A Mechanistic Explanation for the Inverted Face Effect

Alex Tahan* Hsin-Yuan Lee* Kira Fleischer* Nikita Kachappilly* Xavier Chen* Garrison W. Cottrell
Department of Computer Science and Engineering, University of California, San Diego
{atahan,kfleisch,nkachappilly,xavchen,hsl023,gary}@ucsd.edu

Abstract

The Inverted Face Effect has been the subject of a great deal of research and controversy in vision science. The nature of the effect, whether it involves holistic processing or not, where in the visual system it resides in the cortex, and whether it is a result of expertise and can apply to other phenomena, has been debated for decades. However, no *mechanistic* explanation for the effect has been put forward. Here we show that a model that includes multiple fixations on the face driven by a simple salience operator, a foveated retina, and the log-polar mapping from the visual field to V1 can explain this effect. We also simulate Yin's 1969 recognition experiment and show good agreement with our model.

Keywords: Inverted Face Effect, Face Processing, Biologically Plausible Models

Introduction

The inverted face effect (IFE) is one of the most studied phenomena in vision science, perhaps beginning with Köhler, 1940. He noticed that when a face is inverted in the image plane, it looked "strange". As a gestalt psychologist, he wondered if this was because the relationship to the surrounding environment was disrupted, or whether it was the projection on the retina. He had an assistant hold a picture of a face right side up while he bent over and looked through his legs. The face still appeared strange, but when inverted, it looked normal. He concluded that the strangeness was the result of the projection on the retina.

Yin, 1969 found in a recognition experiment that faces were special in this way: subjects performed worst on inverted faces, followed by houses (a mono-oriented stimulus), followed by airplanes. Finally, Diamond and Carey, 1986 found that other stimuli, in particular, dogs, had a similar effect when viewed by dog show judges, indicating that the phenomenon was a result of visual expertise, that is, fine-level discrimination of homogeneous categories. In particular, the dog show judges performed worst on the breeds of dogs that they were experts in, while naïve subjects showed a much smaller effect.

Since then, there has been a great deal of controversy over whether inverted faces are processed holistically or not. Holistic processing means that observers are sensitive to the configuration of the features, rather than the features themselves, and their perception of the whole is influenced by the parts. For example, subjects have difficulty recognizing that the top half of a face is the same when the bottom half is different, indicating configural processing (Young et al., 2013). When faces are inverted, which disrupts perception of the configuration, it is sometimes said that subjects resort to feature processing, although this has been disputed (Murphy et al., 2020; Richler et al., 2011; Rossion, 2008; Sekuler et al., 2004).

Figure 1: The Model 2.0 (TM2.0). The model includes sampling from the image via fixations (here, on the nose), driven by a salience map, followed by foveation, followed by log-polarization. This image is fed into a Convolutional Neural Network (CNN). To elucidate the role of the log-polarization, we compare this to the same model without it.

Yovel and Kanwisher, 2005 found using fMRI that the neural basis for the IFE is in the Fusiform Face Area (FFA), not other temporal lobe brain areas. On the other hand, Rock places the problem with the Thatcher illusion (where inverted features are not detected in inverted images) at the retina (Rock, 1988; Thompson, 1980).

We have simulated the IFE before with a similar model Gahl et al., 2020. In this paper, we advance this work in two ways. First, we add a salience operator and use it to simulate multiple fixations. This allows us to simulate the study and test phases of Yin's experiment, by giving the model a number of fixations corresponding to the study and test time he gave his subjects, assuming three fixations a second.

The work closest to our own is Dobs et al., 2023 (although this was preceded by our work with similar results Gahl et al., 2020). They showed that a "vanilla" CNN trained to identify faces showed an inverted face effect. However, the size of the drop in performance for the CNN was twice that of the human subjects (~20% vs.~10%), and was similar to the performance of a non-face network. This excessive drop in performance is reflected in our results below for a "vanilla" CNN.

