HUMAN VISUAL PERFORMANCE FOR IDENTIFYING LETTERS AFFECTED BY PHYSIOLOGICALLY-INSPIRED SCRAMBLING

A bioRχiv PREPRINT

Xingqi R Zhu*
McGill Vision Research
Department of Ophthalmology & Visual Sciences
McGill University
Montreal, Quebec, Canada
xingqi.zhu@mail.mcgill.ca

Robert F Hess
McGill Vision Research
Department of Ophthalmology & Visual Sciences
McGill University
Montreal, Quebec, Canada
robert.hess@mcgill.ca

Alex S Baldwin
McGill Vision Research
Department of Ophthalmology & Visual Sciences
McGill University
Montreal, Quebec, Canada
alexander.baldwin@mcgill.ca

March 27, 2024

ABSTRACT

In human vision, the retinal input is transformed into internal representations through a series of stages. In earlier stages, the signals from a particular visual field locus are passed in parallel from one visual processing area to the next. The connections at each stage may therefore introduce "error", where incorrect or convergent projections result in a loss of spatial precision. Psychophysical and physiological studies have implicated spatial scrambling of this sort as a cause of the visual deficits in amblyopia. Several methods to measure scrambling (both in amblyopia and in healthy vision) have been developed in recent decades. In this work, we introduce a new approach. We consider two stages of visual processing where scrambling may occur: either at the input to or the output from the simple cell stage in V1. We refer to these as "subcortical" and "cortical" scrambling respectively. We investigated the impact of these two types of scrambling on a letter identification task. A physiologically-inspired decomposition and resynthesis algorithm was used to generate letter stimuli that simulate scrambling at each of these two stages. To establish a performance benchmark, we trained separate Convolutional Neural Networks (CNNs) to perform the task with each scrambling type. Comparing CNN performance against that of eight humans with normal healthy vision, we found humans exhibited greater resilience to subcortical scrambling compared to cortical scrambling. We further investigated performance by comparing confusion matrices. Compared to a simple template matching model, we found the human strategy to be more consistent with our CNNs. We conclude: i) the human resilience for subcortical scrambling suggests this may be the stage at which a greater degree of scrambling is introduced in the visual hierarchy, and ii) humans employ flexible strategies for identifying scrambled stimuli, more sophisticated than a simple template match to the expected target.

Keywords equivalent input noise · efficiency · noise masking · letter identification · letter acuity · scrambling · convolutional neural network · confusion matrix

*corresponding author

1 Introduction

Through our vision, we perceive a vivid external world. This rich percept is the output of a hierarchical system, where representations are built up through multiple stages of visual processing. Physiology puts limits on both the quantity and quality of information that gets transmitted between each stage. For example, information from the retina gets only partially transmitted to the cortex via the optic nerve (Kelly, 1962; Nirenberg et al., 2001). Information bottlenecks also exist in cortical processing, starting as early as V1. These are determined by the number of neurons, their energy consumption, and also by noise and uncertainty in the system (Li, 2014). A dominant topic within vision science has been the question of *how* the brain synthesises a noisy, uncertain input into an internal representation that drives our behaviour. To this end, it is argued that the brain performs attentional selection in the form of a saliency map (Li, 2002) and predictive coding (Clark, 2013) to efficiently encode information (Barlow, 1956).

Psychophysical methods can be used to investigate the effects of noise and uncertainty on visual behaviour. These factors have been shown to place limits on performance both in healthy vision (Pelli, 1981; McAnany and Alexander, 2010; Baldwin et al., 2016), and in disease states (Kersten et al., 1988; Pelli et al., 2004; McAnany et al., 2013). The *equivalent noise paradigm* is a technique which has been used to quantify the quality of the input to visual processing, reflecting the effective "noisiness" of the input. It is based on measuring the impact of added external noise on performance in psychophysical tasks. This allows one to find the *equivalent input noise*: the magnitude of external noise that has an effect equivalent to the limiting internal noise in the system (Pelli and Farell, 1999). In the first instance, equivalent input noise was conceptualized as an effective "dark light" in the eye that induced a "background" level of noise that would otherwise be multiplicative with the light level (Barlow, 1956). Subsequently, many studies have explored an effective *contrast* noise that is additive (Pelli, 1981) or multiplicative (Lu and Dosher, 1999) with the stimulus contrast energy. At the same time, the *equivalent input noise* concept has been extended beyond luminance or contrast to investigate the coding of other stimulus properties such as orientation (Dakin, 2001), disparity (Wardle et al., 2012), or motion (Pavan et al., 2023). In this case, in addition to variability added to a signal fed through the system (e.g. information about the stimulus orientation which is then made noisy), a related consideration may be the uncertainty of the causal stimulus property given the response of some tuned mechanism (e.g. a mechanism tuned to a specific orientation but that responds to a range of similar orientations). The impact of any noise or uncertainty in the input can be analysed in terms of the performance of a theoretical "Ideal Observer" that would make the best possible use of that input to perform the task (Geisler, 2004). The Ideal Observer establishes a benchmark performance against which behaviour can be compared. Human performance relative to that ideal establishes the efficiency with which they perform the task. This allows for a comparison of efficiency across tasks and conditions on an objective scale.

In this study, we adapt these methods for an investigation of positional uncertainty in vision. The visual system can be considered as a hierarchy of maps, in which specific projections are required between successive areas to maintain the accurate preservation of precise positional information (Alonso et al., 2001). However, it is well-established that along the cortical hierarchy, there is a loss of spatial precision that accompanies the increasing complexity of the encoded features (Gattass et al., 1981, 1988; Dumoulin and Wandell, 2008; Weinberg, 1997; Kaas, 1997; Groen et al., 2022). This loss of spatial precision may result from many-to-one correspondence between areas or a one-to-one miswiring that is not accounted for in how the signal is interpreted (Hess and Field, 1993).

An extreme version of this spatial scrambling (or "uncalibrated disarray") has been argued to occur in normal peripheral (Hess and Field, 1993) and amblyopic central vision (Hess et al., 1978; Watt and Hess, 1987; Hess et al., 1990; Hess and Field, 1994). In amblyopia, it has been suggested that the miswirings in connections from the amblyopic eye cannot be compensated for, due to the overriding imposition of a more dominant map from the fellow eye (Hess et al., 1990). This has been proposed to explain the scrambled percepts that can occur in amblyopia (Hess et al., 1978; Sireteanu et al., 1993; Barrett et al., 2003).

Previous psychophysical studies investigating scrambling have applied on-screen positional noise to their stimuli at the level of pixels (Watt and Hess, 1987) or affecting individual wavelets (Hess and Field, 1993, 1994; Hess et al.,

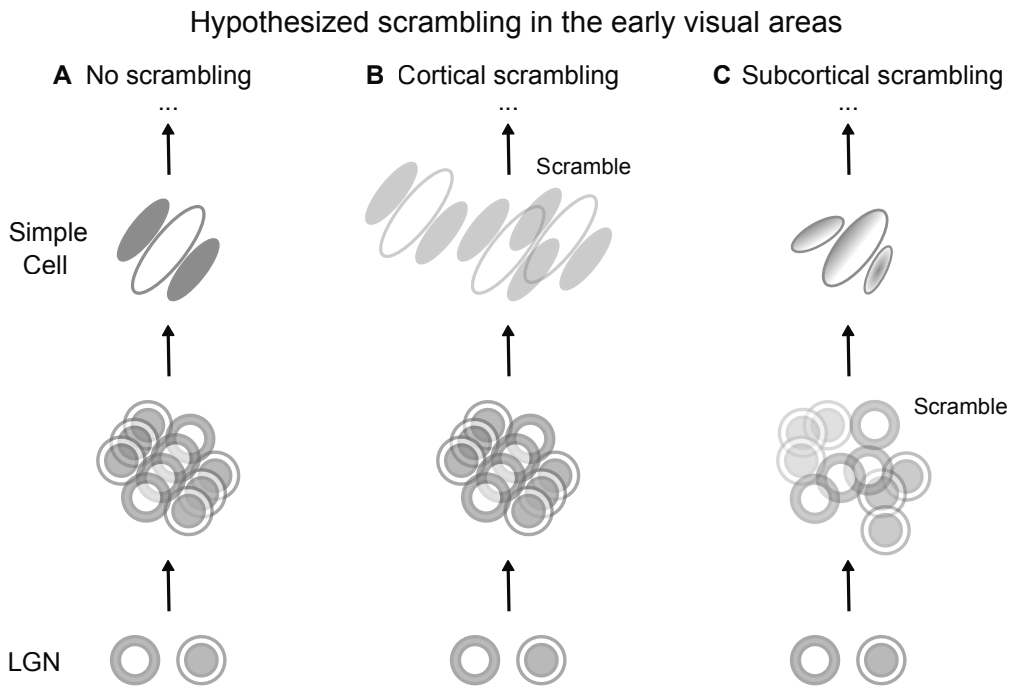

Figure 1: Cortical vs. subcortical scrambling: Panel **A** shows a canonical view of V1 simple cell receptive field formation, being composed of inputs from the lateral geniculate nucleus (LGN) subunits. Panel **B** demonstrates the impact of scrambling simple cell projections to other cortical regions ("cortical scrambling", or **CS**). Panel **C** shows the impact of scrambling the projections received from LGN subunits ("subcortical scrambling", or **SCS**).

1997; Levi et al., 2000, 1999; Baldwin et al., 2017). For instance, Baldwin et al. (2017) developed a paradigm for investigating the uncertainty in position and orientation information used in contour integration (Field et al., 1993) by applying random offsets to the positions or orientations of wavelets forming the contours. In this work, we set out to develop a paradigm to distinguish different possible forms of scrambling in the early visual cortex and investigate how scrambling may limit normal visual function.

The basic model underlying our approach is illustrated in Figure 1. In Figure 1A, we show the classic model of oriented simple cell receptive field formation from the isotropic lateral geniculate nucleus (LGN) "subunits" (Hubel and Wiesel, 1962; Hirsch et al., 1998; Alonso et al., 2001). Complex cells then pool afferents from these simple cell subunits, resulting in their receptive field properties, such as phase-invariance (Hubel and Wiesel, 1962). Subsequent receptive field studies have found spatial heterogeneity in the spatial frequency and orientation tunings of subfields in the receptive fields of early visual areas (Nishimoto et al., 2006; Tao et al., 2012; Zhang et al., 2013). In amblyopic macaques, this heterogeneity is elevated (Tao et al., 2014). Inspired by these findings in animal models, we consider two stages where scrambling could occur in humans. Scrambling could happen at the output stage from V1 simple cells (to higher cortical areas like V2 but also to complex cells and other simple cells through intracortical connections), which we refer to as "cortical scrambling" (CS) (Figure 1B). Scrambling could also occur at the LGN input to the V1 simple cells, which we refer to as "subcortical scrambling" (SCS) here (Figure 1C).

