Crowding in humans is unlike that in convolutional neural networks

Ben Lonnqvist[a,*], Alasdair D. F. Clarke[b], Ramakrishna Chakravarthi[c]

[a] *Business School, University of Aberdeen*
[b] *Department of Psychology, University of Essex*
[c] *School of Psychology, University of Aberdeen*

Abstract

Object recognition is a primary function of the human visual system. It has recently been claimed that the highly successful ability to recognise objects in a set of emergent computer vision systems—Deep Convolutional Neural Networks (DCNNs)—can form a useful guide to recognition in humans. To test this assertion, we systematically evaluated visual crowding, a dramatic breakdown of recognition in clutter, in DCNNs and compared their performance to extant research in humans. We examined crowding in three architectures of DCNNs with the same methodology as that used among humans. We manipulated multiple stimulus factors including inter-letter spacing, letter colour, size, and flanker location to assess the extent and shape of crowding in DCNNs. We found that crowding followed a predictable pattern across architectures that was different from that in humans. Some characteristic hallmarks of human crowding, such as invariance to size, the effect of target-flanker similarity, and confusions between target and flanker identities, were completely missing, minimised or even reversed. These data show that DCNNs, while proficient in object recognition, likely achieve this competence through a set of mechanisms that are distinct from those in humans. They are not necessarily equivalent models of human or primate object recognition and caution must be exercised when inferring mechanisms derived from their operation.

Keywords: convolutional neural networks, object recognition, crowding

1. Introduction

Recognising objects is a central function of the human visual system and the mechanisms underlying this ability have been extensively studied (DiCarlo et al., 2012; Ullman, 2007). One approach to studying human object recognition is to examine situations where it fails in order to determine the constraints for successful recognition. Visual crowding is one such failure of object recognition

*Corresponding author

in human vision (Bouma, 1970; Levi, 2008; Manassi & Whitney, 2018) where objects that are otherwise recognisable in the visual periphery are rendered unrecognisable when surrounded by similar clutter. Studies on visual crowding have given rise to multi-stage models of object recognition (Pelli et al., 2004).

In computer vision, deep convolutional neural networks (DCNNs) have proven to be extremely successful, reaching high accuracy rates in many object recognition and classification tasks (Simonyan & Zisserman, 2014; Szegedy et al., 2014; He et al., 2015a; Huang et al., 2016). DCCNs are loosely inspired by the human visual system and have been argued to be compelling models of primate object recognition (Cadieu et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Güçlü & Gerven, 2015; Yamins & DiCarlo, 2016; Bonner & Epstein, 2017). However, interpreting both the decision process and the relationship between inputs and layers' outputs is difficult, and many approaches to interpreting and understanding DCNNs have been taken (Zeiler & Fergus, 2013; Zhang et al., 2017). The goal of our paper is not to interpret the low-level details of the DCNN decision process, but rather to investigate if DCNNs suffer from human-like crowding patterns, and if so, whether examining these breakdowns in DCNNs can shed light on the mechanisms of object recognition. If DCNNs are to serve as fruitful models of human neural computations, it is crucial to determine the similarities and differences between human and computer vision models. That is, if DCNNs recognise objects using mechanisms analogous to that in humans, then they too should be subject to the flanker-induced interference observed in humans. It is important to understand the behaviour of crowding in DCNNs not only to help us better understand the human visual system, but also to be able to design more efficient computer vision systems.

The phenomenon of crowding in humans displays certain distinctive features. Here, we highlight the most salient and relevant aspects, which form by no means an exhaustive list of its properties. The most striking observation in crowding is that closer flankers interfere with the identification of a target more than distant flankers; that is, the spacing between targets and their flankers strongly modulates identification performance (Bouma, 1970; Toet & Levi, 1992; Pelli et al., 2004). Further, for a fixed spacing between a target and its flankers, crowding (interference) is stronger at larger target eccentricities (distance from fixation; Toet & Levi (1992); Pelli et al. (2004)). Crucially, the flankers interfere with the target over a limited region of space that scales with eccentricity. Under standard circumstances, flankers further than half the target's eccentricity do not crowd the target. This relationship has been called the Bouma Law (Pelli & Tillman, 2008). The relationship seems to hold true for a wide range of objects, from simple features such as oriented gratings and colour to complex real-world objects (Berg et al., 2007; Wallace & Tjan, 2011). Additionally, the size of the objects does not seem to affect crowding: small objects crowd each other as much as large objects do (Pelli et al., 2004). Hence, it was proposed that the distance between the centres of the objects is more relevant than the distance

between edges[1]. Another interesting characteristic of crowding, alluded to above, is that crowding occurs between similar objects but not dissimilar ones (Kooi et al., 1994; Kennedy & Whitaker, 2010). For example, a black letter is strongly crowded by other black letters, but less so by white letters or filled black circles. Finally, visual crowding displays various asymmetries. The most prominent of these asymmetries is the radial-tangential asymmetry: flankers that are in the radial direction (along the axis connecting the fovea and the target) lead to more interference than flankers that are in the tangential direction (Toet & Levi, 1992; Petrov & Meleshkevich, 2011).

Whereas visual crowding has been rigorously tested in humans over the past five decades (Bouma, 1970; Pelli et al., 2004), relatively little is known about crowding in DCNNs. In the first study, Volokitin et al. (2017) argued for the existence of crowding in DCNNs. However, their experiments do not conclusively establish crowding in DCNNs or test their similarity to humans, as their results might be explained by their method to achieve acuity loss, whereby the centres of stimuli are repeatedly sampled with increasingly higher resolution. That is, the models may have exhibited an unnatural preference to process the most central object, which reduced its ability to identify a flanked target. The models used in their research are small-scale and not capable of human-like performance, and might as such not reliably exhibit complex behaviour, such as crowding. Additionally, the methodology used in their research is different from most human crowding research. More recently, Doerig et al. (2019) showed that capsule networks (Sabour et al., 2017) combined with a grouping mechanism are capable of many human-like crowding effects. In addition, Doerig et al. (2020) argued that feedforward DCNN architectures are incapable of producing human-like crowding. To determine whether these claims hold under different conditions (different architectures and data), and to establish a conclusive and comparable picture of crowding in DCNNs, more research is needed.

In this paper we take various successful architectures of DCNNs, including ones that have been previously claimed to be comparable to the human visual system (Cichy et al., 2016; Güçlü & Gerven, 2015; Kheradpisheh et al., 2016), and investigate the the presence and characteristics of visual crowding using methodology inspired by human crowding research. We will assess the effect of the following on target identification:

- The distance between the target and the flankers
- The position of the target and the flankers
- The size and contrast polarity of the target and the flankers
- Different targets and flanker identities

The last two test the effect of similarity. To preview our results, we find that the strength of crowding, defined as flanker-induced reduction in target

[1] Although, there are several caveats to this 'law' (Herzog et al., 2015; Livne & Sagi, 2007).

identification, in DCNNs varies according to the kind of network. However, the results show a peculiar pattern that appears to be independent of the topology of the network. This pattern is in many ways dissimilar from that in humans. Finally, we discuss how these findings affect our understanding of object recognition in humans and DCNNs, and raise concerns that those employing DCNNs in object recognition tasks should keep in mind.

2. Methods

2.1. Models

We investigated three sets of DCNNs of increasing complexity (and chronology). First, we examined a network that has been widely claimed to possess characteristics similar to that of the human visual system (Cichy et al., 2017; Güçlü & Gerven, 2015; Kheradpisheh et al., 2016). That is, the various layers of this network are thought to capture the basic computational processes implemented by the layers of the primate visual system (from V1 to Infero-Temporal Cortex or IT). This network is a variant of the successful AlexNet (Krizhevsky et al., 2012) with 5 convolutional layers and 3 fully connected layers, followed by an activation layer.. We also investigated the VGG-16 network (Simonyan & Zisserman, 2014), which is a more successful 16-layer DCNN that uses small (3x3) filters and achieves a deeper network compared to other similar networks of its time. The family of VGG-networks achieved state-of-the-art or near state-of-the-art performance in 2014 image classification and localisation challenges. Finally, we also tested DenseNet-121 (Huang et al., 2016), a 121-layer DCNN that takes advantage of two recent advancements in deep learning: batch normalisation (Ioffe & Szegedy, 2015) and skip connections. The DenseNet-family of networks achieved state-of-the-art performance in many competitive image classification benchmarks while being parameter-efficient. While it is much more successful than the previous two networks, and has a much deeper architecture, it is important to note that the DenseNet-121 has fewer trainable parameters than the VGG-16. We tested these different architectures, and particularly the DenseNet-121, for two reasons. We wanted to test whether networks that are highly successful in recognising objects are in general susceptible to clutter, or if certain networks recognise objects in such a way that they are robust to flanker presence. Second, we wanted to test if networks considered to be similar to the primate visual system also show characteristics of humans, which include crowding. It has been claimed that even deep networks such as DenseNet and ResNet (He et al., 2015b), are comparable to the primate visual system. In fact, recent investigations demonstrate that such networks are superior to older networks such as AlexNet and VGG-16 in terms of correspondence to the primate system (Schrimpf et al., 2018). Hence, it is appropriate to test a range of networks to determine if they suffer from crowding.

The DenseNet was of particular interest to us, as it includes skip connections, which are also believed to be present in the human visual cortex (Essen & Maunsell, 1983). Here, a layer's feature maps are connected to the filters of all

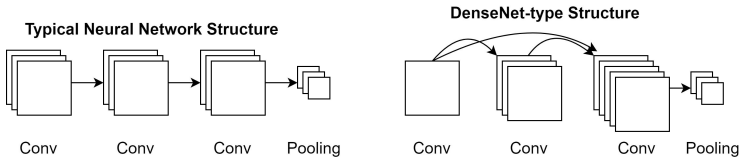

Figure 1: The skip connections of the DenseNet family of architectures. 'Conv' refers to a 2D Convolutional layer.

layers that follow it within a given 'dense block' (described below). For example, for n layers, layer 1's feature maps are connected to all layers' inputs up to the nth layer. This process is repeated for all n layers. DenseNet-121 implements this architecture within 'dense blocks', where a set of layers is densely connected (skip connected) to each other, and at the final layer of the block, the feature maps are pooled using max pooling.

In our research, we changed the rectified linear unit (ReLU) activations of the DenseNet and VGG to Leaky ReLU activations to avoid 'dying neurons' (neurons which do not allow a gradient to flow through them and end up in a perpetually inactive state) (Xu et al., 2015).

We focused our primary attention not on small (either in number of parameters or depth of layers) models, such as those tested by (Volokitin et al., 2017), as we wanted to investigate the behaviour of complex networks that have proved to be successful at identifying and categorising real-world images, in order to understand the patterns of crowding that could emerge from such networks. Additionally, while some have experimented with eccentricity-dependent models (Mnih et al., 2014), we limited the scope of our research to better-established DCNN classes.

2.2. Stimuli and Experimental Setup

Two types of stimuli were used in the experiments. The first type was images of places from the Places2 dataset (Zhang et al., 2017), which we will refer to as backgrounds. Two classes of backgrounds were used: ruins and neighbourhoods. We used these classes because they have relatively similar shapes, requiring the networks to construct more general types of filters that might mimic general scene recognition filters. With this approach we hope to avoid egregious overfitting of our next type of stimuli.