Nevertheless, they failed to provide a mechanism for the IFE in terms of the *computations* involved; i.e., there was no *mechanistic* explanation. In this work, using an anatomically-inspired model (see Figure 1), we predict that the cause of the IFE is at the level of V1. While Yovel and Kanwisher

*Joint first authors

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

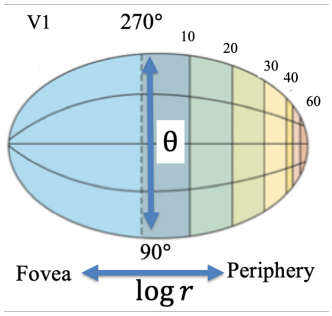

Figure 2: The log-polar approximation of the representation of the image in V1. The numbers on the right upper side are degrees of visual angle. Slightly over half of V1 is taken up by just 10 degrees of visual angle.

located the IFE in the response of the FFA, the FFA is driven, obviously, by V1. It may be difficult given current technology to detect differences at that level of processing using brain imaging, but with more accurate measurements, our model can be falsified.

The important aspect of V1 for our work is that the mapping from the visual field to V1 can be closely approximated by a log-polar transformation (see Figure 2). The log-polar representation results in scale being a left-right shift, and rotation being an up-down shift on V1. When this representation is fed into a standard convolutional neural network, which is relatively shift-invariant, we obtain scale and rotation invariance, modulo the considerations below.

How can we explain the IFE with a rotation-invariant model? The answer is that there is a topological difference between scale and rotation in the log-polar plane. Scale is a continuous value, while rotation is a circle group. As can be seen in Figure 2, due to the fact that the cortex is not a torus, there is a discontinuity between 90°and 270°. This results in a disruption in the configuration of the features, which humans are very sensitive to in face recognition.

This disruption of configural is illustrated in Figure 3. In the second column from the left, note that a slight rotation shifts the image down, while an inverted face shifts it much more (third row). In the fourth column, note that in the log-polar representation, the nose is next to the left eye in the first two rows, while in the third row, it is next to the *right* eye due to the discontinuity. Meanwhile, the features themselves remain the same, just shifted (columns 5 and 6). Hence, by this account, rotation disrupts the *configuration* of the features at the level of V1, but not the features themselves. This disruption of the relations between the features is the aforementioned caveat to true rotation invariance.

In previous work Gahl et al., 2020, we have used this model to explain the IFE, but we used four fixation points on the eyes, nose, and mouth. In this paper, we investigate using a simple salience operator, sampling from the induced probability density distribution. We compare the model to a modified CNN with foveation, a salience map, and sampling fixations, but without the log-polar transform. In this way, we isolate the

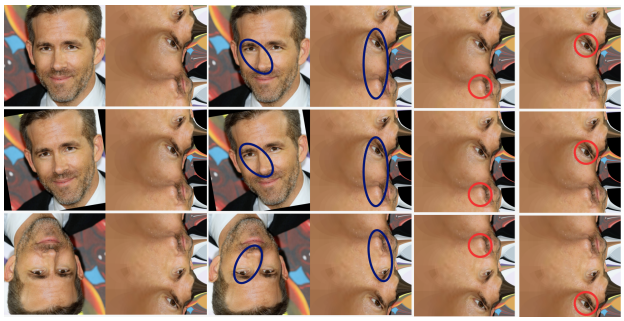

Figure 3: Configuration Shifts Under Face Inversion. When a face is inverted, the relative spatial configuration of features changes: feature pairs that are closest and most informative in the upright orientation (e.g., left eye –nose) are no longer the same after inversion, although the features are the same.

behavior of the model that is due to this transform. We show that with increasing expertise (acquiring more face identities), the IFE in face recognition increases. We also show that the inversion effect is much reduced, as in humans, for non-mono-oriented objects. Finally, we simulate Yin's (1969) face recognition experiment, using a very simple memory model, and find good agreement.