Based on our model (Figure 1), we developed software implementing an algorithm that generates stimuli which mimick expected effects of cortical and subcortical scrambling. We applied our scrambling algorithm to letters, a stimulus class chosen due to its relevance both to clinical measurements and to invariant visual object recognition in humans (see a review by Grainger et al., 2008). We restricted our letter stimuli to a single spatial frequency band to simplify stimulus generation and our interpretation of the scrambling effects (Figure 2). Previous studies have found an optimal spatial frequency band for letter identification, which was the basis for our choice of spatial frequency here (Majaj et al., 2002; Chung et al., 2002; Solomon and Pelli, 1994). The first purpose of this study is to investigate the

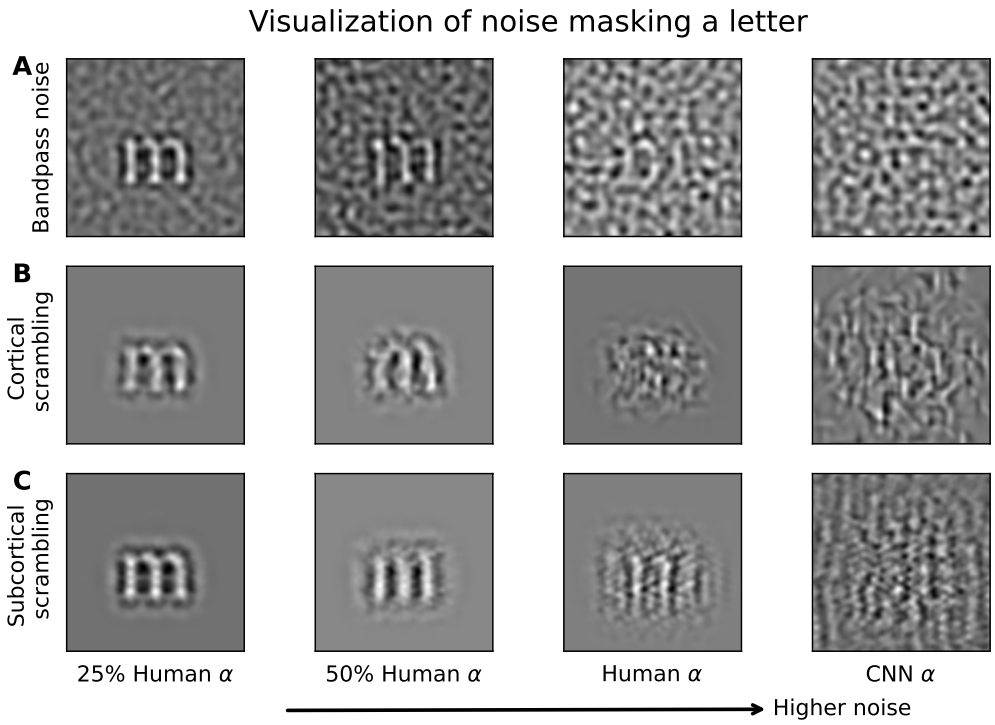

Figure 2: Example bandpass letter "m", showing the effect of increasing noise from left to right for our three conditions: bandpass noise, cortical scrambling, and subcortical scrambling. Noise levels shown here are determined from the thresholds (α) obtained in the noisy letter identification experiment for humans and Convolutional Neural Networks (CNNs). For bandpass noise (**A**), noise magnitude is quantified in the relative RMS contrast of the noise compared to the target. For cortical scrambling (**B**) and subcortical scrambling (**C**), noise magnitude is quantified as the standard deviation of the normal distribution from which offset positions of the wavelets or subunit components are drawn.

extent to which the two types of scrambling we investigate could lead to distinct stimuli that cause effects distinguishable in a psychophysics study. We investigated the impact of scrambling in two such experiments: i) a matching task (Figure 3), and ii) a letter identification task (Figures 4-5).

We were interested in comparing how the different types of noise affected human efficiency in performing letter identification. As the effects of our scrambling algorithms would make the derivation of an analytic Ideal Observer model very challenging, we turned to Convolutional Neural Networks (CNNs) for our benchmark (Kanwisher et al., 2023). Separate networks were trained de novo on each noise type to classify scrambled letters. We could then obtain a human efficiency relative to the CNN by comparing the thresholds obtained when performing the task.

Beyond using our CNNs as a benchmark for efficiency calculations, we were also interested in further comparisons between human and CNN performance. There has been a growing interest in using CNNs to model human vision (Majaj and Pelli, 2018). For object recognition tasks, deep feedforward CNNs trained on large datasets have shown lower classification error rates than humans (Krizhevsky et al., 2012). These CNNs are argued to share similar internal representation to those of V4 and IT in the macaque (Yamins et al., 2013, 2014; Khaligh-Razavi and Kriegeskorte, 2014). Behaviourally, CNNs are prone to make errors with adversarial stimuli to which humans are resilient (Goodfellow et al., 2015; Nguyen et al., 2015), likely as a result of the divergence in representation deeper into the network (Feather et al., 2023). Recent works using images that were scrambled (Farahat et al., 2023) or distorted (Berardino et al., 2017) have found significant divergence in behaviour between humans and CNNs. In this study, we make a detailed comparison (Figure 6-7) between human and CNN performance by constructing letter confusion matrices (similar to approaches used previously by Dodge and Karam, 2017; Borji and Itti, 2014).

2 Results

2.1 Scrambling matching experiment

Before investigating their impact on letter identification, we first wanted to investigate the perceived "noisiness" of our two scrambling manipulations. To measure this, we performed a matching experiment. On each trial, the participant adjusted the scrambling of a "match" letter to appear equivalent to that of a "test" letter. Both stimuli were dynamic, continuously cycling through a pre-generated set. The test letter was fixed at one of five scrambling levels. In all conditions we found the matching relationship to be well-characterized by a simple linear fit on double-log axes (Table 1). When both the match and test letter were of the same scrambling type, the matches were veridical, demonstrating that participants were able to make precise holistic judgements of scrambling magnitude. The results for comparison between the two types of scrambling are shown in Figure 3. We performed the match in both directions, adjusting CS letters to match SCS letters (Figure 3A) and the other way around (Figure 3B). At the lower scrambling levels we tested, more subcortical scrambling (SCS) was needed to match the perceived noisiness of a cortically scrambled (CS) letter. However, the perceived noisiness of the SCS letters increased more rapidly with the nominal noise level so that at the higher magnitudes tested the perceived "noisiness" of the two manipulations became similar.

We would expect that the rule governing the equivalent perceived "noisiness" would be the same regardless of which direction the matching was performed in (i.e. which type was used as the match and which as the test). To verify this, we performed an analysis to test whether the parameters describing the two best-fitting lines in Figure 3 could be reduced to a single slope and offset. We performed a nested model hypothesis test. The reduced model was a single linear model for SCS test CS match with its parameters simultaneously transformed to account for the opposite matching condition (CS test SCS match). Surprisingly, we found the full model (fitting separate linear models for each direction: SCS test CS match and CS test SCS match) explained the data better than the reduced model and rejected the null hypothesis ($p < 0.001$, nested model hypothesis test; details in Table A1 of the Appendix). From Figure 3, we see the reduced model agrees well with the full model for the majority of the scrambling range tested except for the highest test level of the CS test SCS match condition. Pre-empting our identification results (the thresholds from which are shown on Figure 3) the two models do not seem to diverge until the scrambling level is increased above the human threshold for letter identification.

Table 1: Details of linear fits to scrambling matching experiment data on double-log axes. Cross-scrambling-type conditions are plotted in Figure 3.

Condition	**Parameters**	**Value**	**95% CI**	**R^2**
SCS test SCS match	Intercept	-0.02	[-0.10 0.05]	0.94
	Slope	1.01	[0.97 1.04]	
SCS test CS match	Intercept	0.02	[-0.07 0.12]	0.93
	Slope	1.16	[1.11 1.20]	
SCS test CS match (reduced)	Intercept	0.02	[-0.04 0.22]	0.99
	Slope	1.12	[1.09 1.21]	
CS test SCS match (reduced)	Intercept	-0.02	[-0.19 0.04]	0.96
	Slope	0.89	[0.83 0.92]	
CS test SCS match	Intercept	-0.26	[-0.36 -0.17]	0.87
	Slope	0.81	[0.76 0.85]	
CS test CS match	Intercept	-0.08	[-0.17 0.01]	0.92
	Slope	0.99	[0.95 1.03]	

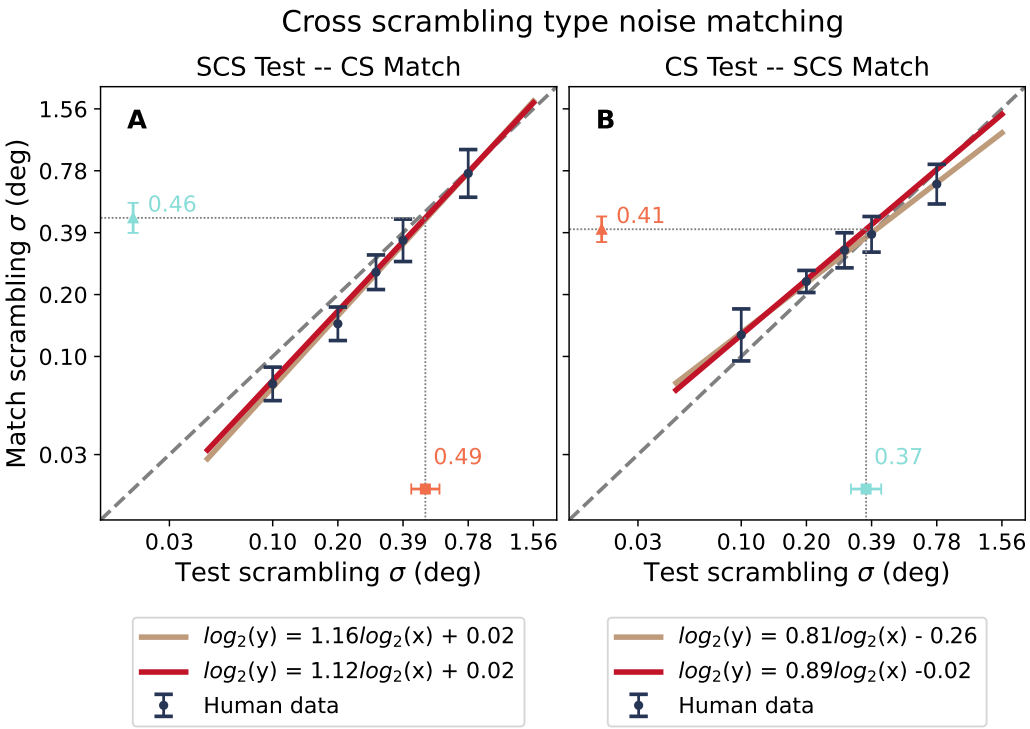

Figure 3: Results from a method-of-adjustment experiment where human participants matched the magnitude of a subcortical scrambled (SCS) letter to a cortical scrambled (CS) letter (**A**) or vice-versa (**B**). The dashed black line shows unity. The best-fit line to the data, fit separately for panels **A** and **B** is shown in gold. The best-fit line using the reduced model (common relationship in both **A** and **B**) is shown in red. Error bars show the standard deviation of the human data. The square marker shows the mean human noise threshold for the letter identification experiment. The triangular marker shows the mean transformed noise threshold using the reduced matching relation (red line). Error bars show the standard deviation of human noise thresholds.

2.2 Letter identification experiment

In our main experiment, we investigated the impact of scrambling on performance in a letter identification task. Along with our two scrambling conditions (CS and SCS), we added a bandpass noise (BN) condition in which the letter was not scrambled but was instead embedded in bandpass noise. This is a more typical stimulus that might be used in a noise-masking experiment (Majaj et al., 2002; Pelli et al., 2004). To obtain a measure of efficiency in performing this task that accounts for differences in the difficulty posed by our different noise types, we compared human performance to that of model CNN observers trained to perform the task with each noise type (details in 4.9 “Convolutional neural network training and testing” in Methods). We also compared performance to that of a template matching (TM) model that simply found the best matching noiseless letter for the noisy stimulus image. In the BN condition, we expect the template matching approach should be optimal. Therefore, we predicted our BN-trained CNN (hereinafter referred to as BN CNN) should approach the performance of our TM model.

The psychometric functions and thresholds characterizing the performance of our computer models are shown in Figure 4. For each noise condition, the best performance was that of the CNN trained on that condition. Performance was worse for CNNs that were trained to perform the task using stimuli generated with other noise types. This is evidence that each noise type encouraged the CNNs to adopt distinct strategies that were not optimal for the others (see 4.9 “Convolutional neural network training and testing” in Methods for a more in-depth analysis of CNN strategies). As expected, the BN CNN performance was similar to that of TM.