The second type of stimuli were uniform grey backgrounds with letters fixed in position, which we will call targets. These stimuli are akin to the stimuli used in psychophysical experiments on crowding (Bouma, 1970; Pelli et al., 2004). There were 8 different target letters: {A, B, C, E, G, M, Y, Q}, and each of them was considered a distinct class, making a total of 10 classes of training stimuli (8 letter image classes and 2 background image classes). We chose this set of letters because they are visually dissimilar from each other, which minimises the error rate, particularly when the acuity reduction procedure was applied to images (see Figure 2 *(a)*), which could have caused confusions between letters. The letters could be of either contrast polarity, near-white and near-black on

Figure 2: Example stimuli. *(a)* shows acuity reduction in images. Acuity is reduced logarithmically between values of acuity = 1 and acuity = 0.2 with linear distance from the centre of the image in 20 steps. *(b)* shows a full acuity letter stimulus with the target letter A, and pair flankers B.

a grey background, and one of two sizes, 20 and 26 points. All stimuli were 224x224 pixels. Each network was trained on these 10 classes of images. When trained on letters, a single letter was presented 56 pixels to the left of the centre of the image, that is, midway between the centre and the left border of the stimulus along the horizontal meridian. Similarly, during testing, a target letter was presented at the location it was trained at. It was flanked by one letter or a pair of letters. When two letters were presented, one letter was placed on each side of the target and were identical to each other. The flankers were selected from a set that included all target letters and two additional letters: {S, H}. The pair of flankers were placed diametrically opposite each other on either side of the target. Each pair of flankers was tested at 10 angular locations around the target, each location separated by 18 degrees of rotation, thus covering the entire region around the target. The centre-to-centre distance between a target and each flanker ranged from 25 to 45 pixels in 2-pixel increments. All combinations of target and flanker letters, contrast polarities and sizes were tested. In total, we tested 70,400 combinations of flankers and targets in each experiment. In experiments where we tested the effect of single flankers, the number of tested combinations doubled (20 angular locations instead of 10).

To study crowding in DCCNs, we wished to model human peripheral vision. This is because crowding in humans occurs most noticeably away from the fovea in peripheral vision, where visual acuity and resolution is much lower than in the centre of the visual field. We wanted to provide the DCNNs the same sort of input as the human visual system would receive. Peripheral input is impoverished relative to central input. To model peripheral vision, we used well established relationships in humans regarding acuity and eccentricity (Anstis, 1974) and reduced acuity logarithmically with distance from the centre of the image in 20 steps, with 1 being full acuity in the centre of the image, and 0.2 being the lowest acuity at the edges of the image. We first took 20 copies of the image and assigned each a value on a logarithmic scale, ranging from 0.2 to 1. We then down-sampled each image by their assigned value, and up-sampled them to their original size using the nearest neighbour algorithm. Finally, we cropped and overlaid the images on top of each other to form a 20-step gradient of acuity reduction (see Figure 2 *(a)* for an example). We did this to strictly lose information, as crowding in humans is not simply blur (Song et al., 2014).

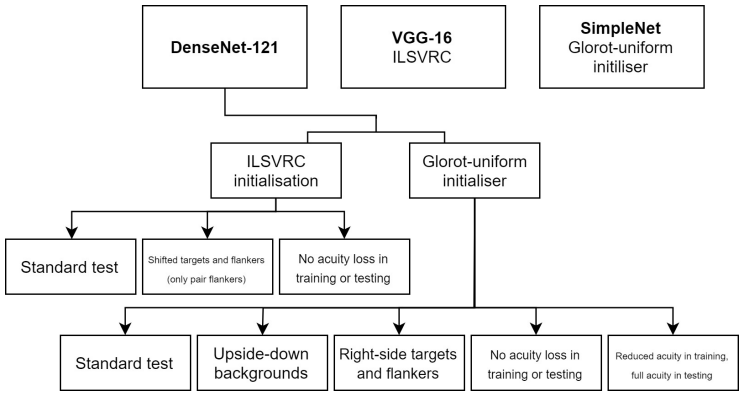

Figure 3: The range of experiments conducted in this study.

We would like to emphasise that training was done on two kinds of backgrounds and 8 target unflanked letters presented in isolation; flankers were introduced only in the testing stage. Model base performance was evaluated on the set of target letters, and a separate set of validation backgrounds.

2.3. Training

All models for all experiments were trained for 24 hours[2] on an NVIDIA Tesla K40c GPU using the Keras library (Chollet et al., 2015). The ADAM-optimiser (Kingma & Ba, 2014) was used with a learning rate of 0.01. Both random initialisation[3] and ImageNet Large Scale Visual Recognition Competition (ILSVRC) initialisation of weights[4] were tested on the DenseNet, random initialisation was tested on the Alexnet and ILSVRC initialisation on the VGG-16. Random initialisation allowed us to test the network's characteristics and performance in the absence of influence from outside sources on the system and controlled for the possibility that any results may have been caused by ILSVRC initialisation of weights. Initialising the network with ILSVRC weights allowed us to mimic the types of environments humans are subjected to on a regular basis in addition to testing an already trained network that has been shown to be successful in image categorisation and identification. It is important to note, however, that the ILSVRC weights had been trained without acuity loss, while our training and testing was primarily conducted on stimuli that had been reduced in acuity. When initialising the network with ILSVRC weights, the following procedure for

[2]This corresponds to roughly 40 epochs on the DenseNet, 80 epochs on the VGG and 200 epochs on the Alexnet. Although this is an arbitrary time limit, given our configuration these models were run for a sufficient number of epochs to enable good recognition performance, similar to what has been implemented in earlier studies (e.g., Simonyan & Zisserman (2014).

[3]As a random initialiser, we use the Glorot-uniform initialiser (Glorot & Bengio, 2010).

[4]ILSVRC initialisation of weights refers to initial weights of the neural network as being set to the weights optimised for the ImageNet Large Scale Visual Recognition Challenge classification task (See Keras documentation; Chollet et al. (2015)).

training was taken to allow stable training and avoid 'gradient nuking'[5] in the upper layers of the network:

1. Freeze all layers above the last one, initialise learning rate = 0.01.
2. When validation loss does not decrease for 2 epochs, open the next layer for training and reduce learning rate by 10^{-2}.
3. When validation loss does not decrease for 2 epochs, open all layers for training and reduce learning rate by 10^{-2}.
4. Training is completed after a total of 24 hours.

3. Results

In our experiments, we did not train the network to recognise targets in the presence of flankers, or letters in the locations where flankers were later placed. Our goal was to present the targets to the models in a specific part of the image, such that it learns to recognise it. We then tested its performance in the presence of flankers[6]. Human crowding has been attributed to either confusing a fully identified flanker for a target or to combining or pooling the features of both the target and flankers (Strasburger & Malania, 2013; Hanus & Vul, 2013). We show that in our experiments, the former does not occur in DCNNs (Section 3.9).

In Figures 4-11, and Supplementary Figures S12-S18, panel *(a)* plots accuracy as a function of target-flanker spacing in pixels, collapsed across all stimulus manipulations. In addition, we show unflanked accuracy, which is near-perfect for almost all experiments. Panel *(b)* plots accuracy for each target-flanker colour combination for both sizes collapsed over all the target-flanker distances. For example, *W/B 20* denotes a white target, a black flanker, with letter size 20 points. Additionally, collapsed data when the letters S and H are excluded is shown. The 95% confidence interval is shown for each data point. Panel *(c)* plots the shape of crowding—accuracy at each position of the flanker, where the origin of the plot is centred on the target, collapsed over all size and contrast polarity combinations. Accuracy is shown with all flankers, accuracy with the flankers S and H excluded, and accuracy using only the flankers S and H. Separated effects of the the letters S and H are shown as they were not a part of training, and therefore serve as 'novel' flankers.

In general, letter recognition performance improved with target-flanker distance as expected from human studies, indicating that networks experience at least some form of crowding. However, unlike in humans, this trend was mild, and we also observed peculiar patterns in many of our experiments. We call these peculiar patterns *anomalies of crowding*, or simply anomalies. An anomaly usually took the form of an unexpected change in performance (e.g. poor accuracy at large target-flanker spacing and better accuracy at short spacing

[5] When using weights optimised for a specific task (e.g. ILSVRC), using them for a different task may cause large gradient updates in the final layers of the network which can cause large changes in the weights of the layers above them.

[6] Full data-frames of results are available at github.com/benlonnqvist/CNNCrowding.

for specific target-flanker configurations). These anomalies were found to be caused primarily by the untrained letters S and H as flankers. However, even after these letters were excluded from analysis, such anomalies persisted. Our findings suggest that only by training and testing a model several times, and by averaging results, can such anomalies be mitigated. Also unlike in humans, the current results showed a strong pattern of crowding along the top-left – bottom-right diagonal in all tests with paired flankers. Interestingly, throughout our experiments no clear pattern of the effect of size or contrast polarity was found. Many models performed better for letter size 20 than for 26, but some exhibited the opposite behaviour. In humans, size has no effect on the strength or extent of crowding, and colour (or similarity) has a strong effect, with different colour flankers causing less crowding than same colour flankers (see Section 1).

3.1. Alexnet with random initialisation

We trained five independent Alexnets with unaltered images (no 'acuity loss') to test the sensitivity of small convolutional networks to different flanker configurations. The primary reason we tested Alexnet was because of the claimed correspondence between such networks and the primate visual system (Cichy et al., 2017; Güçlü & Gerven, 2015; Kheradpisheh et al., 2016). If the two architectures (DCNNs and biological systems) achieve object recognition in comparable ways, then we should see evidence of human-like crowding in Alexnet. We found that while the networks learned to perform the letter identification task with high accuracy, they suffered greatly from flanker presence. In other words, such networks do suffer from crowding: flankers substantially degrade target identification performance. However, there are noticeable differences between the crowding observed in humans and in the Alexnets.

Unflanked targets were identified with high accuracy (96.97%), but the presence of a single flanker even at at a large distance from the flanker reduced performance substantially (flanked performance was 35% or lower). In contrast, crowding in humans is quite weak in the presence of a single flanker and is rapidly alleviated with spacing between the target and the flanker (Petrov & Meleshkevich, 2011). However, for Alexnet, the overall reduction was dramatic with hardly any improvement with spacing. In fact, extrapolating from the data, the target-flanker distance at which there would be no crowding (where performance is the same as in the unflanked condition) would be 218 pixels, which is approximately the entire width of the image. That is, the model is strongly crowded at almost all distances. In addition, when the flankers presented to the model were untrained (the letters S and H), the pattern of crowding became unpredictable; at certain angular locations, flankers further away caused more crowding than those closer. These are examples of anomalies of crowding, described above. This effect does not occur in humans (Huckauf et al., 1999)[7]. These results indicate that DCNNs suffer from crowding in the periphery, that

[7]However, as mentioned above, we believe that it is possible that such anomalies disappear entirely when a model is trained a large number of times and the results averaged.

