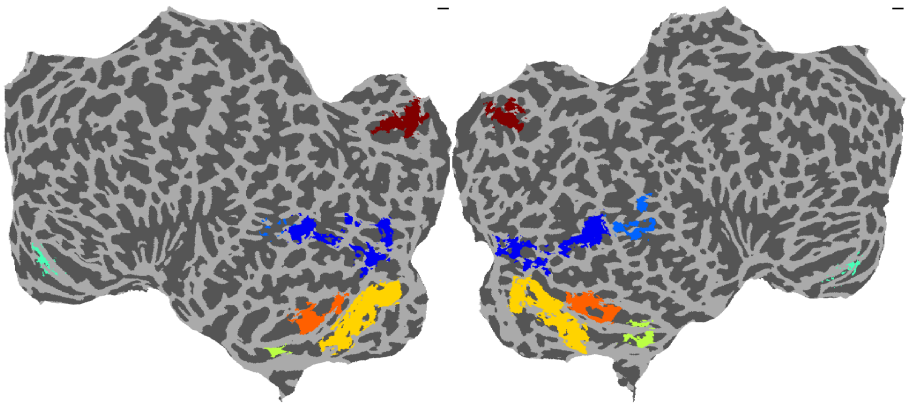


(d) Subject 2: thresholded ROIs ($|t| \geq 2$).

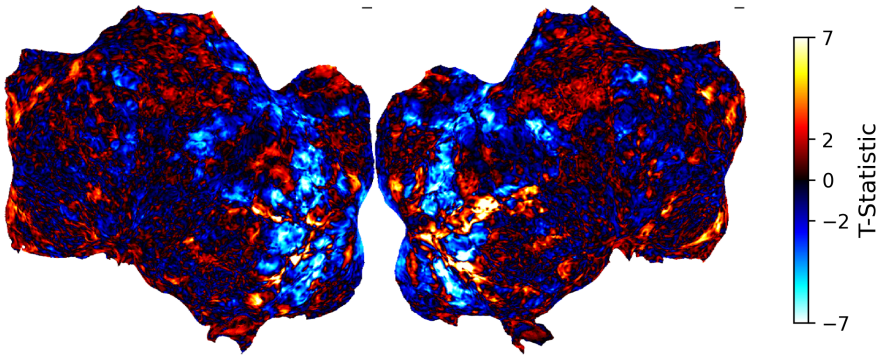
Figure A.1: Subjects 1–2: voxel-wise t-statistics (top of each pair) and thresholded ROIs (bottom of each pair).



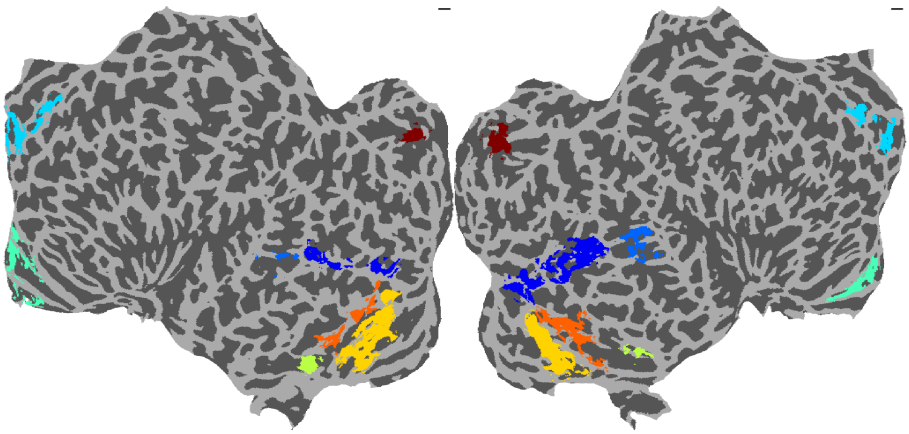
(a) Subject 3: voxel-wise t-statistic.



