Computational and Systems Biology
Neuroscience

A computationally informed comparison between the strategies of rodents and humans in visual object recognition

Anna Elisabeth Schnell ✉, **Maarten Leemans, Kasper Vinken, Hans Op de Beeck**

Department of Brain and Cognition & Leuven Brain Institute, KU Leuven, Leuven, Belgium • Department of Neurobiology, Harvard Medical School, Boston, Massachusetts, United States of America

Reviewed Preprint

Revised by authors after peer review.

About eLife's process

Reviewed preprint version 2
October 27, 2023 (this version)

Reviewed preprint version 1
June 27, 2023

Sent for peer review
April 17, 2023

Posted to bioRxiv
March 15, 2023

Abstract

Many species are able to recognize objects, but it has been proven difficult to pinpoint and compare how different species solve this task. Recent research suggested to combine computational and animal modelling in order to obtain a more systematic understanding of task complexity and compare strategies between species. In the present study, we created a large multidimensional stimulus set and designed a visual discrimination task partially based upon modelling with a convolutional deep neural network (CNN). Experiments included rats (N = 11; 1115 daily sessions in total for all rats together) and humans (N = 45). Each species was able to master the task and generalize to a variety of new images. Nevertheless, rats and humans showed very little convergence in terms of which object pairs were associated with high and low performance, suggesting the use of different strategies. There was an interaction between species and whether stimulus pairs favoured early or late processing in a CNN. A direct comparison with CNN representations and visual feature analyses revealed that rat performance was best captured by late convolutional layers and partially by visual features such as brightness and pixel-level similarity, while human performance related more to the higher-up fully connected layers. These findings highlight the additional value of using a computational approach for the design of object recognition tasks. Overall, this computationally informed investigation of object recognition behaviour reveals a strong discrepancy in strategies between rodent and human vision.

eLife assessment

Schnell et al report **important** differences between the strategies used by rodents and humans when discriminating different visual objects. The evidence supporting these findings is **convincing**, showing that rat performance was influenced far more by low-level cues compared to humans. It is, however, unclear to what extent these differences can be explained by the lower visual acuity of rats. This work will be of general interest to vision and cognition researchers, particularly those studying object vision.

1 Introduction

Humans show high proficiency in invariant object recognition, the ability to recognize the same objects from different viewpoints or in different scenes. This ability is supported by the ventral visual stream, the so-called *what stream* (Logothetis & Sheinberg, 1996). A question that is repeatedly addressed in vision studies is whether and how we can model this stream by means of animal models or computational models to further examine and quantify the representations along the ventral visual stream. Computationally, researchers have recently modelled this stream by using convolutional deep neural networks (CNNs), as for example done by Avberšek and colleagues (2021), Cadieu and colleagues (2014), Duyck and colleagues (2021), Güçlü & Gerven (2015), Kalfas and colleagues (2018), Kar and colleagues (2019), Kubilius and colleagues (2016), Pospisil and colleagues (2018) and Vinken & Op de Beeck (2021). Lately, the rodent model has become an important animal model in vision studies, motivated by the applicability of molecular and genetic tools rather than by the visual capabilities of rodents. Past studies have examined behavioural (Alemi-Neissi et al., 2013; De Keyser et al., 2015; Djurdjevic et al., 2018; Schnell et al., 2019; Tafazoli et al., 2012; Vermaercke & Op de Beeck, 2012; Vinken et al., 2014; Zoccolan, 2015) (for a review see Zoccolan, 2015) as well as neural (Matteucci et al., 2019; Tafazoli et al., 2017; Vermaercke et al., 2014; Vinken et al., 2016) data of rodents (rats and mice) performing in visual pattern recognition tasks. The behavioural findings suggested that rats are capable of learning complex visual discrimination tasks. Here we plan to integrate computational and animal modelling approaches, by using data about information processing in artificial neural networks when designing the animal experiments.

One aspect that almost all rodent studies have in common is that the exact task and stimuli are chosen based on what we know from human and monkey studies. Earlier research showed that the intuition of researchers about the complexity of visual tasks can be misleading (Vinken & Op de Beeck, 2021). Through computational CNN modelling of the tasks from previous studies, they showed that behavioural strategies that seem complex at first hand might be best modelled through relatively early levels of processing in CNNs. They recommended that future studies could obtain more direct information about the complexity of visual tasks and behavioural strategies by incorporating neural network models in the design phase of the experiment. One way of implementing this is to train rodents in a challenging and multidimensional visual task and use CNNs to select stimulus examples targeting strategies with different levels of complexity.

In the present study, we implemented this approach and created a large stimulus set that can be used for a variety of visual experiments. We decided to create the stimuli in a way that they are adaptable to different types of tasks, such as a “simple” discrimination task or non-linear tasks (e.g. Bossens & Op de Beeck, 2016). We then took a subset of these stimuli and performed a visual discrimination experiment in rats (see **Figure 1** for the design). The task itself was defined in a stimulus space with two dimensions, here referred to as concavity and alignment. The stimuli consisted of a base shape that varied in concavity, with three spheres attached to it that were either horizontally aligned or misaligned. The task was then further complicated by transforming the stimuli along several dimensions that preserve the identity of the object. We started by training the animals in a base stimulus pair, with the target being the concave object with horizontally aligned spheres. Once the animals were trained in this base stimulus pair, we used the identity-preserving transformations to test for generalization. After a number of transformation phases, we selected a final stimulus set by choosing a combination of transformations based on the outcomes of a trained CNN. Using the neural network as a (basic) model for the different stages of ventral visual stream processing, we chose stimulus pairs that require either higher or lower levels of processing and thus allow us to maximally differentiate between the task strategies used by the animals. As a final part of the current study, we performed

an online human experiment with the same stimuli and design as the experiment for the rats, providing us with a rich three-way comparison of rat behavioural data with human behavioural data and with CNN data.

2 Results

In this study, we trained and tested 11 rats and 45 humans on a complex two-dimensional discrimination task (see **Figure 1** for the design of the rat study, and **Supplemental Figure 7** for the design of the human study). Rats and humans were first trained in a base pair. Next we tested their ability to generalize across several image transformations. In the last two protocols of the design, we used a computational approach to select stimuli that require different visual strategies.

2.1 Animal study

Training

We first checked the variation in performance across phases and stimulus pairs during training. In the first *Training Phase*, animals were trained in the base stimulus pair, which were the maximally different target and distractor in a concavity x alignment stimulus space where each dimension was varied with 4 values (4 x 4 space). This training was successful for all twelve animals and lasted on average for 8.62 sessions (SD = 1.61). Animals were trained until they reached 80% performance for two consecutive sessions.

Once the animals were successfully trained, we examined whether they use both dimensions (concavity and alignment) by presenting them with two additional stimuli pairs where the target and distractor differ in only one dimension (see **Figure 1**, *Dimension learning).* Performance on the old pair was similar to training performance (85.83%). The animals performed well with the stimuli that differ only along the concavity dimension (78.79%), although it was significantly lower than the performance on the base pair (paired t-test on rat performance, $t(11) = 3.77$, $p = 0.003$). Performance dropped to 67.83% for the alignment-only pair, yet also significantly higher than chance level (one-sample t-test, $p < .0001$). Overall, the *Dimension learning* protocol provides evidence that the animals have picked up each of the two dimensions. This finding already excludes trivial explanations in terms of simple visual dimensions. For example, while concavity is correlated with horizontal size (distractor wider) and with overall brightness (distractor brighter, thus the opposite relevance as in the shaping phase), these simple dimensions cannot explain above-chance performance on the alignment dimension.

