# IDENTIFICATION OF LETTERS DISTORTED BY PHYSIOLOGICALLY-INSPIRED SPATIAL SCRAMBLING

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

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

March 13, 2026

## ABSTRACT

In the geniculostriate pathway of the human visual system, neuronal projections carry signals from a particular retinal locus in parallel from one anatomical area to the next. Imprecision in the fidelity of these projections would place constraints on the ability of the system to perform tasks requiring positional information. We investigated the impact that "spatial scrambling" between stages would have on visual performance. We consider two stages in a simple canonical model of the early visual cortex where scrambling might occur: either the input to the first orientation-tuned mechanisms (analogous to V1 simple cells), or the output from those mechanisms. These are referred as "subcortical" (SCS) and "cortical scrambling" (CS). We developed a wavelet decomposition and resynthesis algorithm to mimic these effects, and measured human performance in letter identification affected by the two types of scrambling. Our results showed SCS and CS have distinguishable effects on both perceived noisiness of letters and letter identification threshold. Comparing human performance against a suite of pre-trained and custom convolutional neural networks (CNNs) that were trained on the scrambled stimuli, relative efficiency (calculated from the ratio of human:CNN thresholds) is higher for CS than SCS. However, in modelling human inefficiency by reducing the proportion of wavelets available to the CNNs, humans are *less* efficient in CS than SCS. These differences in efficiencies show humans are better at processing orientation redundant stimuli (CS) than orientation noisy stimuli (SCS). We hypothesize this reflects differences in integration properties at the input and output stages of simple cells in the cortex.

***Keywords*** human vision · spatial vision · letter identification · efficiency · ideal observer · distortion · internal noise · positional uncertainty · scrambling · convolutional neural network · confusion matrix

*corresponding author

## Author Summary

The brain makes sense of the input from our eyes through a system where features are extracted and combined in successive stages. Our study concerns the spatial fidelity of the connections between visual areas. Previous behavioural and physiological evidence has suggested a scrambling of neuronal projections is present in biological visual systems. In our study, we investigate the ability of the human visual system to perform letter identification with stimuli affected by different types of on-screen distortions. These distortions simulate internal scrambling occurring at two early stages in the visual hierarchy. We used convolutional neural network (CNN) models as a benchmark, against which we compared human performance to find human efficiency in handling the distortions. We found that the type of scrambling in which humans were determined to have greater "efficiency" (relative to the CNNs) depended on the analysis used. The threshold magnitude of scrambling at which the letters could no longer be identified was greater for letters scrambled *after* the oriented features were extracted. Conversely, when efficiency was calculated by starving the CNNs of samples until their performance declined to the human level we instead found that the effective "number of samples" used by our humans was much higher for stimuli simulating scrambling *before* the oriented feature stage. These differences reflect how information is pooled and combined for these two stages.


## 1 Introduction

The rich external world we perceive through vision is the output of a hierarchical system, where representations are built up through multiple stages of visual processing (Hubel and Wiesel, 1962; DiCarlo et al., 2012; Riesenhuber and Poggio, 1999). Physiology puts limits on both the quantity and quality of information that gets transmitted at 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). Similar bottlenecks also exist in cortical processing, starting as early as V1. These are determined by the number of neurons, their energy consumption, and noise in the system (Li, 2014). A dominant topic within vision science has been the question of *how* the brain decodes a noisy 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; Rao and Ballard, 1999) to efficiently encode information (Barlow, 1956).

In addition to being hierarchical, the visual system is also topographical with visual maps corresponding to the outside world (Luo and Flanagan, 2007; Chklovskii and Koulakov, 2004; Gattass et al., 1981, 1988; Dumoulin and Wandell, 2008; Weinberg, 1997; Kaas, 1997; Groen et al., 2022). Deviations in the spatial organisation of projections between regions can result in a loss of spatial precision in the encoding of a feature. For example, in higher cortical areas, encoded representations of more complex stimulus features may be *less* specific in the positions of the features they encode (Gattass et al., 1981, 1988; Dumoulin and Wandell, 2008; Weinberg, 1997; Kaas, 1997; Groen et al., 2022).

Physiological studies have shown scatter in neuronal projections (Nishimoto et al., 2006; Tao et al., 2012; Zhang et al., 2013; Ringach, 2004). In animal models of amblyopia, this scatter is worse leading to more heterogeneous receptive fields (Tao et al., 2014; Swindale and Mitchell, 1994; Bi et al., 2011). Studies using fMRI found enlarged

population receptive fields in amblyopic humans (Clavagnier et al., 2015; Hussain et al., 2015; Szinte et al., 2024; Li et al., 2007). A long history of behavioral studies have investigated the perceptual consequences of proposed intrinsic "scrambling" and "positional uncertainty" in visual tasks. These tasks investigated the effect of position noise on detection/localization (Michel and Geisler, 2011; Hess and Field, 1993, 1994; Morgan et al., 2012), bisection (Levi et al., 1987; Levi and Klein, 1986), shape/form discrimination (Watt and Hess, 1987; Keeble and Hess, 1999; Levi et al., 2000, 1999, 1997), contour integration (Hess et al., 1997; Baldwin et al., 2017), and discrimination of natural images (Christensen et al., 2019). In amblyopia, it has been shown that the amblyopic eye's percept is affected by a subjective "distortion" (Hess et al., 1978, 1990; Sireteanu et al., 1993; Barrett et al., 2003; Maruya et al., 2025). In normal vision, metameric stimuli have been developed that are indistinguishable when shown in the periphery. This phenomenon is thought to reflect a loss of spatial precision in the summary statistic encoding of features at greater eccentricities (Freeman and Simoncelli, 2011; Broderick et al., 2023; Balas et al., 2009).

One approach for characterising internal scrambling or positional uncertainty of a system is the equivalent noise paradigm (Pelli and Farell, 1999; Pelli, 1981). These experiments are most typically conducted in the contrast domain by introducing task-relevant "external" noise to a stimulus and measuring thresholds in a range of different external noise levels: a "noise-masking function". Models fit to these data quantify the external noise magnitude required to exceed the limiting effect of the system's "internal noise" on performance (Pelli and Farell, 1999; Pelli, 1981). Separate from internal noise, efficiency (as a ratio between 0% to 100%) measures the gap in performance between humans and the theoretical ideal observer (a benchmark for the best performance that can be achieved on a particular task; Geisler, 2004). Comparing measurements of performance in high levels of external noise between humans and ideal observer also allows for direct measurement of efficiency (Pelli et al., 2006). These methods have revealed insights into how humans link contours (Hess and Demanins, 1998; Hess et al., 1997; Baldwin et al., 2017), integrate motion direction (Dakin et al., 2005; Pavan et al., 2023), and extracting orientation (Dakin, 2001).

In the current study, we want to investigate how internal scrambling may limit human efficiency in letter identification. For the loci of internal scrambling, we chose to focus on the geniculostriate pathway as it is arguably the most understood and the route through which visual information is transmitted from the retina to primary visual cortex (V1) via the lateral geniculate nucleus (LGN) (Hubel and Wiesel, 1962; Alonso et al., 2001). **Fig 1A** shows the classic model of oriented simple cell receptive field formation, pooling from subunits representing isotropically-tuned LGN afferents (Hubel and Wiesel, 1962; Hirsch et al., 1998). Here, we investigated the extent to which the simulated effects of scrambling could be differentiated when applied at two stages in this simple feed-forward hierarchical model. We developed an algorithm that decomposes a stimulus into a set of wavelets. To mimic internal scrambling, the stimulus can then be reconstructed either veridically, or with scrambling that simulates either: i) random jitter of the spatial location of oriented receptive fields (analogous to simple cells in V1), which we label "cortical scrambling" or CS (**Fig 1B**), or ii) random jitter of the location of the isotropic subunits which form those oriented receptive fields (analogous to LGN afferents) which we label "subcortical scrambling" or SCS (**Fig 1C**). Finally, we also investigated performance for non-scrambled stimuli presented in bandpass noise (BN). This allows us to connect our findings to more traditional contrast noise-masking

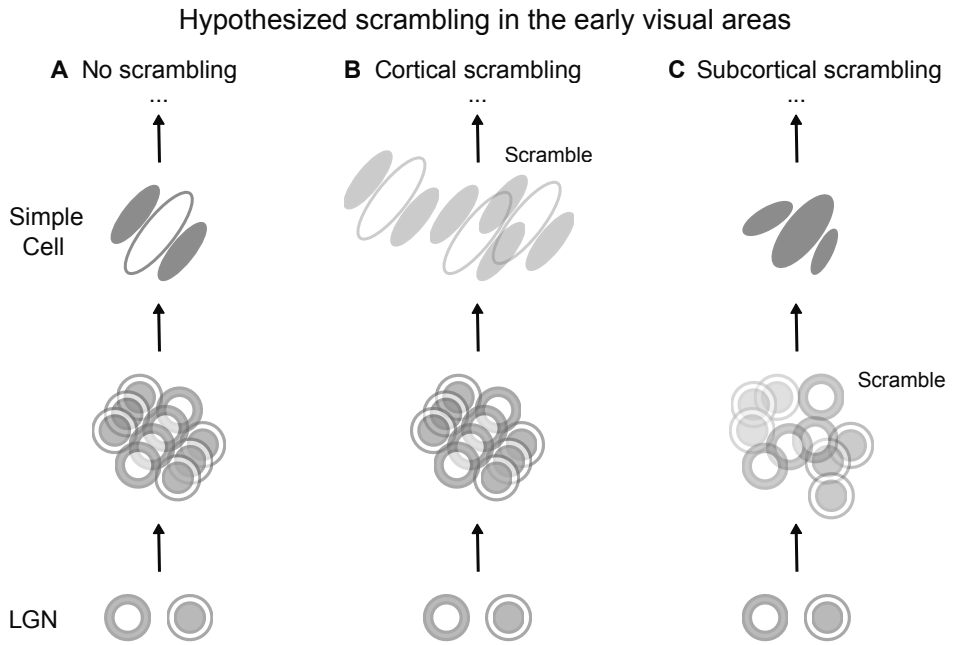


**Fig 1.** Hypothesized cortical vs. subcortical scrambling: Panel (**A**) shows a canonical view of V1 simple cell receptive field formation, being composed of inputs from the lateral geniculate nucleus (LGN) subunits. Panel (**B**) demonstrates the impact of scrambling simple cell projections to other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

approaches (Majaj et al., 2002; Pelli et al., 2004, 2006). **Fig 2** shows our BN, CS, and SCS stimuli for a range of noise and scrambling levels.

Letter identification has clinical relevance in the measurement of letter acuity (Chung et al., 2002b; Pelli et al., 2004, 2006; Chung et al., 2002a), a long-standing importance to the field of computer vision in Optical Character Recognition (Zhang et al., 2015; Bottou et al., 1994), and plays an important role in advancing theories and models of invariant object recognition (Grainger et al., 2008). Conveniently human psychophysics research has identified the optimal spatial frequency band (around 3 cycles/letter) that supports letter identification (Majaj et al., 2002; Chung et al., 2002b; Solomon and Pelli, 1994). This helped simplify our stimulus generation through presenting spatially-bandpass letters at that optimal spatial frequency, which should limit the contribution of complex cross-frequency gain control effects (Heeger, 1992; Bonds, 1989; Brouwer and Heeger, 2011; Eckstein et al., 1997; Foley, 1994; Meese and Holmes, 2006).

As letter identification is a classification task, this approach also facilitated the use of CNN models in our study. These are emerging as promising models against which human object recognition can be compared (Majaj and Pelli, 2018; Lindsay, 2021; Wichmann and Geirhos, 2023; Yamins and DiCarlo, 2016). It is difficult to develop an analytical ideal observer to predict the impact our CS and SCS manipulations would have on letter identification. Without this ideal observer, we also cannot present the efficiency for CS and SCS letter identification on an absolute scale. Instead, we chose to compare human performance against that of a range of CNN models (which, after Kanwisher et al., 2023, may be "optimised" but likely not "optimal"). To have networks optimised specifically for the task of interest, we trained a set of twenty networks from scratch following an architecture search (Francl and McDermott, 2022). We also

investigated the behaviour of a varied set of networks pre-trained on ImageNet: AlexNet (standard feedforward CNN), ResNet50 (with skip connections), VGG19 (stacking small filters only), and CORnetS (with recurrent connections to mimic human vision) (Krizhevsky et al., 2012; Kubilius et al., 2019; He et al., 2015; Simonyan and Zisserman, 2015). Benchmarking human performance against that of CNNs allowed us to calculate relative efficiency for both CS and SCS on an objective scale.

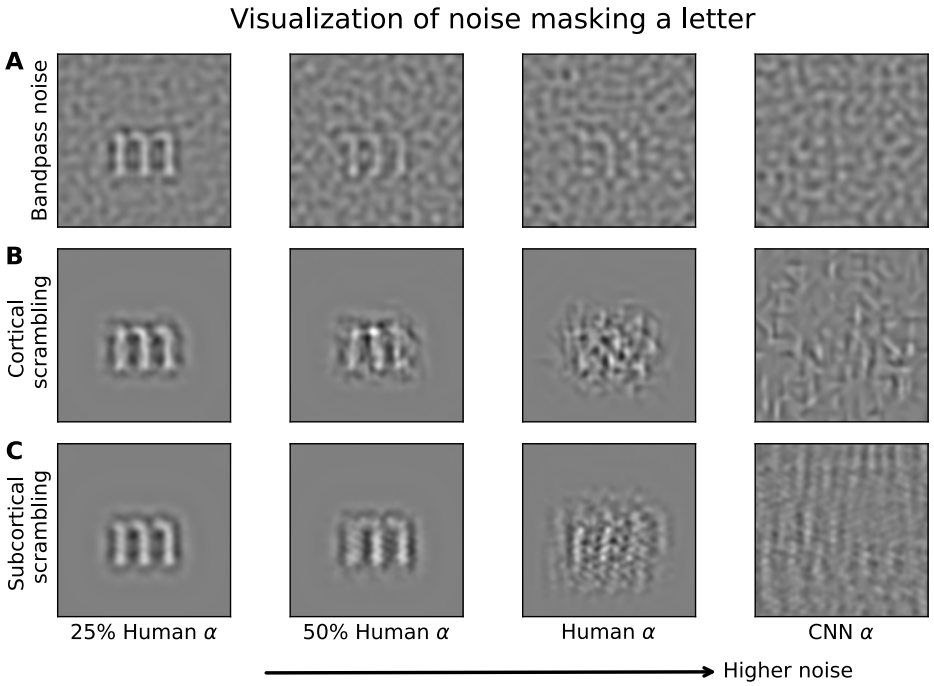


**Fig 2.** Visualizing the effect of masking noise used in this study. Spatial frequency of the letter was set to the optimal object spatial frequency for letter identification at 3 cycles/letter. The example uses a bandpass letter "m", showing the effect of increasing noise magnitude from left to right for our three conditions: (**A**) bandpass noise (BN), (**B**) cortical scrambling (CS), and (**C**) subcortical scrambling (SCS). For illustration, noise levels are shown at fractions of human threshold ($\alpha$), representing 62% correct in the letter identification task (**Fig 6**) for humans. Stimuli at $\alpha$ of convolutional neural networks (CNNs) (**Fig 5**) are also shown for reference. For BN (**A**), noise magnitude is quantified in the relative RMS contrast of the noise compared to the target. For CS (**B**) and SCS (**C**), noise magnitude is quantified as the standard deviation of the Gaussian from which position offsets of the wavelets or subunits are drawn.