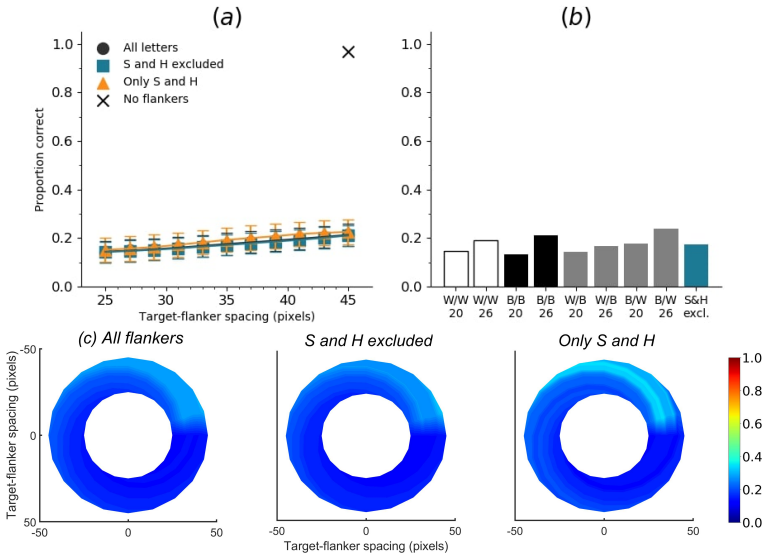

Figure 4: Accuracy of letter identification of the randomly initialised small 5-layer convolutional network with single flankers ('Alexnet'). A total of five independent training sessions and test sessions are averaged in this figure. Results of individual runs are shown in Supplementary Figures S1-S5. Training and testing of all five models in this figure was done without unaltered images. Average accuracy without flankers was 96.97%, as shown in panel *(a)*. 95% confidence intervals are shown for each data point. The letters S and H are excluded in *(c) middle* and shown exclusively in *(c) right*, as the network had not seen the letters in testing. We also tested the Alexnets with acuity-reduced stimuli—Figure S6 shows the averaged results, and Figures S7-S11 show individual models' results. We found that accuracy reached near-chance levels (12.5%), and as such little can be inferred from these results.

they suffer from crowding up to a much greater distance than humans, and that the effect of target-flanker spacing is weaker than in humans, at least in the range of distances we tested. We attempted to fit psychometric curves (see the Appendix, Figure S13), but the fits were unsuccessful in producing meaningful results, primarily due to the anomalies and a lack of a clear upper asymptote. We also note that all five instantiations of Alexnet displayed the same pattern of crowding, indicating that these findings were reproducible and not an artefact of the initial settings.

3.2. VGG-16 with ILSVRC initialisation

We also trained a different architecture of network, the VGG-16 (Simonyan & Zisserman, 2014), to test whether our results are specific to the Alexnet (and to the DenseNet-121, see below) architecture. We found that while the VGG-16 performed somewhat better in our task (Figure 5), it exhibited the same general patterns and behaviour of crowding as the Alexnet. This implies that the presence and characteristics of crowding, and by implication, object recognition in DCNNs, is a property of the basic building blocks of DCNNs and not caused by a particular network architecture. Note that in our study VGG-16 was initialised in a completely different way compared to the Alexnets. Yet, the pattern was the same, with slightly better robustness to flankers[8]. The VGG-16 with ILSVRC initialisation can be considered to be more 'experienced' with visual stimuli. However, both VGG-16 and our Alexnet were highly sensitive to the presence of clutter and were insensitive to a large extent to the spacing between the target and the clutter.

3.3. DenseNet-121 with random initialisation

The previous two architectures of models we have tested contained only simple combinations of convolutional, max-pooling, and densely connected layers. To test whether our results are specific to such configurations, or apply more generally to more sophisticated architectures, we tested the DenseNet-121 (Huang et al., 2016), a recent architecture that takes advantage of batch normalisation and skip connections. Further, as noted above, DenseNets and ResNest (from which DenseNets are derived) have been argued to have a higher correspondance to the primate visual system than earlier networks such as AlexNet and VGG (Schrimpf et al., 2018; nkr, 2017). We trained and tested the DenseNet network extensively under various network and stimulus configurations in order to assess if a highly successful model suffers from crowding and if this crowding is comparable to that in humans, given the claimed correspondence.

Figure 6 shows the results when the network was initialised with random weights and stimuli were degraded to match perceptual input to the human visual system. We found that the DenseNet-121 is much more robust to clutter than the VGG-16 or the Alexnet. The presence of a single flanker reduces target

[8]We attempted two runs of this model, but only one converged.

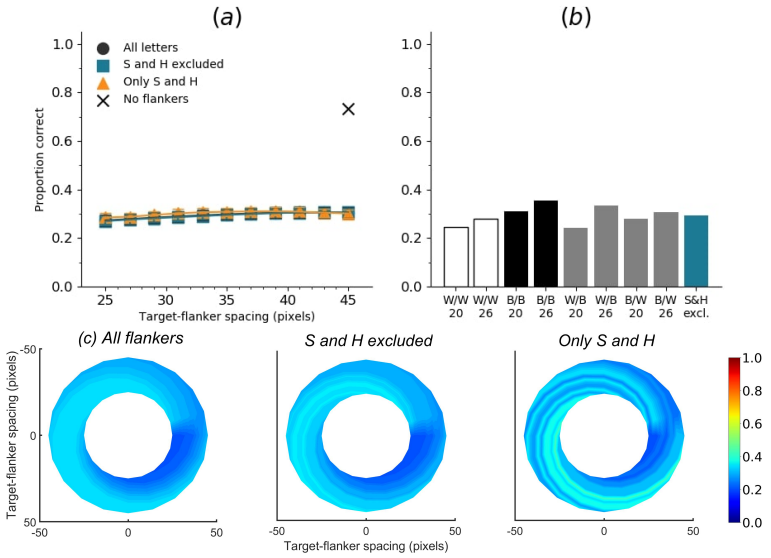

Figure 5: Accuracy of letter identification for the ILSVRC-initialised VGG-16 model with single flankers. Even though the accuracy for unflanked targets in the VGG-16 was lower than for the Alexnets, it appears to be more robust to clutter; that is, flanked performance is higher. Nevertheless, the general pattern of crowding remains the same. VGG-16 accuracy without flankers was 73.46%.

identification performance, but the drop is not dramatic. Further, increasing the spacing of the flanker from the target ameliorates crowding to the extent that far flankers do not interfere with target identification. The performance of this model is reminiscent of human performance. Interestingly, however, as with the Alexnet and VGG-16, the strongest interference by a single flanker is not where its acuity is the lowest (the outermost position on the left of the background image along the horizontal meridian), but instead remains on the bottom-diagonal of the target towards the centre of the image. Note that in humans, the strongest interference is observed when the flanker is placed at this outermost location, and not by a flanker closer to the centre of the image (Petrov & Meleshkevich, 2011).

To determine if the higher interference by a flanker placed along the top-left to bottom-right diagonal and close to the image centre was an artefact of the stimuli used and the training procedure, we trained a new network with the same parameters as before but with the letter stimuli presented on the right side instead of the left. As can be seen in Figure 7, the shape of crowding flips across the vertical axis. That is, it is not the absolute top-left to bottom-right axis that matters, but the presence of a flanker close to the centre of the image, but in the lower visual field that causes the greatest disruption. To replicate these findings, we trained a new model using identical configuration (Figure S19). The general characteristics of crowding in this model remained the same, but it appears that

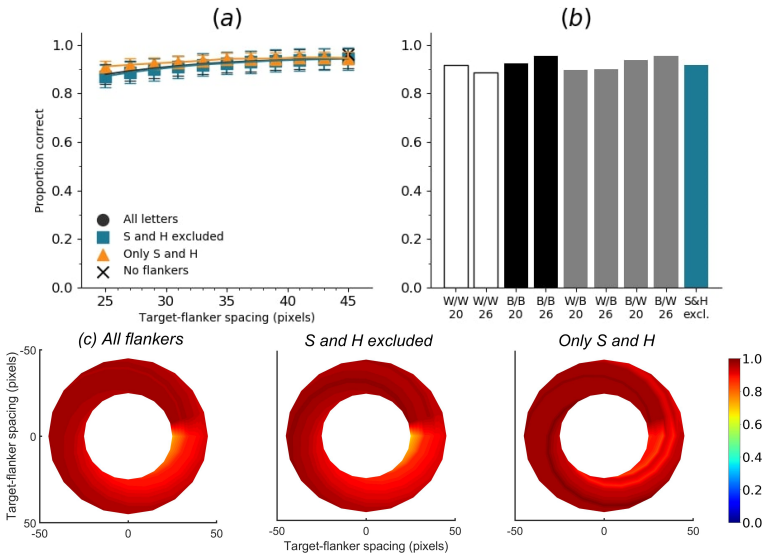

Figure 6: Accuracy of letter identification of the randomly initialised DenseNet-121 in the presence of single flankers. Training and testing was done with acuity loss. The DenseNet-121 is much more robust to the presence of flankers than the VGG-16 or Alexnet models. Accuracy without flankers was 96.11%.

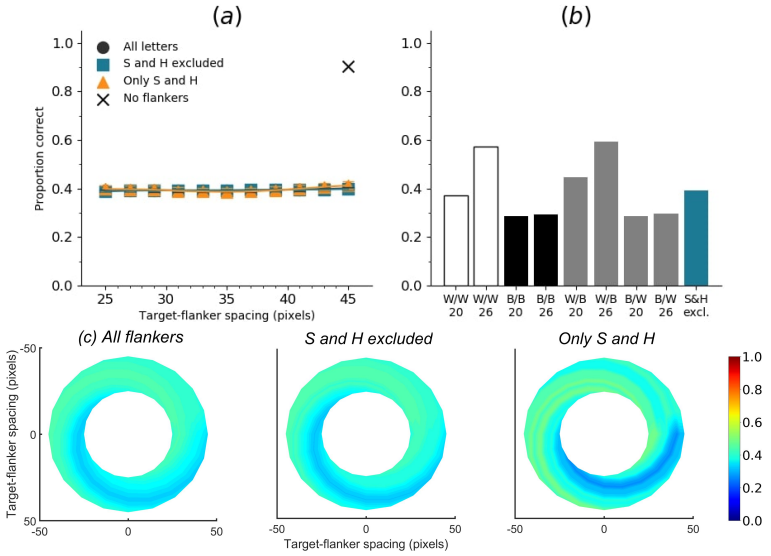

Figure 7: Accuracy of letter identification for the DenseNet-121 with single flankers when the target and flankers are placed on the right-hand side of the image, instead of the left-hand side. Training and testing was done with acuity loss. The area of most crowding on average shifts to the left-hand side of the target, towards the centre of the image, showing evidence that a higher acuity flanker will crowd the target more than a lower acuity flanker. Additionally, distance has little effect on crowding. Model accuracy without flankers was 90.37%.

this instantiation of the model (see Figure S19) is more robust to clutter and exhibits minimally reduced performance with increased target-flanker spacing. The reason for this discrepancy is unclear to us. Importantly, however, it is clear from these models and models trained with unaltered images that the pattern of crowding remains the same even with large changes in network and stimulus characteristics, even if the magnitude changes. This magnitude difference can be partially attributed to the image manipulations ('acuity loss'), but not the pattern of results.

In addition, we trained two DenseNet-121 networks without acuity loss (Figures S16 and S17). The results demonstrate that our acuity reduction procedure does not appear to change the pattern of crowding in the networks, but only its magnitude. The results also suggest that, as with AlexNets, the anomalies of crowding appear to be caused by individual networks, and that the anomalies may not persist when training several models and combining the results.

3.4. DenseNet-121 with random and ILSVRC initialisations with paired flankers

Psychophysical experiments in humans on crowding are often performed with a pair of flankers, one on either side of the target, rather than a single one (e.g., Bouma (1970), Freeman et al. (2012)). Hence, we also tested the DenseNet-121 with paired flankers. In these experiments we trained the DenseNet-121 initialised with random and ILSVRC weights, separately. Results are shown in Figures 8 and S12, respectively. We found that the bottom-right flanker that dominates crowding in single-flanker experiments causes the general pattern of crowding to replicate across the horizontal axis (along the top-left to bottom-right axis). It is interesting to note that the model is crowded more by paired flankers at all distances than by single flankers, and does not reach near-unflanked accuracy even at the furthest target-flanker distance. In humans, paired flankers are more effective in interfering with performance than single flankers, and have a larger range of interference. That is, they are more effective even at larger distances. DenseNet-121 appears to mirror that characteristic. Nevertheless, in humans, crowding is eliminated, under similar circumstances, if the distance between the target and the flankers is greater than half the target eccentricity (the distance between the centre of the image and the target), as codified in the 'Bouma Law' (Pelli & Tillman, 2008). This is equivalent to a spacing of about 29 pixels in our setup. That is, performance should be the same as in the unflanked condition if flankers are separated from the target by about 29 pixels. This is not the case with the DenseNet model.