The third training protocol consisted of a number of small transformations, as visualized in **Figure 1** (*Transformations*). Rats learned these transformations very well, with an average performance of 83.05% (see **Figure 2**). The pairwise percentage matrix in **Figure 2** shows that the distractor with the Size transformation (most right column in the matrix) affected the rat performance the most.

The variation in performance across targets and distractors can be due to a variety of factors. This can include simple dimensions such as brightness. In the base pair, the distractor is brighter than the target. While this is the opposite from the shaping task of detecting a shape versus a black screen, visual inspection of **Figure 2** suggests that the animals perform poorer on trials in which the distractor display is not so much brighter (e.g., when it is small). To quantify this effect of brightness, we calculated the correlation between the performances in the matrix and the difference in pixel values (and thus brightness) of the stimulus pairs. This resulted in a (Pearson) correlation of −0.59 ($p < 0.01$), suggesting that there is indeed an effect of brightness. Yet,

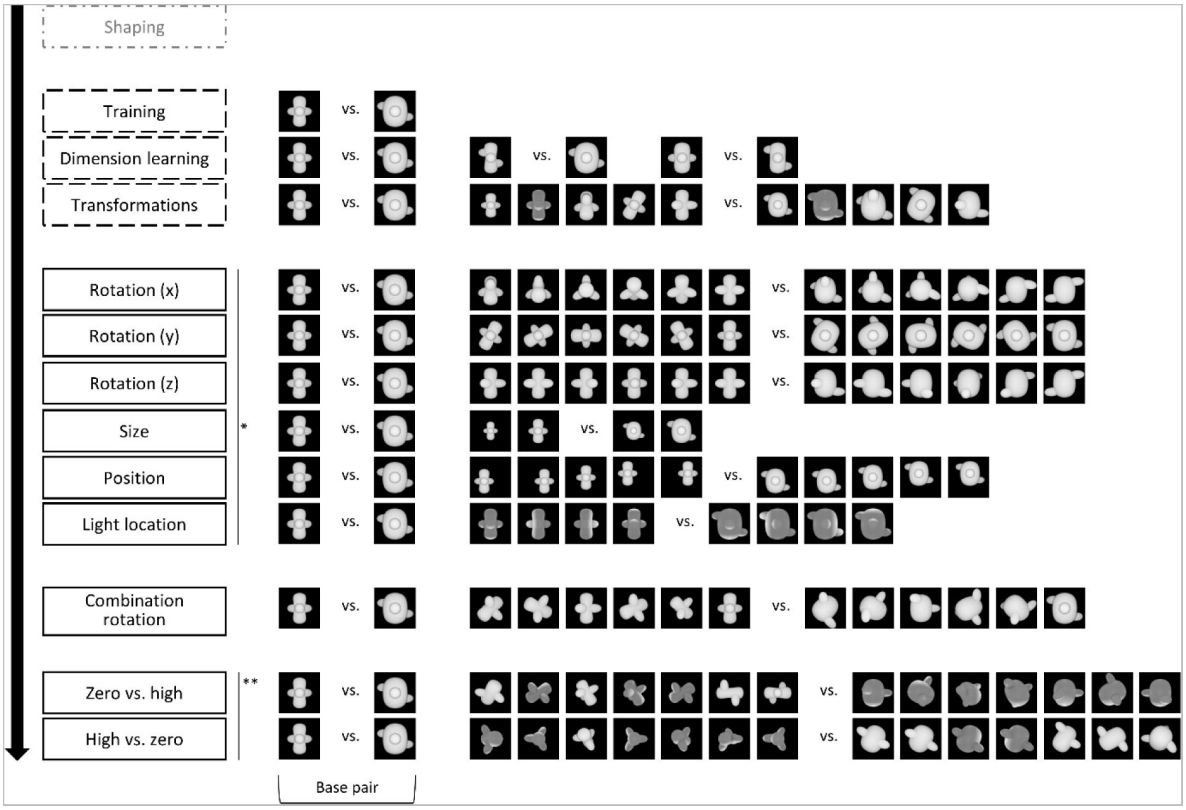


Figure 1

The design of the animal study, including the stimuli.

Animals started with a standardized shaping procedure, followed by three training protocols, as indicated by the dashed outline. In these protocols, animals received real reward, i.e. reward for touching the target. The target corresponds to the concave object in all training protocols. The rats received correction trials for incorrect answers, i.e. touching the convex object. After the three training protocols, the animals went through a number of testing protocols. The order of the first six protocols (*) and the last two testing protocols (**) was counterbalanced between the animals. During testing protocols, animals received one third old trials, and two third new trials. In the new trials, they received random reward in 80% of the trials whereas in the old trials, they received real reward and correction trials if necessary. Again, the target in the testing protocols correspond to the concave objects whereas the distractors correspond to the convex objects.

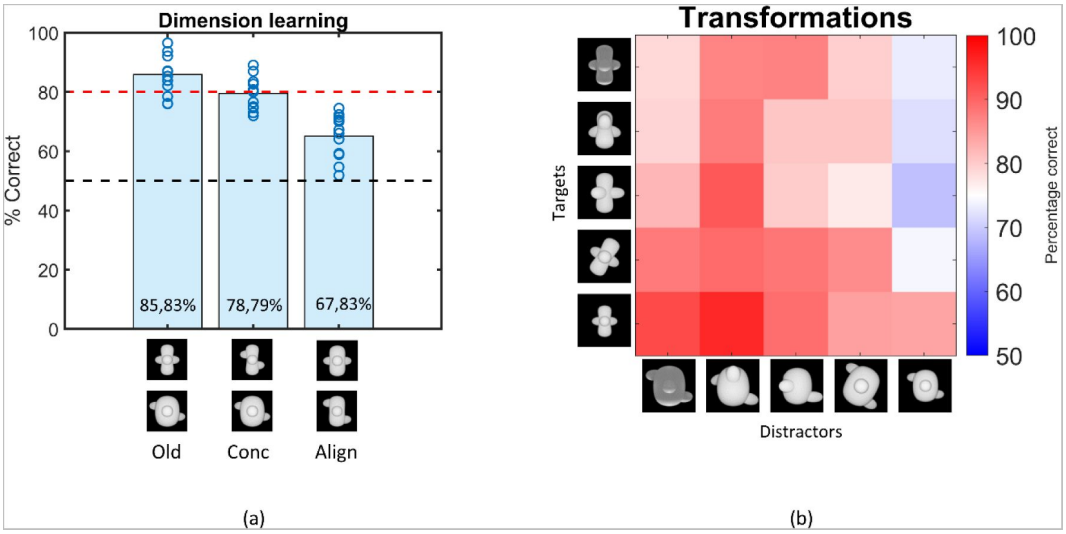


Figure 2

(a) Results of the Dimension learning training protocol. The black dashed horizontal line indicates chance level performance and the red dashed line represents the 80% performance threshold. The blue circles on top of each bar represent individual rat performances. The three bars represent the average performance of all animals on the old pair (Old), the pair that differs only in concavity (Conc) and on the pair that differs only in alignment (Align). **(b) Results of the Transformations training protocol.** Each cell of the matrix indicates the average performance per stimulus pair, pooled over all animals. The columns represent the distractors, whereas the rows separate the targets. The colour bar indicates the performance correct.

brightness is at best a partial explanation because all percentages in the matrix are above chance, with the lowest percentage in the matrix being 68.83%, even though in some pairs the difference in pixel values is abolished or even opposite from the base pair.