To preview our results, humans exhibited higher relative efficiency for cortical scrambling (CS) than subcortical scrambling (SCS) (**Fig 6D-F**). Defining sampling efficiency as the proportion of wavelets made available to the CNNs to match human performance, we instead find more samples were needed in SCS than CS (**Fig 6G-I**). These results suggest difference in processing constraints between input and output stages of simple cells in the cortex.

## 2 Results

### 2.1 Experiment 1: Matching perceived scrambling magnitude

We performed a matching experiment to compare the subjective magnitude of perceived "scrambling" that resulted from our Cortical Scrambling (CS) and Subcortical Scrambling (SCS) algorithms. Although both kinds of scrambling were

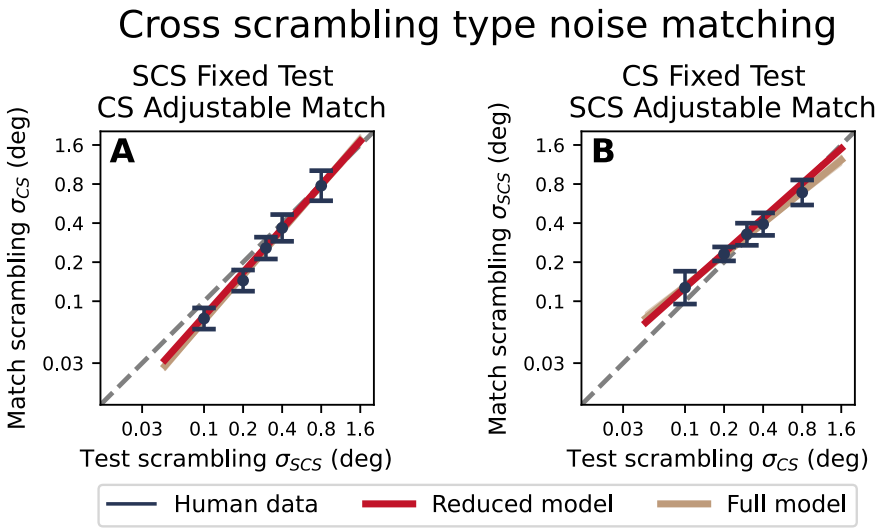


**Fig 3.** Matching cortical scrambling and subcortical scrambling (and vice versa). (**A**) shows results of matching a cortical scrambled (CS) letter (the match) to a subcortical scrambled (SCS) letter (the test). (**B**) shows the opposite match. The dashed grey line shows the unity line. The best-fit line to the data, fit separately for panels (**A**) and (**B**) is shown in gold. The best-fit line using the reduced model (common relation with a single set of slope and intercept) is shown in red. Markers show mean with standard deviation of the human data pooled across five participants.

quantified by the standard deviation ($\sigma$) of the Gaussian process (in degrees of visual angle or "deg") used to perturb the locations of the features, we nevertheless wished to determine whether one type might result in a greater subjective percept of "scrambling". At the same time, we also sought to determine whether either type of scrambling might saturate, such that increases in standard deviation would have a diminishing effect on perceived scrambling beyond some critical magnitude. To answer these questions, we performed an experiment where participants matched the perceived scrambling of two letters affected by either CS (with scrambling magnitude $\sigma_{\text{CS}}$) or SCS (with scrambling magnitude $\sigma_{\text{SCS}}$).

We measured matches both between the same type of scrambling (matching CS to CS, or SCS to SCS) and between the two different types of scrambling (see **Methods** Section 4.7). In all comparisons, we found the matching relationship to be well-characterized by a simple linear fit on double-log axes (**Table 1**). When matching the scrambling magnitude of two letters affected by the same type of scrambling, a veridical match would have a slope of unity and an intercept of zero ($y = x$). We found participants were able to make matches close to this veridical relationship for both CS and SCS. This validates their ability to make accurate matches of perceived scrambling magnitude.

For matches between the two types of scrambling, the relationships were still linear (on our log-log axes, see Figure 3) however the slope and intercept values varied. We performed the match in both directions, adjusting scrambling magnitude of a CS letter to match a SCS letter (**Fig 3A**) and the other way around (**Fig 3B**). We found more SCS was needed to match the perceived noisiness of CS at lower scrambling levels. However, the perceived noisiness of the SCS increased more rapidly with the nominal noise level so that at the higher magnitudes tested the perceived "noisiness" of the two manipulations became similar for stimuli generated with the same $\sigma$.

**Table 1.** Details of linear fits to scrambling matching experiment data on double-log axes. Cross-scrambling-type conditions are plotted in **Fig 3**. Reduced refers to fitting one linear model to matching data in both directions.

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

We would expect that the equivalent SCS magnitude to match an CS stimulus could be predicted by the inverse of the rule for matching a CS stimulus to an SCS stimulus. We performed an analysis to determine whether the same rule could account for matches in both directions (CS to SCS and vice-versa) in a "reduced" model or if the two directions were better accounted for by two different linear relationships (the "full" model). A nested model hypothesis test found the full model to be superior ($p < 0.001$, nested model hypothesis test; details in **Table 1** of Supporting Information). Inspecting **Fig 3**, we see that the full model differs from the reduced model by better accounting for the condition where an SCS stimulus was matched to a reference CS stimulus of the highest scrambling magnitude tested. It is worth noting that the same stimulus magnitude was matched in the opposite direction with $\sigma_{\text{CS}} \approx \sigma_{\text{SCS}}$ so we conclude this deviation is likely an artefact of our testing procedure (for example, it may be that the linear matching relationship *does* break down for magnitudes beyond our chosen range). On that basis, and justified by the overall strong performance performance of our reduced model in accounting for all other data points, we propose that the two types of scrambling (within the range used in this study) follows the relationship:

$$\log_2\left(\sigma_{\text{SCS}}\right) = 0.89\log_2\left(\sigma_{\text{CS}}\right) - 0.02, \tag{1}$$

this is approximated by a simple power law where

$$\sigma_{\text{CS}} = \sigma_{\text{SCS}}^{0.9}. \tag{2}$$

### 2.2 Experiment 2: Identification of letters affected by noise or scrambling

In our main experiment, we investigated how the two types of scrambling (CS and SCS) affected performance on a letter identification task, and how this compared to the identification of letters in bandpass noise (BN). We measured the

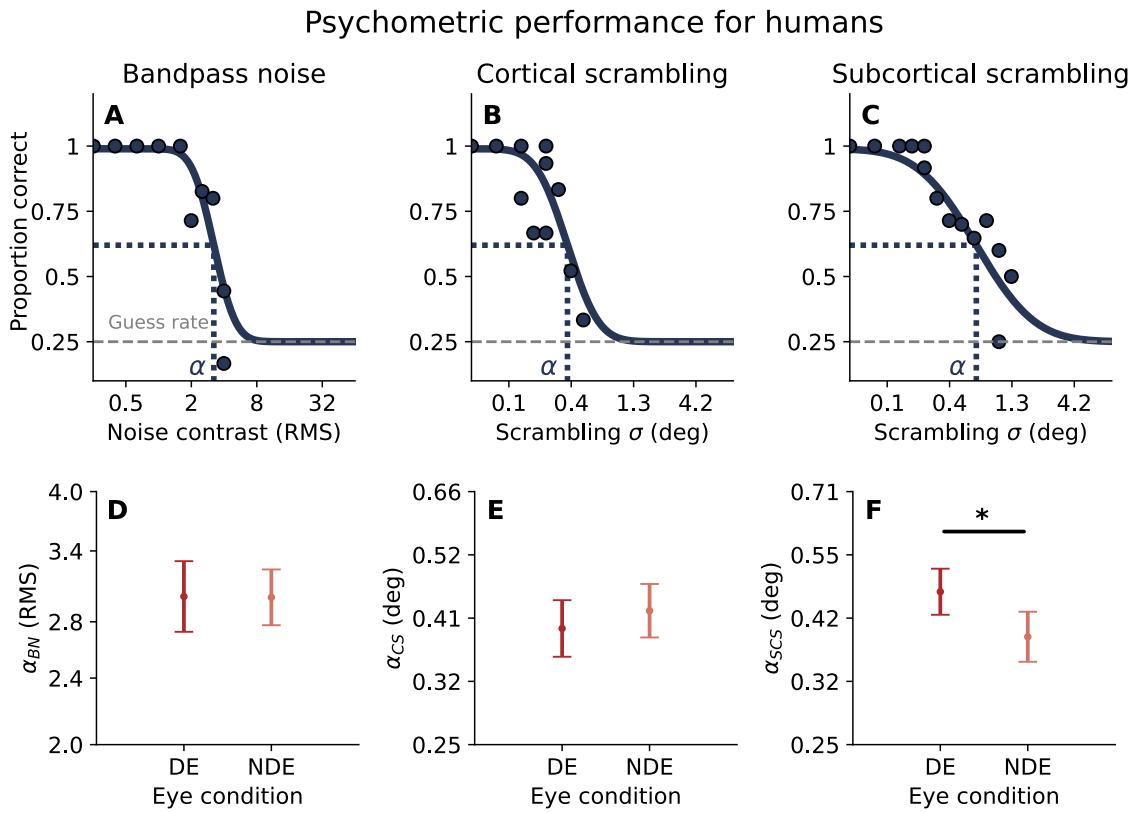


**Fig 4.** The top row (**A**-**C**) show example data from one human participant with the psychometric function fit by our maximum-likelihood method. The bottom row (**D**-**F**) shows mean noise / scrambling thresholds across 20 human subjects (with error bars showing 95% CI).

magnitude of noise (BN) or scrambling (CS or SCS) that would reduce letter identification performance to our threshold level (62% correct). Example psychometric functions from a single participant are shown in **Fig 4A-C**, with the average thresholds calculated over 20 participants shown in **Fig 4D-F**. Separate thresholds were calculated for participants using their dominant (D) and non-dominant (ND) eyes. We found participants were significantly more resilient to SCS (higher threshold) using their D eye (t-test; $t_{19}$=2.18, p=0.042). We did not find the same dominant eye advantage for BN ($t_{19}$=-1.06, p=0.30) or CS ($t_{19}$=0.05, p=0.96). These results are intriguing, as our SCS manipulation would correspond to type of scrambling we would expect to occur *before* binocular combination. To simplify the remainder of our analyses we will combine data from the two eyes of each participant by taking their average, though we will return to this point in the **Discussion**.

### 2.3 Performance of CNN models

The thresholds measured in our three conditions (BN, CS, and SCS) are not directly comparable, as the parameters controlling the magnitude of noise or scrambling are different in each condition. By finding the performance of CNN models optimised to perform the identification task under each stimulus condition, we sought to use them as a benchmark against which human performance could be compared. The relative efficiency, defined as the ratio of

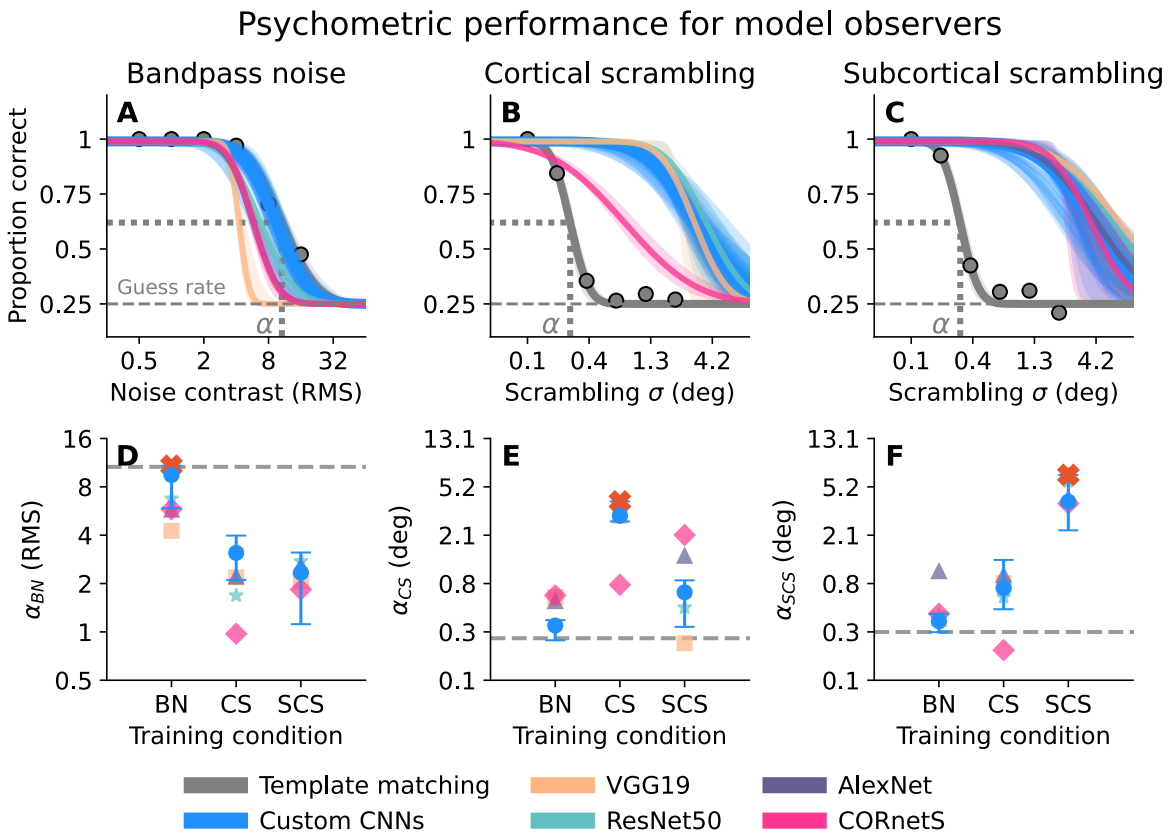


**Fig 5.** Performance of models in the letter identification task. The top row shows the psychometric functions for models performing the task in the bandpass noise (BN) (**A**), cortical scrambling (CS) (**B**), and subcortical scrambling (SCS) (**C**) conditions. Twenty CNNs had custom architectures randomly generated from an architecture search. Transfer learning was done on four pre-trained networks: VGG19, AlexNet, ResNet50, CORnetS. Shaded region shows 95% CI obtained from 200 bootstrap fits. The bottom row (**D-F**) shows threshold $\alpha$ for all CNN models. Threshold $\alpha$ is defined as the point on the psychometric function that corresponds to 62% correct (shown for TM model in the top row). Dotted grey line shows the $\alpha$ of TM model. Blue marker with error bars denotes median with 95% CI for the custom trained CNNs. Orange cross shows the $\alpha$ of the best-performing CNN of all tested models. For each noise condition, it was a custom trained CNN. AlexNet trained and tested on CS is removed due to large fitting error of the psychometric function.

human performance to the performance of a model observer (Tanner and Birdsall, 1958; Pelli et al., 2006), allows for an objective comparison between different stimulus conditions. The models used in these comparisons are often "ideal observers" that make optimal use of all available information to perform the task (Geisler, 2004, 1989). For the two scrambling conditions tested in this study we did not have ideal observers, and so instead use our CNN models as a reference (while bearing in mind they likely fall short of the theoretical "best possible" performance).