Methods

Data

We use two different datasets to test the effects of visual expertise: faces and objects (see Figure 5). The corresponding labels are identities and object categories. Faces are generally mono-oriented, and recognizing them corresponds to visual expertise, where we expect inversion effects. We expect much less of an inversion effect with objects that are not mono-oriented and are learned at the basic level (Rosch et al., 1976). Each dataset is randomly split into training, validation, and test sets with a 80/10/10 ratio.

The faces dataset is taken from CelebA, with 128 celebrity identities in a variety of contexts, backgrounds, and angles. The face category contains approximately 200 unique images for each identity. This is a sufficient number of examples for faces because they are a relatively homogeneous class, with most stimuli presented face-front.

The objects dataset is from ImageNet (Deng et al., 2009).

Figure 4: Foveated patches centered at salient fixation points (red dots, top row). Their corresponding log-polar versions (bottom), are very different from each other.

Figure 5: Example pictures from the faces and objects datasets.

We chose 128 ImageNet objects. These images are variable in pose, scale, and context, a much harder problem. Hence, we used 800 images per object category for the training set, with 100 holdout images. This dataset is four times larger than the face dataset and has 1.6M images (800 examples×128 categories×16 fixations), so training is compute-heavy. The objects dataset uses objects that are generally seen in a variety of orientations and are mapped to basic category labels.

Transformations

Here we describe the details of the transformations applied to the images before presenting them to the network (see Figure 6).

Figure 6: Data Transformation Pipeline. Fixation points are first gathered from each image. Then, centered on a sampled fixation point, each image is cropped, rotated in ±15°, foveated, and log-polarized for TM2.0 only.

Salience Map A Gaussian filter is applied to training images to place greater weight on the center of the images. Then, because in the cortex, salience maps (e.g., in the monkey lateral intraparietal sulcus, or LIP) are computed based on the log-polarized input, we apply the salience operator to the log-polarized image. We used a simple salience operator developed previously (Yamada and Cottrell, 1995): The image is passed through a bank of Gabor filters at every pixel with resolution $\lambda = [4, 8]$ and phase $\psi = [0, \pi/2]$, each at 4 orientations $\theta = \{0, \pi/4, \pi/2, 3\pi/4\}$. These will have 0 response in "flat" parts of the image, and high variability where there is "busyness" in the image. We compute the variance of the magnitudes of the 8 filters at each point, resulting in our salience map. We then sampled 16 fixation points from the resulting distribution. These points are then mapped back to the (x, y) coordinates to get the fixation points on the original image. In preliminary work, we determined that the performance of TM2.0 saturated with 16 fixation points. Thus, each base image is augmented to create 16 images, each centered at a different fixation point.

Cropping and Rotation Each image is cropped from 224x224 pixels to 180x180 pixels, using the 16 fixation points as the center of each crop. Each cropped image in the training set is randomly rotated between -15° and 15°. No rotation is applied to images in the holdout set. The model is tested on both upright and inverted images from the holdout set.

Foveation The human retina is highly foveated (Curcio et al., 1990). We simulate the foveation using the algorithm by (Jiang et al., 2015). This method constructs a multi-scale Gaussian pyramid centered at the fixation point, and downsamples at each level to create progressively blurrier versions. The resulting image is highest resolution at the central fixation point; the resolution drops off the farther from that point.

Log-Polarization The representation of the visual scene in primary visual cortex (V1) can be approximated by a log-polar mapping (Polimeni et al., 2006), in which retinal eccentricity (distance from the fovea) is encoded on a logarithmic scale while polar angle is encoded linearly, so the coordinates go from Euclidean (x, y) to $(\log r, \theta)$ (see Figure 2).