Overall the findings from the training phase and the above-chance performance on a variety of dimensions and transformations suggest that the rats have learned a pattern classification task with a level of complexity that might be competitive with other tasks in the rodent literature.

Testing across transformations

The six protocols that test generalization to various transformations with new, untrained images are associated with performances lower than 80% (binomial test, see **Supplemental Table 5** (lower table) for detailed table with results), but significantly higher than chance level (see **Supplemental Table 5** (lower table)). The pairwise percentage matrices of the animals in **Figure 3** provide a more detailed view of what is happening in every test, and **Supplemental Figure 2** shows the individual accuracy for each animal. The distractor has a higher impact on performance than the target in some tests. **Supplemental Table 6** shows the marginal means and standard deviation for each target and distractor for these two test protocols. From these means it is clear that there is a higher variation in the performance between distractors in *Rotation X* (52%-65%) and *Rotation Z* (56%-73%) than between targets (55%-60% resp. 60%-66%). The same happens in the size test protocol.

After these first six test protocols, the animals were presented with a schedule where all three rotations are combined (see **Figure 1**). On the new stimuli, the animals performed 58.56%, which is rather low, but still significantly different from chance level (binomial test on pooled performance of all animals: $p < 0.0001$; 95% CI [0.57;0.60]).

Testing computational levels of complexity

For the final two test protocols, we used a CNN to find image pairs that would contrast strategies based upon a different stage in visual processing, with either early layers having lower performance than high layers (*Zero vs. high*), or early layers having better performance than high layers (*High vs. zero*). Rat performance was particularly low for *Zero vs. high* (56.47%), yet still significantly different from chance level when averaged across all stimulus pairs (binomial test on pooled performance of all animals; $p < 0.0001$; 95% CI [0.55;0.58]). In contrast, rats were able to solve the *High vs. zero* pairs not only better than chance (average: 64.84%; binomial test on pooled performance of all animals; $p < 0.0001$; 95% CI [0.63;0.66]), but also significantly better than *Zero vs. high* (paired t-test on rat performance, $t(10) = -4.49$, $p = 0.0012$). This suggests that rats align with lower levels of processing when we purposely select image pairs that are optimized to contrast different levels of the visual processing hierarchy.

Next we checked how much individual CNN layers can predict the variation in behavioural performance across image pairs when we take all test protocols together. We calculated the correlation of the generalization across image pairs between the CNN classifier (summarized in **Figure 8**) and the rat performance of all nine test protocols. This correlation includes a total of 287 image pairs, i.e. all image pairs of all nine test protocols together. We did this by concatenating all performances of the animals into one array and all Classification Scores of the network into another array, and calculating the correlation between these two arrays to retrieve a correlation for each network layer. The results are displayed in **Figure 4**. Overall, we see quite low correlations, but several convolutional layers nevertheless show a significant positive correlation (permutation test) with the behavioural pattern of performance at the image pair level.

Even though some of the correlations are significant, they are low. This could indicate that no CNN layer is able to capture what rats do. Alternatively, it could be caused by a very low reliability of the behavioural data. To test the reliability of the variations in behavioural performance between

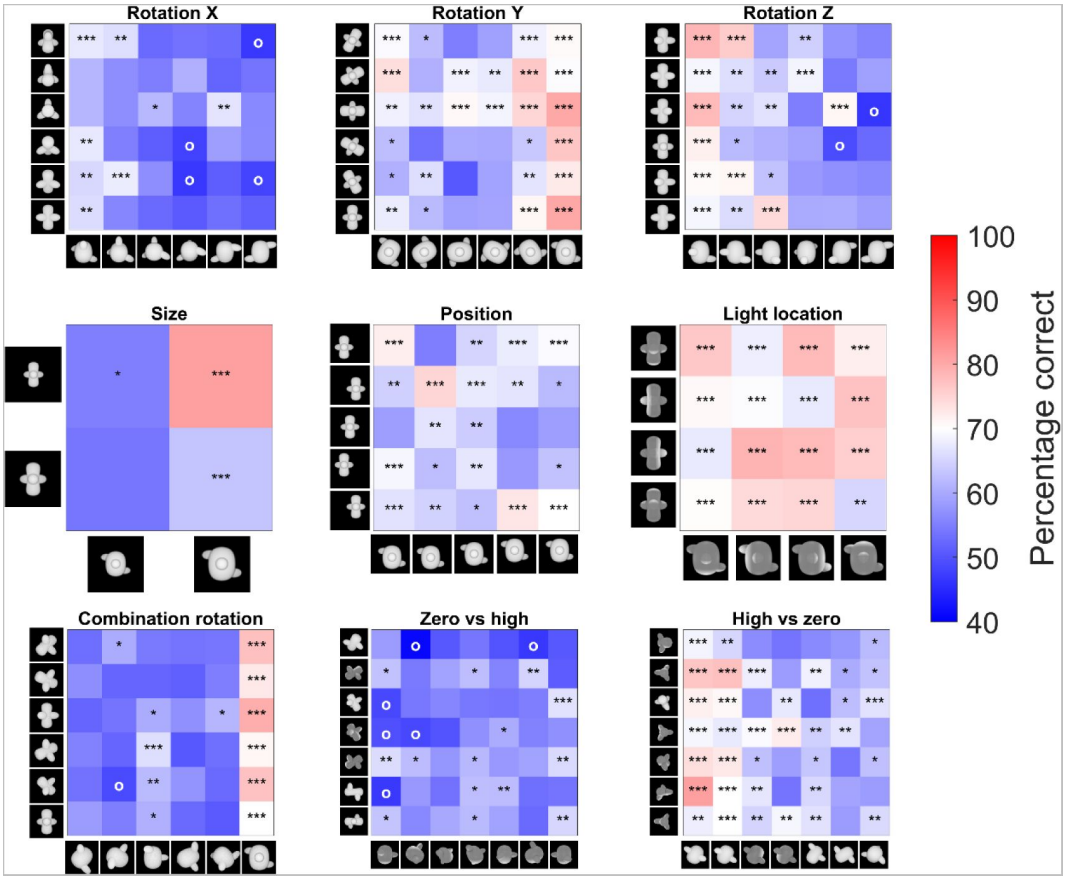


Figure 3

Pairwise percentage matrices of all nine testing protocols for the rat data.