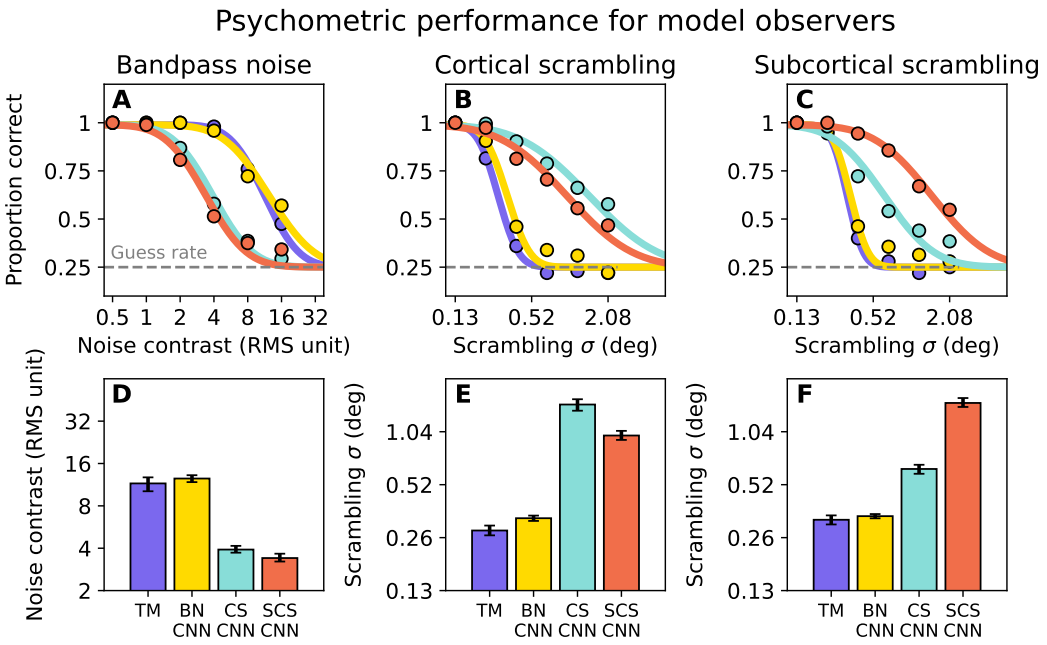

Figure 4: Performance of model observers in the letter identification task. The top row shows the psychometric functions for each model observer (colour-coded according to the bars in the bottom row) performing the task in the BN (**A**), CS (**B**), and SCS (**C**) conditions. The corresponding bar plots in the bottom row show the thresholds (median of 250 bootstraps, error bars show 95% CI) obtained from these psychometric functions. In every noise condition, the best performance (highest threshold in this paradigm) is obtained from the CNN that was trained to perform the task under that noise condition.

These computer models were used as a benchmark against which human data were compared. Psychometric functions from an example participant are shown in Figure 5A-C. Our analysis was based on the thresholds calculated from psychometric functions such as those given as examples here. Thresholds from eight humans are presented as a ratio of those from the computer models in Figure 5D-E. Humans performed worse (lower thresholds) than the CNN models that had been trained on the noise type used in the measurement. Interestingly, humans outperformed TM and BN CNN in the SCS condition (purple and yellow bars in F), indicating they used a strategy more complex than simply matching a template.

Without reference to CNNs, we saw humans were able to perform the task with more SCS than CS ($p < 0.001$, two-tailed repeated measure t-test) (Figure 5G). We were curious whether this difference in letter identification thresholds could be accounted for by the difference in perceived noisiness of our scrambling noises alone. From our matching experiment, we obtained the relation of the perceived noisiness between CS and SCS. We used this perceived scrambling relation (reduced model) to transform the noise threshold in one scrambling condition to the other condition and compare thresholds in the same perceived unit. We found there was still a significant difference between scrambling thresholds after the transform ($p < 0.01$, two-tailed repeated measure t-test for both transform directions; Comparing the triangular marker in one plot with the square marker in the other plot in Figure 3). This means that even for stimuli that would be perceived as being equally "noisy" (after scrambling levels were transformed by the matching relationship), humans were more tolerant to the SCS scrambling in their letter identification performance. When looking at performance relative to the best-performing CNNs (Figure 5H), there was a significant main effect of noise condition (one-way repeated measure ANOVA, $F(2,14) = 9.61$, $p = 0.002$). Post-hoc tests using Bonferroni correction revealed that efficiency relative to the CNN was significantly higher in the SCS condition than in the CS condition ($p < 0.001$).

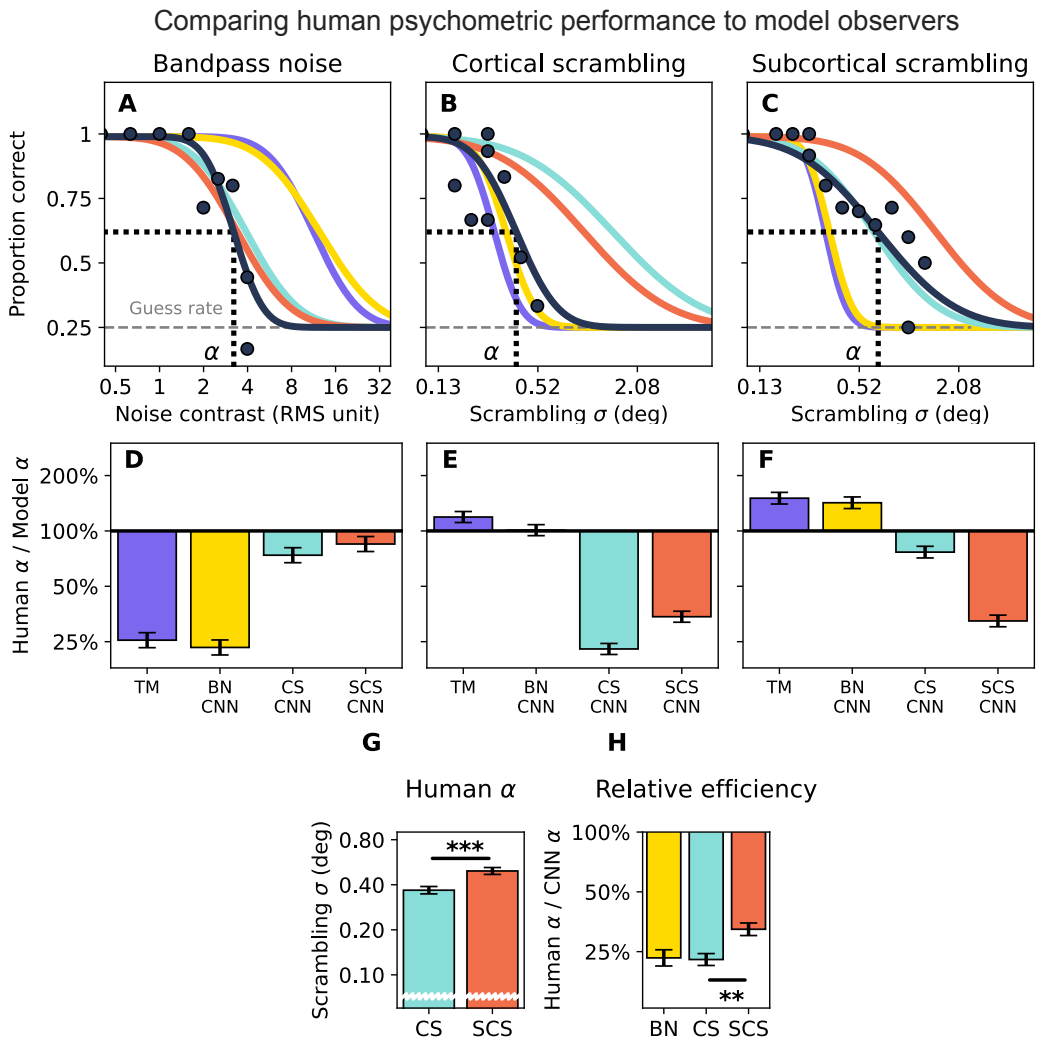

Figure 5: The top row (**A**-**C**) shows example data from one human participant (black circles) with the psychometric function fit to that data (black line). The performance of each computer model is replotted from Figure 4 for comparison with the human data. The human threshold (α) is extracted at 0.62 proportion correct. The middle row (**D**-**F**) shows the mean threshold of eight participants for each condition. These are given as a ratio relative to the performance of each model observer on that task. For values over 100%, humans are outperforming the computer models. The bottom row (**G**-**H**) shows the comparison between human thresholds in the CS and SCS conditions based on the scrambling magnitude (**G**), and the comparison between all three noise conditions based on efficiency relative to the best-performing CNN (**H**, replotting a subset of the data also shown in the middle row). All error bars show standard errors.

2.3 Comparing confusion matrices between humans and models

Beyond just comparing thresholds for performing the letter identification task, we investigated the patterns in the letter identification responses obtained from humans and model observers. We constructed confusion matrices of "response letter" against "presented letter". The entries along the diagonal on this matrix (e.g. response "m" to letter "m") give the proportion of correct responses for each letter stimulus (see an example in Figure A1 of the Appendix). From these matrices, we extracted two probability vectors: mistake frequencies (the off-diagonal elements) and letter response biases (summing across rows through each column). Alongside the overall performance (analysed above), these two

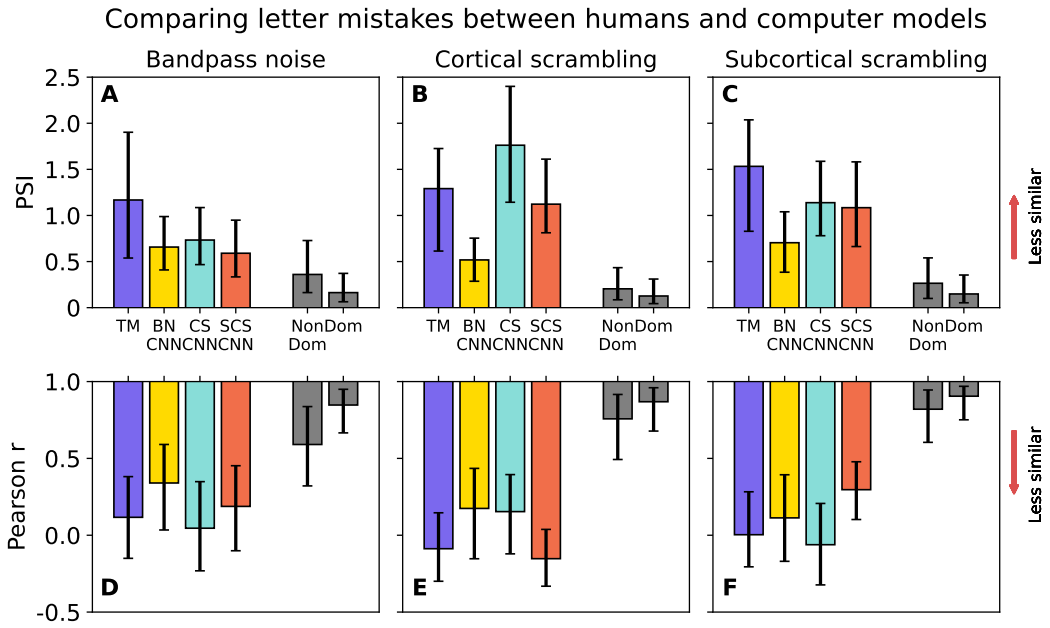

Figure 6: Comparing the mistakes made (off-diagonal elements of confusion matrices) between humans and computer models. In each case, the comparison is being made to the dominant eye human data. The figure includes a comparison of data obtained with the non-dominant eye (NonDom), and with randomly resampled dominant eye viewing data (Dom). Bar height represents the median of 250 bootstrap samples (with 95% CI).

vectors provide a more complete description of the behaviour of each observer. We compared the probability vectors between human and model observers to assess their similarity. As these vectors contained probabilities, we used the Population Stability Index (PSI). This is a relative entropy measure commonly used in model monitoring. It measures the similarity between two probability distributions (Yurdakul and Naranjo, 2019). A higher PSI means a lower likelihood that one distribution is the result of samples taken from the other. We also computed the Pearson correlation coefficient, to have an analysis comparable with previous work that studied confusion matrices in humans and CNNs (Dodge and Karam, 2017; Borji and Itti, 2014). Mistake frequencies (off-diagonal elements of the confusion matrices) are compared in Figure 6. In each case, behaviour is compared to that of the dominant eye of the (pooled) human participants (Dom). A Dom-Dom comparison was possible (without being at 100% by definition) by comparing two bootstrapped samples from the same dominant eye viewing dataset. Note that these two samples were not independent. For an independent comparison, we further include a comparison against data collected with the non-dominant eye (NonDom). We expect humans (with normal vision) to behave similarly regardless of which eye they use. This condition therefore gives a useful indicator of what PSI or Pearson r values would indicate a “ceiling” level of similarity.