3.5. Effect of acuity loss manipulation

As the DenseNet-121 was more robust to flanker interference, we trained and tested the DenseNet-121 with the same hyperparameters on images that had not been reduced in acuity (unaltered images). We found that the general shape of crowding remained the same in all tests but one (pair flankers with ILSVRC initialisation: Figure S15), and barring that experiment the effect of flankers was

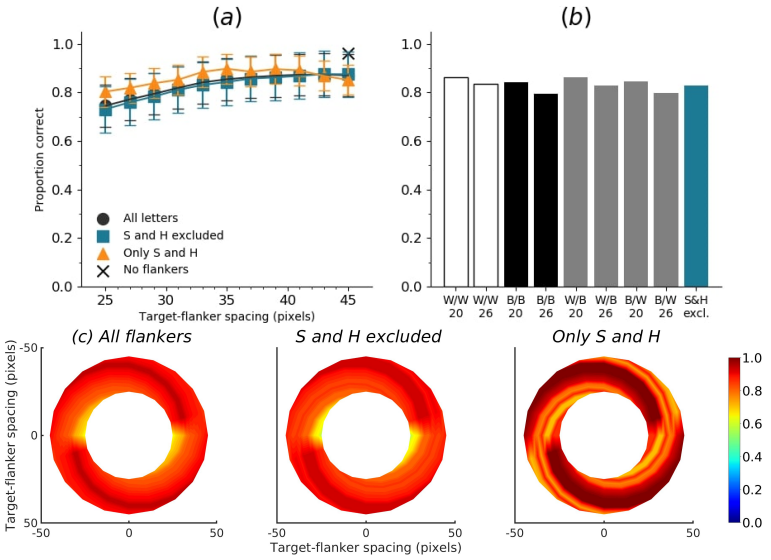

Figure 8: Accuracy of letter identification for the randomly initialised DenseNet-121 tested on paired flankers. The model was trained and tested with acuity loss, and its accuracy without flankers is 96.11%.

dramatically reduced. This suggests that the image manipulation cannot explain the pattern of our results, apart from its magnitude. These results also point to the proposal that the pattern of results observed here is inherent to DCNNs.

In the special case of ILSVRC initialisation with a pair of flankers, the performance was much lower than expected (roughly 40%), whereas for most other experiments with acuity loss this ranged from 60-85%. This strange behaviour may have been caused by differences in convergence of the network. In addition to poor performance in the test, the axis of crowding flipped compared to all other experiments. These results closely mimic the effects seen in the Alexnet tests; robustness to clutter of the DenseNet is higher, and regardless of whether the acuity loss procedure is used, the general characteristics of crowding remain.

We found that while using unaltered images in training and testing can lead to some unpredictable results, such as massive performance drops or improvements with flankers, the general shapes of crowding tended to stay the same. We also found that in the case of experiments trained and tested with unaltered image, anomalies of crowding, described in Section 3.4, largely disappeared when the flankers S and H were excluded from analysis.

Finally, we tested the randomly initialised DenseNet that was trained with acuity loss to see how behaviour changes when the network gains access to full acuity without additional training. Results are shown in Figure 10.

We found that there was no large difference in the general characteristics of crowding regardless of whether the network had reduced acuity during training

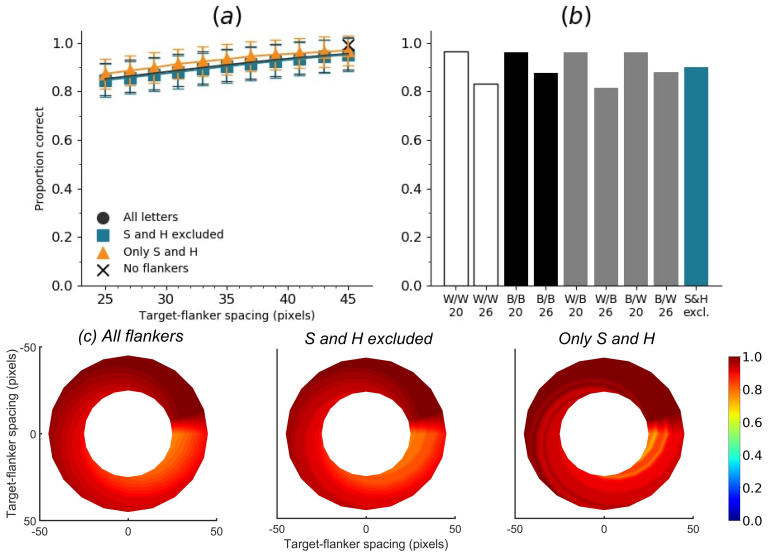

Figure 9: Accuracy of letter identification of the DenseNet-121 for stimuli that were not reduced in acuity. ILSVRC initialisation with single flankers. Testing and training were done without acuity loss. Model accuracy without flankers was 99.34%.

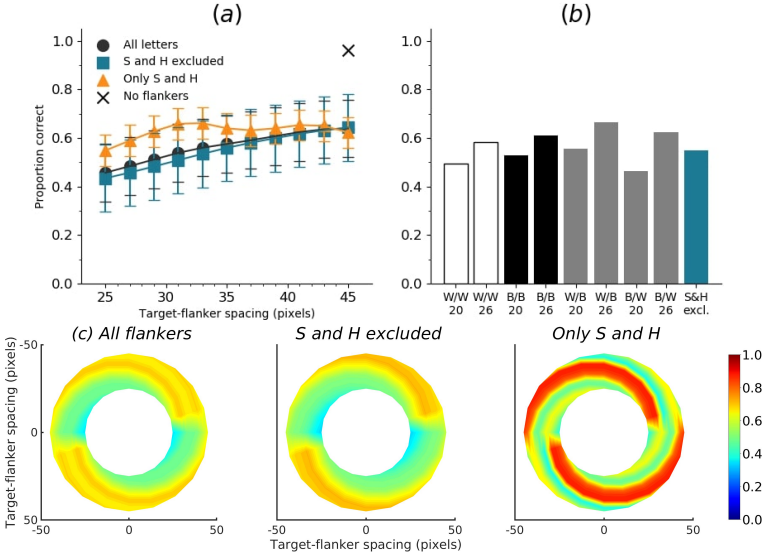

Figure 10: Accuracy of letter identification of the DenseNet-121 with randomly initialised weights trained with acuity loss and tested with full acuity stimuli. The network was shown pair flankers and did not exhibit a large change in behaviour with access to full acuity. Model accuracy without flankers was 96.11%.

or during testing. Acuity affected crowding primarily in magnitude, but not in shape or general characteristics, such as the effect of distance on crowding, or the effect of similarity (contrast polarity and size) between the target and flankers.

3.6. Effect of the amount of useful information in a local region

Because our models suffered from crowding to a greater degree in the lower half of the images than in the upper half, we tested whether flipping our background images vertically in training would also flip crowding vertically. It is possible that natural images have diagnostic information in the lower visual field and the network is more sensitive to clutter in that part of the visual field. Results are shown in Figure 11. We found that a relatively large portion of crowding does shift to the upper half of the image, practically equalising the amount of crowding on both halves of the image (59.66% accuracy on the top-half, 59.42% accuracy on the bottom-half). This suggests that the amount of useful information in local regions of a stimulus plays a contributing role in crowding in DCNNs. This effect is the opposite of what is observed in humans. Humans have greater attentional resolution and lower crowding in the lower half of the visual field (Intriligator & Cavanagh, 2001). Our DCNN models do not. We hypothesise that the networks developed to have greater preference for regions with a higher density of useful information for classification and hence flankers placed in such locations caused more crowding. In other words, the data used to train the networks likely contributes to this effect.

It is interesting to note that while in the randomly initialised single-flanker model with upright background images the models exhibited a greater degree of crowding in the lower portion of the image (89.47% accuracy in the lower half, 95.31% in the upper half), this effect was not entirely reversed when the background images were vertically flipped; the accuracy between the two halves of the stimuli only equalises. We are unable to explain this phenomenon.

3.7. Radial-tangential asymmetry

We examined if the networks demonstrate the radial-tangential asymmetry, where flankers along the radial direction (flankers along the axis connecting the target to the centre of the image) are more influential than tangential flankers. We plotted the accuracy of identification when it was surrounded by the nearest letters along these two axes (Figure 12). We found that crowding is asymmetric in the expected direction—radial flankers crowd more than tangential flankers do, like in humans (Toet & Levi, 1992; Petrov & Meleshkevich, 2011). We also find that if the overall accuracy is higher, crowding is more asymmetric than if it is lower.

We also note, as described above, that the networks demonstrate an in-out asymmetry, where there is a difference in effect of the 'inner' flanker, or flanker closest to the image centre and 'outer' flanker, or the farthest flanker, The inner flanker appears to be more powerful in disrupting target identification than the outer one. However, the effect is the opposite of what is observed among humans, where the outer flanker is more influential than the inner one.

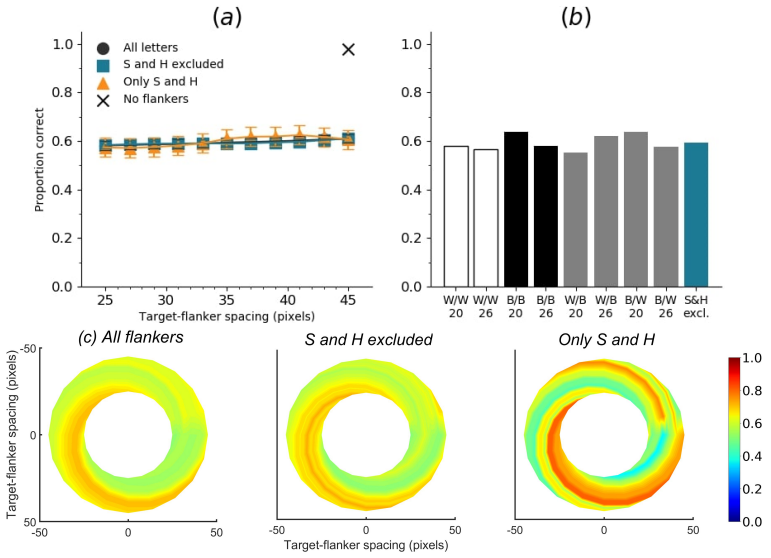

Figure 11: Accuracy of letter identification for a randomly initialised DenseNet-121 with background stimuli vertically flipped. Training and testing were done with acuity loss. We found that the degree of crowding decreases with distance less than in most of our experiments. Model accuracy without flankers was 97.98%.

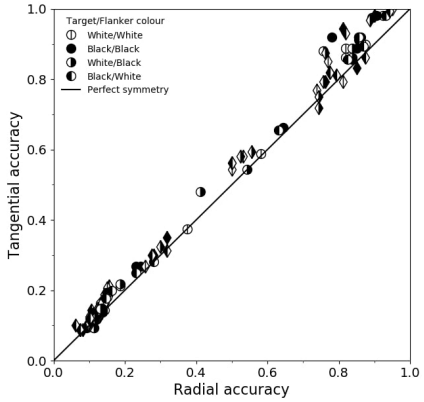

Figure 12: Radial-tangential accuracy for all single-flanker models. Only the horizontal and vertical flankers at 25px centre-centre distance from the target are plotted. Circle-shaped markers denote the letter size 20, and diamond markers denote the letter size 26. The asymmetry is relatively mild—in humans at certain distances effects have greater magnitude.

