Face-Off: Pitting Computational Models of the Fusiform Face Area Against Each Other with Controversial Face Stimuli

Wenxuan Guo[*] **(wg2361@columbia.edu)**
Department of Psychology, Columbia University
New York, NY 10027, USA

Tal Golan[*] **(golan.neuro@bgu.ac.il)**
Department of Cognitive and Brain Sciences, Ben-Gurion University of the Negev
Be'er Sheva, Israel

Heiko H. Schütt (hs3110@columbia.edu)
Zuckerman Mind Brain Behavior Institute, Columbia University
New York, NY 10027, USA

Nikolaus Kriegeskorte (n.kriegeskorte@columbia.edu)
Departments of Psychology, Neuroscience, and Electrical Engineering, Columbia University
New York, NY 10027, USA

[*] The first two authors contributed equally to the work.

Abstract

Hundreds of studies have characterized the response properties of the fusiform face area (FFA), but we have yet to reveal the computational mechanisms underlying its representations. A methodological challenge is that distinct computational models can make indistinguishable predictions for randomly sampled faces. This fMRI study employs synthetic controversial face stimuli designed to elicit distinct predictions of six candidate neural network models of face representation in FFA. We present preliminary data from one participant scanned in four sessions. The controversial faces revealed many significant differences among the models in terms of their ability to predict FFA representational dissimilarity matrices (RDMs), whereas randomly sampled faces did not enable reliable adjudication among models. A neural network trained on inverse rendering—mapping face images to a latent space of a 3D face model—outperformed alternative models sharing the same architecture but trained on identification, classification, or autoencoding. Our results support the view that face recognition involves representations that reflect the physical structure of faces and demonstrate the need for optimized controversial stimuli to adjudicate among brain-computational models with neuroimaging experiments.

Keywords: neural networks; representational geometry; representational similarity analysis; model comparison; controversial stimuli; fusiform face area

Introduction

While functional neuroimaging studies in the past few decades have identified several occipitotemporal regions selective for different functional aspects of human face processing (Kanwisher, McDermott, & Chun, 1997; Haxby, Hoffman, & Gobbini, 2000), brain mapping does not directly probe the underlying computational mechanism of face processing. A deeper computational understanding requires testing quantitative predictions of brain-computational models against behavioral and neural data (Kriegeskorte & Douglas, 2018). Recent studies of face representation have begun testing quantitative predictions of brain-computational models against neural responses elicited by face stimuli (e.g., Carlin & Kriegeskorte, 2017; Grossman et al., 2019; Ratan Murty, Bashivan, Abate, DiCarlo, & Kanwisher, 2021). However, these studies used randomly sampled faces, and their results illustrate that distinct models can make indistinguishable predictions for these stimuli. In this fMRI study, we compare six computational hypotheses about face representation in the right FFA using controversial stimuli (Golan, Raju, & Kriegeskorte, 2020; Golan, Guo, Schütt, & Kriegeskorte, 2022) synthesized to elicit distinct representational predictions from the models.

Methods

We trained the VGG16 architecture on six distinct objectives: (1, 2) face identification for Basel-Face-Model (BFM; Gerig et al., 2018) and natural faces, (3) object categorization, (4, 5) autoencoding of BFM and natural faces, (6) inverse rendering (see Yildirim, Belledonne, Freiwald, & Tenenbaum, 2020; Golan, Guo, et al., 2022). The trained neural networks provide six computational hypotheses for human right FFA.