For PSI (Figure 6A-C), we found TM PSI was significantly higher than the comparison to the NonDom performance in the two scrambling conditions, but not in the BN condition (purple bars in Figure 6A-C). This suggests humans likely did not employ a template matching strategy, at least when letters were scrambled. For the networks, SCS CNN and CS CNN PSIs were high in both scrambling conditions (cyan and orange bars in Figure 6B, C), suggesting a divergence in their internal representations from humans. In the BN condition, all CNN models did not significantly diverge from the NonDom comparison (yellow, cyan and orange bars in Figure 6A). Two things were surprising to us. First, we saw the CNN trained on the same noise condition as the test did not have higher similarity with humans compared to other CNNs. So, whatever strategy the networks were adopting based on the task was not more similar to the human approach, at least by this metric. Second, BN CNN was more similar to humans than the scrambling trained CNNs for the scrambling conditions, especially for CS (yellow bar in Figure 6B). For Pearson correlation (Figure 6D-E), we saw again the significant divergence between scrambling-trained CNNs and controls in the two scrambling conditions.

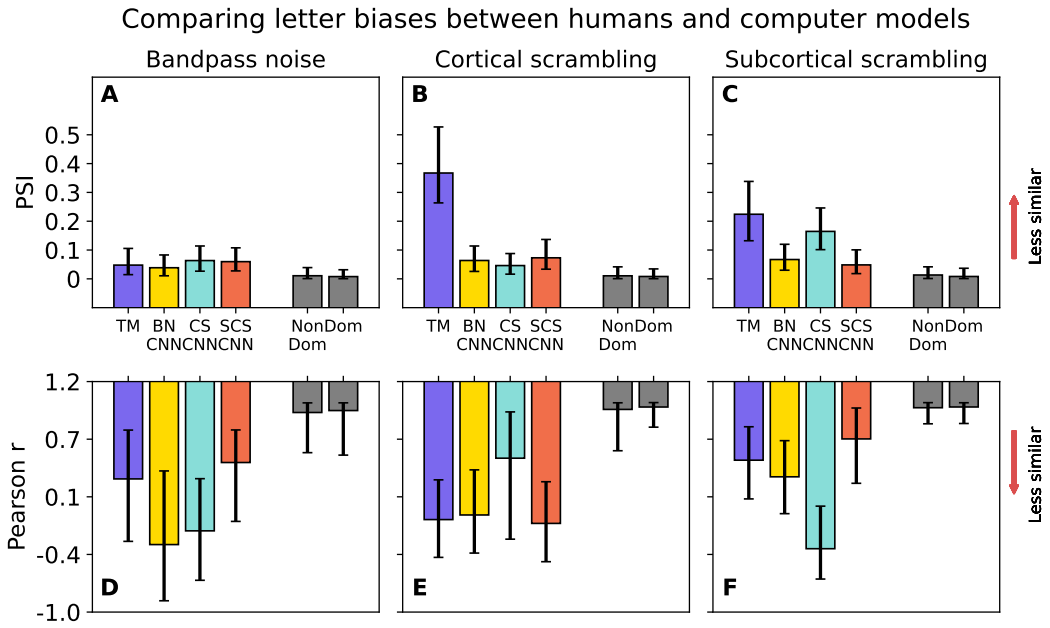

Figure 7: Comparing letter biases between humans and computer models. As in Figure 6, each bar indicates the comparison with that obtained from the dominant eye in our pooled human data. For further information see also the Figure 6 caption.

While this was initially confirmatory of our finding using PSI, we found Pearson correlation was a poor indicator of the similarity between humans and models upon closer examination of the scatter plots (Figure A2 of the Appendix). This is most effectively illustrated in the comparison between BN CNN and CS CNN for the CS noise condition (Figure A2 [CS row]), where their correlation distributions were similar, but the probabilities were drastically different. Due to the type of data under comparison, we proposed PSI as the more accurate method for analysing these confusion matrices compared to Pearson correlation.

The comparisons of the biases for the letter response are shown in Figure 7. The PSI of TM again significantly differed from the NonDom "control" in the scrambling conditions but not the BN condition (purple bars in Figure 7A-C). This means that the scrambling led to this model favouring certain letter responses over others. Although the probability of each letter being shown was equal, the TM model was not able to incorporate this information to adjust its propensity to choose each letter. For the CNNs, only the CS CNN significantly differed from the NonDom benchmark (cyan bar in Figure 7C). The failure of the Pearson correlation in measuring similarity was again demonstrated (Figure A3 of the Appendix). For example, for the CS condition, TM provided overall poorer predictions of human letter biases than BN CNN (Figure A3 [CS row]), evident in the high letter "m" probability compared to humans. This was not reflected in the correlation (purple and yellow bars in Fig 7E) but was shown using PSI (purple and yellow bars in Fig 7B).

3 Discussion

Our results on human letter identification performance in situations of cortical and subcortical scrambling yield new insights into the perceptual effects that may occur due to scrambling of the projections in the early visual cortex. Our main conclusion is that humans are more resilient to subcortical scrambling (SCS) than to cortical scrambling (CS). This holds either when considering the "perceived noisiness" resulting from each manipulation, or on the basis of comparison to convolutional neural networks trained on each task. The possible explanation we propose for this would be if there were more positional uncertainty in the LGN afferents to the cortex than there is in the simple cell afferents

in the cortex. The consequence of that could be an adaptation of the human visual system to handle a greater degree of scrambling of the type represented in our SCS stimuli.

Although our scrambling algorithms are physiologically inspired, we do not assert that they give a complete and accurate account of the physiological basis of scrambling in the human visual system. However, the fact that our two forms of scrambling do result in discriminable effects suggests that they tap into distinct aspects of the underlying physiology. Previous works studying scrambling have mainly employed a task design that measured the just noticeable difference between two stimuli: one with scrambling and one without scrambling (or sometimes with a fixed level of pedestal scrambling). In this work we took a different approach where we explored the configurational space of scrambled stimuli in the high noise limit where humans were still able to recognize a letter at some fixed percentage correct.

Two stimuli that are physically distinct but perceptually invariant (in a relaxed sense of being classified as the same category) are known as metamers in the vision science literature (Freeman and Simoncelli, 2011; Feather et al., 2023). Our results suggest the metameric space of subcortical scrambling is greater than that for cortical scrambling. This implies during inference, the “more optimal” stimulation of V1 simple cells in the CS condition (with intact Gabors) in fact constraints the solution space compared to the “less optimal” stimulation of simple cells through the mixing of subunits. Complex feature combination has been argued to be a gradual process starting in the early visual cortex with the presence of heterogeneity in spatial frequency and orientation tunings of the subfields (Nishimoto et al., 2006; Tao et al., 2012). The effect of our subcortical scrambling noise primarily being that of orientation mixing (see 4.4 “Stimulus details and further scrambling algorithms” in Methods) may be able to exploit this early feature combination better than cortical scrambling noise, leading to better psychophysical performance. Our results here for letters are also consistent with previous studies using position and orientation noise to scramble their stimuli where they found the efficiency was higher for orientation than position for coding of contours (Baldwin et al., 2017) and images of natural stimuli (Christensen et al., 2019).

Comparing confusion matrices and noise thresholds between computer models and humans allows us to generate hypotheses about the strategy used by humans. Past work modelling letter identification has primarily used template-based models (Beckmann and Legge, 2002; Chung et al., 2002; Gold et al., 1999; Parish and Sperling, 1991; Watson and Ahumada, 2008, 2012, 2015; Georgeson et al., 2023). In our bandpass noise (BN) condition, since we used noiseless templates (which should be optimal for this task), the template matching (TM) model was able to greatly outperform, as expected. Nevertheless, we found the confusion matrices of the TM model are similar to humans. To our surprise, the scrambling-trained CNNs share just as much similarity with humans in confusion matrices while at the same time more closely mimicking human threshold than the TM model, making them better models for letter identification in BN noise than the TM model. Previous work using CNNs to model letter identification in white noise (Ziskind et al., 2014) showed CNNs trained on restricted resources (reduced number of layers and filters) approximate human thresholds better than deeper counterparts. Our BN condition results show that merely training CNNs on scrambling noises without changing the architecture can capture human performance both in terms of threshold as well as mistakes and biases. In the scrambling conditions, we see the TM model shows a significant divergence from humans in confusion matrices and gets outperformed by humans, suggesting humans are using a more sophisticated strategy than template matching. Interestingly, the human strategy is more similar to the BN CNN which can better predict both the thresholds and the confusion matrices (especially the CS condition). This finding suggests that the internal neural computations preserve the relative positions of individual features like in the noiseless bandpass letters (a requirement for the BN CNN to perform well; also see BN CNN row in Figure 11 in 4.9 "Convolutional neural network training and testing” in Methods) even when letters are scrambled.

Our use of CNNs offers a use case for deep neural networks in human cognitive science research. In this study, they allowed us to distinguish the two forms of stimulus manipulation generated through our scrambling algorithms. Furthermore, they served as a benchmark against which human performance could be compared across experimental conditions. While our CNN models were meant to give a high level of performance for the task, we also incorporated

features like successive convolutional layer and pooling layer pairs to deliberately make the CNN architecture reflect that of early visual computations (Yamins and DiCarlo, 2016). In future work, we will apply these methods to investigate the role of scrambling both in peripheral vision and in amblyopia.

4 Methods

4.1 Equipment

Participants viewed a gamma-corrected CRS Display++ LCD screen (Cambridge Research Systems, United Kingdom) from a distance of 62 cm. At this distance, there were 31 pixels per degree of visual angle. The mean luminance of the display was 58 cd/cm^2. The experiment software ran on a Dell XPS laptop (Dell Technologies, Inc.). It was written in Python, making use of the PsychoPy library (Peirce, 2007). All experiments were performed indoors under ambient lighting.

4.2 Participants

Eight participants (aged 23-29, 4 female, 4 male) participated in the letter identification experiment. A subset of four participants and one additional participant (aged 23-26, 2 female, 3 male) participated in the noise-matching experiment. All had normal or corrected-to-normal vision. All experiments were carried out with informed written consent in accordance with the tenets of the Declaration of Helsinki. Study procedures were approved by the Centre for Applied Ethics at the McGill University Health Centre.

4.3 Stimulus synthesis overview

Our letter stimuli were designed to be similar to those used in previous letter identification studies. We chose four lowercase letters: o, m, d, and z. These were selected on the basis that they were found to be perceptually distinct from each other in previous work measuring letter similarities (Courrieu et al., 2004; Janini et al., 2022). The Times New Roman font was chosen, for consistency with previous studies that investigated the optimal spatial frequency channel using bandpass-filtered letters (Chung et al., 2002). The letter height (typographic x-height) was set to 1.5 degrees of visual angle. At this letter height, the optimal spatial frequency would be 1.5 cycles/deg (Majaj et al., 2002; Chung et al., 2002; Solomon and Pelli, 1994).

We generated our scrambled stimuli using a novel stimulus decomposition and resynthesis algorithm. This was inspired by previous studies of the activation properties of cortical receptive fields (Daugman, 1985; Field, 1987) and multi-resolution wavelet transform (Würtz, 1995; Fischer et al., 2007). Our decomposition and resynthesis scheme involved a transformation of basis using log-Gabor kernels. A log-Gabor has a spatial frequency tuning that is Gaussian over log-transformed spatial frequency (as opposed to a Gabor which is Gaussian over spatial frequency). Compared to Gabors, they have the convenience of being D.C.-balanced in any phase (Field, 1987).

In Figure 8 we show the stimulus generation pipeline for the three noise conditions. The raw letter images were 200 pixels square. We constructed log-Gabor filters in frequency space with a peak spatial frequency of 9.8 cycles/image (corresponding to 1.5 cycles/degree) with a bandwidth (full-width at half-magnitude) of 1.5 octaves. Our goal here was not to reconstruct the original image but rather to generate stimuli that allow us to investigate performance using the spatial frequency channel of interest (that should be optimal in humans for this task). Filters were created at eight preferred orientations equally spaced between 0° and 360° (orientation bandwidth ±30°). At each orientation, two filters were created in quadrature phase (0 and 90 degrees). This gave a total of 16 filters. We applied sparse sampling, obtaining a weight at every half-wavelength of the log-Gabor. We then thresholded these weights, retaining only the top 1% of the entire stack.