The colour bar indicates the percentage correct of the pooled responses of all animals together. The more red a cell is, the higher the average performance. Values below 40% accuracy are indicated in the highest intensity of blue. Cells with an 'o' marker indicate a below chance performance, whereas cells with an *, ** or *** marker indicate a performance that is significantly higher than chance level (p-value < 0.05, < 0.01 or < 0.001 respectively). This was calculated with a binomial test on pooled performance of all animals.

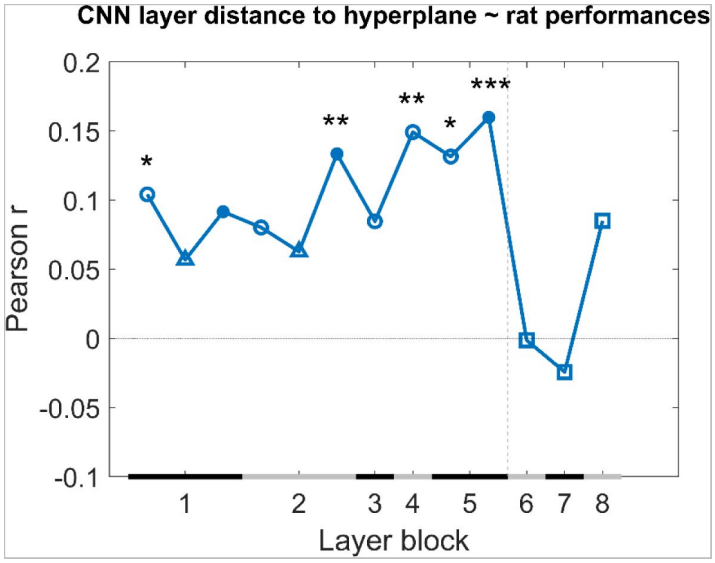


Figure 4

Correlation of the Classification Score for single target/distractor pairs between single CNN layers and the rat performance, for all nine test protocols together.

The black and grey horizontal lines on the x-axis indicated the layer blocks (block 1 consisting of conv1, norm1, pool1; block 2 consisting of conv2, norm2, pool2; block 3-4 corresponding to conv3-4 (respectively); block 5 consisting of conv5, pool5; block 6-7-8 corresponding to fc6-7-8, respectively. The vertical grey dashed line indicates the division between convolutional and fully connected layer blocks. The horizontal dashed line indicates a correlation of 0. The different markers indicate different sorts of layers: circle for convolutional layers, triangle for normalization layers, point for pool layers, and squares for fully connected layers. The asterisks indicate significant correlations according to a permutation test (* < 0.05, ** < 0.01 and *** < 0.001).

stimulus pairs in all nine test protocols, we calculated the split-half reliability, as previously done in (Schnell et al., 2023), resulting in a correlation of 0.40. By applying the Spearman-Brown correction, we obtain a full-set reliability correlation of 0.58. This correlation is much higher than the correlations with individual CNN layers.

It is possible that rat performance would be based upon multiple levels of processing, in which case we would need a combination of layers in order to explain the variation in performance across stimulus pairs. Given the low correlation between neighbouring layers (**Supplemental Table 7**), a multiple linear regression was calculated with the Classification Scores of the 13 layers as 13 regressors, and the rat performances as response vector. The results of this regression indicate a significant effect of the Classification Scores ($F(287,273) = 2.22$, $p = 0.00907$, $R^2 = 0.10$). Further investigating the 13 predictors showed that the later convolutional layers 8, 9 and 10 of the network were significant predictors in the regression model (see **Supplemental Table 8** for results of the regression model). The $R^2 = 10$ of the full model would correspond to a correlation of around 0.32. This is better than the correlation of single layers, but still clearly smaller than the reliability of the rat data of 0.58. In conclusion, the CNN model provides a partial explanation of how the performance of rats varies across image pairs.

Given the relevance of convolutional layers, we can expect that relatively basic visual features might partially explain the behavioural strategy of rats. This includes dimensions such as brightness and pixel-based similarity. To get a first indication of the relevance of these features, we calculated the correlation across image pairs between rat performance and brightness and pixel similarity. Here we found a correlation of 0.34 for pixel similarity and 0.39 for brightness, suggesting that these two visual features partially explain our results when compared to the full-set reliability of rat performance (0.58).

2.2 Human study

A final part of this study was to include an online human study that follows the same design as the animal part. **Figure 5** shows the average performance of humans (dark blue) versus rats (light blue) for all nine test protocols, as well as their performance on the old stimuli that were added in (or during) the testing protocols as quality control. Overall, humans performed better on all tests protocols than rats, with an average performance over all tests of 94.34% (humans) and 62.29% (rats). There was already a difference in terms of training performance (humans: 92.86% vs. rats: 77.84%), but the difference on the test protocols is larger. We subtracted the training performance of humans or rats from the testing performance of humans or rats, respectively, and even with this normalization for training performance there is still a significantly higher test performance in humans compared to rats ($t(16) = -6.47$, $p < 0.0001$). Thus, not surprisingly, the degree of invariance in this object classification task is higher for humans compared to rat.

The variation in performance across test protocols and across image pairs can give an indication of the strategies that each species follows. Overall, humans and rats show a mild correspondence in terms of which image pairs are more difficult, with a human-rat correlation of 0.18 across all image pairs of the nine test protocols ($p < 0.001$ with permutation test). Albeit significant, this correlation is clearly lower than the maximum value that could be obtained given the reliability of the data. The split-half reliability of the human data was 0.46, corresponding to a full-set reliability of 0.63. We reported above that full-set reliability is 0.58 for the rat data, resulting in a combined reliability of 0.60 (calculated as described in Op de Beeck et al., 2008). Thus, after taking data reliability into account there remains a pronounced discrepancy between rats and humans in terms of how performance varies across image pairs.

The main question of the present study is how this discrepancy relates to computationally informed strategies. If we take a closer look specifically at the two CNN-informed test protocols (*Zero vs. high* and *High vs. zero*), we see an opposite behaviour between animals and humans. Humans performed significantly better in the *Zero vs. high* protocol, i.e. where we used stimuli

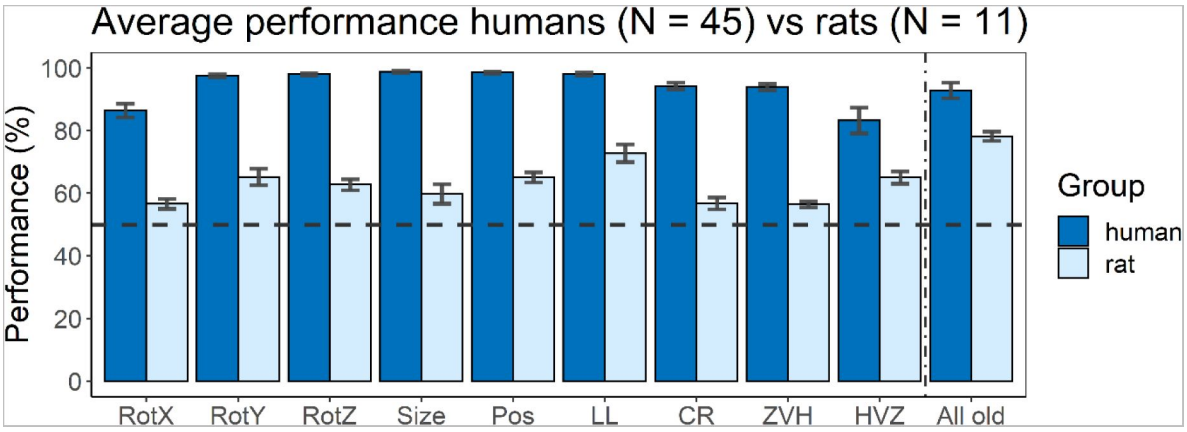


Figure 5

Average performance of humans versus rats.