We looked at the performance of both custom CNNs and pre-trained networks popular in the literature (that underwent transfer learning on our task). For our custom CNNs, to generate an optimised set for each noise condition we first performed an architecture search by generating 100 architectures with random hyper-parameters and picked the top twenty best-performing architectures (histograms of hyperparameters shown in **Fig 4** of Supporting Information). These twenty architectures were then trained from scratch until convergence. For the pre-trained networks, we chose VGG19, AlexNet, ResNet, CORnetS pre-trained on ImageNet. In our set of models, we also include a template

matching (TM) model. This is the ideal observer model for identifying letters in white noise Van Trees (1968); Pelli et al. (2006); Ziskind et al. (2014). We therefore expect it to be optimal for the BN condition (and so would expect CNNs trained to perform the BN task to reach its performance level).

The psychometric functions for our computer models are shown in **Fig 5A-C**. The psychometric threshold $\alpha$ for various models are shown in **Fig 5D-F**. In each column, we see the best-performing custom CNNs for each stimulus condition were those that were *trained* on that same condition, i.e. the best-performing custom CNNs for the BN condition were those that were trained to identify BN stimuli. This is evidence that identification of BN stimuli is best supported by a distinct strategy. This is unsurprising, as the BN task (identification of a letter in-place without distortion) differs fundamentally from that of either of our scrambling conditions. We learn more from the comparison between the CS and SCS conditions. Although the generation of these stimuli involves distinct operations, it would nevertheless be possible that a similar strategy would be optimal to identify scrambled letters under either manipulation. Such an equivalence, had we found it, would limit our expectations for any ability to distinguish the two types of scrambling through behavioural measurements. Instead, we find CS-trained networks outperform SCS-trained networks when the task is to identify CS stimuli, and vice-versa. These differences were tested using a Mann-Whitney U test ($p < 0.0001$ for pairwise comparisons among BN CNN, CS CNN, and SCS CNN for all three noise conditions). These results echo previous works that have shown CNNs do not generalize well to out-of-distribution samples (Geirhos et al., 2018).

For the BN condition, we found custom CNN thresholds are not significantly different from TM threshold (Mann-Whitney U test, U=19, $p = 0.19$). This suggests our architecture search and training procedure produced CNNs that approximated the ideal observer performance for the BN stimulus condition. For the CS and SCS conditions we cannot know how close our custom CNNs are to the theoretical performance limit. In comparisons against previously-published CNNs, we do find that the performance of our custom CNNs reaches or exceeds that of the pre-trained models.

### 2.4 Comparing human performance against that of CNN models

In **Fig 6D-F**, we show the relative efficiency $\eta$ calculated between the human data and the different CNN models. A higher relative efficiency indicates human performance is closer to the performance of the CNNs. The dashed black line shows the mean efficiency across models. For BN, the efficiency was 51%. This means that the human thresholds were approximately half that of the CNN models. For CS stimuli, the efficiency was 13% (AlexNet and CORnetS models excluded as outliers) and for SCS it was 9%. Marginalizing over models, a one-way repeated measures ANOVA was conducted to examine the effect of noise condition on relative efficiency in eighteen subjects. The analysis revealed a statistically significant main effect of noise condition ($F_{(1.61,27.32)}$= 879.68, $p < 0.001$, Greenhouse-Geisser corrected). Post-hoc pairwise comparisons revealed a significant difference between BN and CS ($t_{17}$ = 27.35, $p < 0.001$), BN and SCS ($t_{17}$ = 54.95, $p < 0.001$), and CS and SCS ($t_{17}$ = 7.45, $p < 0.001$). Compared to the best model (dotted orange line), efficiency was 29% for BN, 10% for CS, and 7% for SCS. One-way repeated measures ANOVA also revealed a statistically significant main effect of noise condition ($F_{(1.61,27.32)}$= 612.94, $p < 0.001$, Greenhouse-Geisser corrected). Post-hoc pairwise comparisons also revealed a significant difference between BN and CS ($t_{17}$ = 20.15, $p < 0.001$), BN

and SCS ($t_{17}$ = 47.50, p < 0.001), and CS and SCS ($t_{17}$ = 10.23, p < 0.001). These results show that the efficiency with which humans identify scrambled letters, when expressed as the magnitude of scrambling required to degrade performance to the threshold level, was significantly higher in CS than in SCS.

### 2.5 Modelling inefficiency as a loss of samples

The human participants in our study never out-performed our model observers ($\eta$ was always $<$100%), demonstrating our CNNs made better use of the information available in the stimulus to perform the task. One possibility is that the humans use an inferior algorithm that leads to information loss. Previous studies have modelled this as the effective number of samples that are being used if the strategy applied to them is optimal (Dakin, 2001; Pelli and Farell, 1999; Dakin et al., 2005). Our wavelet-based stimulus generation algorithm allows us to explore this, with the numerical findings of such an analysis giving insight into the differences between human and CNN performance.

We simulated a loss in the sampling of the wavelets forming our letter stimuli by generating variants that were composed of a reduced number of wavelets (randomly sampled). These were then fed into our CNN models, allowing us to find the degree of under-sampling where CNN performance fell to the human level. This gave an indirect measure of the proportion of stimulus information humans effectively used. The modelling results are shown in **Fig 6A-C**. For all models and noise conditions except for one (VGG19 in BN), the CNN $\alpha$ monotonically decreased below $\alpha_{human}$ as we decreased the number of wavelets in the stimulus (visualization of stimuli).

We fit a four-parameter (slope, vertical shift, horizontal shift, and amplitude) sigmoid function or a simple linear model to the data depending on which model gave lower scores of Bayesian Information Criterion (BIC) (Schwarz, 1978) and Akaike Information Criterion (AIC) (Akaike, 1998). We then used the fit relation to find $\rho$, the proportion of wavelets sampled that would make CNNs have the same threshold as $\alpha_{human}$. $\rho$ for different CNN models are shown in **Fig 6D-F**. A one-way repeated measures ANOVA revealed a significant main effect of noise condition, ($F_{1.76,29.91}$ = 1135.78, p < 0.001, Greenhouse-Geisser corrected). Post-hoc pairwise comparisons revealed that BN and CS conditions did not differ significantly ($t_{17}$ = -2.11, p = 0.15), while both BN and CS differed significantly from SCS (BN vs.SCS: $t_{17}$ = -50.89, p < 0.001, CS vs. SCS: $t_{17}$ = -40.61, p < 0.001).

It is interesting to compare the two types of efficiency: $\eta$ and $\rho$. In **Fig 6D-F**, $\eta$ is quantified as the ratio between the thresholds of the humans and the CNN models. This measures how resilient humans were in their ability to identify letters under the different types of noise or scrambling. By that analysis, humans are closer to the CNN performance for BN (50%) than they are for the other noise types, with SCS being the lowest (9%). For $\rho$ shown in **Fig 6G-I**, we were instead asking how much information (in terms of the proportion of wavelets sampled) needed to be removed from the CNN's input before its performance fell to human levels. This change in approach reversed our findings. For BN stimuli, CNNs reach human performance with only 4% of wavelets (i.e. removing 96% of the wavelets from the input). This is the lowest efficiency we find, though for CS it was only a little higher. For SCS, the performance of the CNNs

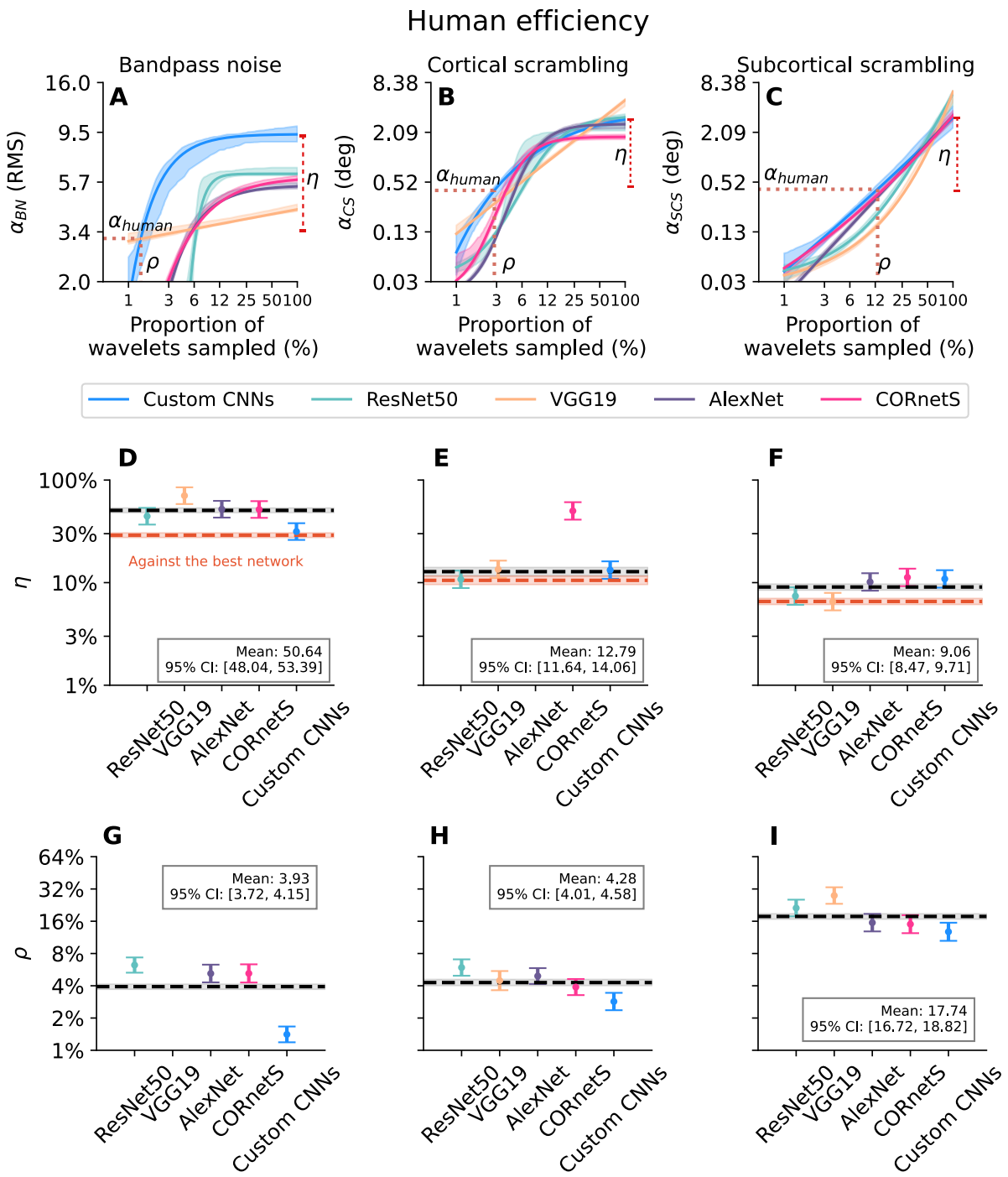


**Fig 6.** Top row shows threshold $\alpha$ of CNNs sampling at different percentages of wavelets in the stimulus for (**A**) BN, (**B**) CS, and (**C**) SCS. Sigmoid or linear function of best fit is shown with 68% CI obtained from 100 bootstrapped fits. Relative efficiency $\eta$ (ranging from 0% to 100%) is defined as the ratio of human $\alpha_{human}$ to CNN $\alpha$ sampling at 100% (**D-F**). Sampling efficiency $\rho$ is the percentage of wavelet sampled by the CNNs to match $\alpha_{human}$ (**G-I**). For the model $\alpha$ of custom CNNs, the mean $\alpha$ of twenty CNNs was used. Markers indicate the mean with 95% CI for eighteen subjects across both eyes (two participants removed after outlier removal). Dotted black line denotes the mean efficiency across CNN models with shaded region indicating 95% CI, values of which are shown in the labeled box. Dotted orange line with shaded region denotes the mean $\eta$ compared to the best performing CNN model with 95% CI. The fitting error for AlexNet in CS at 100% wavelets was large and $\eta$ is not shown (**E**). For VGG19 in BN (**G**), there were not enough valid thresholds below human performance for $\rho$ to be extracted and is not shown. All function fits with data points are shown in **Fig 5** of Supporting Information.

dropped to the human level at 18%. This indicates that the identification of SCS letters is far more dependent on the number of samples than for CS stimuli (as is also evident from the slope of the decline).

# 3 Discussion

## 3.1 Identification of scrambled letters by humans and CNNs

In this study we investigated the impact of physiologically-inspired scrambling distortion on letter identification. We employed an algorithm through which stimuli could be generated that exhibited two possible types of scrambling. These were "cortical scrambling" (CS) where the encoded positions of oriented features are randomly perturbed, and "subcortical scrambling" (SCS) where the wavelets used to reconstruct the image are themselves perturbed. Although the distinction between these two types of scrambling is inspired by different possible sites in a back-pocket model of the visual hierarchy (either at the output from simple cells in V1, or at the input forming those simple cell receptive fields) we do not draw any direct conclusions about the physiological nature of scrambling from this behavioural study. Instead, our results demonstrate these two manipulations can have distinct perceptual impact. CS and SCS degrade different information in the image because networks trained on one type of stimuli performed not as well when faced with the other. We further showed difference in CS and SCS can affect perceived scrambling magnitude in a noise matching experiment as well as give rise to different efficiencies for letter identification.