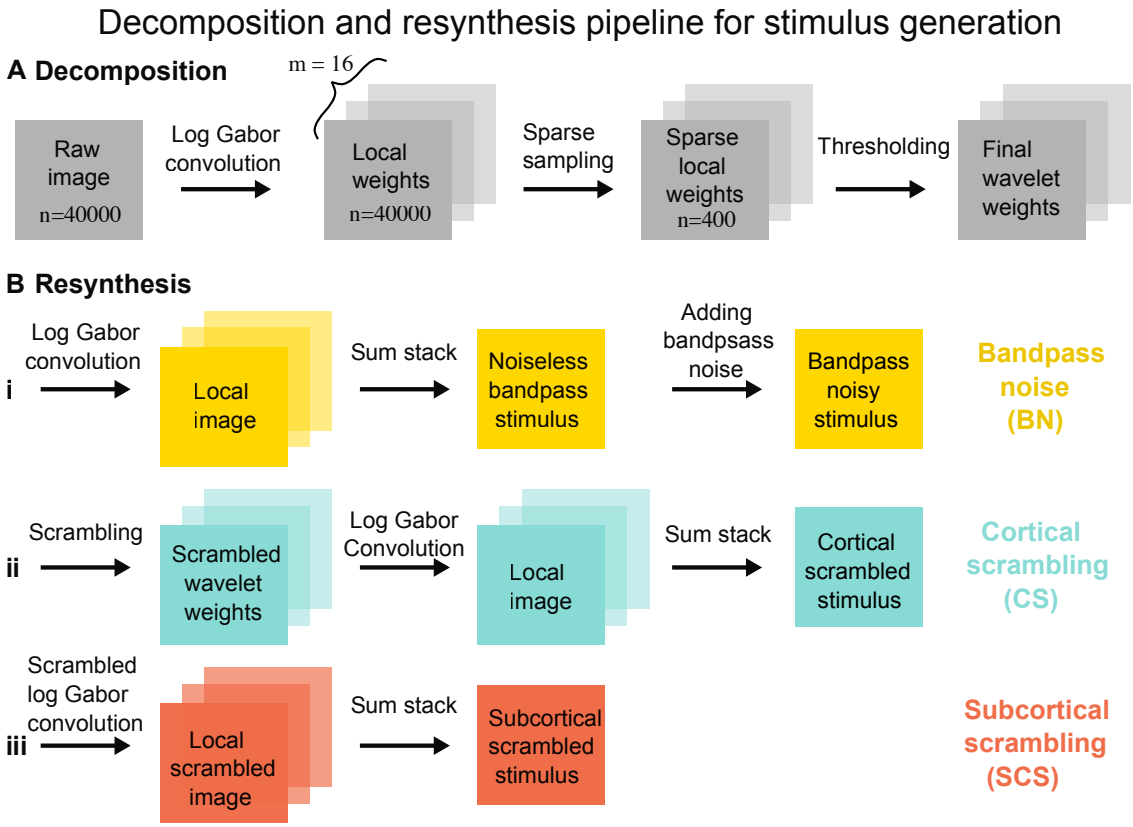

Figure 8: To generate our stimuli, we first decompose the letter image (**A**) into its local wavelet weights (n = 10,000 sensors; m = 16 log-Gabor filters, consisting of 8 orientations and 2 phases). The sparse sampling rate is half the log-Gabor wavelength (at 9.8 cycles/image, this leads to a reduction of n by a factor of 100). Thresholding is applied to the entire stack, keeping the top 1%. We then resynthesize (**B**) a stimulus image from the final wavelet weights: **i**) for BN stimuli, the letter is reconstructed (no scrambling) and bandpass noise is added, **ii**) for CS stimuli, the positions of the weights are scrambled, and then convolved with log-Gabors, **iii**) for SCS stimuli, the weights (without scrambling) are convolved with scrambled log-Gabors (Figure 9).

In the resynthesis, we shifted the phase of the 0 degree components to 180 degrees. This accounts for the flipped sign in complex conjugation as we convert back to representing the image in the spatial domain. All conversions to and from the Fourier domain were performed using the Fast Fourier Transform module `fft` of the NumPy package in Python (Harris et al., 2020). In the bandpass noise condition, bandpass noise samples were generated by feeding white noise samples separately to the pipeline and subsequently added to the resynthesized letter image (Figure 8B.i). In the cortical scrambling condition, we drew random position offset values for each wavelet in both the horizontal and vertical directions. These were obtained from a Gaussian probability distribution, the standard deviation of which controlled the magnitude of scrambling (Figure 8B.ii). For subcortical scrambling, we scrambled the positions of the subunits forming the log Gabor filters used to resynthesize the image (Figure 8B.iii).

4.4 Further stimulus details and scrambling algorithms

Our decomposition and resynthesis pipeline used Cartesian-separable log-Gabors. These have an orientation bandwidth that is perpendicular to the radial (spatial frequency) axis in the Fourier domain (Meese, 2010; Baker et al., 2022). Compared to the more commonly-used polar-separable log-Gabor, the spatial-domain filter kernels defined by the Cartesian-separable log-Gabor function do not splay outwards. This gives a form that is more consistent with the single-cell physiology of V1 (Jones and Palmer, 1987).

In the decomposition, we first convolved a square image $\mathbf{X}$ with a set of m log-Gabor filters $\mathbf{G}_{j(j\in\Omega)}$ where $\Omega = \{1, ..., m\}$, each indexing a filter defined in Fourier space with a different orientation preference and phase preference. This was done via multiplication in the Fourier domain:

$$\mathbf{c}_j = \mathcal{F}^{-1}\left(\mathbf{G}_j \cdot \mathcal{F}\left(\mathbf{X}\right)\right), \tag{1}$$

where $\mathbf{c}_j$ represented the local weights of a single (jth) channel. $\mathcal{F}$ and $\mathcal{F}^{-1}$ refer to the discrete Fourier and inverse Fourier transform respectively. Doing this for all the channels gave $\mathbf{c}_{j(j\in\Omega)}$ which was a stack or "pyramid" of local weights. The decomposition, defined as a wavelet transform, is:

$$W^{\text{decomp}}\left(\mathbf{X}\right) = \mathbf{c}_{j(j\in\Omega)}. \tag{2}$$

Each $\mathbf{c_j}$ was then sparsely sampled and the entire stack was subject to thresholding to keep the most active weights. This gave the final wavelet weights $\widetilde{\mathbf{c}}_{j(j\in\Omega)}$. For the resynthesis with no scrambling, we simply convolved $\widetilde{\mathbf{c}}_{j(j\in\Omega)}$ with complex conjugated versions of the filters $\overline{\mathbf{G}}_j$ and summed across channels:

$$W^{\text{resynth}}\left(\widetilde{\mathbf{c}}_{j(j\in\Omega)}\right) = \sum_{j\in\Omega} \mathcal{F}^{-1}\left(\overline{\mathbf{G}}_j \cdot \mathcal{F}\left(\widetilde{\mathbf{c}}_j\right)\right) \tag{3}$$

For cortical scrambling, local weights $\widetilde{\mathbf{c}}_j$ were scrambled using the scramble function S:

$$W^{\text{resynth}}_{\text{cortical}}\left(\widetilde{\mathbf{c}}_{j(j\in\Omega)}\right) = \sum_{j\in\Omega} \mathcal{F}^{-1}\left(\overline{\mathbf{G}}_j \cdot \mathcal{F}\left(S\left(\widetilde{\mathbf{c}}_j\right)\right)\right) \tag{4}$$

For subcortical scrambling, the filters $\overline{\mathbf{G}}_j$ were scrambled using the scramble function $S^{'}$:

$$W^{\text{resynth}}_{\text{subcortical}}\left(\widetilde{\mathbf{c}}_{j(j\in\Omega)}\right) = \sum_{j\in\Omega} \mathcal{F}^{-1}\left(S^{'}\left(\overline{\mathbf{G}}_j\right) \cdot \mathcal{F}\left(\widetilde{\mathbf{c}}_j\right)\right) \tag{5}$$

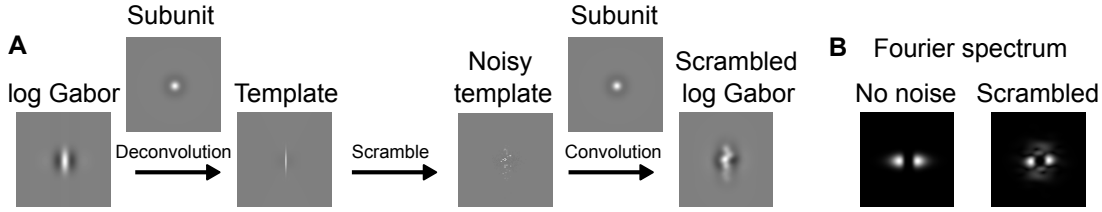

Figure 9: Panel (**A**) shows the generation of a log-Gabor affected by subcortical scrambling. The log-Gabor is first deconvolved with an isotropic subunit to obtain a "wiring diagram". The wiring diagram is then scrambled to simulate miswiring in the assembly of the modelled V1 receptive field, giving our scrambled log Gabor. Panel (**B**) shows the amplitude spectrums of the log-Gabor before and after scrambling.

The pipeline for generating the stimuli with subcortical scrambling is shown in Figure 9. We first found the "wiring diagram" that would compose the oriented log Gabor from isotropic log Gabor subunits (which have a donut-shaped Fourier spectrum) through deconvolution. The subunit function was:

$$S(f,\theta) = \exp\left(\frac{-(log2(f/f_0)^2}{2(0.424\omega)}\right), \tag{6}$$

where f_0 is the peak spatial frequency and ω is the spatial frequency bandwidth (full-width at half-height) in octaves. We then kept the top 1% of this wiring diagram and scrambled it in the same manner as the cortical wavelet weights were scrambled in the cortical scrambling condition. After the wiring diagram became scrambled it was convolved with the subunit to construct a scrambled log-Gabor. A single scrambled wavelet was generated for each orientation and phase used in the reconstruction. In the physiology we would expect the scrambling affecting each receptive field to be distinct.

In Figure 10, we show how our subcortical scrambling affects the Fourier spectrum, introducing energy into neighbouring orientations. In the Fourier domain, we fitted spatial frequency and orientation tuning functions to 200 example log-Gabors at three different scrambling levels. We see scrambling significantly decreased the energy within the original spatial frequency and orientation channels, and this energy was spilled into neighbouring orientation channels (resulting in increasing orientation bandwidths for higher scrambling levels as the orientation spectrum becomes flat in the limit). Intriguingly, we also noticed a small downward shift in the peak spatial frequency and the spatial frequency bandwidth for some low scrambling levels.

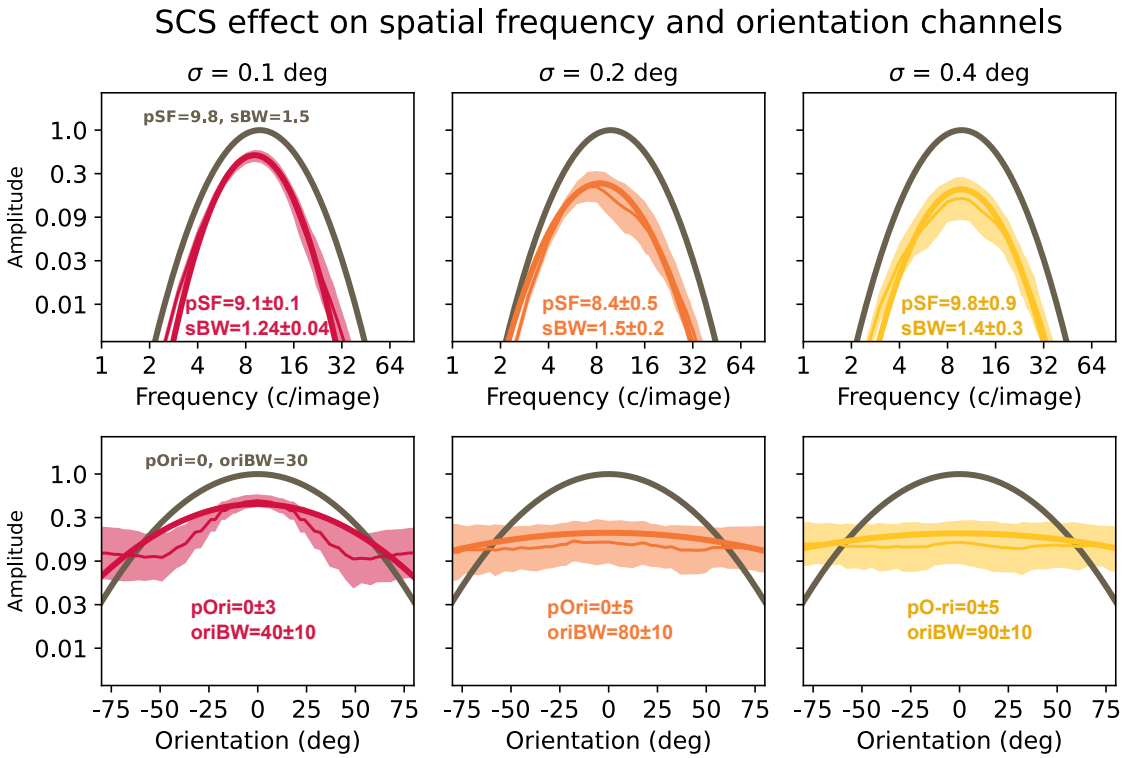

Figure 10: The impact of the subcortical scrambling on the spatial frequency and orientation tuning of the resulting filters. Original tuning functions are shown in black. Spatial frequency tuning data are obtained by taking a radial slice of the scrambled log Gabor at the optimal orientation (before scrambling). Orientation tuning data are obtained by taking a curved slice at the radius equal to the optimal spatial frequency. We performed 200 iterations at three scrambling levels and plot the mean amplitude values with their standard deviations (shaded regions). Four parameters of the tuning functions were fit to each of the 200 samples: peak spatial frequency pSF in cycles/image, spatial frequency bandwidth sBW in octaves, peak orientation pOri in degrees, and orientation bandwidth oriBW in degrees. We report the mean and standard error of the parameters and plot the tuning functions that they would give rise to (coloured curves).