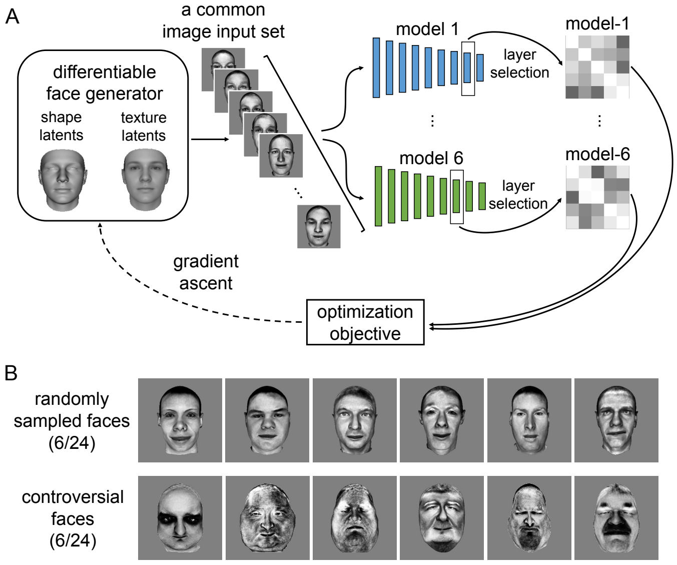

Figure 1: **Controversial face synthesis**. **(A)** Maximizing model discriminability by face-latent optimization. We implemented a differentiable version of the 3D morphable Basel Face Model (Gerig et al., 2018) as our face generator. The generator initializes a set of faces parameterized by shape and texture latents. Given the rendered face images, each candidate model generates its prediction of the representational geometry in the right FFA, measured as a representational dissimilarity matrix (RDM). Using a variant of the differentiable model discriminability objective proposed in our previous behavioral study (Golan, Guo, et al., 2022, [Eq.1]), we iteratively adjusted the face latents to increase the discriminability of the RDMs predicted by candidate models.

To efficiently discriminate among these models, we synthesized a controversial stimulus set by optimizing the face latents of 24 BFM faces. The latents were iteratively adjusted to maximize model discriminability based on their representational geometries (Figure 1). Given a stimulus set ξ, we used each layer l of each model m in turn as the ground truth, data-generating model and measured the discriminability of its predicted representational geometry from each of the other candidate models. The stimulus optimization procedure maximized the expected value of this discriminability measure across potential data-generating models and layers.

Specifically, for each layer l and model m, we computed the whitened Pearson correlation r_w (Diedrichsen et al., 2021) between the ground truth RDM $\mathbf{d}_{m,l}$ and the RDM of each alternative model m', taking the highest correlation across the layers of model m' as the performance of model m'. The discriminability is the difference between the performance of the true model (here, r_w is always 1.0, since we did not account for noise or multiple model realizations) and the mean performance of the alternative models (which should be as low as possible). We thus define the global utility U of a stimulus set ξ as the following objective to be maximized:

$$U(\xi) = \sum_{m} p(m) \sum_{l} p(l|m) \left(1 - \operatorname*{mean}_{m'} \max_{l'} r_w(\mathbf{d}_{m,l}, \mathbf{d}_{m',l'})\right), \tag{1}$$

where $p(m)$ is our prior belief in model m as the best candidate model, and $p(l|m)$ is our belief that layer l is the best data-generating representation in model m. We assumed uniform distributions for both models and layers before data collection. See Golan, Guo, and colleagues (2022) for a thorough description of stimulus synthesis method for disentangling model representational geometries. As a baseline condition, we generated 24 faces that maximized the sum of latent variances (i.e., first computing the faces' variance in each latent dimension and then summing across dimensions). For both conditions, each face's shape and texture latents were separately constrained within an L2-norm ball.

One subject participated in four fMRI sessions. First, we located the right FFA using a functional localizer (Stigliani, Weiner, & Grill-Spector, 2015) and cortical landmarks (Rosenke, van Hoof, van den Hurk, Grill-Spector, & Goebel, 2020). We then recorded 14 runs of fMRI responses to the face stimuli over three sessions. Each stimulus was shown at least 22 times in total. We presented each stimulus with a size of 8 degrees visual angle, using a three-flash (3 cycles of 800 ms ON, 200 ms OFF) design (Allen et al., 2022), followed by a 1000-ms interstimulus interval. Eye movements were tracked to verify fixation, and a memory task was conducted at the end of each run to encourage attention.