(b) Subject 3: thresholded ROIs ($|t| \geq 2$).

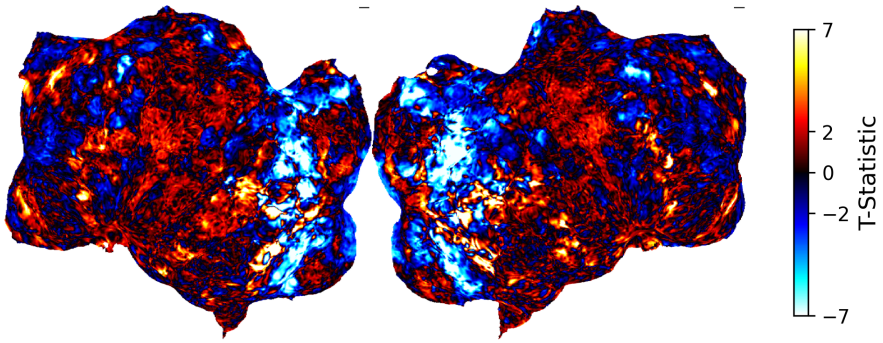


(c) Subject 4: voxel-wise t-statistic.

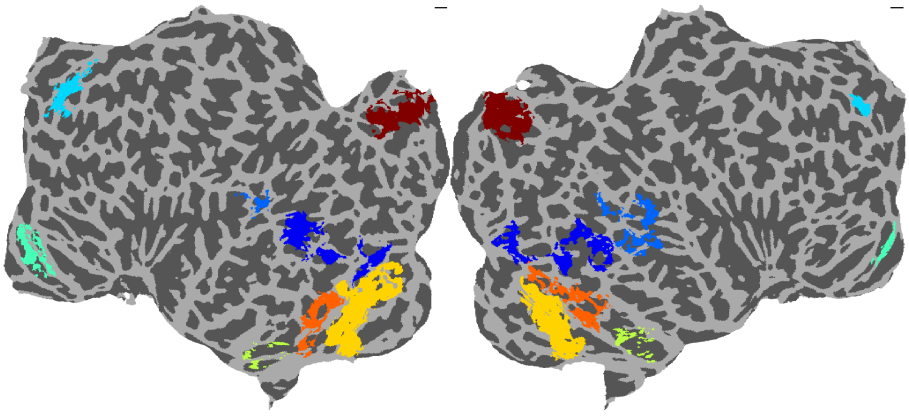


(d) Subject 4: thresholded ROIs ($|t| \geq 2$).

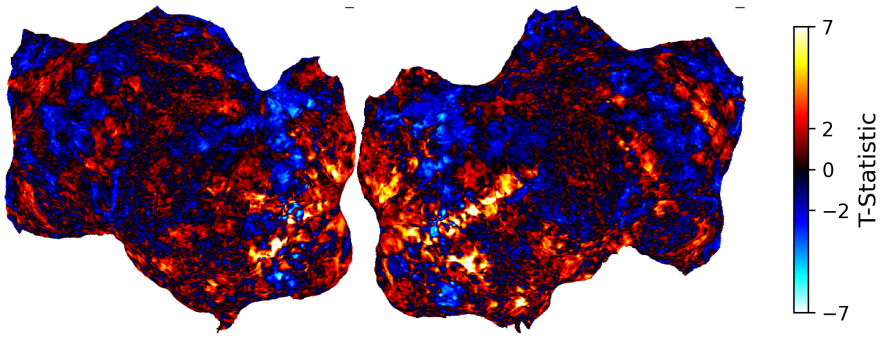
Figure A.2: Subjects 3–4: voxel-wise t-statistics (top) and thresholded ROIs (bottom).