We leveraged recent advances in deep learning to train CNN models to substitute for the role an optimal "Ideal Observer" would play in conventional noise masking studies to extract efficiency (Yamins and DiCarlo, 2016; Kanwisher et al., 2023; Lindsay, 2021). Our approach relies on training the CNNs to exploit the information in our letter stimuli. We trained both custom models (selected in an architecture search then trained from scratch) and four "off-the-shelf" pre-trained networks. We validated our approach (in part) using a condition where the identification of undistorted letters was made difficult by embedding them in spatially bandpass noise (BN), for which the ideal observer strategy is known (template matching) (Van Trees, 1968; Pelli et al., 2006; Ziskind et al., 2014). Our custom CNNs achieving similar levels of performance as template matching in BN is a sign that the training procedure is optimized for the task for the variety of architectures we used.

Combining our psychophysical data with predictions from the CNN models allowed us to investigate the different task demands and relative efficiency for the three stimulus conditions (BN, CS, and SCS). We quantified human performance relative to that of the CNNs, and present this as our first measure of efficiency $\eta$. This efficiency measured in bandpass noise (51% across all models; 29% comparing to the best model) was significantly higher than the 10% measured in a similar task by Pelli et al. (2006), who instead used spatially broadband noise. We suggest this difference may be due to our use of bandpass stimuli (reducing cross-channel suppressive interactions) and/or our use of a smaller set of letter identities to be discriminated (4 in our study vs. 26 in Pelli et al., 2006). Comparing the two scrambling manipulations, we find a small but significant advantage for CS stimuli (13% vs. 9% across all models; 10% vs. 7%

comparing to the best model). This means that, relative to the CNNs, humans are more tolerant to random translations of the position of oriented features than they are to distortions of the wavelets representing those features.

One limitation of the above approach for our scrambling conditions is that the CNN thresholds are around ten times higher than the human thresholds. This places them so far beyond the human limit for recognising "a letter" (see right column of **Fig 2**) that the computation involved may be quite different from that involved in identifying letters near the human threshold (for example, it may rely more on global statistics than local structure). For that purpose, we performed a second analysis to calculate an alternate measure of efficiency $\rho$. Instead of asking how much more noise or scrambling the CNN could tolerate than a human participant, we instead set the human threshold as our benchmark and asked how much we needed to disadvantage the CNN (by randomly removing samples from its input) before its performance was degraded to the human level. This produced an entirely different pattern of results. We find that humans are remarkably efficient at processing SCS letters: CNNs need 18% of the wavelets to be preserved if they are to reach the human performance level. In comparison, CNNs need only around 4% of the stimulus wavelets to match human performance for BN or CS stimuli. Visualising the stimuli with different noise or scrambling magnitudes and with different proportions of wavelets preserved (**Fig 6-8** in the Supporting Information) gives the subjective impression that the SCS stimuli tend to remain more "letter-like" whereas the BN stimuli are reduced to parts of a letter and the CS stimuli appear as a scattering of wavelets.

Inspecting the top row of **Fig 6**, it seems that our three stimulus conditions may differ in how much redundancy exists in the input. For SCS, any loss of wavelets causes an immediate decline in performance for any of our CNN models. For the other two stimulus conditions we most often see a plateau as the proportion of wavelets declines from 100%, followed by a decline once the proportion drops below some critical value. The only exception is VGG 19, for which we saw immediate decline in performance once wavelets were reduced. We suspect this is due to a defining feature of the network that is it only contains 3 by 3 small filters. Despite transfer learning, the pre-trained weights of the network might still bias it to fully exploit the redundancy in our stimuli. Investigating the redundancy in our stimuli, and whether similar plateaus exist in human performance as a function of the proportion of wavelets preserved, is a topic for future study.

### 3.2 Error similarity between humans and CNNs

Despite the presumed similarity in the fundamental architecture, human and CNNs can be shown to behave differently (Wichmann and Geirhos, 2023). For example, CNN models can be worse than simpler models of early visual processing in predicting the visibility of distortions (Berardino et al., 2017). Previous works have also shown a divergence in the pattern of classification errors between CNNs and humans (Dodge and Karam, 2017; Borji and Itti, 2014; Geirhos et al., 2020; Rajalingham et al., 2018). To investigate this, we carried out a supplementary analysis examining error pattern similarity between our CNNs and humans (**Fig 1-3** of Supporting Information). We normalised our data so that this analysis would be independent of the overall performance of the participant (and so not confounded with the main results). We found errors are more similar among CNNs than humans, reinforcing previous findings (Geirhos

et al., 2020). We investigated whether our reduced sampling procedure (that lowered CNN performance to the human level) would also bring the pattern of errors into alignment between CNNs and humans. We found only partial evidence, suggesting this effect for SCS but not for CS. Therefore, even though this manipulation made the CNNs more similar to humans when considering the threshold stimulus magnitude, there remains a gap in fully modelling human responses using CNNs.

### 3.3 Internal scrambling and positional uncertainty in human spatial vision

In noise masking studies, it is typically assumed that the dimensions of the external noise applied to the stimulus correspond to the task-relevant internal noise in the visual system. This is necessary if the main effect of that noise is to increase response variability rather than suppressing the response to a target stimulus through nonlinear interactions (Baldwin et al., 2016; Baker and Meese, 2012). Previous noise masking studies have investigated scrambling or positional uncertainty using relatively simple stimuli synthesised from oriented wavelets (Hess et al., 1997; Baldwin et al., 2017). Others measuring Just-Noticeable-Difference (JND) as a function of pedestal position noise level have used regularly spaced lattice (Morgan et al., 2012) and natural images (Christensen et al., 2019). In all cases, they demonstrated internal sensory thresholds for the coding of position in human spatial vision.

Compared to stimuli used in previous works, our study differs in two major aspects. First, we applied our manipulations to letter stimuli. Letters are moderately complex and understanding how visual system performs letter identification informs us processes underlying recognition of more complex classes of objects (Grainger et al., 2008). While Levi et al. (1997, 1999) used a tumbling letter E, their task was an orientation discrimination task with the stimulus consisted of only 17 Gabors arranged with equal spacing. This is different from our stimulus generation scheme where the letters contain many more wavelets whose position, orientation, and phase are more consistent with internal representation based on the activities of oriented feature detectors in the cortex (see stimulus generation in 4.3). Second, we also investigated the effects of a positional noise affecting the subunits from which oriented receptive fields are formed in our subcortical scrambling noise (SCS). This approach is inspired by projection scatter measured in physiology studies (Tao et al., 2012; Nishimoto et al., 2006). In Section 4.3, we show using simulations that the effect of our SCS noise is primarily spreading energy of the Gabor into nearby orientations, increasing the orientation bandwidth while keeping in the spatial frequency passband. This differs from the simple in-place orientation jitter implemented in previous studies (Morgan et al., 2008; Christensen et al., 2015; Dakin, 2001). We do not assert our SCS noise describes a complete and accurate account of the physiological basis of scrambling in the human visual system, however the fact that CS and SCS have discriminable effects suggests they may be used to interrogate distinct aspects of the underlying physiological mechanisms. For example, the advantage we find for the dominant eye in the current study (**Fig 4**) is present only with SCS stimuli. Under our simplified model of the visual hierarchy, the proposed site of SCS would be just prior to binocular combination. As the processing of SCS stimuli appears to be more vulnerable to a reduction in sampling, we theorise that the dominant eye advantage may indicate a physiological bias (where more mechanisms are responsive to the dominant eye). Different sampling properties have been proposed to account for higher position

uncertainty in peripheral and amblyopic vision (Levi and Klein, 1986; Levi et al., 1997, 1987). Studies of the effect of reducing the number of samples available to *human* participants (comparing the dominant and non-dominant eyes) may give further insight into this possibility.

Where our results with SCS stimuli hint at possible differences in projections between dominant and non-dominant eye processing, far greater differences have been argued to exist in amblyopia (Hess et al., 1978; Watt and Hess, 1987; Hess et al., 1990; Hess and Field, 1994; Maruya et al., 2025). It has been suggested that the mis-wirings 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). Physiology and fMRI studies have shown that amblyopic eye receptive fields are larger and more heterogeneous (Tao et al., 2014; Hussain et al., 2015; Clavagnier et al., 2015; Szinte et al., 2024; Li et al., 2007). This disruption of amblyopic eye filters has been proposed to explain the distorted percepts that can occur in amblyopia (Hess et al., 1978; Sireteanu et al., 1993; Barrett et al., 2003; Maruya et al., 2025; Molaei et al., 2025). Recent work by Maruya et al. (2025) showed the amblyopic eye precept can be modeled as having a different filter set in the steerable pyramid. In ongoing work, we have applied our methods to the study of spatial distortions in amblyopia and found amblyopic eye makes different mistakes than the fellow eye in the classification of letters affected by SCS, and this divergence in mistake patterns correlates with the visual acuity difference (Zhu et al., 2025).

In normal vision, it is proposed that information is "pooled" within larger spatial windows at increasing visual field eccentricities. Humans may have access only to summary statistics of the content within each pooled region (Freeman and Simoncelli, 2011; Balas et al., 2009). Past work has validated this hypothesis by generating model-driven perceptual metamers, images that are physically distinct but perceptually invariant (Freeman and Simoncelli, 2011; Broderick et al., 2023). Future work measuring the limit at which images can be distorted while remaining metameric to both a given (scrambled) model and humans will shed light on scrambling or internal position uncertainty in peripheral vision.

# 4 Methods

## 4.1 Participants

We recruited twenty one adult human participants with normal or corrected-to-normal vision. Five participants (aged 23-26, 2 female, 3 male) were tested in Experiment 1 (Section 4.6), in which stimuli were viewed binocularly. Twenty participants (aged 20-29, 10 female, 10 male) took part in Experiments 2 (Sections 4.7) in which they were tested monocularly by having them wear an eye patch. Eye dominance was determined by the participant's sighting dominance (hole-in-the-card test).All testing was carried out with informed written consent in accordance with the tenets of the Declaration of Helsinki, and with the approval of the Centre for Applied Ethics at the McGill University Health Centre.

## 4.2 Equipment

Stimuli were presented on a gamma-corrected Display++ LCD screen (Cambridge Research Systems, United Kingdom) with a mean luminance of 58 cd/cm$^2$. At the viewing distance of 62 cm, there were 31 pixels per degree of visual

angle. The experiment software ran on a Dell XPS laptop (Dell Technologies, Inc.) and was written in Python using the PsychoPy library (Peirce, 2007). All experiments were performed indoors under dimmed lighting. All CNN training was conducted using PyTorch (Paszke et al., 2019) on CUDA-enabled GPUs.

### 4.3 Stimulus decomposition and re-synthesis algorithm

We based our stimuli on those used in previous studies of letter identification in humans. We used the Times New Roman font, consistent with previous studies investigating the optimal spatial frequency channel using bandpass-filtered letters (Chung et al., 2002b). The letter height (typographic x-height) was set to 1.5 degrees of visual angle, at which the optimal spatial frequency would be 1.5 cycles/deg (Majaj et al., 2002; Chung et al., 2002b; Solomon and Pelli, 1994). Four lowercase letters were chosen as the stimulus set: **o**, **m**, **d**, and **z**; based on prior work showing they can be easily distinguished from each other (Courrieu et al., 2004; Janini et al., 2022).

We generated our scrambled stimuli using a novel decomposition and re-synthesis algorithm, inspired by work on the activation properties of cortical receptive fields (Daugman, 1985; Field, 1987) and the multi-resolution wavelet transform (Würtz, 1995; Fischer et al., 2007). Our goal here was not to reconstruct the original image, but rather to generate stimuli that allow us to investigate the impact of scrambling on performance for the spatial frequency channel of interest (chosen as that which should optimally support letter identification).

Our decomposition and re-synthesis scheme involved a transformation of basis using Cartesian-separable log-Gabor kernels (Meese, 2010; Baker et al., 2022), which have the convenience of being D.C.-balanced in any phase (Field, 1987). Compared to the more common "polar-separable" log-Gabor, the spatial-domain 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 **Fig 7** we present an overview of the stimulus generation pipeline. This can be divided into two stages: We first decomposed the original image at different orientations and phases using a log-Gabor filter bank. Then, we re-synthesised the image from the output of that first stage. It is during re-synthesis that we applied our simulated scrambling. For the Bandpass Noise (BN) condition, no scrambling was applied. For cortical scrambling (CS) or subcortical scrambling (SCS) conditions we applied scrambling at different stages in the re-synthesis.

In the decomposition stage, we started with raw letter images $\mathbf{X}$ that were square (200 $\times$ 200 pixels). These were filtered by a bank of $m$ log-Gabor filters $\mathbf{G}_{j(j\in\Omega)}$. Each of $\Omega = \{1, ..., m\}$ indexed a filter defined in Fourier space with a different orientation preference and phase preference. All filters had a peak spatial frequency of 9.8 cycles/image (1.5 cycles/degree) with a bandwidth (full-width at half-magnitude) of 1.5 octaves. There were eight preferred orientations equally spaced between 0° and 360° (orientation bandwidth ±30°). At each orientation, two filters were included in quadrature phase (0 and 90 degrees). This gave a total of $m = 16$ filters. Filtering was performed by multiplication in the Fourier domain

$$\mathbf{c}_j = \mathcal{F}^{-1}\left(\mathbf{G}_j \cdot \mathcal{F}\left(\mathbf{X}\right)\right), \tag{3}$$