On the x-axis, the nine test protocols in addition to the performance on all old stimuli are presented in the following order: rotation x (RotX), rotation y (RotY), rotation z (RotZ), size, position (Pos), light location (LL), Combination rotation (CR), Zero vs. high (ZVH), High vs. zero (HVZ) and All Old. The dashed horizontal line indicates chance level. The error bars indicate standard error over humans/rats.

where the earlier layers of the network perform worse than the higher layers, than in the *High vs. zero* protocol (paired t-test: t(44) = 2.85, p = 0.0067). Rats, however, show the opposite (see above for statistics). There even is a significant interaction between species and test protocol (unpaired t-test: t(54) = 2.50, p = 0.016). This suggests a different strategy between animals and humans: rats use strategies that are captured in the lower layers of the network, and thus correspond more to low level visual processing. Humans, however, tend to rely more on strategies captured by the higher layers of the network, and thus we are looking at more high-level visual processing.

As a next step, we calculated the correlation between the generalization across image pairs between the CNN classifier and the human performance of all nine test protocols in an identical manner as for the rat performance (**Figure 4**). The results are displayed in **Figure 6**. Overall, we see quite high correlations, especially in the higher layers. This pattern across layers is very different from the pattern in rats where the highest layers showed no correlations, which again suggests that, despite successful generalization, rats rely on decisively lower-level strategies than humans in the same discrimination task.

A multiple linear regression was calculated in an identical manner as we did with the rat performance. The results of this regression indicate a significant effect of the Classification Scores ($F(287,273) = 6.8$, $p < 0.0001$, $R^2 = 0.25$). Further investigating the 13 predictors showed that in particular the fully connected layers 11, 12 and 13 of the network were strong predictors in the regression model (see **Supplemental Table 9** for results of the regression model).

Given the high correlations with fully connected layers, we would expect to not find much evidence for an influence of basic visual dimensions such as brightness and pixel-based similarity. Indeed, we find small correlations with variation in human behavioural performance across image pairs for pixel similarity (0.12) and for brightness (−0.12).

3 Discussion

In the current study, we trained and tested rats and humans in a discrimination task using two-dimensional stimuli, with the two dimensions being concavity and alignment. We tested generalization across a range of viewing conditions. For the last two testing protocols, we used a computational approach to select the stimuli in terms of specifically dissociating low and high stages of processing. Rats were able to learn both dimensions (concavity and alignment) and showed a preference for concavity. Their performance on the testing protocols revealed a wide variety in percentage correct: for some test protocols they performed just above chance level, e.g. *Zero vs. high*, whereas for others they could easily reach about 70% correct (*Position*). Humans, on the other hand, performed better overall, with performances of 80% or higher on the testing protocols. Addressing the question of the complexity of the underlying strategies, rats performed best on the test protocol designed to specifically target lower levels of processing whereas humans performed best on the high-level processing protocol. Likewise, direct comparisons with artificial neural network layers showed that the variation of rat performance across images was best explained by late convolutional layers, whereas human performance was most associated with representations in fully connected layers.

All animals started by being trained in three training protocols. The first *Training* protocol only included one image pair, the base pair, containing the most different target and distractor without any further transformations. Learning of the individual dimensions of concavity and alignment was investigated through the *Dimension learning* protocol. The results from this *Dimension learning* protocol indicate that our rats have more difficulties learning the alignment dimension as opposed to the concavity dimension. One possible explanation for the superior performance on the concavity dimension could be that the animals were partially solving the task such that the brighter stimulus, i.e. the convex base shape, is the distractor and that their strategy is to pick the

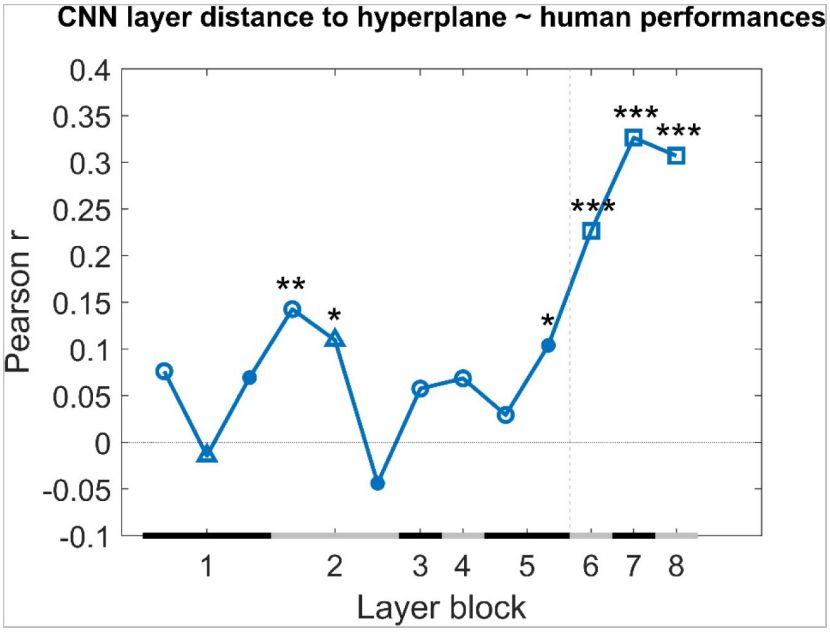


Figure 6

Correlation of the Classification Score for single target/distractor pairs between single CNN layers and the human performance, for all nine test protocols together.

The naming convention on the x axis corresponds to the layers of the network, identical as in **Figure 4**. The black and grey horizontal lines on the x-axis indicated the layer blocks (block 1 consisting of conv1, norm1, pool1; block 2 consisting of conv2, norm2, pool2; block 3-4 corresponding to conv3-4 (respectively); block 5 consisting of conv5, pool5; block 6-7-8 corresponding to fc6-7-8, respectively. The vertical grey dashed line indicates the division between convolutional and fully connected layer blocks. The horizontal dashed line indicates a correlation of 0. The different markers indicate different sorts of layers: circle for convolutional layers, triangle for normalization layers, point for pool layers, and squares for fully connected layers. The asterisks indicate significant correlations according to a permutation test (* < 0.05, ** < 0.01 and *** < 0.001).

stimulus with the lowest brightness. This was confirmed by analyses on the third training protocol (*Transformations*) that included small transformations along various dimensions. Nevertheless, the rats still performed above chance level for trials in which the brightness differences were reversed, indicating that other dimensions are involved and overrule a contribution from brightness. Similar findings have been obtained in human behaviour and neuroscience. For example, despite the clear category selectivity in regions such as the fusiform face area, the selectivity in these regions is also modulated very strongly by various low-level dimensions (Yue et al., 2011). With regard to the size and position transformations it is important to keep in mind that the animals were freely moving in the touchscreen chambers, and so even for the original base pair was already undergoing changes in retinal size and retinal position. What we manipulate, is rather the size and position relative to the rest of the set-up (e.g., relative to screen position and size).