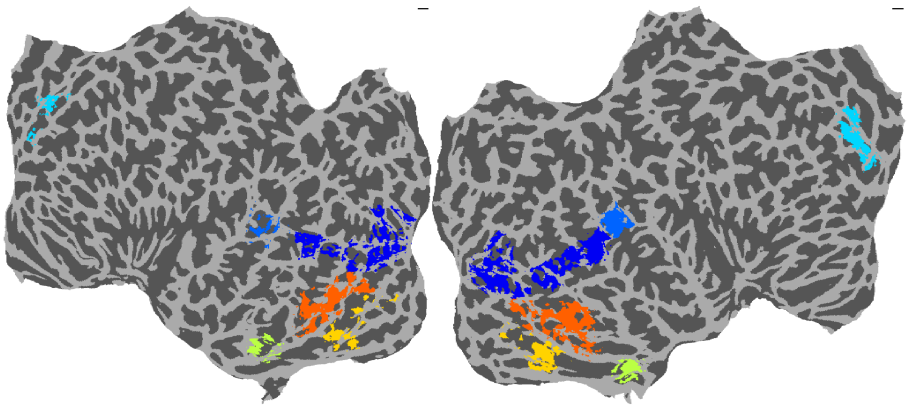
(a) Subject 5: voxel-wise t-statistic.



(b) Subject 5: thresholded ROIs ($|t| \geq 2$).

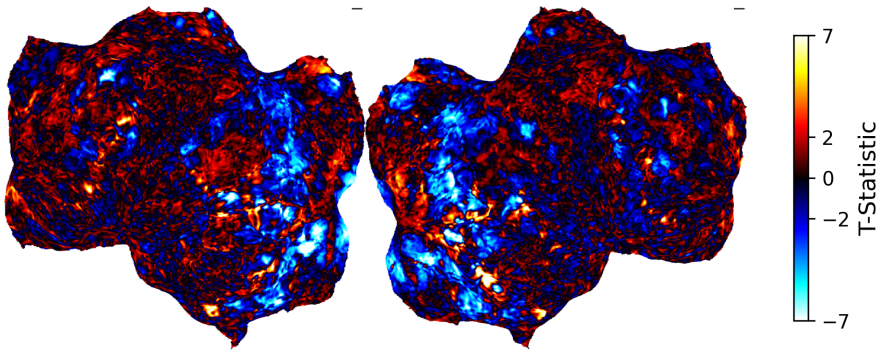


(c) Subject 6: voxel-wise t-statistic.

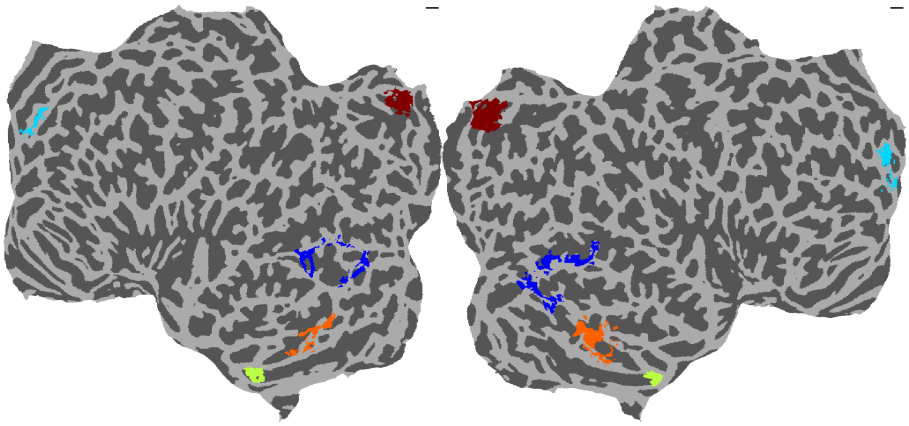


(d) Subject 6: thresholded ROIs ($|t| \geq 2$).