Model

For all experiments, we use a pretrained, modified ResNet-18 backbone (He et al., 2016) as a convolutional feature extractor. The network's final average pooling and fully connected layers are removed, retaining only the convolutional stages up to the last residual block. The resulting 512-channel feature map is passed through an adaptive average pooling layer to produce a global feature vector, which is then processed by two fully connected layers with a ReLU nonlinearity followed by a softmax layer for classification. All models are fine-tuned with the X-Ent objective function, 0.5 dropout, Adam optimization, minibatch size of 256, and an initial learning rate of 0.001.

We compare two different models: a standard CNN and TM2.0. As described above, we apply the same transformations to the data, but only TM2.0 gets the log-polar transform (the last transformation in Figure 6), allowing us to isolate the effect of the log-polarization. Then these two sets of stimuli are given as input to the modified ResNet-18.

Experiments and Results

Experiment 1: Faces

Experimental Setup Using the preprocessing pipeline described above, we set the number of fixation points at 16, and evaluated TM2.0's effectiveness on increasing numbers of identities.

During training, we gradually increase the number of identities the network is trained on to mimic development, reflecting how a child is exposed to more and more people. The first 40 epochs train on 4 identities as outputs, the next 40 double

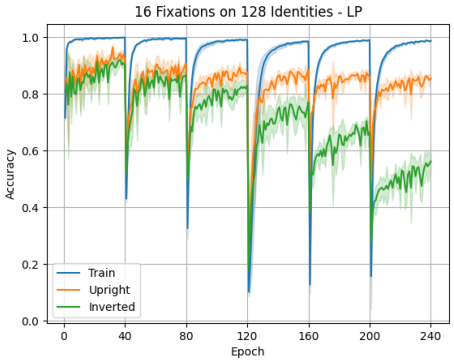

Figure 7: Accuracy throughout training faces using TM2.0. Results are averaged over 4 trials, and ±1 standard deviation is shaded. Blue: training accuracy; orange: validation accuracy on upright images; green: validation accuracy on the same images as the orange line, rotated 180°.

that number, and so on. In total, there are 240 epochs, with 40 epochs each for 4, 8, 16, 32, 64, and 128 identities.

During inference, we use a simple voting mechanism to combine the results of the multiple fixations: we sum the 16 logits for each identity and take the argmax over the identities. It should be noted that this voting mechanism is only implemented during inference; during training, each processed image is treated independently.

Results We ran the model four times with different initial random seeds and averaged the results. Figures 7 and 8 show the accuracies throughout the experiment for both the training and validation sets for the TM2.0 and CNN models, respectively. At the introduction of new identities after every 40 epochs, the accuracy initially drops as the model learns these new identities.

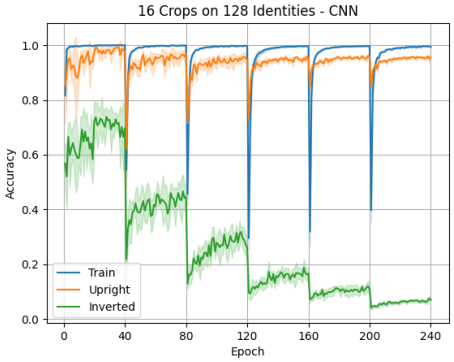

Figure 8: Accuracy throughout training on faces using a standard CNN, with the same parameters as in Figure 7.

The CNN and TM2.0 models see similar trends between the two validation sets. At 4 identities for TM2.0, the upright and inverted set accuracies are similar. As the number of identities increases, the upright set maintains roughly the same accuracy, while the inversion effect (we define the IFE as the

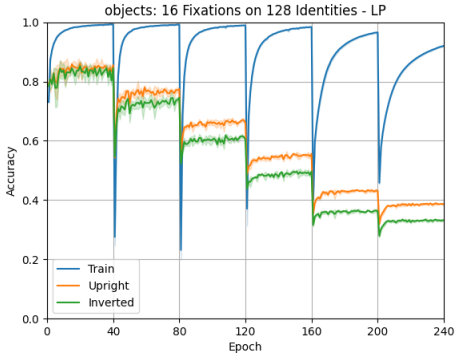

Figure 9: TM2.0 results on Objects with 800 examples per category and 16 fixations per image, averaged over four runs.