After these three training protocols, the animals were tested for generalization in a variety of testing protocols, each testing a separate transformation on the stimuli. The first six test protocols included rotation along all the three axes, size, position and light location, following by a test protocol in which we combined the rotation along the three axes. Overall, we found that the performance of the animals on these test protocols is affected by these transformations, but still significantly above chance in each protocol. Studies in the literature would often stop here, or proceed by systematically testing even larger transformations. Stimulus choices are based upon intuitions of what strategy animals might be using, and upon theories of how visual perception works. However, in some cases, a further computational modelling of the task and stimuli finds that what intuitively seems like a task of a particular complexity might not be so complex after all. The first tests of invariant object recognition seemed impressive, but were found to be easily solved with earlier layers of processing (Minini & Jeffery, 2006; Vinken & Op de Beeck, 2021). This was recently also highlighted by relatively simple pixel-based analyses (Kell et al., 2020). As another example, Vinken & Op de Beeck (2021) have used a computational approach to further investigate the levels of information processing in rodents by comparing three hallmark studies that provided evidence for higher order visual processing in rodents (Djurdjevic et al., 2018; Vinken et al., 2014; Zoccolan et al., 2009) with CNNs. They found that for all three studies, the low and mid-level layers captured the rat performances best, providing thus evidence against the previously concluded high level visual processing in rodents.

For these reasons, we decided to directly test image pairs through computational modelling with CNNs and select pairs that are particularly suited of dissociating different levels of processing. Stimuli were chosen by a CNN from a very large set of possible stimuli and combinations, such that the higher layers and the lower layers of the network make distinct errors on classifying the stimuli (*Zero vs. high* and *High vs. zero* protocol), and thus are diagnostic of the level of underlying visual strategies. We chose to work with Alexnet, as this is a network that has been used as a benchmark in many previous studies (e.g. (Cadieu et al., 2014; Groen et al., 2018; Kalfas et al., 2018; Nayebi et al., 2023; Zeman et al., 2020)), including studies that used more complex stimuli than the stimulus space in our current study. The stimuli of the *Zero vs. high* protocol included stimuli where the higher layers of the network performed better than the lower layers, and thus they address higher level visual processing. The opposite can be said for the *High vs. zero* protocol, which includes stimuli that specifically target lower level visual processing, given that the lower layers of the network perform best on these stimuli. After presenting these stimuli to the animals, we found that our rats performed best in the *High vs. zero* protocol, suggesting that they focus on low level visual cues to solve this discrimination task. We found the opposite CNN pattern for humans, indicating that they use high level visual processing. These findings provide more direct information about the level of processing that underlies the behavioural strategies compared to overall performance or to effects of image manipulations.

This is a new promising way to design experiments in a way that is computationally informed rather than based on researcher intuitions or qualitative predictions. It is in line with the literature that a typical deep neural network, AlexNet and also more complex ones, can explain human and animal behaviour to a certain extent but not fully. The explained variance might differ among DNNs, and there might be DNNs that can explain a higher proportion of rat or human behaviour. Most relevant for our current study is that DNNs tend to agree in terms of how representations change from lower to higher hierarchical layers, because this is the transformation that we have targeted in the Zero vs. high and High vs. zero testing protocols. (Pinto et al., 2008) already revealed that a simple V1-like model can sometimes result in surprisingly good object recognition performance. This aspect of our findings is also in line with the observation of Vinken & Op de Beeck (2021) that the performance of rats in many previous tasks might not be indicative of highly complex representations. Nevertheless, there is still a relative difference in complexity between lower and higher levels in the hierarchy. That is what we capitalize upon with the Zero vs. high and High vs. zero protocols. Thus, it might be more fruitful to explicitly contrast different levels of processing in a relative way rather than trying to pinpoint behaviour to specific levels of processing.

Partially thanks to these computationally inspired tests, our total dataset finds a marked dissociation between how humans and rats solve this object recognition task. Even in the sessions where only the old pairs are shown, the animals performed lower than humans. This was most likely due to motivation and/or distractibility. Our analyses show dissociation between humans and rats most convincingly by correlating the variation in performance across image trials with the predictions of CNN layers. There were significant correlations with multiple layers in both species. In humans, the most pronounced correlations were present for the highest, fully connected layers, while in rats correlations were limited to low and middle convolutional layers. This is the most direct evidence available in the literature that rats resolve object recognition tasks through a very different and computationally simpler strategy compared to humans. The CNN approach does not inform us how we can verbalize this simpler strategy, but based upon earlier work (Schnell and colleagues, 2023); Vermaercke & Op de Beeck, 2012) we would hypothesize that rats rely upon visual contrast features (e.g., this area is darker/lighter than that other area). Such contrast features are also used by humans and monkeys, e.g. for face detection (Ohayon et al., 2012; Sinha, 2002), but in addition humans have access to more complex strategies that e.g. refer to complex shape features such as aspect ratio and symmetry (Bossens & Op de Beeck, 2016). Tests in the present study reveal that other features that partially explain rat performance include basic dimensions such as brightness and pixel-based similarity, the latter being a proxy for retinotopically based computations that are expected to be present in convolutional layers.

Our analyses of rat behaviour and DNN modelling do not take into account potential trial-to-trial variability in the distance and position of the rat's head. From earlier work we can derive that rats typically make their decision from about 12 cm from the stimulus display (Crijns & Op de Beeck, 2019), but we have no information on trial-to-trial variability. We can hypothesize about the possible effect. If such variability would exist, then it would artificially increase the variation in distance and position during training, and thus help the animals to achieve higher levels of invariance during testing. As a consequence, the difference between rat and human performance in terms of inferred level of processing might even increase under more controlled circumstances.

For future studies, it will be highly valuable to use this computational informed strategy on a wider battery of behavioural tasks, as well as a wider range of species such as tree shrews and marmosets (Callahan & Petry, 2000; Kell et al., 2020, 2021; Meyer et al., 2022; Petry et al., 2012; Petry & Bickford, 2019). One step further, we can use the information from computational modelling together with behaviour and how it differs among stimuli to further select stimuli for neurophysiological investigations of neuronal response properties along the

visual information processing hierarchy, in this way following experimental designs that are optimized for highlighting the primary differences between processing stages and between species.

4 Methods

4.1 Animal study

4.1.1 Animals

A total of twelve male outbred Long Evans rats (Janvier Labs, Le Genest-Saint-Isle, France) started this behavioural study. Out of these twelve animals, two were tested extensively in a first pilot study, and were included in the remainder of the study as well. All animals were 11 weeks old at the start of shaping and were housed in groups of four per cage. Each cage was enriched with a plastic toy (Bio-Serv, Flemington, NJ), paper cage enrichment and wooden blocks. Near the end of the experiment, one animal had to be excluded because of health issues. During training and testing, the animals were food restricted to maintain a body weight between 85% and 90% of their underprived body weight. They received water ad libitum. All experiments and procedures involving living animals were approved by the Ethical Committee of the University of Leuven and were in accordance with the European Commission Directive of September 22, 2010 (2010/63/EU).