3.8. Effect of size and contrast polarity

Among humans, contrast polarity, and other cues of similarity, strongly modulates target identification. Objects similar to each other crowd more than dissimilar objects (Kooi et al., 1994; Kennedy & Whitaker, 2010). On the other hand, crowding is not sensitive to object size (Pelli et al., 2004). The strength and extent of crowding is comparable for objects of different sizes. Here, we examined if we observe the same pattern of results. In fact, we found no clear effects of size or contrast polarity. Performance was not higher for different polarity letters than it was for same polarity letters (for example, see Figure S3 and Figure S5). However, they showed differences in performance for the two letter sizes. Many networks identified smaller letters more successfully than larger letters, however, the opposite pattern was noticeable in other networks.

Some networks, such as the DenseNet-121 with random initialisation (Figure 6 were poor at detecting black targets, but were not modulated by target-flanker similarity. Other networks, such as the DenseNet-121 with random initialisation and with the targets and flankers on the right side (Figure S19) had clear difficulty distinguishing letters of a specific size and colour combination compared to the others. In general, we found no discernible patterns that applied to all models (even within a specific architecture).

3.9. Confusion between targets and flankers

Finally, we analysed the DenseNets' and VGG's reported output to examine whether targets were confused with flankers more often than they were confused with other letters. We found that for all single-flanker results there is little difference between the model reporting another target 'at random' and the model reporting the flanker letter. On error trials, the flanker is reported 0.0125 percentage points more often than any other single letter, on average.

This finding implies that in our experiments, the DCNNs were highly sensitive to the position of the target and that they were not prone to confuse the flanker as the target. This also rules out the hypothesis that flanker substitution contributes to crowding in DCNNs when DCNNs are trained in a simplistic manner, like it does in humans (Freeman et al., 2012; Hanus & Vul, 2013).

4. Discussion

We investigated crowding in DCNNs and found that they follow a predictable pattern regardless of network topology, size or colour of flankers, or whether images have been reduced in acuity. Overall, we do find that flankers reduce target identification performance, demonstrating that all the networks we tested suffer from crowding. On the other hand, importantly, we found that object recognition in humans has distinctly different characteristics from those exhibited by DCNNs. The pattern of crowding found follows a combination of several factors:

- **Robustness to flanker interference:** We found that the Alexnet and VGG-16 models were much more susceptible to flanker interference than the DenseNet-121. This suggests that some characteristics unique to the DenseNet (e.g., skip connections, batch normalization, or simply the prsence of more layers) causes the model to be more robust to clutter.

- **Distance of the flanker from the target:** In almost all experiments recognition performance for a target surrounded by known flankers strictly follows a positive relationship with distance between them. This suggests that crowding is, at least in part, caused by local pooling of information. This relationship is mild, however, and is in two models reversed (Figures S19 and S15).

- **Flanker substitution versus pooling:** Flankers near the image centre (in the lower visual field) cause more crowding than 'outer' ones, for a given spacing. These letters tend to be less subject to image degradation in our acuity loss manipulation. We suspect that this asymmetry in crowding is therefore due to local pooling, as suggested by Volokitin et al. (2017). In our experiments we also found that target-flanker confusion does not contribute much to crowding, further supporting the hypothesis that local pooling causes crowding in DCNNs under simplistic training regiments. This reason may partly explain why there is more crowding with more foveal flankers than peripheral flankers, unlike in humans (Petrov & Meleshkevich, 2011; Petrov et al., 2007). A caveat to keep in mind is that we also found that acuity loss does not drastically change the patterns of crowding, but instead its magnitude.

- **Amount of useful information in stimuli:** The bottom-corner position of the flanker towards the centre of the image caused most crowding in our experiments. In humans, the bottom-half of the visual field has higher resolution and lower crowding (Intriligator & Cavanagh, 2001). The images of ruins and neighbourhoods we used in training and testing have a sizeable portion of their top-half contain "useless" information, possibly contributing to this effect. Additionally, when the backgrounds were vertically flipped, this bias towards the bottom-half of the image was neutralised. Further support for this argument is given by the fact that in our experiments, the ILSVRC-initialised models were subject to a higher degree of crowding. We suspect that this effect is primarily caused by the training data.

- **Unrecognised clutter:** When the networks are subject to flankers they do not recognise, these flankers cause effects that are unpredictable in terms of classification of the target in individual models. Often these stimuli cause a reduction in accuracy in positions and distances which do not follow a clear pattern. However, these effects may be mitigated by training and testing a model several times, and averaging results.

We also observed other dissimilarities in machine and human crowding. In many of our experiments, we found differences in the degree of crowding

with differently-sized letters, violating the Bouma Law (Pelli & Tillman, 2008). Additionally, black letters are not crowded more by other black letters than they are by white letters, and vice versa. In humans, this effect is clear (Kooi et al., 1994; Kennedy & Whitaker, 2010). Despite these differences, crowding in DCNNs and humans share some similarities. For example, the degree of crowding in both DCNNs and humans decreases with increased spacing between a target and its flankers (Bouma, 1970; Toet & Levi, 1992; Pelli et al., 2004). The radial-tangential asymmetry also shares a resemblance with human crowding asymmetry, with radial flankers crowding the target more (Toet & Levi, 1992; Petrov & Meleshkevich, 2011).

We conclude that crowding is present in DCNNs regardless of whether a network is trained on unaltered images or acuity reduced input, and that its magnitude can be reduced by employing a more sophisticated architecture that does not rely only on convolutional, max pooling and densely connected layers. Based on the current evidence, we conjecture that local pooling is the primary source of crowding in DCNNs, and that the position in which crowding occurs is caused by the data the network has been subject to in training. As such, we suggest those who train networks to use data augmentation (Perez & Wang, 2017) in order to minimise the effect of crowding.

While DCNNs are loosely based on human models of object recognition, and have indeed been considered comparable, they exhibit patterns of behaviour that are substantially different from those in humans. At first glance, both demonstrate flanker induced interference. However, a closer look shows a myriad of differences. We suggest that these differences in behaviour of object recognition between humans and DCNNs are caused by one or several of many neural differences. For example, in the human visual cortex there are many different types of neurons which serve different purposes. The presented DCNNs also do not use recurrent connections—in the human visual cortex, there are many recurrent connections, and these recurrent connections contribute enormously to visual processing (Bullier et al., 2001). For example, Doerig et al. (2019) was able to reproduce many characteristics of human crowding using capsule networks, which utilise recurrent connections. Others have also shown that adding recurrence to deep neural networks may make them more human-like (Kar et al., 2019; Kietzmann et al., 2019; Linsley et al., 2018; Spoerer et al., 2017; Tang et al., 2018). The way in which the human visual system and DCNNs are built are fundamentally different, and our experiments show that they exhibit fundamentally different behaviour in object recognition tasks.

Acknowledgements

We would like to acknowledge the use of a Tesla K40 GPU card that has been donated to Dr M. S. Baptista by Nvidia. We would also like to thank Dr Micha Elsner for helpful discussions.

(2017). Do coarser spatial patterns represent coarser categories in visual cortex? `https://nikokriegeskorte.org/2017/10/19/do-coarser-spatial-`

patterns-represent-coarser-categories-in-visual-cortex/. Accessed: 2020-02-10.

Anstis, S. M. (1974). Letter: A chart demonstrating variations in acuity with retinal position. *Vision Research*, 14(7), 589–592.

Berg, R. v. d., Roerdink, J. B. T. M., & Cornelissen, F. W. (2007). On the generality of crowding: Visual crowding in size, saturation, and hue compared to orientation. *Journal of Vision*, 7(2), 14–14.

Bonner, M. F. & Epstein, R. A. (2017). Coding of navigational affordances in the human visual system. *Proceedings of the National Academy of Sciences*, (pp. 201618228).

Bouma, H. (1970). Interaction Effects in Parafoveal Letter Recognition. *Nature*, 226(5241), 177–178.

Bullier, J., Hupé, J. M., James, A. C., & Girard, P. (2001). The role of feedback connections in shaping the responses of visual cortical neurons. *Progress in Brain Research*, 134, 193–204.

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., & DiCarlo, J. J. (2014). Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLOS Computational Biology*, 10(12), e1003963.

Chollet, F. et al. (2015). Keras. `https://keras.io`.

Cichy, R. M., Khosla, A., Pantazis, D., & Oliva, A. (2017). Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. *NeuroImage*, 153, 346–358.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6(1), 1–13.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, 73(3), 415–434.

Doerig, A., Bornet, A., Choung, O. H., & Herzog, M. H. (2020). Crowding reveals fundamental differences in local vs. global processing in humans and machines. *Vision Research*, 167, 39 – 45.

Doerig, A., Schmittwilken, L., Sayim, B., Manassi, M., & Herzog, M. H. (2019). Capsule networks as recurrent models of grouping and segmentation. bioRxiv pre-print.

Essen, D. C. V. & Maunsell, J. H. R. (1983). Hierarchical organization and functional streams in the visual cortex. *Trends in Neurosciences*, 6, 370–375.

Freeman, J., Chakravarthi, R., & Pelli, D. G. (2012). Substitution and pooling in crowding. *Attention, Perception & Psychophysics*, 74(2), 379–396.

Glorot, X. & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In Y. W. Teh & M. Titterington (Eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research* (pp. 249–256). Chia Laguna Resort, Sardinia, Italy: PMLR.

Güçlü, U. & Gerven, M. A. J. v. (2015). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *Journal of Neuroscience*, 35(27), 10005–10014.

Hanus, D. & Vul, E. (2013). Quantifying error distributions in crowding. *Journal of vision*, 13(4), 17.

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

He, K., Zhang, X., Ren, S., & Sun, J. (2015b). Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

Herzog, M. H., Sayim, B., Chicherov, V., & Manassi, M. (2015). Crowding, grouping, and object recognition: A matter of appearance. *Journal of Vision*, 15(6).

Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2016). Densely Connected Convolutional Networks. *arXiv:1608.06993 [cs]*. arXiv: 1608.06993.

Huckauf, A., Heller, D., & Nazir, T. A. (1999). Lateral masking: Limitations of the feature interaction account. *Perception & Psychophysics*, 61(1), 177–189.

Intriligator, J. & Cavanagh, P. (2001). The spatial resolution of visual attention. *Cognitive Psychology*, 43(3), 171–216.

Ioffe, S. & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. arXiv: 1502.03167 [cs.LG].

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. 22(6), 974–983.

Kennedy, G. J. & Whitaker, D. (2010). The chromatic selectivity of visual crowding. *Journal of Vision*, 10(6), 15.

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

Kheradpisheh, S. R., Ghodrati, M., Ganjtabesh, M., & Masquelier, T. (2016). Deep Networks Can Resemble Human Feed-forward Vision in Invariant Object Recognition. *Scientific Reports*, 6, 32672.

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K. A., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, 116(43), 21854–21863.

Kingma, D. P. & Ba, J. (2014). Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*. arXiv: 1412.6980.

Kooi, F. L., Toet, A., Tripathy, S. P., & Levi, D. M. (1994). The effect of similarity and duration on spatial interaction in peripheral vision. *Spatial Vision*, 8(2), 255–279.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12 (pp. 1097–1105). USA: Curran Associates Inc.

Levi, D. M. (2008). Crowding–an essential bottleneck for object recognition: a mini-review. *Vision Research*, 48(5), 635–654.