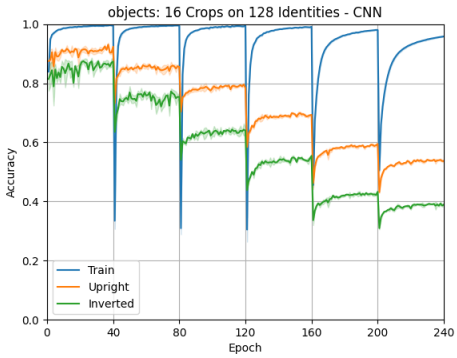

Figure 10: CNN results on Objects dataset with the same parameters as in Figure 9.

difference between the orange and green curves) increases as the model becomes a face expert. While both models see an increased IFE as more identities are learned, the standard CNN model struggles much more with the inverted faces, achieving an accuracy of around 10% for 128 identities. While still performing better than chance, this low accuracy demonstrates the CNN's inability to adapt to rotations, making its IFE much larger than that of human subjects. TM2.0's log polar transformation allows the model to have more rotational invariance, with the model achieving a final accuracy near 60%.

Experiment 2: Objects

Here we repeat the experiment with objects. For this experiment, we chose objects that can be in any orientation, such as cell phones and pens (see Figure 5). The results in Figures 9 and 10 show a much smaller inversion effect, consistent with Yin's (1969) results. TM2.0 and the CNN perform similarly, because they have both been trained on objects at multiple orientations.

Experiment 3: Simulating Yin (1969)

Yin's (1969) study is one of the foundational studies in face recognition, demonstrating that human memory for faces is disproportionately impaired by image inversion relative to memory for other visual categories such as houses, airplanes,

and stick figures. This selectivity became a cornerstone of the argument that face recognition relies on specialized processing mechanisms rather than generic object recognition strategies.

In Yin's original experiment, participants were presented with a study set series of 40 pictures, shown individually at a rate of 3 seconds per image by the experimenter. This was followed by a test series of 24 pairs of images; each pair contained exactly one old picture (an exact duplicate of an image from the inspection series) and one new picture (never previously shown). Participants indicated which image in each pair was the old one, proceeding at their own pace. In Experiment I, Yin tested upright-upright and inverted-inverted matching (consistent orientation across study and test). In Experiment II, he tested cross-orientation transfer - subjects studied upright faces and then were tested on inverted (Up-Down), or studied inverted faces and then were tested on upright (Down-Up).

In order to simulate this experiment, we used a very simple memory model. First, we adapted the model to use a binary representation at the last hidden layer by replacing the ReLU units with logistic units. We then treated each neuron as a binomial distribution and sampled from it using the logistic activation as a probability to get a binary representation. A very small amount of fine tuning allowed the model to use this stochastic representation and recover its accuracy. To simulate imperfect episodic memory, we added noise to the resulting vector via XOR-masking:

$$\mathbf{h}_{\text{noisy}} = \mathbf{h} \oplus \mathbf{m}, \quad \mathbf{m}_i \sim \text{Bernoulli}(p_{\text{noise}})$$

This operation stochastically flips individual bits with probability p_{noise}, simulating the unreliability of neural encoding (and hence, memory). As Yin's subjects were given 3 seconds to inspect the study images, and humans saccade roughly 3 times a second, we sampled 10 fixations from our salience map to simulate the study phase. Hence each of the forty faces is represented by 10 noisy binary vectors.