Results

We used fMRIPrep (Esteban et al., 2018) for preprocessing and GLMsingle (Prince et al., 2022) for estimating voxelwise BOLD response amplitude for each trial. The controversial face set showed a higher noise ceiling than the randomly sampled baseline face set (Figure 2, left panels), indicating that its face stimuli elicited response patterns that were more distinct from each other compared to those of the baseline set.

To compare the models, we evaluated model performance on each run using whitened Pearson RDM correlation, where the best layer for each model was chosen based on the other runs (Figure 2, right panels). Paired t-tests among the Fisher-transformed correlation coefficients indicated that in the controversial condition, the VGG16 network trained on inverse rendering (i.e., predicting face latents and scene properties of synthetic BFM faces) significantly outperformed all other candidate models. Four of the six models were significantly less accurate than the lower bound on the noise ceiling. The baseline condition failed to uncover a winning model and found no significant gaps between the models and the noise ceiling.

Conclusion

We demonstrated that face stimuli synthesized to maximally discriminate among alternative models of cortical representational geometry facilitate model adjudication with fMRI. Among our neural networks optimized with distinct objectives, the single-subject results here favor a model of the right FFA that is optimized to recover the parameters defining the 3D shape, texture, and external scene properties of the human face perceived.

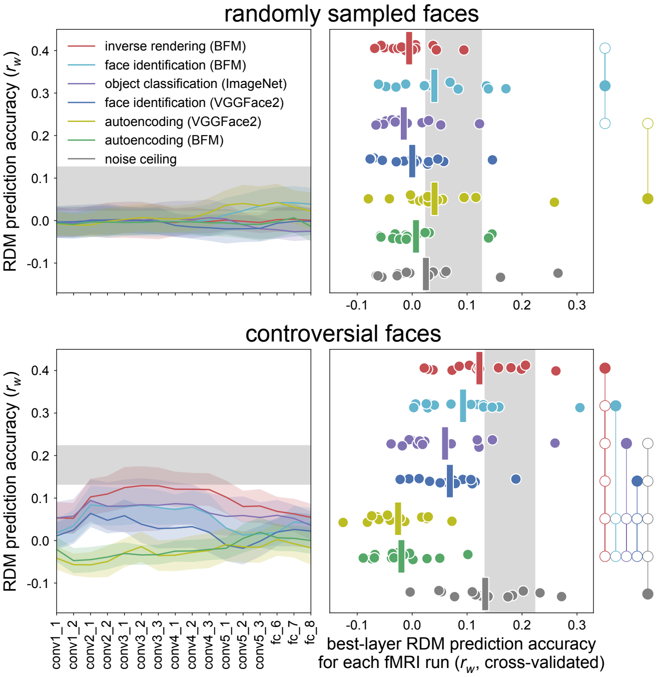

Figure 2: **Accuracy of model predictions of the representational geometry of right-FFA fMRI response.** The **left panels** show the average layer-wise model prediction accuracy across 14 fMRI runs for the randomly sampled faces (top) or controversial faces (bottom). Each colored line indicates one model's prediction accuracy (whitened Pearson RDM correlation, r_w) averaged across runs, and the corresponding shaded region indicates a 95% confidence interval. The grey region marks the noise ceiling bounds. The lower bound was estimated as the average performance of predicting each run by the mean of all other runs; for the upper bound, we used the average performance of predicting each run by the mean of all runs, including the predicted run itself. The **right panels** show cross-validated best-layer RDM prediction accuracy on each fMRI run. Each dot depicts the performance of one candidate model for one fMRI run. On the right are the results of paired t-tests between the Fisher-transformed correlation coefficients of each pair of candidate models. Each solid dot connecting to a set of open dots indicates that the model aligned with the solid dot significantly outperforms the candidate models aligned with the open dots (FDR controlled at $q = 0.05$). The randomly sampled faces lead to a low noise ceiling and similar model performance. The controversial face stimulus set lifts the noise ceiling and elicits significantly distinct representational geometries in different candidate models. The results indicate that the VGG16 network trained on inverse rendering of synthetic faces outperforms all other models.