Linsley, D., Kim, J., Veerabadran, V., & Serre, T. (2018). Learning long-range spatial dependencies with horizontal gated-recurrent units. arXiv: 1805.08315 [cs.CV].

Livne, T. & Sagi, D. (2007). Configuration influence on crowding. *Journal of Vision*, 7(2), 4.1–12.

Manassi, M. & Whitney, D. (2018). Multi-level Crowding and the Paradox of Object Recognition in Clutter. *Current biology: CB*, 28(3), R127–R133.

Mnih, V., Heess, N., Graves, A., & Kavukcuoglu, K. (2014). Recurrent models of visual attention. *Advances in Neural Information Processing Systems*, 3.

Pelli, D. G., Palomares, M., & Majaj, N. J. (2004). Crowding is unlike ordinary masking: Distinguishing feature integration from detection. *Journal of Vision*, 4(12), 12–12.

Pelli, D. G. & Tillman, K. A. (2008). The uncrowded window of object recognition. *Nature neuroscience*, 11(10), 1129–1135.

Perez, L. & Wang, J. (2017). The Effectiveness of Data Augmentation in Image Classification using Deep Learning. *arXiv:1712.04621 [cs]*. arXiv: 1712.04621.

Petrov, Y. & Meleshkevich, O. (2011). Asymmetries and idiosyncratic hot spots in crowding. *Vision Research*, 51(10), 1117–1123.

Petrov, Y., Popple, A. V., & McKee, S. P. (2007). Crowding and surround suppression: Not to be confused. *Journal of vision*, 7(2), 12.1–12.9.

Sabour, S., Frosst, N., & Hinton, G. E. (2017). Dynamic routing between capsules. arXiv: 1710.09829 [cs.CV].

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Schmidt, K., Yamins, D. L. K., & DiCarlo, J. J. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv preprint*.

Simonyan, K. & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556. [cs.CV].

Song, S., Levi, D. M., & Pelli, D. G. (2014). A double dissociation of the acuity and crowding limits to letter identification, and the promise of improved visual screening. *Journal of Vision*, 14(5), 3.

Spoerer, C. J., McClure, P., & Kriegeskorte, N. (2017). Recurrent convolutional neural networks: A better model of biological object recognition. *Frontiers in Psychology*, 8, 1551.

Strasburger, H. & Malania, M. (2013). Source confusion is a major cause of crowding. *Journal of Vision*, 13(1), 24–24.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2014). Going Deeper with Convolutions. arXiv:1409.4842 [cs.CV].

Tang, T., Shindell, D. T., Samset, B. H., Boucher, O., Forster, P. M., Hodnebrog, O., Myhre, G., Sillmann, J., Voulgarakis, A., Andrews, T., & Faluvegi, G. S. (2018). Dynamical response of mediterranean precipitation to greenhouse gases and aerosols. *Atmos. Chem. Phys.*, 18, 8439–8452.

Toet, A. & Levi, D. M. (1992). The two-dimensional shape of spatial interaction zones in the parafovea. *Vision Research*, 32(7), 1349–1357.

Ullman, S. (2007). Object recognition and segmentation by a fragment-based hierarchy. *Trends in Cognitive Sciences*, 11(2), 58–64.

Volokitin, A., Roig, G., & Poggio, T. A. (2017). Do Deep Neural Networks Suffer from Crowding? In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, NIPS'17.

Wallace, J. M. & Tjan, B. S. (2011). Object crowding. *Journal of Vision*, 11(6).

Xu, B., Wang, N., Chen, T., & Li, M. (2015). Empirical Evaluation of Rectified Activations in Convolutional Network. *arXiv:1505.00853 [cs, stat]*. arXiv: 1505.00853.

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

Zeiler, M. D. & Fergus, R. (2013). Visualizing and Understanding Convolutional Networks. *arXiv:1311.2901 [cs]*. arXiv: 1311.2901.

Zhang, Q., Wu, Y. N., & Zhu, S.-C. (2017). Interpretable Convolutional Neural Networks. *arXiv:1710.00935 [cs]*. arXiv: 1710.00935.

Supplementary Materials

Supplementary figures

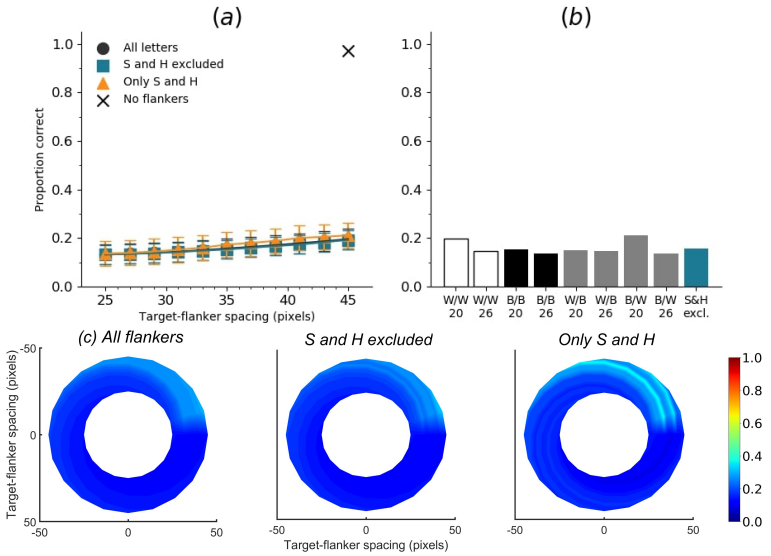

Figure S1: Accuracy of letter identification of the first randomly initialised small 5-layer convolutional network with single flankers ('Alexnet 1'). Training and testing was done without acuity loss. Average accuracy without flankers was 97.16%.

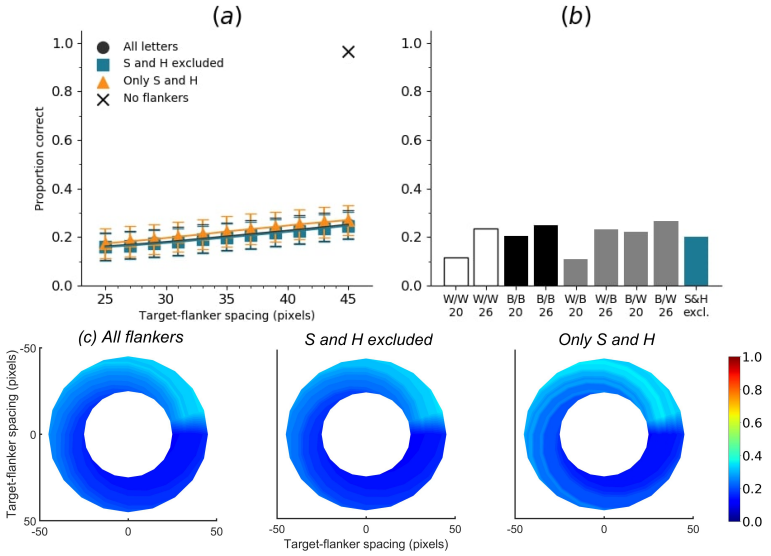

Figure S2: Accuracy of letter identification of the second randomly initialised small 5-layer convolutional network with single flankers ('Alexnet 2'). Training and testing was done without acuity loss. Average accuracy without flankers was 96.62%.

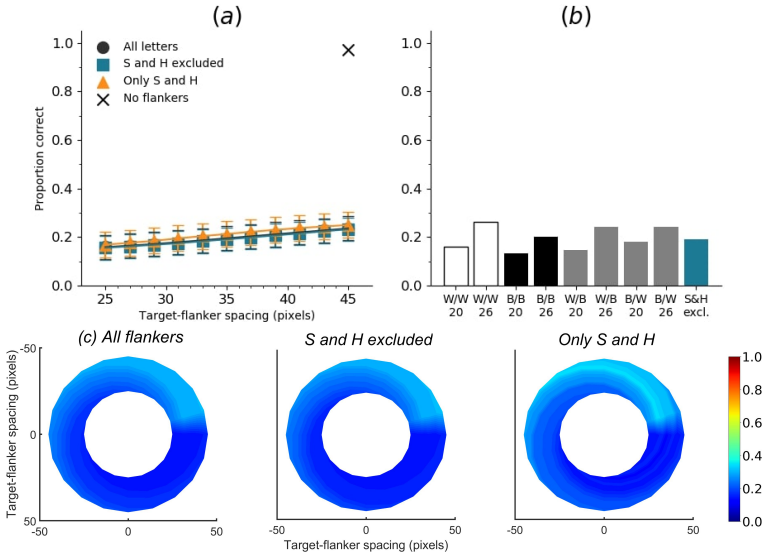

Figure S3: Accuracy of letter identification of the third randomly initialised small 5-layer convolutional network with single flankers ('Alexnet 3'). Training and testing was done without acuity loss. Average accuracy without flankers was 97.25%.

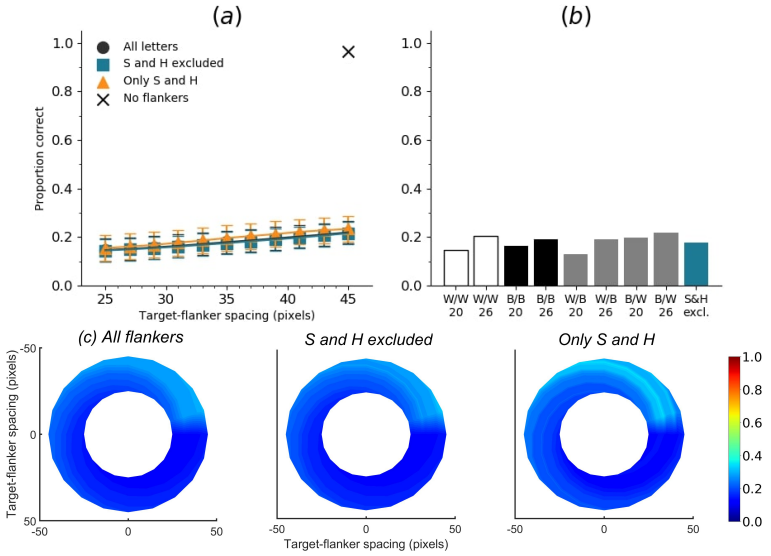

Figure S4: Accuracy of letter identification of the fourth randomly initialised small 5-layer convolutional network with single flankers ('Alexnet 4'). Training and testing was done without acuity loss. Average accuracy without flankers was 96.59%.

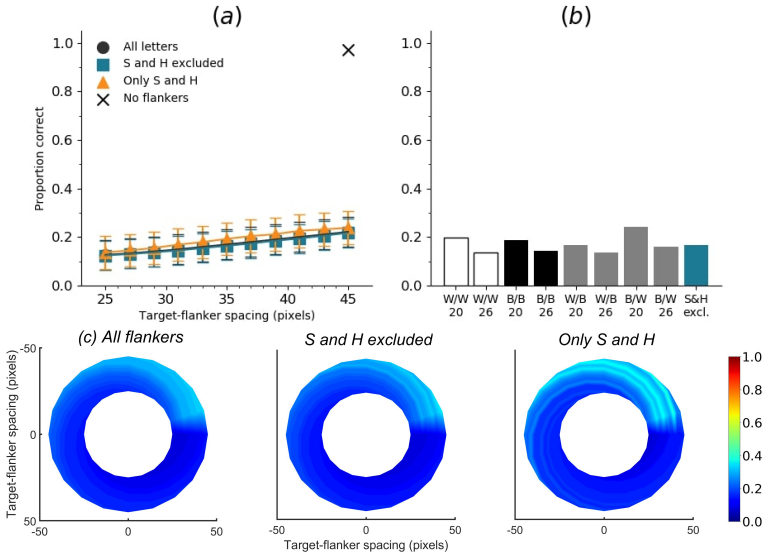

Figure S5: Accuracy of letter identification of the fifth randomly initialised small 5-layer convolutional network with single flankers ('Alexnet 5'). Training and testing was done without acuity loss. Average accuracy without flankers was 97.22%.