We then treated these vectors as a kernel density representation, as in previous work (Barrington et al., 2008). For the test phase, the model generated binary representations for both the old and new face in each pair. Binomial noise was again applied to these retrieval representations, modeling the additional unreliability inherent to memory retrieval. A familiarity score was then computed for each test face using a the kernel density estimation (KDE) framework drawn from the Barrington et al. model of saccade-based visual memory:

$$p(\mathbf{f} \mid c) = \frac{1}{|M_c|} \sum_{\mathbf{m} \in M_c} \exp\left(-\frac{\|\mathbf{f} - \mathbf{m}\|^2}{2\sigma^2}\right)$$

Here, $\mathbf{f}$ is the binary feature vector extracted from a single fixation patch of the test face, and M_c is the set of stored memory traces for identity c accumulated during the study phase, with $|M_c|$ denoting the number of such traces.

Matching Yin's 24-pair test series, 24 old identities were randomly sampled (without replacement) from the 40 studied identities. These were paired with 24 new (unknown) identities that were not presented during the study phase. For each identity at test, 32 fixations on each pair were used, reflecting the slower, self-paced viewing in Yin's test phase. We therefore generated 32 binary representations for both the old and new faces in each pair. Binomial noise was again applied to these retrieval representations, modeling perceptual noise.

Using the kernel density model, the log-likelihood of a test face (comprising multiple fixation patches) against all 40 stored memory traces was computed, and the identity receiving the highest log-likelihood was assigned the higher familiarity score. If the old face received a higher familiarity score than the new face, the trial was scored as correct, directly mirroring Yin's "indicate which is the old one" instruction.

We then adjusted p_{noise} to fit the upright-upright performance of Yin's subjects. This procedure yielded $p_{\text{noise}} = 0.30$ (the model achieved 95.83% at this level, versus 96.29% for human subjects). All four conditions were then run with this fixed noise parameter. The results are shown in Table 1.

Table 1: Yin (1969) Human vs. Model Accuracy

Study	Test	Human	Model
Upright	Upright	96.29%	95.83%
Inverted	Inverted	81.88%	87.50%
Upright	Inverted	84.13%	83.33%
Inverted	Upright	78.58%	75.00%

The model successfully reproduces the qualitative ordering of conditions observed in Yin's data. Upright–Upright performance was highest in both human data and the model, and Inverted–Upright yielded the lowest accuracy in both cases. One deviation from this pattern is that the model performs too well in the Inverted–Inverted condition. It appears that there is an order effect in the subject data; initially having to store an inverted face may be more difficult. Using another parameter to add more noise when storing inverted faces at study would correct this; for now we leave our results as is.

Discussion

We have presented The Model 2.0 (TM2.0), an anatomically-inspired model of the primate visual system that includes multiple fixations on a stimulus, a foveated retina, and the log-polar mapping from the visual field to V1. Due to the nature of this transformation, there is a discontinuity in polar angle between 90°and 270°in area V1. We hypothesized that due to this discontinuity, features that were in one arrangement when faces are upright are transformed into a different arrangement when inverted, as shown in Figure 3. The careful reader will have noticed that we have ignored the hemispheric bifurcation of the visual system (Hsiao et al., 2013); we leave that wrinkle for future work.

As it learns an increasing number of faces, we hypothesize that the model begins to rely more on the configuration of features to distinguish identities. Indeed, CNNs trained on face recognition have been shown to be sensitive to alterations

of configurations (Strehle et al., 2024), The behavior of the model reflects this - over the learning trajectory, the inversion effect grows for faces, but it does not for objects. This difference is the result of the task: fine-level discrimination for faces, which requires separating representations of identities, versus basic-level categorization of objects, which requires mapping within-category objects to similar representations ("clumping") while separating category representations.

The decline in performance for inverted faces is substantially more severe for the standard CNN, showing CNNs can't handle inversion. TM2.0 maintains a significantly higher inverted accuracy and relative performance, more closely matching the observed performance of humans on inverted faces (Diamond and Carey, 1986; Yin, 1969).

To verify this, we then used a model of saccade-based memory (Barrington et al., 2008) in order to simulate Yin's classic recognition experiments. We fit the model's performance to human subjects' performance in the upright-upright condition by adjusting a single noise parameter in the memory model, and then tested the model on the other three conditions in Yin's experiments. The model demonstrated a reasonably good fit to the data.