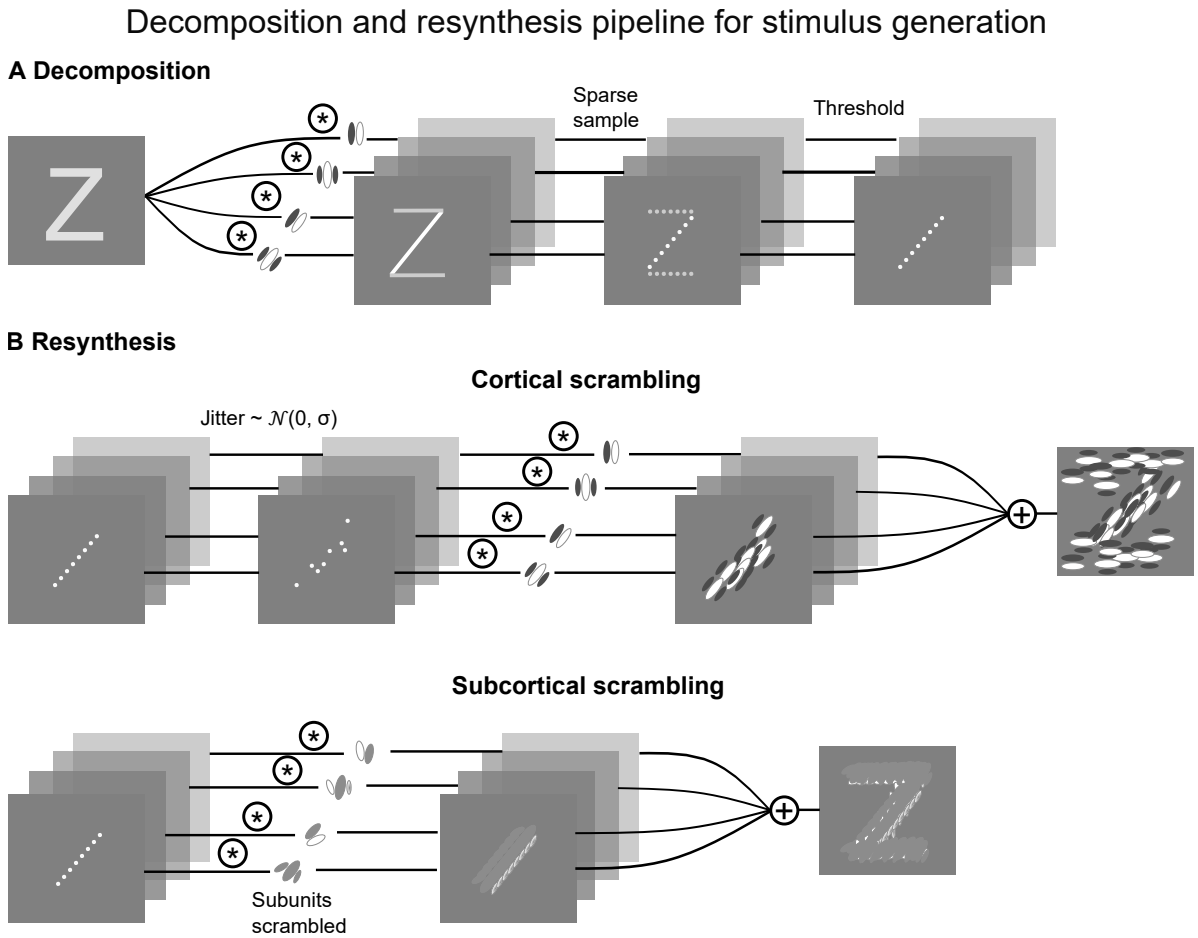


**Fig 7.** Overview of stimulus generation pipeline. In the decomposition (**A**), the image was decomposed by log-Gabors at different orientations and phases. In the re-synthesis (**B**), the weights were used to synthesize the stimulus made of log-Gabors. For cortical scrambling, we jittered the positions of the weights. For subcortical scrambling, we jittered the positions of the wiring diagram that composed log-Gabor from its isotropic subunits.

where $\mathbf{c}_j$ represented the local weights of a single (j$^{\text{th}}$) channel. $\mathcal{F}$ and $\mathcal{F}^{-1}$ refer to the discrete Fourier and inverse Fourier transform respectively.[2] 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:

$$W^{\text{decomp}}(\mathbf{X}) = \mathbf{c}_{j(j\in\Omega)}. \tag{4}$$

After filtering, we took a sparse sample of each $\mathbf{c_j}$ to give a weight value at every half-wavelength of the log-Gabor. We then applied a threshold to retain only the 1% most active weights in the entire stack. This gave the final wavelet weights $\widetilde{\mathbf{c}}_{j(j\in\Omega)}$.

A letter could then be re-synthesised from these weights using the same filter bank (shifting the 0 degree phase components to 180 degrees, corresponding to complex conjugation in the wavelet transform). For the BN stimuli (without scrambling), we simply convolved $\widetilde{\mathbf{c}}_{j(j\in\Omega)}$ with complex conjugated versions of the filters $\overline{\mathbf{G}}_j$ and summed

[2]Conversion to and from the Fourier domain was performed using the Fast Fourier Transform module `fft` of the NumPy package in Python (Harris et al., 2020).

across channels

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

In the bandpass noise (BN) condition, 2D white noise was first fed through the same decomposition and resynthesis pipeline to generate bandpass noise, which was then added to the bandpass filtered undistorted letter.

To simulate cortical scrambling (CS), stimuli were re-synthesised with random position offsets applied to the position of each log-Gabor wavelet. The offsets were drawn from a Gaussian probability distribution, the standard deviation of which controlled the magnitude of scrambling (**Fig 7B, cortical scrambling**). These were applied to the local weights $\widetilde{\mathbf{c}}_j$, defining the scramble function $S$

$$W^{\text{resynth}}_{\text{cortical}}\left(\widetilde{\mathbf{c}}_{j(j\in\Omega)}\right) = \sum_{j\in\Omega} \mathcal{F}^{-1}\left(\overline{\mathbf{G}}_j \cdot \mathcal{F}\left(S\left(\widetilde{\mathbf{c}}_j\right)\right)\right) \tag{6}$$

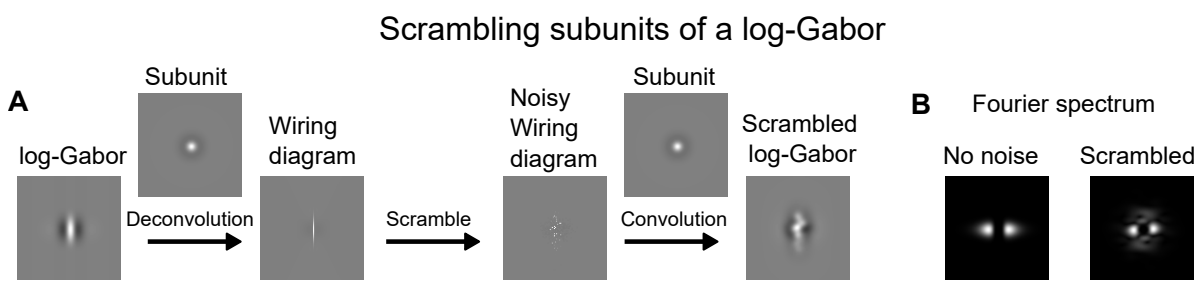


**Fig 8.** Panel (**A**) shows the generation of a scrambled log-Gabor that was used for subcortical scrambling condition. The log-Gabor was first deconvolved with an isotropic subunit to obtain a wiring diagram. The wiring diagram was then scrambled to simulate miswiring in the subunits. Panel (**B**) shows the amplitude spectrum of the log-Gabor before and after scrambling.

For subcortical scrambling (SCS), scrambling was applied at an earlier stage in the presumed hierarchy by scrambling the log-Gabor wavelets themselves (**Fig 7B**). We first found the wiring diagram that would compose the oriented log-Gabor from isotropic log-Gabor subunits (which have a donut-shaped amplitude spectrum in the 2D Fourier domain) through deconvolution (**Fig 8**). The subunit function was

$$\text{Subunit}(f,\theta) = \exp\left(\frac{-(log2(f/f_0)^2}{2(0.424\omega)}\right), \tag{7}$$

where $f_0$ is the peak spatial frequency and $\omega$ is the spatial frequency bandwidth (full-width at half-height) in octaves. For computational efficiency, we kept the top 1% of this wiring diagram and drew position offsets from a Gaussian distribution, the standard deviation of which determined the magnitude of scrambling. The scrambled wiring diagram was convolved with the isotropic subunit to construct a scrambled log-Gabor. This defines the scramble function $S'$ which was applied to the filters $\overline{\mathbf{G}}_j$ in the SCS re-synthesis

$$W^{\text{resynth}}_{\text{subcortical}}\left(\widetilde{\mathbf{c}}_{j(j\in\Omega)}\right) = \sum_{j\in\Omega} \mathcal{F}^{-1}\left(S'\left(\overline{\mathbf{G}}_j\right) \cdot \mathcal{F}\left(\widetilde{\mathbf{c}}_j\right)\right) \tag{8}$$

A single scrambled wavelet was generated for each orientation and phase used in the reconstruction for computational efficiency (Maruya et al., 2025). Physiologically however, we would expect the scrambling affecting each receptive field to be distinct.

In **Fig 9**, we summarise the spatial frequency and orientation tuning functions of SCS log-Gabors at three different scrambling levels (based on 200 simulations per plot). Scrambling had the effect of spreading energy from the original orientation channel into neighbouring channels (increasing orientation bandwidth for higher scrambling levels, with the orientation tuning function becoming flat in the limit). We also note a small downward shift in peak spatial frequency and spatial frequency bandwidth.

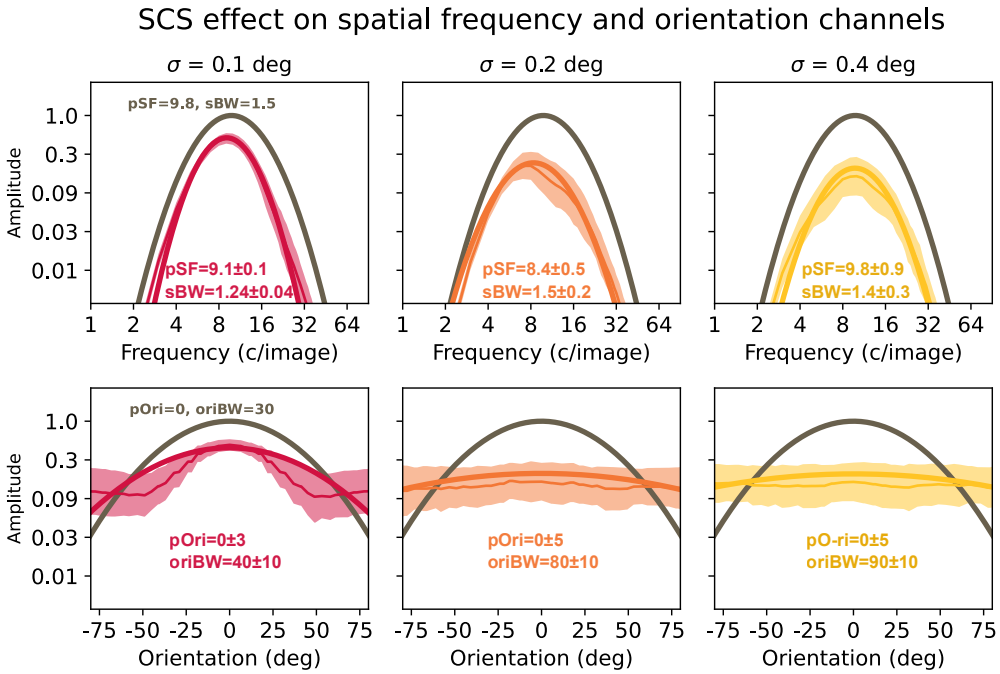


**Fig 9.** Impact of Subcortical Scrambling (SCS) on the spatial frequency and orientation tuning of log-Gabors. Original tuning functions are shown in grey. Spatial frequency tuning was found by taking a radial slice (in the Fourier domain) of the scrambled log-Gabor (at its optimal orientation before scrambling). Orientation tuning was found by taking a curved slice at the radius equal to the optimal spatial frequency. We scrambled 200 log-Gabors at three scrambling levels and here 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 (cycles/image), spatial frequency bandwidth sBW (octaves), peak orientation pOri (degrees), and orientation bandwidth oriBW (degrees). We report the mean and standard error of the parameters and plot the tuning functions that they would give rise to (coloured curves).

### 4.4 Psychometric function fitting

For baseline contrast sensitivity experiment (Sections 4.5 and 4.7), threshold stimulus magnitudes were found by fitting cumulative normal psychometric functions. For each function, we fit the threshold (for 62% percent correct) and psychometric slope as free parameters. The guess rate and lapse rate were fixed at 25% and 1% respectively. 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). We calculated the standard

error of each fitted threshold using non-parametric bootstrapping. We visualized the histogram of fitting errors for both humans and CNNs and removed outliers beyond the outer fence using Tukey's method (Tukey, 1977).

### 4.5 Baseline contrast sensitivity experiment: Measuring threshold contrast for bandpass letter identification

To account for individual differences in contrast sensitivity, the stimulus contrast in our main experiment (Experiment 2, see Section 4.7) was set to a multiple of the threshold contrast measured in each eye for each participant. To measure these thresholds, we had participants perform a single-interval four-alternative forced-choice letter identification task. The stimuli were our spatially bandpass letters re-synthesised without scrambling.

For a stimulus $\mathbf{L}$ with local luminance $L_i$ at each of $n$ points and mean luminance $\overline{L}$, we can calculate its contrast either as the root-mean-square (RMS) value (Moulden et al., 1990; Bex et al., 2009):

$$C_{\text{RMS}} = \sqrt{\frac{\sum_i \left(L_i - \overline{L}\right)^2}{n}} \Big/ \overline{L}\,, \tag{9}$$

or as the peak Weber contrast (Peli, 1990)

$$C_{\text{Weber}} = \max\left(|L_i - \overline{L}|\right)/\overline{L}\,. \tag{10}$$

the two differ in that $C_{\text{Weber}}$ is determined by the contrast of a single point in the image (the greatest deviation from the mean luminance), whereas $C_{\text{RMS}}$ is proportional to the square root of the contrast energy of the entire stimulus (for a fixed size $n$). They can therefore be thought of as measures of "local" and "global" contrast respectively (Meese et al., 2017). Linear contrast scaling of any specific stimulus image will have equivalent effects on the two contrast measures, though the overall offset between the two will vary on a per-image basis. The stochastic nature of our stimulus generation algorithm for Experiments 1 and 2 could randomly introduce contrast peaks that would have large effects on Weber contrast without making the task easier. On that basis, we chose to measure and control RMS contrast in our experiments. Where convenient, we also report the expected Weber contrast of a stimulus.

The stimuli had a duration of 250 ms above half-magnitude (stimulus contrast controlled by a raised-cosine envelope with 50 ms ramp and 200 ms plateau). The participant identified which of the four letters was shown, and were then given feedback as to whether they were correct. The RMS contrast of the letters was controlled by a two-down one-up staircase (starting level equivalent to 7% Weber contrast, with step sizes set to a factor of 1.6 for the first 4 reversals and 1.3 thereafter). The staircase terminated after a minimum of 30 trials and 8 reversals. The measurement was repeated three times for each eye, with thresholds then obtained through psychometric function fitting (Section 4.4).

### 4.6 Experiment 1: Measuring perceived magnitude of scrambling

We quantified the relative subjective magnitude of the two types of scrambling (CS and SCS) using the matching paradigm. Participants were shown three scrambled letters, positioned in a horizontal row. The letters were spanned 20 degrees of visual angle, with the central letter placed halfway (10 deg) between the left and right letters. All three letters

had the same identity (**o**, **m**, **d**, or **z**). The central letter was not scrambled: its purpose was to ensure the participant knew the identity of the three letters. The "reference" letter, presented on the left, had a fixed scrambling magnitude from a set of five logarithmically-spaced levels. The "match" letter, presented on the right, had an adjustable scrambling magnitude. On each trial, the scrambling applied to the match letter was adjusted by the participant until they were satisfied that it appeared "equally scrambled" as the reference.