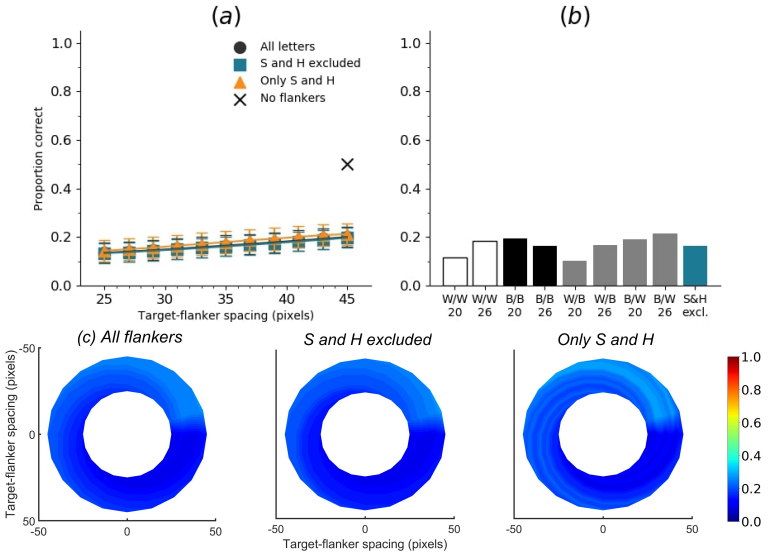

Figure S6: Accuracy of letter identification of the randomly initialised small 5-layer convolutional network with single flankers ('Alexnet'). A total of five independent training sessions and test sessions are combined in this figure. Training of all five models in this figure was done without acuity loss, and testing was done with acuity loss. Average accuracy without flankers was 50.02%.

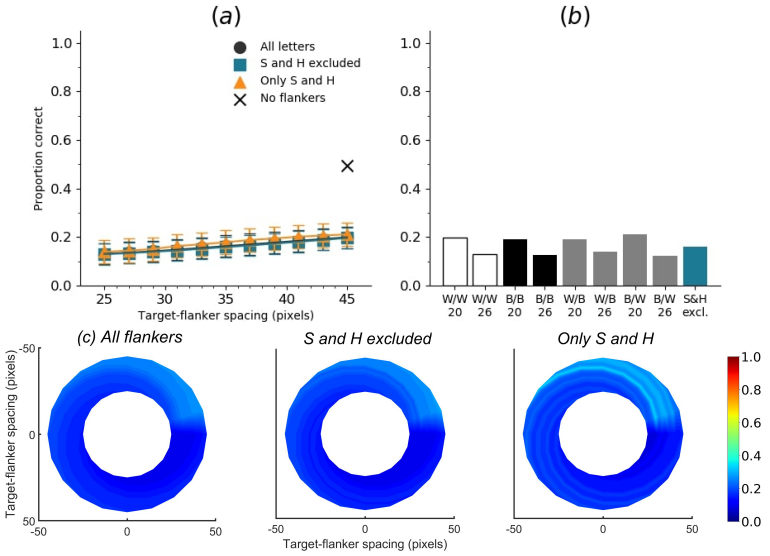

Figure S7: Accuracy of letter identification of the first randomly initialised small 5-layer convolutional network with single flankers ('Alexnet 1'). Training was done without acuity loss, and testing with acuity loss. Average accuracy for acuity-reduced data without flankers was 49.33%.

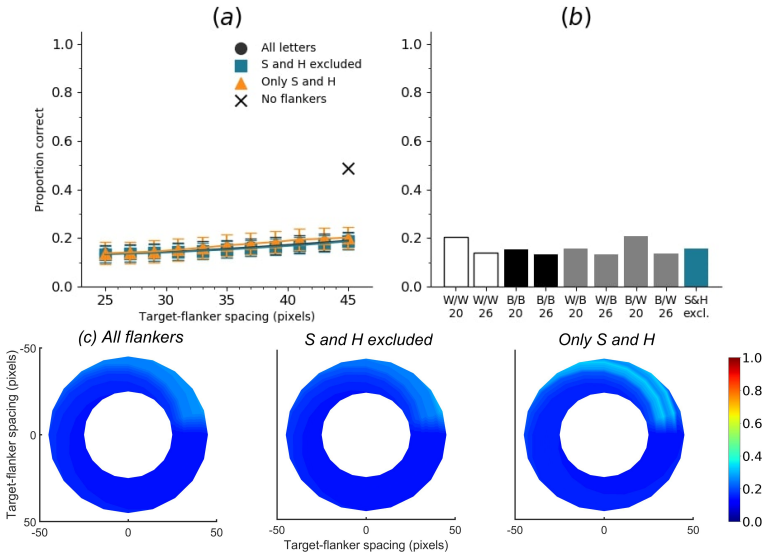

Figure S8: Accuracy of letter identification of the second randomly initialised small 5-layer convolutional network with single flankers ('Alexnet 2'). Training was done without acuity loss, and testing with acuity loss. Average accuracy for acuity-reduced data without flankers was 48.91%.

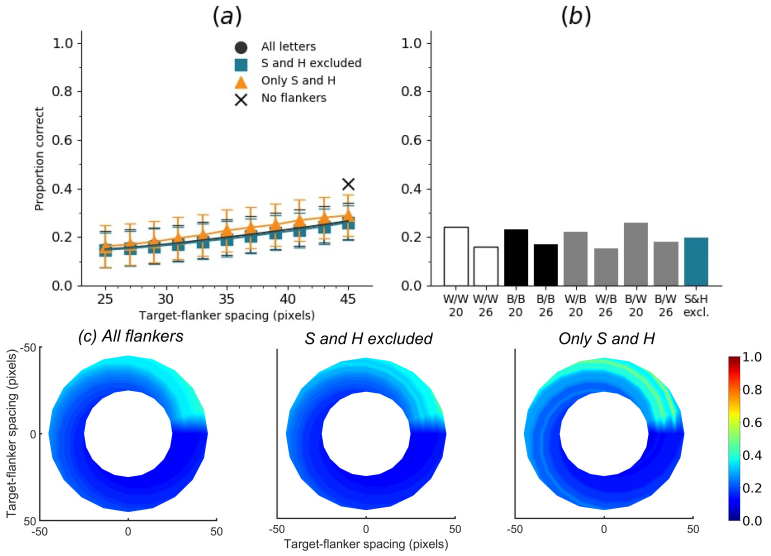

Figure S9: Accuracy of letter identification of the third randomly initialised small 5-layer convolutional network with single flankers ('Alexnet 3'). Training was done without acuity loss, and testing with acuity loss. Average accuracy for acuity-reduced data without flankers was 42.06%.

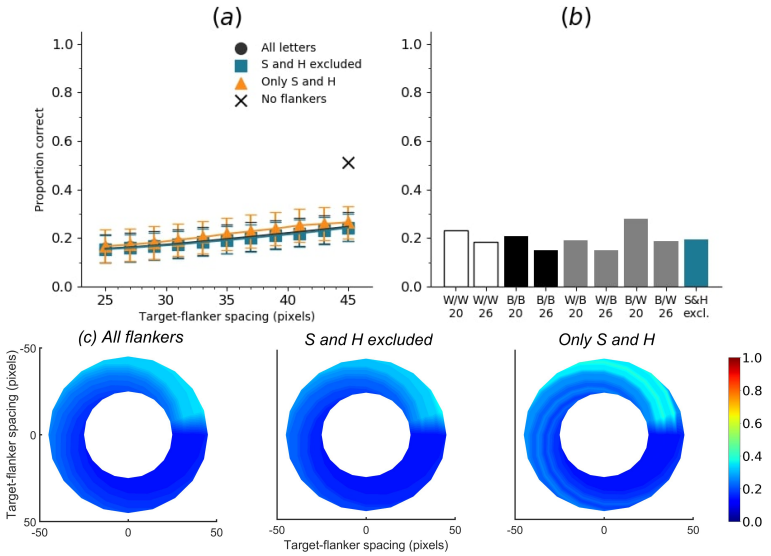

Figure S10: Accuracy of letter identification of the fourth randomly initialised small 5-layer convolutional network with single flankers ('Alexnet 4'). Training was done without acuity loss, and testing with acuity loss. Average accuracy for acuity-reduced data without flankers was 51.21%.

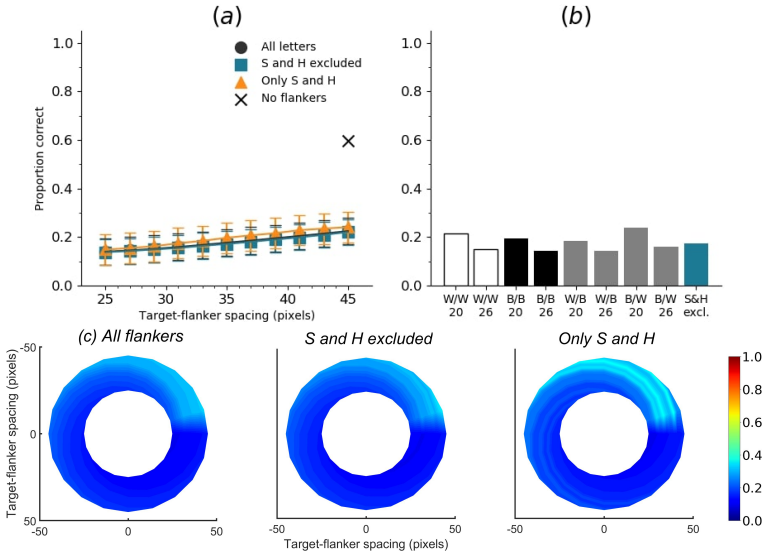

Figure S11: Accuracy of letter identification of the fifth randomly initialised small 5-layer convolutional network with single flankers ('Alexnet 5'). Training was done without acuity loss, and testing with acuity loss. Average accuracy for acuity-reduced data without flankers was 59.51%.

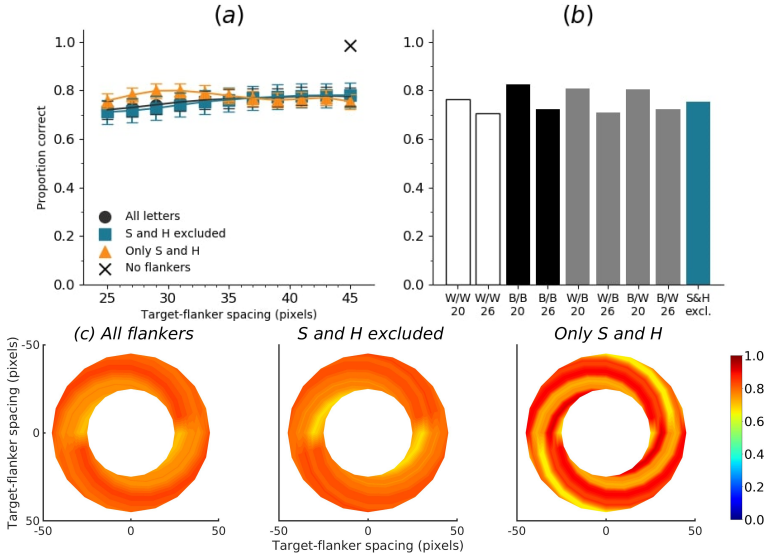

Figure S12: Accuracy of letter identification for the ILSVRC-initialised DenseNet-121. Training and testing was done with acuity loss. Model accuracy without flankers was 98.52%.