4.5 Psychometric function fitting

For both our human and model performance data, we calculated threshold stimulus levels by fitting cumulative normal functions to the raw psychometric function data. The slopes and thresholds were fit as free parameters, with the lapse rate and guess rate fixed at 0.01 and 0.25 respectively for our four-alternative forced-choice tasks (obtaining a threshold for 62% percent correct). Fitting was performed through a maximum-likelihood fitting procedure (after Kingdom and Prins, 2016) using the `scipy.optimize.minimize` function from Python's SciPy package (Virtanen et al., 2020).

4.6 Contrast threshold measurements

To standardize the visibility of our letter stimuli, we measured the contrast threshold for performing a single-interval four-alternative forced-choice letter identification task for our noiseless letters. The contrast of the letters ramped on over 50 ms following a raised-cosine envelope, remained constant for 200 ms, and then ramped off over 50 ms. The peak RMS contrast of the letters was controlled by a 2-down 1-up staircase. The starting Weber contrast was 7%. The step size for the contrast change was a factor of 1.58 for the first 4 reversals and 1.26 for the last 4 reversals. The staircase terminated after a minimum of 30 trials and 8 reversals. Data were combined over 3 repetitions of the task for analysis. Contrast thresholds were obtained by fitting a psychometric function (see 4.5 "Psychometric function fitting").

4.7 Scrambling matching experiment

To make an initial comparison of the effects of our cortical and subcortical scrambling algorithms, we performed a simple matching experiment. Three scrambled letters (with the same letter identity) were shown simultaneously on the screen. They were positioned side-by-side, with 10 degrees of visual angle between them. The central noiseless letter made the participant certain of what letter was being shown. The participant adjusted the scrambling level of the match letter (positioned on the right) to that of the test (on the left). Participants proceeded through the experiment at their own pace with no time limit. Four matching conditions were tested: CS to CS, SCS to SCS, CS to SCS, and SCS to CS. For each condition, we tested five logarithmically-spaced scrambling levels for the same four letters as before, each twice, one in which the match noise level was initialized at a value above the test noise level, one below, for a total of 40 trials that were randomized in order. On each presentation, participants adjusted the scrambling of the match letter to match that of the test letter before proceeding to the next trial. During the stimulus presentation, both the standard and match letters were updated at 5 Hz, being replaced by another random stimulus having the same letter identity and scrambling magnitude. The purpose of making the stimuli dynamic in this way was to encourage the participants to make a holistic judgement of the scrambling magnitude.

The participants performed the task through natural binocular viewing. The stimulus RMS contrast was fixed at ten times the mean contrast threshold of the five participants. The total range of Weber contrast for all stimuli was 57-72% for CS and 57-70% for SCS. Other psychophysical and experimental settings were the same as those in the letter identification task.

4.8 Letter identification experiment

For our scrambled letter identification task, we also used a single-interval four-alternative forced-choice design. The testing procedure was the same as in the contrast threshold measurements task, however stimuli were shown at a fixed contrast (four times the participant's contrast threshold, measured above) with the task difficulty being controlled by the noise magnitude. This requires a flipped staircase, as increasing the contrast of the bandpass noise in the BN task, or the scrambling magnitude in the CS or SCS task makes the task more difficult (where increasing contrast in the previous task made it easier). A 2-up 1-down staircase controlled the noise level of the stimulus, terminating after a minimum of 8 reversals and 30 trials.

Two viewing conditions were tested for each noise type: dominant eye viewing and non-dominant eye viewing. This gave six conditions in total, which were each repeated three times. Monocular viewing was achieved by wearing an eye patch. The testing order of all conditions was randomized. Although the RMS contrast was constant across the different noise types, the relation between RMS contrast and expected Weber contrast varied. Overall, the expected Weber contrast calculated across participants, viewing stimuli at the noise threshold with their dominant eye was 5% for BN stimuli, 12% for CS, and 9% for SCS. The total range of Weber contrast for all stimuli was 5-8% for BN stimuli, 7-12% for CS, and 6-11% for SCS.

4.9 Convolutional neural network training and testing

We separately trained three CNNs to each classify letters for one of our three noise conditions. For each condition, we generated a dataset containing 200 stimuli at each of the six logarithmically-spaced noise magnitudes covering the entire range of noise levels shown to humans, for a total of 1200 stimuli. We then followed a split of 70%, 20%, and 10% for training, validation, and testing (832, 224, and 144 images respectively). The four letters were shown with equal probability. Stimuli in all datasets were max-min normalized to have the highest contrast. The expected quantization in the grey level that would result from humans viewing the stimuli at a lower contrast on our display was implemented to prevent the CNN having access to more information than the human participants.

Our CNN was built with Keras and TensorFlow (Abadi et al., 2015) in Python. It comprised three convolution layers each followed by a max-pooling layer, one fully-connected layer, and one output layer with four output units (one per letter). The network had 2 million trainable weights in total. The activation function was the rectified linear unit (`ReLU`) for the convolutional layers and the first dense layer. For the last output layer, the `SoftMax` activation function was used. We used the `SparseCategoricalCrossentropy` objective loss function (Chollet, 2021) and the Adam optimizer (Kingma and Ba, 2014) to minimize this loss function with a fixed learning rate of 0.001. For each epoch during training, we visualized the validation loss and training loss. Early stopping was used to avoid overfitting with the patience parameter (number of epochs to wait before terminating if validation loss does not decrease) set to 10. After the training was complete, we tested to ensure the models achieved a minimum of 90% accuracy on the test set (in reality most networks achieved at least 95% accuracy). After this, we applied the trained CNN to classify letters on a larger new test set (200 images for each noise level, 1200 in total). This training and prediction cycle was performed five times in total for each noise condition and the prediction data from the five iterations were pooled.

To further understand the strategies used by CNNs, we conducted a Gradient-weighted Class Activation Mapping (Grad-CAM) analysis (Selvaraju et al., 2020) to understand which parts of the stimulus a CNN uses to make its prediction. To do this we first visualized the activation maps of the last convolutional layer for test stimuli of the three noise conditions. We then computed the gradient of the top predicted class output neuron activation with respect to the mean activations of the 16 activation maps. The individual partial derivatives gave a measure of the relative importance of the different feature activations in the CNN making its prediction. We then weighed each activation map by its gradient and summed them together to obtain a heatmap. This was repeated 50 times for a dataset of test stimuli and the final heatmap was taken as the average (Figure 11). We see that the CNNs adopt distinct strategies. In the scrambling conditions, we see the BN CNN places more importance on the centre of the scrambled letters compared with the scrambling-trained CNNs.

4.10 Template matching model

A simple template matching strategy was investigated by performing an in-place pixel-by-pixel multiplication between each noiseless resynthesized letter (as a template) with the noisy stimulus. This generated four response images. A "max" rule was applied to pick the response image with the highest standard deviation. The identity of the letter template that gave rise to this response image was then chosen by the model. We tested this template matching model on the same test set as the CNNs.

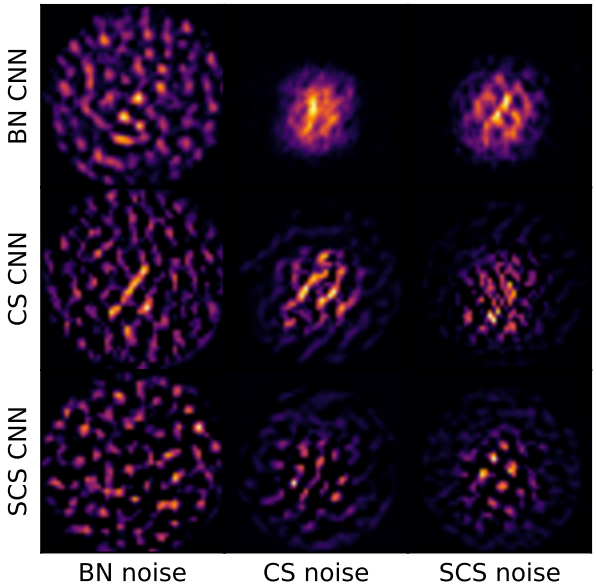

Figure 11: Heatmaps of activations in the last convolutional layer for an input stimulus with identity z. Each map shows the combined activations across features where each feature map is weighted by the gradient of the predicted class output neuron with respect to the mean activation of that feature map. Noise levels are chosen to be the human thresholds for the letter identification task. Each map is the average of 50 maps obtained from generated test stimuli at the same noise level.

4.11 Measuring response similarity using confusion matrices

For each observer that performed the letter identification task (human or computer model), we bootstrapped from their letter response data to construct confusion matrices. Normalized confusion matrices were generated using the `sklearn.metrics.confusion_matrix` function in Python (Pedregosa et al., 2011). For human data, we pooled letter response data across participants with each participant contributing the same length of data to a new dataset (90 trials from each participant for a total of 720 trials). We did this for dominant eye viewing and non-dominant eye viewing data. These datasets were not perfectly balanced, meaning some letters were shown slightly more often than others. This was due to the random selection of the letter to be shown on any given trial. To ensure the comparison with models was a fair one, we bootstrapped a balanced dataset first (150 trials for each letter for a total of 600 letters) and then bootstrapped from this new set to construct a confusion matrix for comparison with the models. This balancing step was done for every single regular bootstrap. The bootstrapping ran 250 times.

From each bootstrap confusion matrix, we calculated letter biases and mistake frequencies (see Figure A1 in "Confusion matrix example" in the Appendix). For letter biases, we added probabilities across rows for each column to get a probability vector of length 4. For mistake frequencies, we first removed the diagonal entries and then normalized the rest of the probabilities. This gives us a pattern of mistakes we can analyse that is independent of the overall performance level. After flattening this gave us a probability vector of length 12. For a single bootstrap, we calculated both the Pearson correlation and the Population Stability Index (PSI) from the probability vectors of the observer condition and the reference condition. As we were interested in comparing human performance against the models, that reference condition was always the dominant eye viewing condition. The PSI was defined as

$$\mathrm{PSI}\left(P, Q\right) = \mathrm{D}_{\mathrm{KL}}\left(P|Q\right) + \mathrm{D}_{\mathrm{KL}}\left(Q|P\right), \tag{7}$$

where D_{KL} is the Kullback–Leibler divergence. In the discrete case, this is defined as

$$\mathrm{D_{KL}}\left(P|Q\right) = \sum_i P\left(i\right) \ln\left(\frac{P\left(i\right)}{Q\left(i\right)}\right) \tag{8}$$

In our case i denoted the entry of the probability vector. We see PSI is a symmetric measure compared to the KL divergence because the metric is agnostic to which distribution is chosen as the reference distribution.

5 Additional information

5.1 Acknowledgements

The funding support for this work comes from Canadian Institute for Health Research (CIHR) (#8139 to RFH), Fonds de Recherche du Québec – Santé (FRQ-S) (Master's Training Award to XRZ), and the Réseau de Recherche en Santé de la Vision (Recruitment Award to XRZ).