Four matching conditions were tested, with matches both between the same type of scrambling (matching CS to CS, or SCS to SCS) and between different types of scrambling (matching CS to SCS, and vice-versa). For each of the four letter identities and the five reference levels, the participant performed two matches: one where the match letter began with a "high" level of scrambling and another where its initial scrambling was relatively "low". This gave 40 trials for each matching condition, which were performed in a randomised order.

To encourage participants to make a holistic judgement of scrambling magnitude, the scrambled letter stimuli in this experiment were dynamic. Both the reference and match letters were updated at 5 Hz, being replaced by another scrambled letter of the same identity and scrambling magnitude. Participants viewed the stimuli binocularly, with stimulus RMS contrast set to ten times the mean contrast threshold of the five participants to ensure visibility. The total range of Weber contrast for all stimuli was 57-72% for CS and 57-70% for SCS.

### 4.7 Experiment 2: Measuring noise threshold for letter identification

Similar to the measurements of contrast threshold in baseline contrast sensitivity experiment (Section 6.2), our main experiment was a single-interval four-alternative forced-choice letter identification task. In this experiment though, the contrast of the letters was fixed at an RMS contrast four-times the threshold contrast measured in each eye for each participant. Converted to Weber contrast, the mean contrasts used were 5% for BN, 11% for CS, and 10% for SCS (for both the dominant and non-dominant eyes). We instead varied the magnitude of the noise or scrambling applied to the letter to find the threshold level where performance reduced to threshold (62% percent correct). 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 made the task more difficult (where increasing contrast in the previous task made it easier). A two-up one-down staircase controlled the noise level of the stimulus, with the same terminating criteria as used in baseline contrast sensitivity experiment. There were three repeats of each eye and noise condition.

### 4.8 Training and selecting custom CNNs

We generated three sets of stimuli to train the CNNs, one for each noise condition (BN, CS, and SCS). Each comprised of stimuli at six log-spaced magnitudes that covered the entire range used in our human testing. At each magnitude, we generated 50 examples for each target letter. This resulted in a training set of 1200 stimuli. During training, we used a split of 70%, 20%, and 10% for training, validation, and testing. Each batch consisted of 32 images. Stimuli in all datasets were max-min normalized following 8-bit quantization in the grey level (preventing the CNN having access to more information than our human participants).

**Table 2.** Discrete prior distributions for the architecture search. Values were drawn from a uniform distribution.

| Parameter | Discrete prior distribution |
|---|---|
| Number of max pooling layers | [3, 4, 5] |
| Number of convolutional layers per max pooling layer | [1, 2] |
| Number of channels | [32, 64, 128] |
| Convolutional kernal size | [(2,2), (3,3)] |
| Dropout rate | [0, 0.1, 0.2] |

To optimise the CNN architecture for each condition, we performed an architecture search (**Table 2**). Values were drawn for each hyperparameter from a discrete uniform prior. We randomly sampled 100 architectures from this grid, and selected the top 20 by ranking their validation accuracy after training for 10 epochs. Histograms of hyperparameters for the generated and selected architectures are shown in **Fig 4** of Supporting Information.

The top 20 architectures we chose were then trained from scratch for 100 epochs. Early stopping was used with the patience parameter set to 10. All architectures had the same max pooling kernel size of (2,2), as larger kernel sizes frequently led to invalid architectures. The first and second convolutional layers always had 32 and 64 channels respectively to ensure good feature extraction early in the network. Each convolutional layer was always preceded by batch normalization (apart from the first) and followed by the rectified linear unit (`ReLU`) activation function. We used one fully connected layer after the last max pooling layer and before the output layer for which 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. All model predictions presented in this paper were made on a a test dataset with 1200 stimuli (50 stimuli per letter and noise level for four letters and six noise levels).

### 4.9 Transfer learning using pre-trained networks

We employed transfer learning with four convolutional neural network architectures: ResNet-50, AlexNet, VGG-19, and CORnet-S pre-trained on ImageNet (Krizhevsky et al., 2012; Kubilius et al., 2019; He et al., 2015; Simonyan and Zisserman, 2015). To accommodate single-channel grayscale input, we modified the first convolutional layer of each network by averaging the pre-trained RGB weights across color channels, preserving the learned edge and texture features while adapting to our stimulus format. The final classification layers were replaced with a new fully-connected layers matching our 4-class letter identification task. Training proceeded in two phases: first, we froze the convolutional backbone and trained only the modified first convolutional layer and new classification head for a minimum of 10 epochs using the Adam optimizer again (Kingma and Ba, 2014) with a learning rate of 0.001; second, we unfroze all network parameters and fine-tuned the entire model with a reduced learning rate of 0.0001. We trained for the same number of epochs as custom CNNs (100 epochs) and early stopping patience was also set to 10 epochs monitoring validation loss to prevent overfitting in each phase.

### 4.10 Template matching model

We also implemented a template matching model, which performed an in-place pixel-by-pixel multiplication between each noiseless bandpass letter with the noisy stimulus. The "max" rule was used to pick the letter that gave the highest standard deviation in the response distribution. The predictions of the template matching model were made on the same test set as the CNNs.

## 5 Additional information

### 5.1 Acknowledgements

This work was supported by an NSERC Discovery Grant (RGPIN-2022-04216) awarded to ASB, a CIHR Vanier Doctoral Scholarship awarded to RXZ, and the Montreal General Hospital Foundation (through their support of the Centre for Digital Brain Therapies). The authors would like to thank Robert Hess for discussion and feedback, and members of RXZ's thesis committee: Curtis Baker, Pouya Bashivan, and Christopher Pack for helpful comments and guidance.

### 5.2 Author contributions in CREDIT format

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

### 5.3 Data availability statement

All relevant human and model data, CNN training testing datasets, CNN models can be found on figshare at 10.6084/m9.figshare.25631925 (at publication). All experiment, simulation, analysis, CNN training code can be found on figshare at 10.6084/m9.figshare.25631934 (at publication).

# References


Akaike, H. (1998). Information Theory and an Extension of the Maximum Likelihood Principle. In Parzen, E., Tanabe, K., and Kitagawa, G., editors, *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer, New York, NY.

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. and Meese, T. S. (2012). Zero-dimensional noise: The best mask you never saw. *Journal of Vision*, 12(10):20–20.

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.

Balas, B., Nakano, L., and Rosenholtz, R. (2009). A summary-statistic representation in peripheral vision explains visual crowding. *Journal of Vision*, 9(12):13.

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.

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

Bex, P. J., Solomon, S. G., and Dakin, S. C. (2009). Contrast sensitivity in natural scenes depends on edge as well as spatial frequency structure. *Journal of Vision*, 9(10):1.1–19.

Bi, H., Zhang, B., Tao, X., Harwerth, R. S., Smith, E. L., and Chino, Y. M. (2011). Neuronal responses in visual area V2 (V2) of macaque monkeys with strabismic amblyopia. *Cerebral Cortex (New York, N.Y.: 1991)*, 21(9):2033–2045.

Bonds, A. B. (1989). Role of Inhibition in the Specification of Orientation Selectivity of Cells in the Cat Striate Cortex. *Visual Neuroscience*, 2(1):41–55.

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.

Bottou, L., Cortes, C., Denker, J. S., Drucker, H., Guyon, I., Jackel, L. D., LeCun, Y., Muller, U. A., Sackinger, E., Simard, P., and Vapnik, V. (1994). Comparison of classifier methods: a case study in handwritten digit recognition. pages 77,78,79,80,81,82–77,78,79,80,81,82. IEEE Computer Society.

Broderick, W. F., Rufo, G., Winawer, J., and Simoncelli, E. P. (2023). Foveated metamers of the early visual system. *eLife*, 12.

Brouwer, G. J. and Heeger, D. J. (2011). Cross-orientation suppression in human visual cortex. *Journal of Neurophysiology*, 106(5):2108–2119.

Chklovskii, D. B. and Koulakov, A. A. (2004). MAPS IN THE BRAIN: What Can We Learn from Them? *Annual Review of Neuroscience*, 27(Volume 27, 2004):369–392.

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

Christensen, J. H., Bex, P. J., and Fiser, J. (2015). Prior implicit knowledge shapes human threshold for orientation noise. *Journal of Vision*, 15(9):24.

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. (2002a). Spatial-frequency characteristics of letter identification in central and peripheral vision. *Vision Research*, 42(18):2137–2152.

Chung, S. T. L., Levi, D. M., Legge, G. E., and Tjan, B. S. (2002b). Spatial-frequency properties of letter identification in amblyopia. *Vision Research*, 42(12):1571–1581.

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

Clavagnier, S., Dumoulin, S. O., and Hess, R. F. (2015). Is the Cortical Deficit in Amblyopia Due to Reduced Cortical Magnification, Loss of Neural Resolution, or Neural Disorganization? *Journal of Neuroscience*, 35(44):14740–14755.

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. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision*, 18(5):1016–1026.

Dakin, S. C., Mareschal, I., and Bex, P. J. (2005). Local and global limitations on direction integration assessed using equivalent noise analysis. *Vision Research*, 45(24):3027–3049.

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.

DiCarlo, J. J., Zoccolan, D., and Rust, N. C. (2012). How Does the Brain Solve Visual Object Recognition? *Neuron*, 73(3):415–434.

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.

Eckstein, M. P., Ahumada, A. J., and Watson, A. B. (1997). Visual signal detection in structured backgrounds. II. Effects of contrast gain control, background variations, and white noise. *JOSA A*, 14(9):2406–2419.

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.

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.

Foley, J. M. (1994). Human luminance pattern-vision mechanisms: masking experiments require a new model. *JOSA A*, 11(6):1710–1719.

Francl, A. and McDermott, J. H. (2022). Deep neural network models of sound localization reveal how perception is adapted to real-world environments. *Nature Human Behaviour*, 6(1):111–133.

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.

Geirhos, R., Meding, K., and Wichmann, F. A. (2020). Beyond accuracy: quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. arXiv:2006.16736 [cs].

Geirhos, R., Temme, C. R. M., Rauber, J., Schütt, H. H., Bethge, M., and Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

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

Geisler, W. S. (1989). Sequential ideal-observer analysis of visual discriminations. *Psychological Review*, 96(2):267–314. Place: US.

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. Number: 7825.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs].

Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, 9(2):181–197.

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 Demanins, R. (1998). Contour integration in anisometropic amblyopia. *Vision Research*, 38(6):889–894.

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.

Hussain, Z., Svensson, C.-M., Besle, J., Webb, B. S., Barrett, B. T., and McGraw, P. V. (2015). Estimation of cortical magnification from positional error in normally sighted and amblyopic subjects. *Journal of Vision*, 15(2):25.

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.

Keeble, D. R. T. and Hess, R. F. (1999). Discriminating local continuity in curved figures. *Vision Research*, 39(19):3287–3299.

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

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. Curran Associates, Inc.

Kubilius, J., Schrimpf, M., Kar, K., Rajalingham, R., Hong, H., Majaj, N., Issa, E., Bashivan, P., Prescott-Roy, J., Schmidt, K., Nayebi, A., Bear, D., Yamins, D. L., and DiCarlo, J. J. (2019). Brain-Like Object Recognition with High-Performing Shallow Recurrent ANNs. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Kullback, S. and Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Levi, D. M. and Klein, S. A. (1986). Sampling in spatial vision. *Nature*, 320(6060):360–362. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 6060 Primary_atype: Research.

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.

Levi, D. M., Klein, S. A., and Yap, Y. L. (1987). Positional uncertainty in peripheral and amblyopic vision. *Vision Research*, 27(4):581–597.

Levi, D. M., Sharma, V., and Klein, S. A. (1997). Feature integration in pattern perception. *Proceedings of the National Academy of Sciences*, 94(21):11742–11746.

Li, X., Dumoulin, S. O., Mansouri, B., and Hess, R. F. (2007). The fidelity of the cortical retinotopic map in human amblyopia: Cortical retinotopy in amblyopia. *European Journal of Neuroscience*, 25(5):1265–1277.

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.

Lindsay, G. W. (2021). Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. *Journal of Cognitive Neuroscience*, 33(10):2017–2031.

Luo, L. and Flanagan, J. G. (2007). Development of Continuous and Discrete Neural Maps. *Neuron*, 56(2):284–300.

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.

Maruya, A., Ramakrishnan, B., Olianezhad, F., Wang, J., Alonso, J.-M., and Zaidi, Q. (2025). Phantom Forms in Amblyopic Vision and what they reveal about the Generative Brain. ISSN: 2692-8205 Pages: 2025.06.24.661078 Section: New Results.

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.

Meese, T. S., Baker, D. H., and Summers, R. J. (2017). Perception of global image contrast involves transparent spatial filtering and the integration and suppression of local contrasts (not RMS contrast). *Royal Society Open Science*, 4(9):170285.

Meese, T. S. and Holmes, D. J. (2006). Spatial and temporal dependencies of cross-orientation suppression in human vision. *Proceedings of the Royal Society B: Biological Sciences*, 274(1606):127–136.

Michel, M. and Geisler, W. S. (2011). Intrinsic position uncertainty explains detection and localization performance in peripheral vision. *Journal of Vision*, 11(1):18.

Molaei, H., Abbas Farishta, R., and Farivar, R. (2025). Letter Distortion Mapping in Amblyopia: Spatial Patterns, Stability, and Relationship to Visual Acuity. *Investigative Ophthalmology & Visual Science*, 66(15):65.

Morgan, M., Chubb, C., and Solomon, J. A. (2008). A 'dipper' function for texture discrimination based on orientation variance. *Journal of Vision*, 8(11):9–9.

Morgan, M. J., Mareschal, I., Chubb, C., and Solomon, J. A. (2012). Perceived pattern regularity computed as a summary statistic: implications for camouflage. *Proceedings of the Royal Society B: Biological Sciences*, 279(1739):2754–2760.

Moulden, B., Kingdom, F., and Gatley, L. F. (1990). The Standard Deviation of Luminance as a Metric for Contrast in Random-Dot Images. *Perception*, 19(1):79–101.

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.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv:1912.01703 [cs].

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.

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, 162(1):8–13.

Peli, E. (1990). Contrast in complex images. *JOSA A*, 7(10):2032–2040.

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

Pelli, D. G., Burns, C. W., Farell, B., and Moore-Page, D. C. (2006). Feature detection and letter identification. *Vision Research*, 46(28):4646–4674.

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.

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., and DiCarlo, J. J. (2018). Large-Scale, High-Resolution Comparison of the Core Visual Object Recognition Behavior of Humans, Monkeys, and State-of-the-Art Deep Artificial Neural Networks. *Journal of Neuroscience*, 38(33):7255–7269.

Rao, R. P. N. and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87.

Riesenhuber, M. and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025. Number: 11.

Ringach, D. L. (2004). Haphazard Wiring of Simple Receptive Fields and Orientation Columns in Visual Cortex. *Journal of Neurophysiology*, 92(1):468–476.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464.

Simonyan, K. and Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 [cs].