Acknowledgments

This publication was made possible with the support of the Charles H. Revson Foundation to TG. The statements made and views expressed, however, are solely the responsibility of the authors. This material is based upon work supported by the National Science Foundation under Grant No. 1948004 to NK. We acknowledge Dr. T. Vetter, Department of Computer Science, and the University of Basel, for the Basel Face Model data.

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., . . . Kay, K. (2022). A massive 7T fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, *25*(1), 116–126. doi: 10.1038/s41593-021-00962-x

Carlin, J. D., & Kriegeskorte, N. (2017). Adjudicating between face-coding models with individual-face fMRI responses. *PLOS Computational Biology*, *13*(7), 1–28. doi: 10.1371/journal.pcbi.1005604

Diedrichsen, J., Berlot, E., Mur, M., Schütt, H. H., Shahbazi, M., & Kriegeskorte, N. (2021). Comparing representational geometries using whitened unbiased-distance-matrix similarity. *Neurons, Behavior, Data Analysis, and Theory*, *5*(3), 2785-2795.e4. doi: 10.51628/001c.27664

Esteban, O., Markiewicz, C., Blair, R. W., Moodie, C., Isik, A. I., Erramuzpe Aliaga, A., . . . Gorgolewski, K. J. (2018). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*. doi: 10.1038/s41592-018-0235-4

Gerig, T., Morel-Forster, A., Blumer, C., Egger, B., Lüthi, M., Schönborn, S., & Vetter, T. (2018). Morphable face models - an open framework. *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 75–82. doi: 10.1109/fg.2018.00021

Golan, T., Guo, W., Schütt, H. H., & Kriegeskorte, N. (2022). Distinguishing representational geometries with controversial stimuli: Bayesian experimental design and its application to face dissimilarity judgments. In *SVRHM 2022 Workshop @ NeurIPS*. Retrieved from https://openreview.net/forum?id=a3YPu2-Mf2h

Golan, T., Raju, P. C., & Kriegeskorte, N. (2020). Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proceedings of the National Academy of Sciences*, *117*(47), 29330–29337. doi: 10.1073/pnas.1912334117

Grossman, S., Gaziv, G., Yeagle, E. M., Harel, M., Mégevand, P., Groppe, D. M., . . . Malach, R. (2019). Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks. *Nature Communications*, *10*(1), 4934. doi: 10.1038/s41467-019-12623-6

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, *4*(6), 223-233. doi: 10.1016/s1364-6613(00)01482-0

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, *17*(11), 4302–4311. doi: 10.1523/JNEUROSCI.17-11-04302.1997

Kriegeskorte, N., & Douglas, P. K. (2018). Cognitive computational neuroscience. *Nature Neuroscience*, *21*(9), 1148–1160. doi: 10.1038/s41593-018-0210-5

Prince, J. S., Charest, I., Kurzawski, J. W., Pyles, J. A., Tarr, M. J., & Kay, K. N. (2022, nov). Improving the accuracy of single-trial fmri response estimates using glmsingle. *eLife*, *11*, e77599.

Ratan Murty, N. A., Bashivan, P., Abate, A., DiCarlo, J. J., & Kanwisher, N. (2021). Computational models of category-selective brain regions enable high-throughput tests of selectivity. *Nature Communications*, *12*(1), 5540. doi: https://doi.org/10.1038/s41467-021-25409-6

Rosenke, M., van Hoof, R., van den Hurk, J., Grill-Spector, K., & Goebel, R. (2020). A Probabilistic Functional Atlas of Human Occipito-Temporal Visual Cortex. *Cerebral Cortex*, *31*(1), 603-619. doi: 10.1093/cercor/bhaa246

Stigliani, A., Weiner, K. S., & Grill-Spector, K. (2015). Temporal processing capacity in high-level visual cortex is domain specific. *Journal of Neuroscience*, *35*(36), 12412-12424. doi: https://doi.org/10.1523/jneurosci.4822-14.2015

Yildirim, I., Belledonne, M., Freiwald, W., & Tenenbaum, J. (2020). Efficient inverse graphics in biological face processing. *Science Advances*, *6*(10), eaax5979. doi: 10.1126/sciadv.aax5979