5.2 Author contributions in CREDIT format

XRZ: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Software, Visualization, Writing - Original Draft. **RFH**: Conceptualization, Funding Acquisition, Methodology, Project Administration, Supervision, Writing - Review & Editing. **ASB**: Conceptualization, Formal Analysis, Funding Acquisition, Methodology, Project Administration, Software, Supervision, Visualization, Writing - Review & Editing.

5.3 Potential Conflicts of Interest

Alex S. Baldwin (ASB) and Robert F. Hess (RFH) are both inventors on patent(s) and other intellectual property concerning disorders of binocular vision including amblyopia. Some of these technologies (unrelated to the area of research here) have been commercially licensed by McGill University to Novartis International AG, and have been developed by ASB and RFH under a research agreement between Novartis and the Research Institute of the McGill University Health Centre.

6 Appendix

6.1 Nested model hypothesis

The results of the nested model hypothesis testing for the matching task are given in Table A1. We find that the "full" model (where matching SCS to CS is not equivalent to matching CS to SCS) gives a better account of our data. When comparing the difference between these two models, we find that assuming the reduced model would result in a less than 10% difference for the majority of the test range. For the first and second highest scrambling levels tested in the CS test SCS match condition, the difference was 12% and 45% respectively.

Table A1: Nested model hypothesis testing. The reduced model is the linear model for the SCS test CS match condition, which is transformed to fit the data for the CS test SCS match condition. The full model refers to fitting two linear models to two conditions separately.

SSE_{full}	$SSE_{reduced}$	Δ SSE	Δ p	MSE_{full}	F statistic	F p-value
0.03	0.17	0.14	2	0.005	13.91	0.0056

6.2 Confusion matrix example

Examples of calculating mistake frequencies and letter biases from a confusion matrix are shown in Figure A1. Both biases and mistakes (after flattening) are normalized probability vectors.

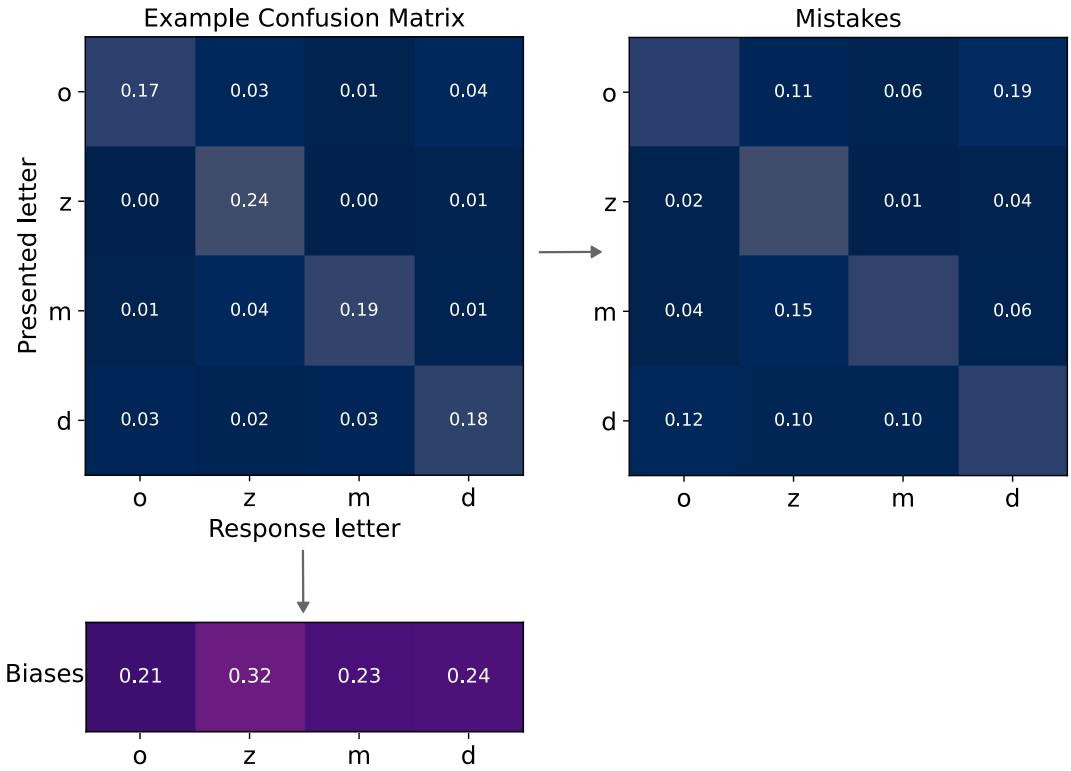

Figure A1: Obtaining letter biases and mistakes from a confusion matrix. The diagonal elements of a confusion matrix are the probabilities of correctly classifying each letter and the off-diagonal elements are the probabilities of incorrect classifications. Letter biases are obtained by simply summing each column. Mistakes are calculated by normalizing the off-diagonal elements.

6.3 Scatter plots of mistakes and biases

Scatter plots of the mistakes and letter biases are shown in Figure A2 and A3. In the Dom-Dom comparison, we know the underlying probability distributions are the same since we are drawing samples from the same dataset. The sizes of the clusters then give a baseline level of variance from the bootstrapping of our method. For the NonDom-Dom condition, we see humans exhibit consistent mistakes and biases regardless of which eye they use to identify letters (clusters lying on the unity line). Visually, we see all model observers diverge from humans to some degree. However some models perform better (or worse) than others. For example, BN CNN in CS provides a better match of human data than the other models. In scrambling noises, TM diverges quite significantly from human data, especially in letter biases. However, this divergence is less in the BN.

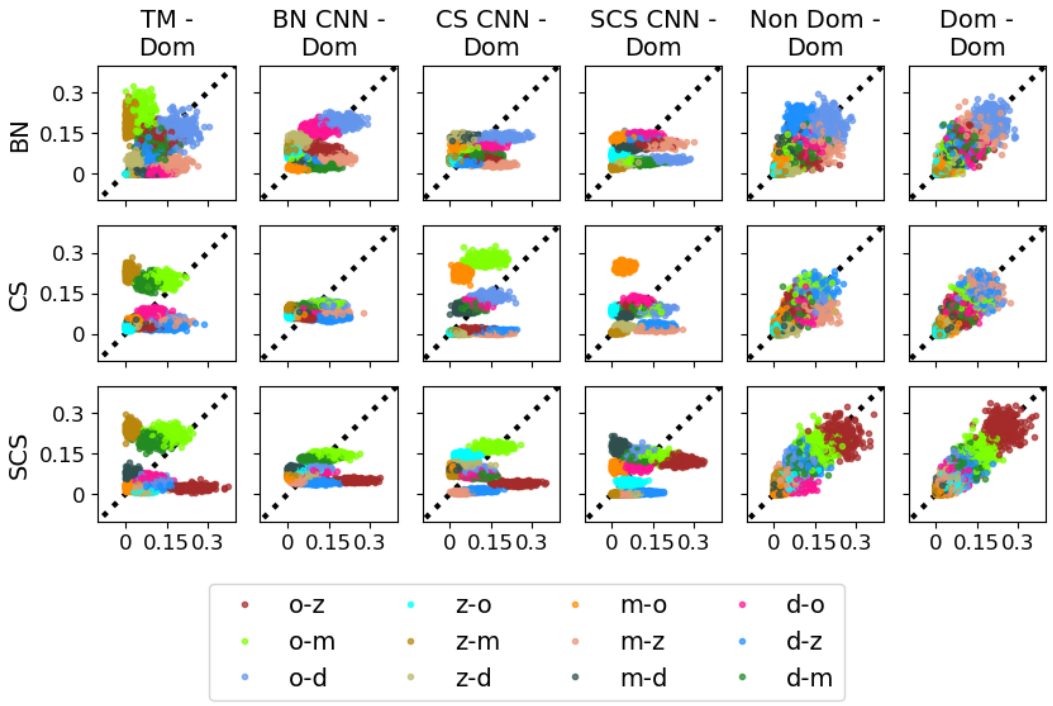

Figure A2: Scatter plots of bootstrap CMs. Each row represents one noise condition and each column represents one observer condition (computer models and humans). For each scatter plot, dominant eye viewing human data are plotted on the x-axis and observer condition data are plotted on the y-axis. The first and second letters of the letter pairs shown in the legend are the presented letter and the response letter respectively.

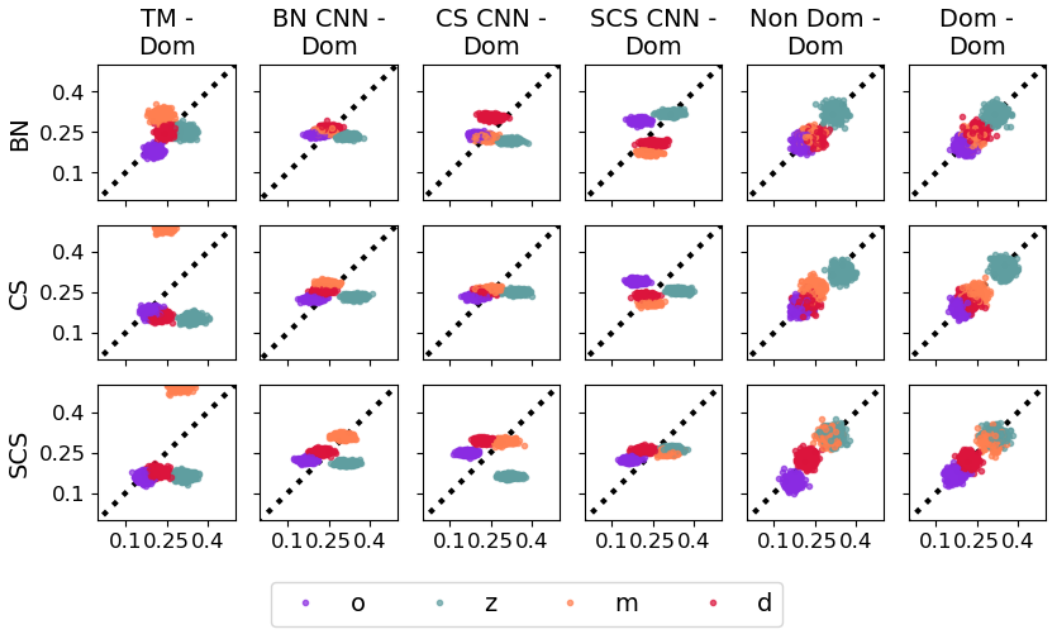

Figure A3: Scatter plots of bootstrap letter biases. Each row represents one noise condition and each column represents one observer condition (computer models and humans). For each scatter plot, dominant eye viewing human data are plotted on the x-axis and observer condition data are plotted on the y-axis.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems.

Alonso, J.-M., Usrey, W. M., and Reid, R. C. (2001). Rules of Connectivity between Geniculate Cells and Simple Cells in Cat Primary Visual Cortex. *Journal of Neuroscience*, 21(11):4002–4015.

Baker, D. H., Summers, R. J., Baldwin, A. S., and Meese, T. S. (2022). A psychophysical performance-based approach to the quality assessment of image processing algorithms. *PLOS ONE*, 17(5):e0267056.

Baldwin, A. S., Baker, D. H., and Hess, R. F. (2016). What Do Contrast Threshold Equivalent Noise Studies Actually Measure? Noise vs. Nonlinearity in Different Masking Paradigms. *PLOS ONE*, 11(3):e0150942.

Baldwin, A. S., Fu, M., Farivar, R., and Hess, R. F. (2017). The equivalent internal orientation and position noise for contour integration. *Scientific Reports*, 7(1):13048.

Barlow, H. B. (1956). Retinal Noise and Absolute Threshold. *JOSA*, 46(8):634–639.

Barrett, B. T., Pacey, I. E., Bradley, A., Thibos, L. N., and Morrill, P. (2003). Nonveridical visual perception in human amblyopia. *Investigative Ophthalmology & Visual Science*, 44(4):1555–1567.

Beckmann, P. J. and Legge, G. E. (2002). Preneural limitations on letter identification in central and peripheral vision. *JOSA A*, 19(12):2349–2362.

Berardino, A., Laparra, V., Ballé, J., and Simoncelli, E. (2017). Eigen-Distortions of Hierarchical Representations. In *Advances in Neural Information Processing Systems*, volume 30.