Sireteanu, R., Lagreze, W. D., and Constantinescu, D. H. (1993). Distortions in two-dimensional visual space perception in strabismic observers. *Vision Research*, 33(5-6):677–690.

Solomon, J. A. and Pelli, D. G. (1994). The visual filter mediating letter identification. *Nature*, 369(6479):395–397. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 6479 Primary_atype: Research.

Swindale, N. V. and Mitchell, D. E. (1994). Comparison of receptive field properties of neurons in area 17 of normal and bilaterally amblyopic cats. *Experimental Brain Research*, 99(3):399–410.

Szinte, M., Lascombes, U., Sheynin, Y., Levi, D. M., Silver, M. A., and Chopin, A. (2024). Modeling retinotopic maps in amblyopia reveals cortical reorganization across the visual hierarchy. *Journal of Vision*, 24(10):551.

Tanner, Jr., W. P. and Birdsall, T. G. (1958). Definitions of d' and $\eta$ as psychophysical measures. *The Journal of the Acoustical Society of America*, 30(10):922–928.

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.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley Publishing Company.

Van Trees, H. L. (1968). *Detection, Estimation, and Modulation Theory, Part I: Detection, Estimation, and Linear Modulation Theory*. John Wiley & Sons. Google-Books-ID: Xzp7VkuFqXYC.

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, I., 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.

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.

Wichmann, F. A. and Geirhos, R. (2023). Are Deep Neural Networks Adequate Behavioral Models of Human Visual Perception? *Annual Review of Vision Science*, 9(Volume 9, 2023):501–524.

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

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.

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.

Zhang, X., Zhao, J., and LeCun, Y. (2015). Character-level Convolutional Networks for Text Classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Zhu, R. X., Hess, R. F., and Baldwin, A. S. (2025). The amblyopic acuity deficit: impact on the identification of letters distorted by spatial scrambling algorithms. ISSN: 2692-8205 Pages: 2025.07.02.662570 Section: New Results.

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.

# IDENTIFICATION OF LETTERS DISTORTED BY PHYSIOLOGICALLY-INSPIRED SPATIAL SCRAMBLING SUPPORTING INFORMATION

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

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

March 13, 2026

## 1 Nested model hypothesis testing

Nested model hypothesis test results are shown in **Table 1**. 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. 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 1.** Comparing full model with reduced model for the matching data using nested model hypothesis test.

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

## 2 Contrast thresholds for bandpass letter identification

We measured monocular contrast thresholds for identifying letters shown to the Dominant (D) and Non-Dominant (ND) eyes of each participant. The stimuli used in these measurements were re-synthesised (spatially-bandpass) letters without any added noise or scrambling. Across our 20 participants, D eye thresholds (mean $C_{\mathrm{RMS}}$ 0.0036 95% CI [0.0033 to 0.0039]) did not significantly differ from those in ND eyes (0.0038 95% CI [0.0034 to 0.0041]) when

*corresponding author

compared using a paired t-test ($t_{19}$=-1.08, p=0.29). These RMS contrast thresholds were equivalent to an expected Weber contrast of 2.5% for D eye and 2.6% for ND eye.

## 3 Comparing error patterns

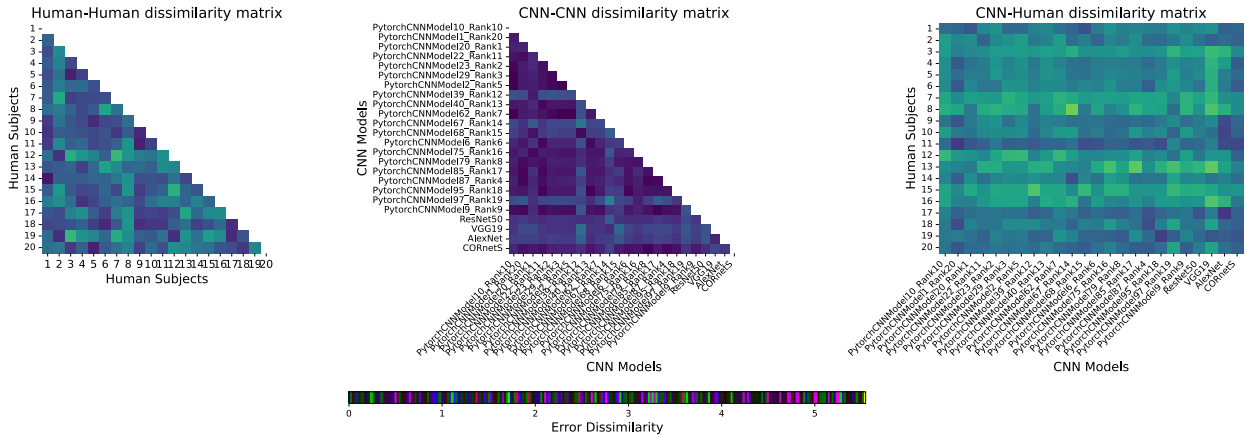


**Fig 1.** Pairwise dissimilarity for humans against humans, CNN against CNNs, humans against CNNs. Noise condition is SCS. Human data are from dominant eye viewing condition. CNNs are sampling 100% of the wavlets.

To investigate the error patterns, we first computed confusion matrices for each human and CNN. To ensure the fairness of the comparison between humans and CNNs, we generated another dataset with noise levels that humans were shown in the staircase. For each pairwise comparison, the dissimilarity of errors was calculated separately for each target letter first and added together. For the metric of dissimilarity, We calculated the Kullback–Leibler (KL) divergence between each probability vector and the average of the two probability vectors before suming them. The higher this value the more *dissimilar* the probability distributions are from each other. This metric is symmetric and circumvents the undefined case of any entry in the true distribution being zero while the approximate distribution entry being non-zero. Our metric has the form:

$$Dissimilarity(\bar{\mathbf{P}}, \bar{\mathbf{Q}}) = \sum_{i=1}^{4} \left[ KL\left(\mathbf{P_i}, \frac{\mathbf{P_i}+\mathbf{Q_i}}{2}\right) + KL\left(\mathbf{Q_i}, \frac{\mathbf{P_i}+\mathbf{Q_i}}{2}\right) \right], \tag{1}$$

where $\bar{\mathbf{P}}$ and $\bar{\mathbf{Q}}$ are confusion matrices excluding the diagonal elements, $\mathbf{P_i}$ and $\mathbf{Q_i}$ are response probability vectors of length 3 for the ith target letter. Our metric is bounded between 0 (same distribution) and 8log(2) (disjoint distribution).

We were interested in whether errors made by CNNs were similar to human errors. Furthermore, we wondered whether constraining CNN performance to human level with reduced sampling could potentially align errors made by humans and models. For reference, we also computed similarity of errors for humans compared to other humans, and CNNs compared to other CNNs. Example dissimilarity matrices are shown in **Fig 1** (100% wavelets kept for CNNs, SCS).

The complete results of comparing error patterns are shown in **Fig 2** and **Fig 3** for humans viewing through their dominant eye (DE) and non-dominant eye (NDE), respectively. We found that across proportions of wavelet sampled, CNNs had higher dissimilarity to humans than humans to other humans for CS (Mann-Whitney U test. D: U=1442, p<0.0001; ND: U=1334, p=0.0009) and SCS (D: U=1206, p=0.043; ND: U=1288, p=0.0058). Interestingly we found a significant linear relationship for SCS (D: slope parameter, $t_6$=3.35, p=0.015, $R^2$: 0.65; ND: slope parameter, $t_6$=3.64, p=0.011, $R^2$: 0.69) where decreasing the percent of wavelet sampled aligned the CNN errors to be closer to humans. For BN, we found CNN errors were not significantly different from human errors (D: U=1045, p=0.87; ND: U=1033, p=1.03). In all cases, we found CNNs were more similar to other CNNs than humans to other humans for both dominant eye (BN: U=53, p<0.0001; CS: U=21, p<0.0001; SCS: U=53, p<0.0001) and non-dominant eye (BN: U=38, p<0.0001; CS: U=58, p<0.0001; SCS: U=67, p<0.0001).

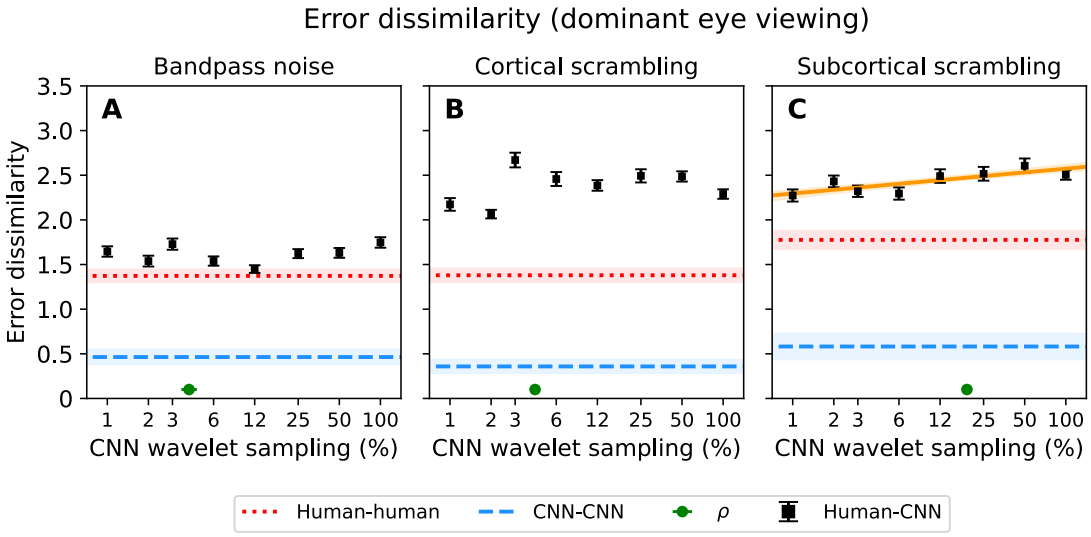


**Fig 2.** Comparing error dissimilarity between CNNs and humans (dominant eye). Dissimilarity across different proportions of sampled wavelets for (**A** BN), (**B**) CS, and (**C**) SCS. The error dissimilarity metric is based on the Kullback-leibler (KL) divergence. Human to CNN dissimilarity is shown in black markers with mean and 95% CI. SCS is the only noise condition for which the slope is significantly different from zero after fitting a simple linear model. Its line of best fit with 95% CI shaded is shown in orange. For reference, human to human dissimilarity (red) and CNN to CNN dissimilarity (blue) are also shown with 95% CI shaded region. For each noise condition, average $\rho$ across models is shown as a green marker with mean and 95% CI.

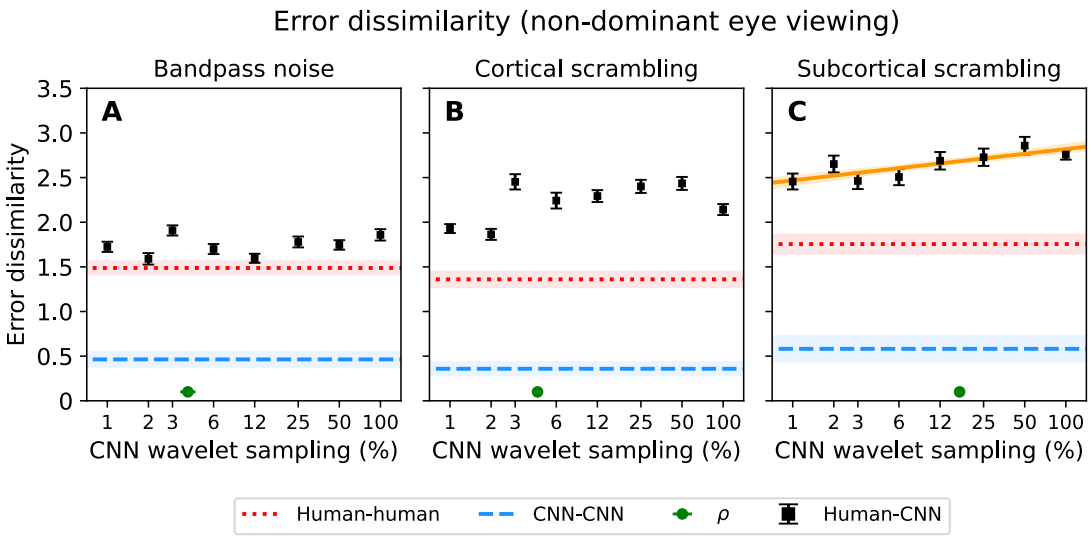


**Fig 3.** Comparing error dissimilarity between CNNs and humans (non-dominant eye). See captions of **Fig 2** for details. SCS is again the only noise condition for which the slope is significantly different from zero after fitting a simple linear model.

## 4 Architecture search hyperparameter histograms

The architecture search results are shown in **Fig 4**. Out of 100 architectures, top 20 architectures were selected based on validation accuracy after training for 10 epochs. We see chosen architectures tend to have higher numbers of pooling (first row) and convoultional layers (third row).

## 5 Individual fits to CNN threshold as function of sampling

Individual function fits are shown in **Fig 5**. CNN performance in bandpass noise and cortical scrambling exhibited a plateau when wavelet sampling was reduced initially and a sharp fall off at lower sampling rate (except for VGG19). In subcortical scrambling, CNN performance decreased immediately once sampling was reduced.

## 6 Visualisation of sparse sampling wavelets in the stimulus

We can directly visualize the stimuli at different sampling rate and noise level to qualitatively judge whether what equates CNNs to human performance looks discriminable to humans. The stimuli are inherently stochastic but here we aim to give a high-level impression of what the stimuli look like at various sampling rate and noise level. The visualizations are shown in **Fig 6** to **Fig 8**. If humans were sampling at the same rate $\rho$ that made CNNs have the same performance as humans $\alpha$, they would look like stimuli shown in the second column from the left. For BN and CS, they look like a smattering of wavelets. In SCS, the contour of the letter is more preserved at the cost of losing orientation tuning of individual wavelets.