4.1.2 Setup

The setup is identical to the one used by Schnell and colleagues (2019) and Schnell and colleagues (2023). A short description will follow here. The animals were trained and tested in four automated touch-screen rat-testing chambers (Campden Instruments, Ltd., Leicester, UK) with ABET II controller software (v2.18, WhiskerServer v4.5.0). The animals performed one session per day and each session lasted for 100 trials or 60 minutes, whichever came first. A reward tray in which sugar pellets (45-mg sucrose pellets, TestDiet, St. Louis, MO) could be delivered was installed on one side of the chamber. On the other side of the chamber, an infrared touchscreen monitor was installed. This monitor was covered with a black Perspex mask containing two square response windows (10.0 x 10.0 cm). A shelf (5.4cm wide) was installed onto this black mask (16.5cm above the floor) to force the animals to attend to the stimuli and to view the stimuli within their central visual fields. Close proximity to the screen was enough to elicit a response because the screens are infrared. As the position of the rats in the touchscreen setup is not fixed, the actual size and position of the stimuli might vary in retinal coordinates. In a previous study we manipulate the cycles per degree of stimuli in an orientation discrimination task, and estimated that the decision distance of rats in this setup lies around 12.5 cm from the screen (Crijns & Op de Beeck, 2019). **Supplemental Figure 1** shows the timeline graphic of a correct and incorrect trial as well as images of the experimental setup.

4.1.3 Stimuli

Stimuli were created using the Python scripting implementation of the 3D modelling software Blender 3D (version 2.93.3) and measured 100 x 100 pixel. In general, the stimuli were objects that consisted of a body (base) with three spheres attached to it. A first step was to alter two dimensions of the object, namely the concavity of the base and the alignment of the three spheres. The base was made either concave or convex by increasing (convex) or decreasing (concave) the base parameter. The alignment of the spheres was altered by changing the placement of the left and the right spheres. These spheres could either be horizontally aligned or misaligned. In the misaligned case, the spheres were placed diagonally from upper left to lower right. **Figure 7a** shows two example stimuli, the ones that later were selected as the so-called “base pair”. Next, additional exemplars were created by uniformly tiling the two-dimensional stimulus space

between these two example stimuli. We decided to create eleven levels of the concavity dimension and four levels of alignment. This already yields 44 stimuli (see **Supplemental Figure 3**). We chose these levels of concavity and alignment based on the pixel dissimilarity of the stimuli (see **Supplemental Figure 4**). The final goal was to construct a 4×4 stimulus grid (**Figure 7b**) by selecting a subset of the 4×11 stimulus grid. We chose a large number of concavity levels, as this ensures flexibility in the calibration of the two dimensions relative to each other.

We added identity-preserving transformations to the stimuli, such as rotation among the x-axis, y-axis and z-axis in six different angles (0° to 180° in steps of 30°), as well as changing the light location (left, under, up, right, front) and finally the size and position. The latter two transformations were implemented using Python (3.7.3). Excluding the size and position transformation, these transformations resulted in a total set of 75460 stimuli (4 (alignment) * 11 (concavity) * 7 (x-axis rotation) * 7 (y-axis rotation) * 7 (z-axis rotation) * 5 (light location) = 75460 stimuli). **Supplemental Figure 5** shows examples of these transformations and **Figure 1** shows an overview of all image pairs that were used in this study.

4.1.4 Protocols

Once the pilot was finished (see supplementary for details), we set up the experiment and chose our stimuli. We started by reducing the 4×11 stimulus grid to a 4×4 stimulus grid (see **Figure 7b**). All stimuli on the diagonal can be seen as ambiguous stimuli (four stimuli in total), as they can be identified as a target as well as a distractor. The six stimuli above this diagonal create the target part of the grid, and the six stimuli below this diagonal resemble the distractor sub-grid.

The different phases of the experiment are shown in **Figure 1** and this figure shows all stimuli that were used. In the main Training phase, we trained the animals in the maximally different stimuli that are placed at the very ends of the corners (**Figure 7a**). We refer to this as the base pair. After this Training phase, the experiment consisted of two further training protocols. In the Dimension learning training phase, we pushed the animals to learn both dimensions (concavity and alignment) by presenting them two additional stimuli pairs from **Figure 7b** in which the target and distractor differ in only one dimension. A third training protocol (*Transformations*) consisted of stimuli with some small transformations, such as 30° rotation along the x-axis, 30° rotation along the y-axis, 30° rotation along the z-axis, light location below, and size reduction of 80%, resulting in a total of 25 possible stimulus pairs (every combination of target-distractor with the 5 transformed stimuli). During these two training protocols, one third of the trials were so-called “old trials” with the base pair. Correction trials were given if an animal answered incorrectly, i.e. the same trial was repeated until the animal answered correctly. These correction trials were excluded from the analyses. In all trials, rats received a reward for touching the correct screen, i.e. the screen with the target.

After these three training protocols, the testing part of the experiment included nine test protocols. The crucial defining difference between these test protocols and the prior training protocols is that rats received a reward randomly in 80% of the trials with new stimulus pairs, and no correction trials were given for an incorrect response. This random reward is important to keep the animals motivated during the testing protocols and to measure real generalization, and not training behaviour. We have used a similar approach in the past, where we rewarded the animals in every testing trial (Schnell et al., 2019; Vinken et al., 2014). One third of the trials in all test protocols consisted of old trials with the base pair, and here, the animals received reward for touching the target and correction trials were shown if necessary. Regularly, we inserted a *Dimension learning* session in between two test sessions to maintain the performance high enough on training stimuli, especially for the animals in which we saw a drop in performance on the base pair. We excluded any test sessions where the performance on the base pair stimuli dropped to below 65% and the performance on this base pair was not included in the accuracy calculations.

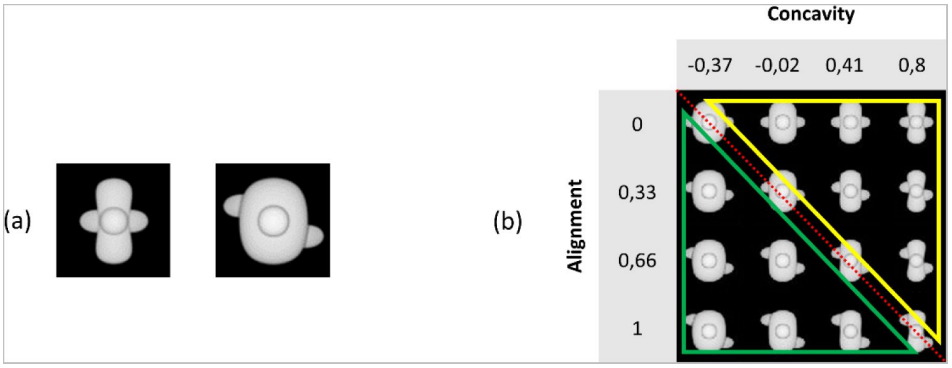


Figure 7

Illustration of the base pair and our stimulus grid.