These results support our conclusion that respecting the anatomy of the visual system can explain the pattern of inversion effects for faces and objects in a mechanistic fashion. There are many ways to try to understand how perception and cognition work - behavioral experiments, neural imaging, and multi-cellular recordings from animal and human brains, but only models that actually implement the behavior can explain the underlying computations that lead to the behavior.

Acknowledgements

This work was supported by NSF grant #2208362. We would also like to thank Siddharth Agrawal for some last-minute simulations and for creating some of the figures in this paper.

References

Barrington, L., Marks, T. K., Hsiao, J. H.-w., & Cottrell, G. W. (2008). Nimble: A kernel density model of saccade-based visual memory. *Journal of Vision*, *8*(14), 17–17.

Curcio, C. A., Sloan, K. R., Kalina, R. E., Hendrickson, A. E., Cac, O., & Science, C. (1990). Human photoreceptor topography. *The Journal of Comparative Neurology*, *292*, 497–523.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*, 248–255.

Diamond, R., & Carey, S. (1986). Why faces are and are not special: An effect of expertise. *Journal of Experimental Psychology: General*, *115*(2), 107–117.

Dobs, K., Yuan, J., Martinez, J., & Kanwisher, N. (2023). Behavioral signatures of face perception emerge in deep neural networks optimized for face recognition. *Proceedings of the National Academy of Sciences*, *120*(32), e2220642120.

Gahl, M., Yuan, M., Sugumar, A., & Cottrell, G. (2020). The face inversion effect and the anatomical mapping from the visual field to the primary visual cortex. *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Hsiao, J. H., Cipollini, B., & Cottrell, G. W. (2013). Hemispheric asymmetry in perception: A differential encoding account. *Journal of Cognitive Neuroscience*, *25*(7), 998–1007.

Jiang, M., Huang, S., Duan, J., & Zhao, Q. (2015). Salicon: Saliency in context. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1072–1080.

Köhler, W. (1940). *Dynamics in psychology*. Liveright Publishing Corporation.

Murphy, J., Gray, K. L., & Cook, R. (2020). Inverted faces benefit from whole-face processing. *Cognition*, *194*, 104105.

Polimeni, J., Balasubramanian, M., & Schwartz, E. (2006). Multi-area visuotopic map complexes in macaque striate and extra-striate cortex. *Vision Research*.

Richler, J. J., Mack, M. L., Palmeri, T. J., & Gauthier, I. (2011). Inverted faces are (eventually) processed holistically. *Vision research*, *51*(3), 333–342.

Rock, I. (1988). On Thompson's inverted-face phenomenon (research note). *Perception*, *17*, 815–817.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382–439.

Rossion, B. (2008). Picture-plane inversion leads to qualitative changes of face perception. *Acta psychologica*, *128*(2), 274–289.

Sekuler, A. B., Gaspar, C. M., Gold, J. M., & Bennett, P. J. (2004). Inversion leads to quantitative, not qualitative, changes in face processing. *Current Biology*, *14*(5), 391–396.

Strehle, V. E., Bendiksen, N. K., & O'Toole, A. J. (2024). Deep convolutional neural networks are sensitive to face configuration. *Journal of Vision*, *24*(12), 6–6.

Thompson, P. (1980). Margaret Thatcher: A new illusion. *Perception*, *9*(4), 483–484.

Yamada, K., & Cottrell, G. W. (1995). A model of scan paths applied to face recognition. *Proceedings of the 17th annual conference of the cognitive science society*, 55–60.

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, *81*, 141–145.

Young, A. W., Hellawell, D., & Hay, D. C. (2013). Configurational information in face perception. *Perception*, *42*(11), 1166–1178.

Yovel, G., & Kanwisher, N. (2005). The neural basis of the behavioral face-inversion effect. *Current biology*, *15*(24), 2256–2262.