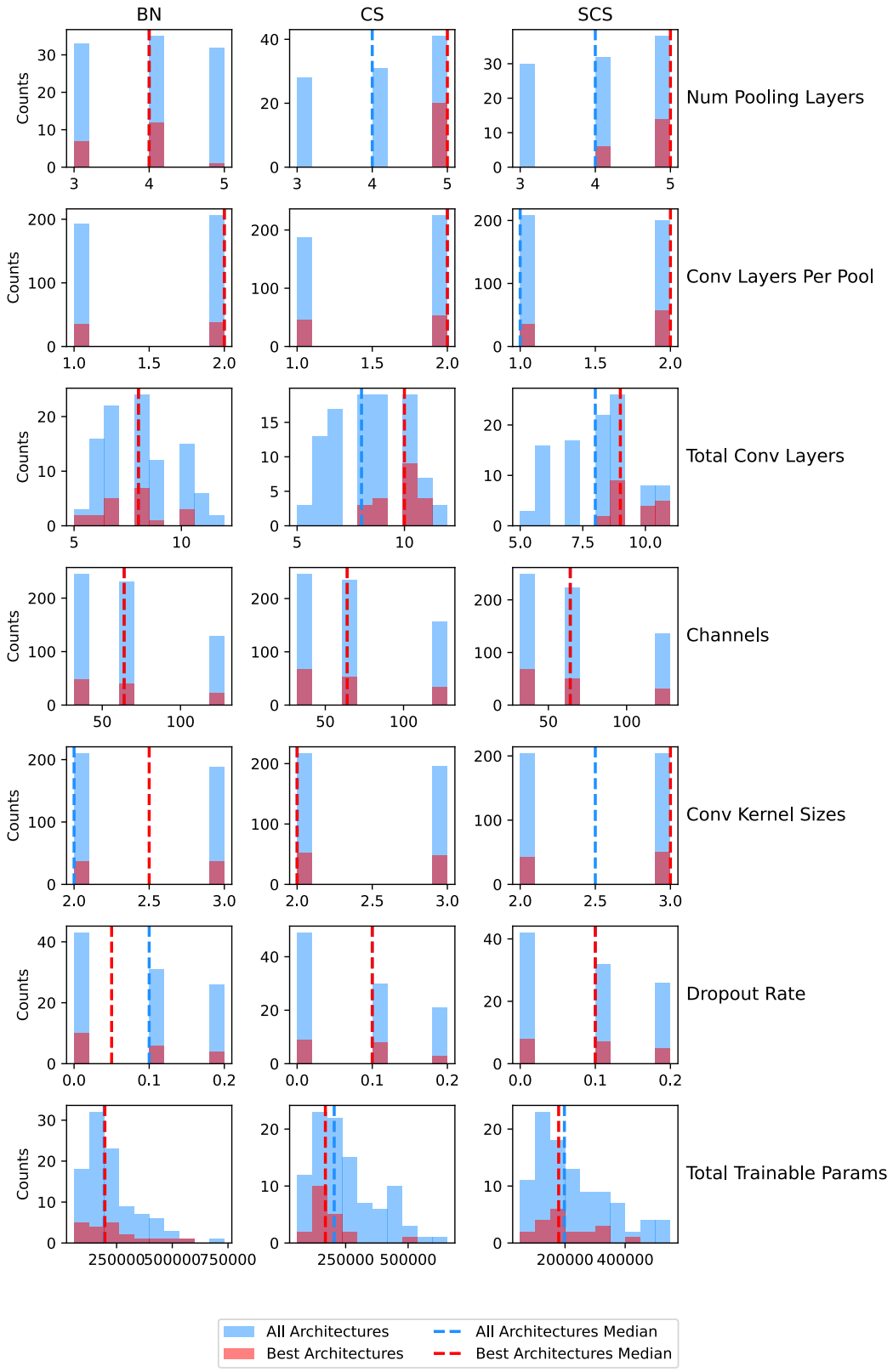


**Fig 4.** Histograms of hyperparameters of CNNs trained on bandpass noise (BN), cortical scrambling (CS), and subcortical scrambling (SCS) for 100 random architectures searched (blue) and 20 selected architectures (red).

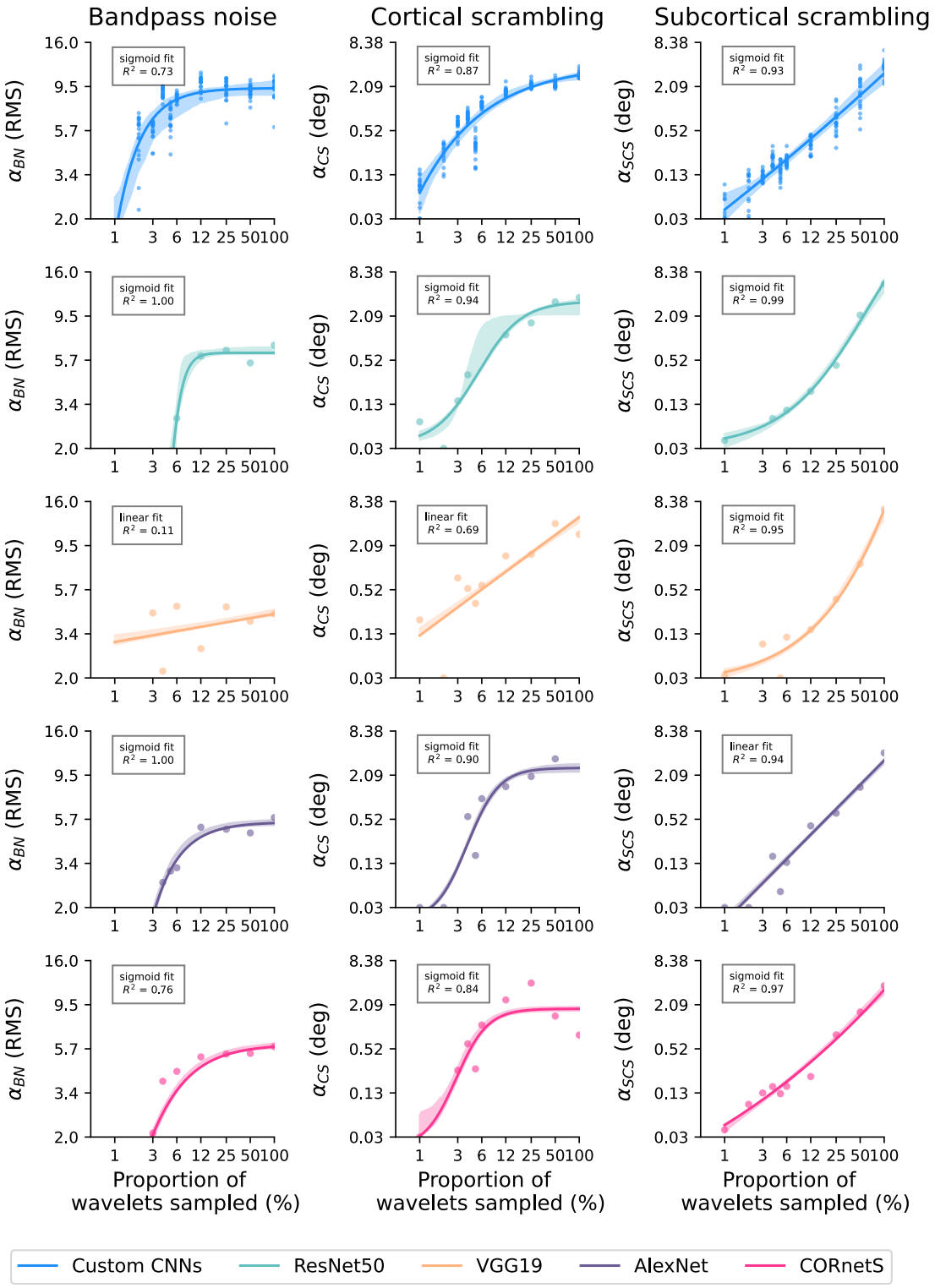


**Fig 5.** Sigmoid or linear fits to CNNs. Sigmoid model contained four free parameters: slope, vertical shift, horizontal shift, and amplitude. Linear function contained two free parameters: slope and intercept. The fit is shown for the better model defined as the one that gives lower score on Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC).

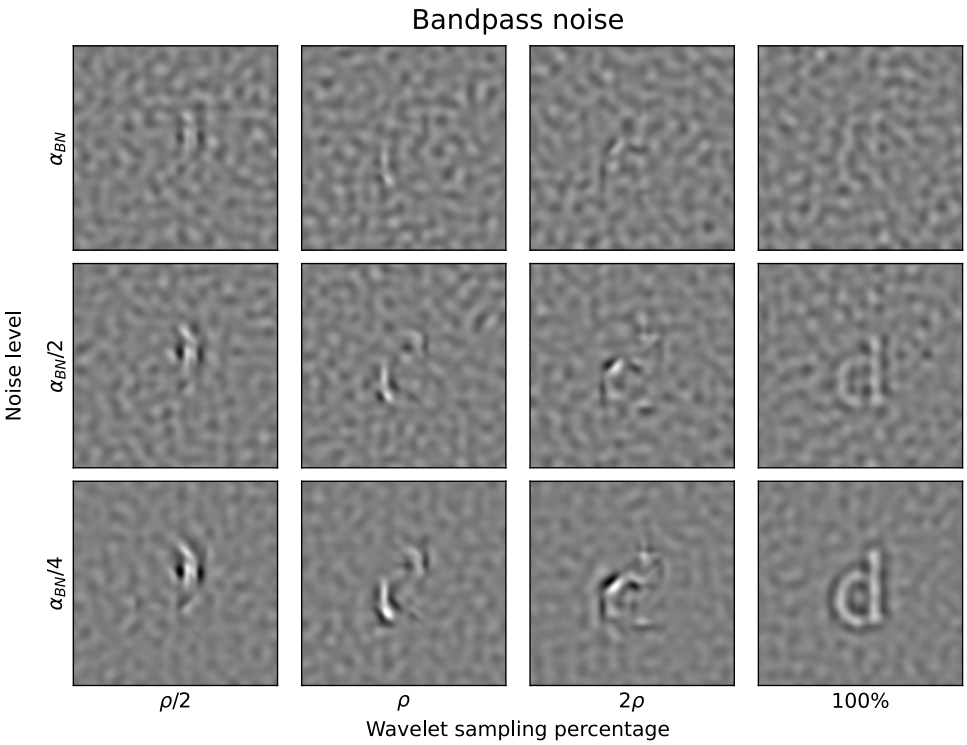


**Fig 6.** Jointly visualizing the effect of noise magnitude and wavelet sampling on a letter "d" for bandpass noise. $\alpha$ indicates the human threshold across dominant and non-dominant eyes. $\rho$ indicates the sampling percentage that equated CNNs performance with human performance.

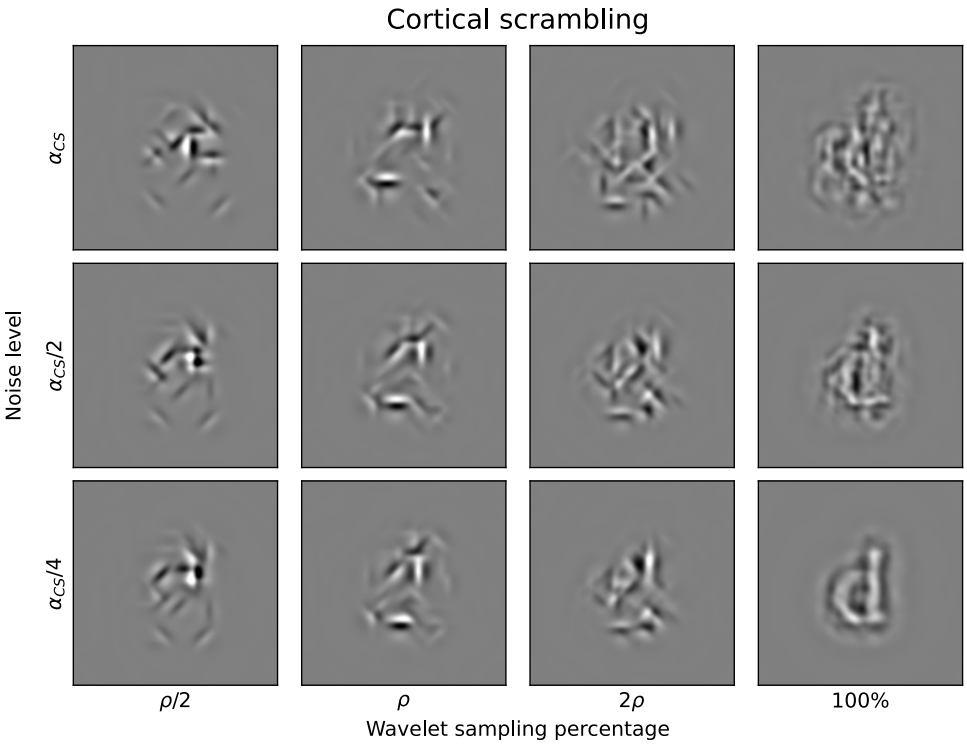


**Fig 7.** Jointly visualizing the effect of noise magnitude and wavelet sampling for cortical scrambling.

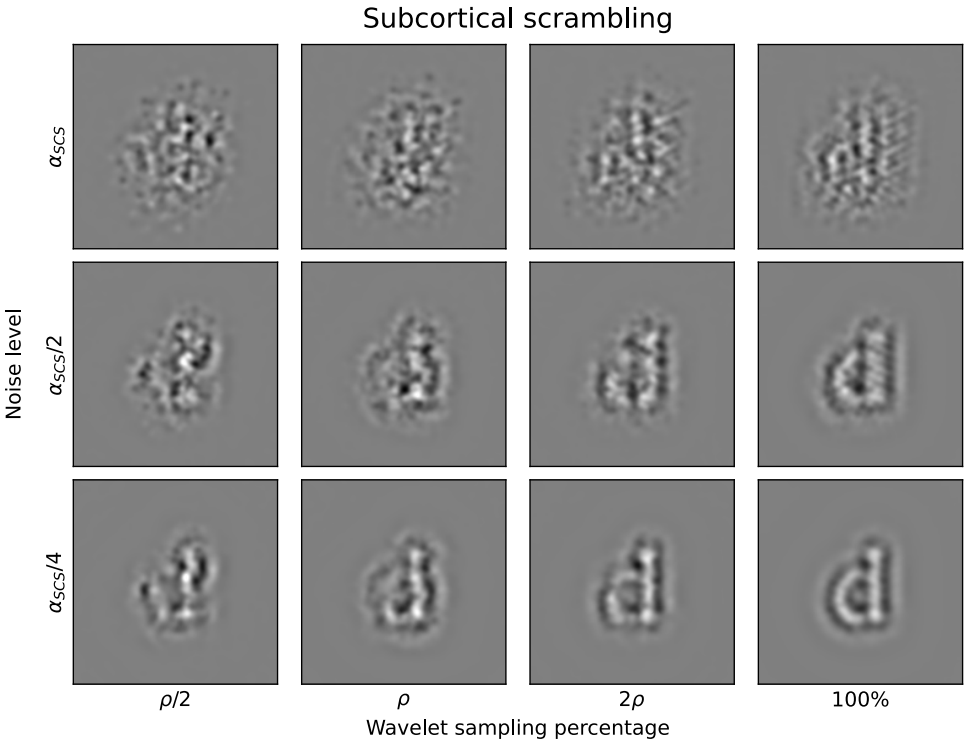


**Fig 8.** Jointly visualizing the effect of noise magnitude and wavelet sampling for cortical scrambling.