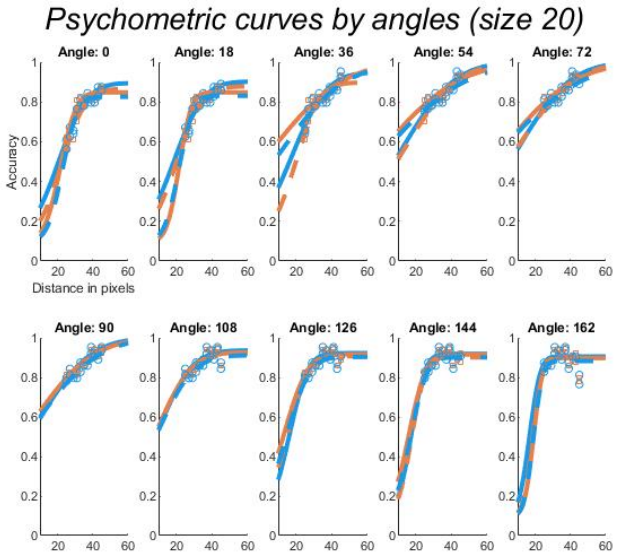

Figure S13: Psychometric curves for the randomly initialised DenseNet-121 with pair flankers without the flanker S and H, trained and tested with acuity loss. Even when the unseen flankers are omitted, the curve fits are not consistent. Markers show data points, while lines show Gauss error function fits.

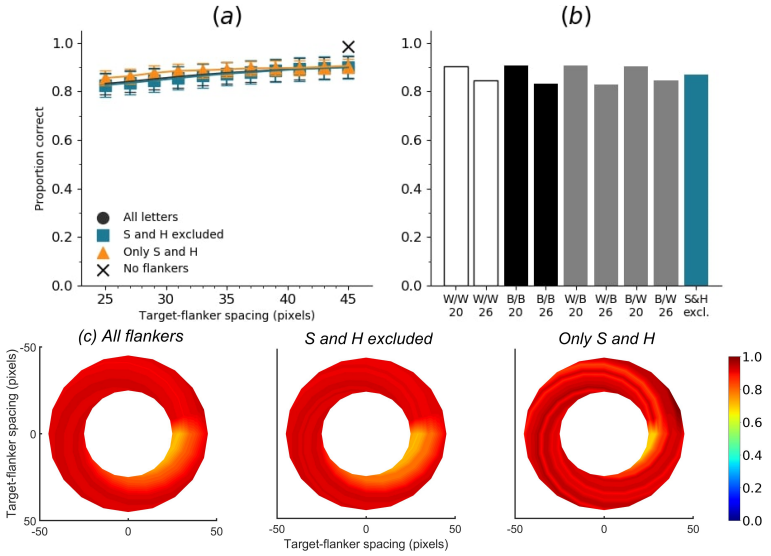

Figure S14: Accuracy of letter identification for the ILSVRC-initialised model with single flankers. We find that regardless of weight initialisation, crowding behaves similarly. Training and testing was done with acuity loss. Accuracy without flankers was 98.52%.

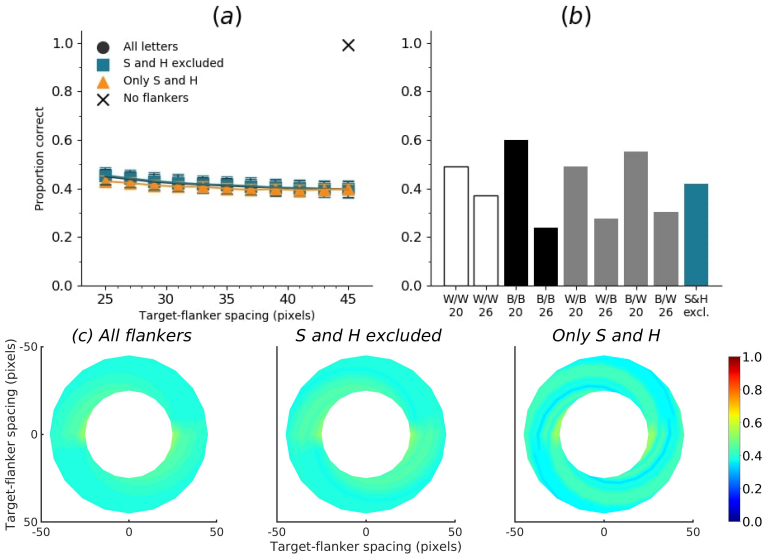

Figure S15: Accuracy of letter identification using the same weights as Figure 9, but with pair flankers. We suspect convergence issues with this test, resulting in unexpected test performance—while some experiments did not show a clear decrease in the degree of crowding with distance, this is one of the two models for which crowding increases with distance.

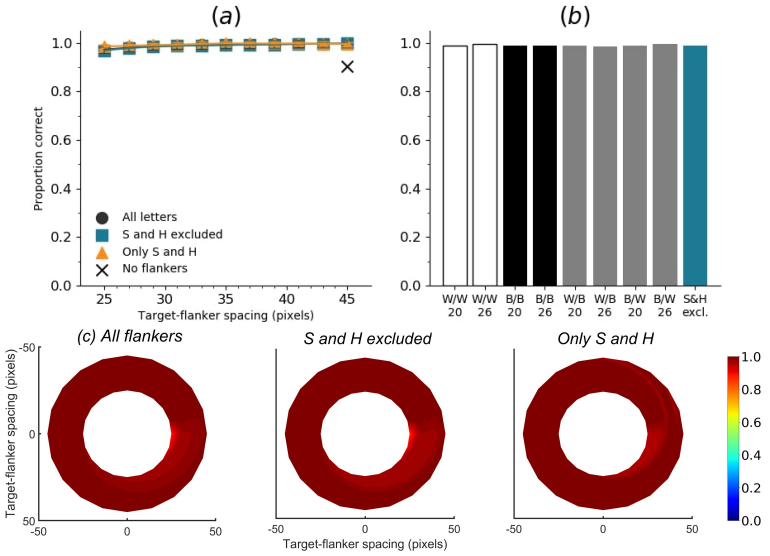

Figure S16: Accuracy of letter identification of the DenseNet-121 with random weight initialisation and no acuity loss with single flankers. Model base accuracy was 90.25%. Note that accuracy is increased by adding a flanker—the only position that does not exhibit this behaviour is the same position that causes the most crowding in almost all of our other tests.

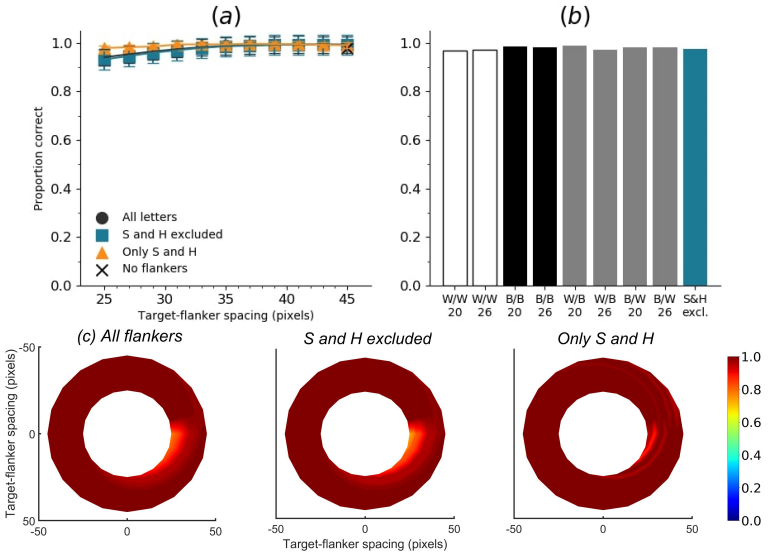

Figure S17: Accuracy of letter identification of a re-run of the DenseNet-121 with random weight initialisation and no acuity loss with single flankers. Model base accuracy was 97.85%. Again, this model exhibits the behaviour that adding a flanker increases performance.

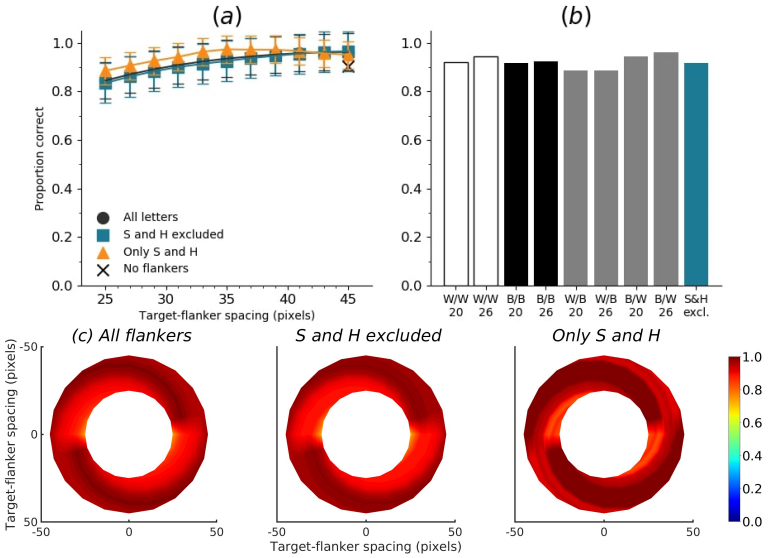

Figure S18: Accuracy of letter identification of the DenseNet-121 by distance and colour for pair flankers with random weight initialisation and no acuity loss. Accuracy without flankers was 90.25%. Note that as this is the same model as presented in Figure S16—some positions of flankers also increase accuracy.

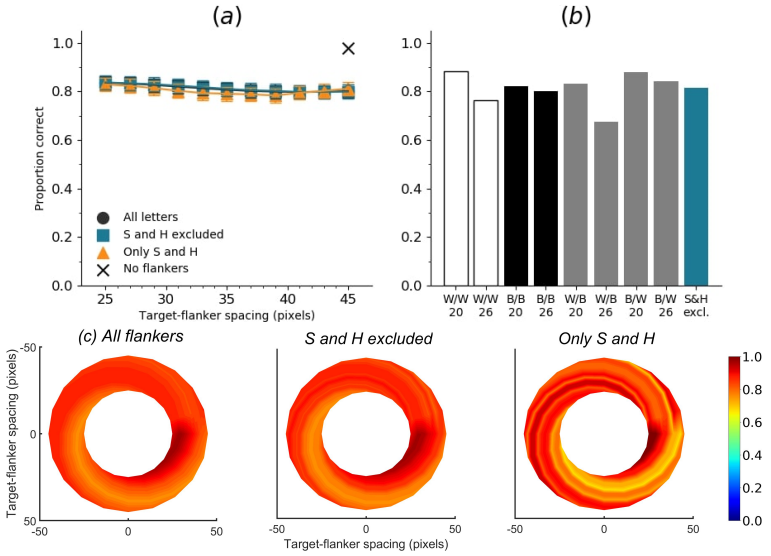

Figure S19: Accuracy of letter identification for the DenseNet-121 with single flankers when the target and flankers are placed on the right-hand side of the image, instead of the left-hand side. Training and testing was done with acuity loss. This is an additional run of the model presented in Figure 7 to verify results. Notice that target-flanker spacing decreases accuracy. Model accuracy without flankers was 97.92%.