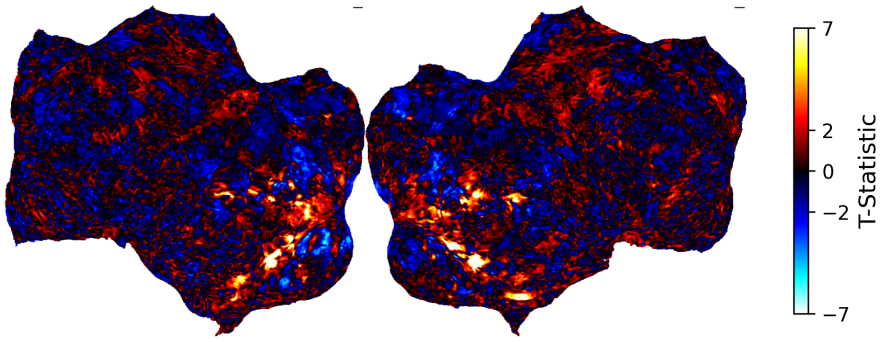
Figure A.3: Subjects 5–6: voxel-wise t-statistics (top) and thresholded ROIs (bottom).



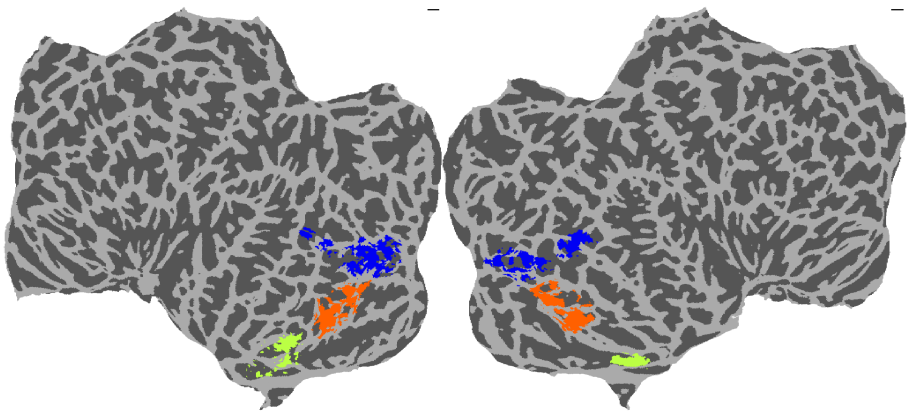
(a) Subject 7: voxel-wise t-statistic.



(b) Subject 7: thresholded ROIs ($|t| \geq 2$).



(c) Subject 8: voxel-wise t-statistic.



(d) Subject 8: thresholded ROIs ($|t| \geq 2$).

Figure A.4: Subjects 7–8: voxel-wise t-statistics (top) and thresholded ROIs (bottom).

Bibliography

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., Nau, M., Caron, B., Pestilli, F., Charest, I., Hutchinson, J. B., Naselaris, T., & Kay, K. (2022). A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, *25*(1), 116–126. https://doi.org/10.1038/s41593-021-00962-x

Bergey, S. (2024). *Modelling neural organization in the human visual system* [Master's thesis]. University of Utrecht.

Cichy, R. M., & Oliva, A. (2020). Am/eeg-fmri fusion primer: Resolving human brain responses in space and time. *Neuron*, *107*(5), 772–781. https://doi.org/10.1016/j.neuron.2020.07.001

Contreras, J. M., Banaji, M. R., & Mitchell, J. P. (2013). Multivoxel patterns in fusiform face area differentiate faces by sex and race. *PLOS ONE*, *8*(7), 1–6. https://doi.org/10.1371/journal.pone.0069684

Deng, J., Guo, J., Ververas, E., Kotsia, I., & Zafeiriou, S. (2020). Retinaface: Single-shot multi-level face localisation in the wild. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5202–5211. https://doi.org/10.1109/CVPR42600.2020.00525