Borji, A. and Itti, L. (2014). Human vs. Computer in Scene and Object Recognition. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 113–120.

Chollet, F. (2021). *Deep Learning with Python, Second Edition*. Manning Publications.

Christensen, J. H., Bex, P. J., and Fiser, J. (2019). Coding of low-level position and orientation information in human naturalistic vision. *PLOS ONE*, 14(2):e0212141.

Chung, S. T. L., Legge, G. E., and Tjan, B. S. (2002). Spatial-frequency characteristics of letter identification in central and peripheral vision. *Vision Research*, 42(18):2137–2152.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *The Behavioral and Brain Sciences*, 36(3):181–204.

Courrieu, P., Farioli, F., and Grainger, J. (2004). Inverse discrimination time as a perceptual distance for alphabetic characters. *Visual Cognition*, 11(7):901–919.

Dakin, S. C. (2001). Information limit on the spatial integration of local orientation signals. *JOSA A*, 18(5):1016–1026.

Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *JOSA A*, 2(7):1160–1169.

Dodge, S. and Karam, L. (2017). A Study and Comparison of Human and Deep Learning Recognition Performance under Visual Distortions. In *2017 26th International Conference on Computer Communication and Networks (ICCCN)*, pages 1–7.

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

Farahat, A., Effenberger, F., and Vinck, M. (2023). A novel feature-scrambling approach reveals the capacity of convolutional neural networks to learn spatial relations. *Neural Networks*, 167:400–414.

Feather, J., Leclerc, G., Mądry, A., and McDermott, J. H. (2023). Model metamers reveal divergent invariances between biological and artificial neural networks. *Nature Neuroscience*, 26(11):2017–2034.

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, 4(12):2379.

Field, D. J., Hayes, A., and Hess, R. F. (1993). Contour integration by the human visual system: Evidence for a local "association field". *Vision Research*, 33(2):173–193.

Fischer, S., Šroubek, F., Perrinet, L., Redondo, R., and Cristóbal, G. (2007). Self-Invertible 2D Log-Gabor Wavelets. *International Journal of Computer Vision*, 75(2):231–246.

Freeman, J. and Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature Neuroscience*, 14(9):1195–1201.

Gattass, R., Gross, C. G., and Sandell, J. H. (1981). Visual topography of V2 in the macaque. *The Journal of Comparative Neurology*, 201(4):519–539.

Gattass, R., Sousa, A. P., and Gross, C. G. (1988). Visuotopic organization and extent of V3 and V4 of the macaque. *Journal of Neuroscience*, 8(6):1831–1845.

Geisler, W. (2004). Ideal observer analysis. In *Chalupa, L.M., Werner, J.S. (Eds.), The Visual Neurosciences*. The MIT Press, Cambridge, MA.

Georgeson, M. A., Barhoom, H., Joshi, M. R., Artes, P. H., and Schmidtmann, G. (2023). Revealing the influence of bias in a letter acuity identification task: A noisy template model. *Vision Research*, 208:108233.

Gold, J., Bennett, P. J., and Sekuler, A. B. (1999). Identification of band-pass filtered letters and faces by human and ideal observers. *Vision Research*, 39(21):3537–3560.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples.

Grainger, J., Rey, A., and Dufau, S. (2008). Letter perception: from pixels to pandemonium. *Trends in Cognitive Sciences*, 12(10):381–387.

Groen, I. I. A., Dekker, T. M., Knapen, T., and Silson, E. H. (2022). Visuospatial coding as ubiquitous scaffolding for human cognition. *Trends in Cognitive Sciences*, 26(1):81–96.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array Programming with NumPy. *Nature*, 585(7825):357–362.

Hess, R. F., Campbell, F. W., and Greenhalgh, T. (1978). On the nature of the neural abnormality in human amblyopia; neural aberrations and neural sensitivity loss. *Pflugers Archiv: European Journal of Physiology*, 377(3):201–207.

Hess, R. F. and Field, D. (1993). Is the increased spatial uncertainty in the normal periphery due to spatial undersampling or uncalibrated disarray? *Vision Research*, 33(18):2663–2670.

Hess, R. F. and Field, D. J. (1994). Is the spatial deficit in strabismic amblyopia due to loss of cells or an uncalibrated disarray of cells? *Vision Research*, 34(24):3397–3406.

Hess, R. F., Field, D. J., and Watt, R. J. (1990). The puzzle of amblyopia. In *Vision: coding and efficiency*. Cambridge University Press.

Hess, R. F., McIlhagga, W., and Field, D. J. (1997). Contour integration in strabismic amblyopia: the sufficiency of an explanation based on positional uncertainty. *Vision Research*, 37(22):3145–3161.

Hirsch, J. A., Alonso, J.-M., Reid, R. C., and Martinez, L. M. (1998). Synaptic Integration in Striate Cortical Simple Cells. *Journal of Neuroscience*, 18(22):9517–9528.

Hubel, D. H. and 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.

Janini, D., Hamblin, C., Deza, A., and Konkle, T. (2022). General object-based features account for letter perception. *PLOS Computational Biology*, 18(9):e1010522.

Jones, J. P. and Palmer, L. A. (1987). An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58(6):1233–1258.

Kaas, J. H. (1997). Topographic Maps are Fundamental to Sensory Processing. *Brain Research Bulletin*, 44(2):107–112.

Kanwisher, N., Khosla, M., and Dobs, K. (2023). Using artificial neural networks to ask 'why' questions of minds and brains. *Trends in Neurosciences*, 46(3):240–254.

Kelly, D. (1962). Information capacity of a single retinal channel. *IRE Transactions on Information Theory*, 8(3):221–226.

Kersten, D., Hess, R. F., and Plant, G. T. (1988). Assessing contrast sensitivity behind cloudy media. *Clinical Vision Science*, 2(3):143–158.

Khaligh-Razavi, S.-M. and Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLOS Computational Biology*, 10(11):e1003915.

Kingdom, F. A. and Prins, N. (2016). *Psychophysics: A Practical Introduction*. Elsevier/Academic Press, Amsterdam, 2 edition.

Kingma, D. P. and Ba, J. (2014). Adam: A Method for Stochastic Optimization.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25.

Levi, D. M., Klein, S. A., and Sharma, V. (1999). Position jitter and undersampling in pattern perception. *Vision Research*, 39(3):445–465.

Levi, D. M., Klein, S. A., Sharma, V., and Nguyen, L. (2000). Detecting disorder in spatial vision. *Vision Research*, 40(17):2307–2327.

Li, Z. (2002). A saliency map in primary visual cortex. *Trends in Cognitive Sciences*, 6(1):9–16.

Li, Z. (2014). *Understanding Vision: Theory, Models, and Data*. Oxford University Press. Google-Books-ID: 9DlsAwAAQBAJ.

Lu, Z.-L. and Dosher, B. A. (1999). Characterizing human perceptual inefficiencies with equivalent internal noise. *JOSA A*, 16(3):764–778.

Majaj, N. J. and Pelli, D. G. (2018). Deep learning—Using machine learning to study biological vision. *Journal of Vision*, 18(13):2.

Majaj, N. J., Pelli, D. G., Kurshan, P., and Palomares, M. (2002). The role of spatial frequency channels in letter identification. *Vision Research*, 42(9):1165–1184.

McAnany, J. J. and Alexander, K. R. (2010). Spatial contrast sensitivity in dynamic and static additive luminance noise. *Vision Research*, 50(19):1957–1965.

McAnany, J. J., Alexander, K. R., Genead, M. A., and Fishman, G. A. (2013). Equivalent Intrinsic Noise, Sampling Efficiency, and Contrast Sensitivity in Patients With Retinitis Pigmentosa. *Investigative Ophthalmology & Visual Science*, 54(6):3857–3862.

Meese, T. S. (2010). Spatially extensive summation of contrast energy is revealed by contrast detection of micro-pattern textures. *Journal of Vision*, 10(8):14.

Nguyen, A., Yosinski, J., and Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 427–436.

Nirenberg, S., Carcieri, S. M., Jacobs, A. L., and Latham, P. E. (2001). Retinal ganglion cells act largely as independent encoders. *Nature*, 411(6838):698–701.

Nishimoto, S., Ishida, T., and Ohzawa, I. (2006). Receptive Field Properties of Neurons in the Early Visual Cortex Revealed by Local Spectral Reverse Correlation. *Journal of Neuroscience*, 26(12):3269–3280.

Parish, D. H. and Sperling, G. (1991). Object spatial frequencies, retinal spatial frequencies, noise, and the efficiency of letter discrimination. *Vision Research*, 31(7):1399–1415.

Pavan, A., Contillo, A., Yilmaz, S. K., Kafaligonul, H., Donato, R., and O'Hare, L. (2023). A comparison of equivalent noise methods in investigating local and global form and motion integration. *Attention, Perception, & Psychophysics*, 85(1):152–165.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Pelli, D. G. (1981). *Effects of Visual Noise (Doctoral Dissertation)*. PhD thesis, University of Cambridge.

Pelli, D. G. and Farell, B. (1999). Why use noise? *JOSA A*, 16(3):647–653.

Pelli, D. G., Levi, D. M., and Chung, S. T. L. (2004). Using visual noise to characterize amblyopic letter identification. *Journal of Vision*, 4(10):904–920.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2020). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, 128(2):336–359.

Sireteanu, R., Lagreze, W.-D., and Constantinescu, D. H. (1993). Distortions in two-dimensional visual space perception in strabismic observers. *Vision Research*, 33(5):677–690.

Solomon, J. A. and Pelli, D. G. (1994). The visual filter mediating letter identification. *Nature*, 369(6479):395–397.

Tao, X., Zhang, B., Shen, G., Wensveen, J., Smith, E. L., Nishimoto, S., Ohzawa, I., and Chino, Y. M. (2014). Early Monocular Defocus Disrupts the Normal Development of Receptive-Field Structure in V2 Neurons of Macaque Monkeys. *Journal of Neuroscience*, 34(41):13840–13854.

Tao, X., Zhang, B., Smith, E. L., Nishimoto, S., Ohzawa, I., and Chino, Y. M. (2012). Local sensitivity to stimulus orientation and spatial frequency within the receptive fields of neurons in visual area 2 of macaque monkeys. *Journal of Neurophysiology*, 107(4):1094–1110.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, , Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., and van Mulbregt, P. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272.

Wardle, S. G., Bex, P. J., Cass, J., and Alais, D. (2012). Stereoacuity in the periphery is limited by internal noise. *Journal of Vision*, 12(6):12.

Watson, A. B. and Ahumada, A. J. (2012). Modeling acuity for optotypes varying in complexity. *Journal of Vision*, 12(10):19–19.

Watson, A. B. and Ahumada, A. J. (2015). Letter identification and the Neural Image Classifier. *Journal of Vision*, 15(2):15.

Watson, A. B. and Ahumada, Jr, A. J. (2008). Predicting visual acuity from wavefront aberrations. *Journal of Vision*, 8(4):17.

Watt, R. J. and Hess, R. F. (1987). Spatial information and uncertainty in anisometropic amblyopia. *Vision Research*, 27(4):661–674.

Weinberg, R. J. (1997). Are Topographic Maps Fundamental to Sensory Processing? *Brain Research Bulletin*, 44(2):113–116.

Würtz, R. (1995). Multilayer dynamic link networks for establishing image point correspondences and visual object recognition.

Yamins, D. L., Hong, H., Cadieu, C., and DiCarlo, J. J. (2013). Hierarchical Modular Optimization of Convolutional Networks Achieves Representations Similar to Macaque IT and Human Ventral Stream. In *Advances in Neural Information Processing Systems*, volume 26.

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

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624.

Yurdakul, B. and Naranjo, J. (2019). Statistical Properties of the Population Stability Index. *Journal of Risk Model Validation*, 14(4).

Zhang, B., Tao, X., Shen, G., Smith, E. L., Ohzawa, I., and Chino, Y. M. (2013). Receptive-Field Subfields of V2 Neurons in Macaque Monkeys Are Adult-Like Near Birth. *Journal of Neuroscience*, 33(6):2639–2649.

Ziskind, A., Hénaff, O., LeCun, Y., and Pelli, D. (2014). The bottleneck in human letter recognition: A computational model. *Journal of Vision*, 14(10):1311.