Dumoulin, S. O., & Wandell, B. A. (2008). Population receptive field estimates in human visual cortex. *NeuroImage*, *39*(2), 647–660. https://doi.org/10.1016/j.neuroimage.2007.09.034

Fischl, B. (2012). Freesurfer. *Neuroimage*, *62*(2), 774–781. https://doi.org/10.1016/j.neuroimage.2012.01.021

Haak, K. V., Winawer, J., Harvey, B. M., Renken, R., Dumoulin, S. O., Wandell, B. A., & Cornelissen, F. W. (2013). Connective field modeling. *NeuroImage*, *66*, 376–384. https://doi.org/10.1016/j.neuroimage.2012.10.037

Henriksson, L., Mur, M., & Kriegeskorte, N. (2015). Faciotopy—a face-feature map with face-like topology in the human occipital face area [The whole is greater than the sum of the parts]. *Cortex*, *72*, 156–167. https://doi.org/10.1016/j.cortex.2015.06.030

Hesse, J. K., & Tsao, D. Y. (2020). The macaque face patch system: A turtle's underbelly for the brain. *Nature Reviews Neuroscience*, *21*(12), 695–716. https://doi.org/10.1038/s41583-020-00393-w

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, *160*(1), 106–154. https://doi.org/10.1113/jphysiol.1962.sp006837

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *J Neurosci*, *17*(11), 4302–4311. https://doi.org/10.1098/rstb.2006.1934

Kanwisher, N., & Yovel, G. (2006). The fusiform face area: A cortical region specialized for the perception of faces. *Philos Trans R Soc Lond B Biol Sci*, *361*(1476), 2109–2128.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*. https://doi.org/10.3389/neuro.06.004.2008

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., & Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, *60*(6), 1126–1141. https://doi.org/10.1016/j.neuron.2008.10.043

Kwong, K. K., Belliveau, J. W., Chesler, D. A., Goldberg, I. E., Weisskoff, R. M., Poncelet, B. P., Kennedy, D. N., Hoppel, B. E., Cohen, M. S., & Turner, R. (1992). Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proc Natl Acad Sci U S A*, *89*(12), 5675–5679. https://doi.org/10.1073/pnas.89.12.5675

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., & Dollár, P. (2015). Microsoft coco: Common objects in context. https://doi.org/10.48550/arXiv.1405.0312

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer research*, *27*(2_Part_1), 209–220.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature neuroscience*, *2*(11), 1019–1025. https://doi.org/10.1038/14819

Roth, Z. N., & Merriam, E. P. (2023). Representations in human primary visual cortex drift over time. *Nature Communications*, *14*(1), 4422. https://doi.org/10.1038/s41467-023-40144-w

Szinte, M., & Knapen, T. (2019). Visual organization of the default network. *Cerebral Cortex*, *30*(6), 3518–3527. https://doi.org/10.1093/cercor/bhz323

Tsantani, M., Kriegeskorte, N., Storrs, K., Williams, A. L., McGettigan, C., & Garrido, L. (2021). Ffa and ofa encode distinct types of face identity information. *Journal of Neuroscience*, *41*(9), 1952–1969. https://doi.org/10.1523/JNEUROSCI.1449-20.2020

Tsao, D. Y., Moeller, S., & Freiwald, W. A. (2008). Comparing face patch systems in macaques and humans. *Proceedings of the National Academy of Sciences*, *105*(49), 19514–19519. https://doi.org/10.1073/pnas.080966210

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. https://doi.org/10.1016/j.neuroimage.2015.12.012

Wandell, B. A., Dumoulin, S. O., & Brewer, A. A. (2007). Visual field maps in human cortex. *Neuron*, *56*(2), 366–383. https://doi.org/10.1016/j.neuron.2007.10.012

Welch, B. L. (1947). The generalization of 'student's'problem when several different population varlances are involved. *Biometrika*, *34*(1-2), 28–35. https://doi.org/10.1093/biomet/34.1-2.28