(a) The base pair of the main experiment. (b) The chosen 4×4 stimulus grid. The red diagonal dotted line indicates the ambiguous stimuli that can be seen as target as well as distractor. All stimuli below this line (green triangle) indicate the distractor sub-grid, whereas all stimuli above this line (yellow triangle) highlight the target sub-grid.

The first six test protocols included one protocol for each transformation, i.e. *Rotation X, Rotation Y, Rotation Z, Light Location, Size* and *Position.* The order in which these first six test protocols were given to the animals was counterbalanced between the animals. The stimuli that were used in these six test protocols can be seen in **Figure 1** and every combination of target-distractor per test protocol was presented to the animals. For the rotation protocols, we used rotation degrees in steps of 30°, ranging from 30° to 180°. This resulted in 36 possible stimulus pairs for each of the three rotation protocols. In the *Light Location* protocol, we used stimuli where the light location was set at four different positions (below, left, right and up), resulting in 16 possible stimulus pairs for this protocol. In the *Size* protocol, we selected targets and distractors that were 80% and 60% reduced in size compared to the original, training pair. This protocol included 4 possible stimulus pairs. And finally, in the *Position* protocol, we changed the position of the 80% reduced in size stimuli and placed the objects in the lower left corner, lower right corner, centre, upper left corner and upper right corner. We have a total of 25 possible stimulus pairs for this protocol.

After these six test protocols, we presented the animals with six targets and six distractors where all three rotations were combined (*Combination rotation*), i.e. x-, y- and z-axis were rotated with the same degree (ranging from 30° to 180°, in steps of 30°). This resulted in a total of 36 new stimulus pairs. Again, no correction trials were included after the trials where rotated stimuli were shown and animals received random reward in 80% of the trials. One third of the trials consisted of the stimulus pair from the first *Training* phase (i.e. the base pair), and here, correction trials were given after an incorrect response and real reward was given to the animals.

In a final set of two test protocols, we created a CNN-informed stimulus set. The details of the computational modelling are explained in the next section. The first protocol (*Zero vs. high*) included stimuli in which the lower layers of the network performed around chance level (i.e. target-distractor difference in Classification Scores (difference in signed distance to hyperplane) of about 0), whereas the higher layers scored high (see **section 4.2**). The second protocol (*High vs. zero)* included stimuli where the network did the opposite. That is, the earlier layers performed well whereas the higher layers performed around chance level. The order of the two test protocols was counterbalanced between the animals. Each of these test protocols included 7 targets and 7 distractors, giving a total of 49 new stimulus pairs.

Animals stayed in each session for 60 minutes or until they reached 100 training trials or 120 testing trials. We used an intertrial interval (ITI) of 20s and a time-out of 5s during training sessions. This time-out was only used in incorrect trials. From another pilot study in the lab, we noticed we could decrease the ITI and time-out without affecting the rats' performance. Therefore, we decided to use an ITI of 15s and time-out of 3s during testing, and to increase the number of trials during a testing session to 120 trials. The stimuli remained on the screen until the animals made a choice and so there was no time limit for the animals.

Each protocol was run for multiple sessions per animal. Given that we were interested in how performance would vary across stimulus pairs, we completed more sessions for the protocols that included more stimulus pairs. **Supplemental Table 1** indicates the average number of trials per test protocol for all rats together.

One animal was not placed in the *Transformations* phase as it was the slowest animal during training. However, its performance on the test protocols did not significantly differ from the other animals. We tested this by calculating the correlation of the variation of performance across stimulus pairs for each rat with the pooled responses of all other rats. The average correlation for each of the other animals with the pooled response was 0.24 (±0.09), and the correlation of this slowest animal with the others was very similar, 0.23.

To further examine the visual features that could explain rat performance, we calculated the correlation between the rat performances and image brightness of the transformations. We did this by calculating the difference in brightness of the base pair (brightness base target – brightness base distractor), and subtracting the difference in brightness of every test target-distractor pair for each test protocol (brightness test target – brightness test distractor for each test pair). We then correlated these 287 brightness values (1 for each test image pair) with the average rat performance for each test image pair. We performed a similar correlation analysis for pixel similarity to investigate the correlation between pixel similarity of the test stimuli in relation to the base stimuli with the average performance of the animals on all nine test protocols. We did this by calculating the pixel similarity between the base target with every other testing distractor (A), the pixel similarity between the base target with every other testing target (B), the pixel similarity between the base distractor with every other testing distractor (C) and the pixel similarity between the base distractor with every other testing target (D). For each test image pair, we then calculated the average of (A) and (D), and subtracted the average of (C) and (B) from it. We correlated these 287 values (one for each image pair) with the average rat performance on all test image pairs.

4.2 Computational modelling

One important goal of this study was to create a CNN-informed stimulus set to present to the animals. To do so, we followed the steps of Schnell and colleagues (2023) and Vinken & Op de Beeck (2021) to train a CNN on the same stimuli on which our animals were trained. The steps of training the network are identical to Schnell and colleagues (2023) and a short description will follow here. We used the standard AlexNet CNN architecture that was pre-trained on ImageNet to classify images into 1000 object categories (MATLAB 2021b Deep Learning Toolbox). Following Vinken & Op de Beeck (2021), we applied principal component analysis to calculate the activations in every layer, to standardize the values across inputs and to reduce the dimensionality. We then trained a linear support vector machine classifier by using the MATLAB function fitclinear, with limited-memory BFGS solver and default regularization. We performed this with the standardized DNN layer activations in the principal component space as inputs, before ReLU, to our 24 training stimuli (see **Figure 1** **Error! Reference s ource not found.**), i.e. all stimuli of the *Training*, *Dimension* learning and *Transformations* protocols. The layers of AlexNet were divided into 13 sublayers, similar as in Schnell and colleagues (2023) and Vinken & Op de Beeck (2021).

Figure 8 shows the performance of the network for each of the test protocols after training classifiers on the training stimuli using the different CNN layers. We added noise to the inputs of the network such that the average training performance, averaged over 100 iterations, lies around 75%. By adding noise in this way, the performance on the training pairs matches overall with rat performance on those pairs, otherwise the performance of the network would be at 100% on the training pairs and this would complicate comparisons with the animal data (see also Vinken & Op de Beeck, 2021). Note that the results for the *Size* test are unreliable given the low number of stimulus pairs in that test. The performance of the network on the tests (green line in **Figure 8**) differs among the tests and across layers, but typically the network had no problems to achieve a training performance of about 85% in all test protocols in at least some layers. The change in performance across layers is variable across test protocols.

To examine the performance of the model for specific image pairs during training and testing in more detail than possible with a binary categorization decision, we calculate the distance to the classifier's hyperplane (decision boundary) of the targets and distractors. We do this by computing the difference in signed distance to the hyperplane between target and distractor (target – distractor). This is referred to as the Classification Score. For each stimulus pair in the test protocols we computed this Classification Score and we have such a score per layer.

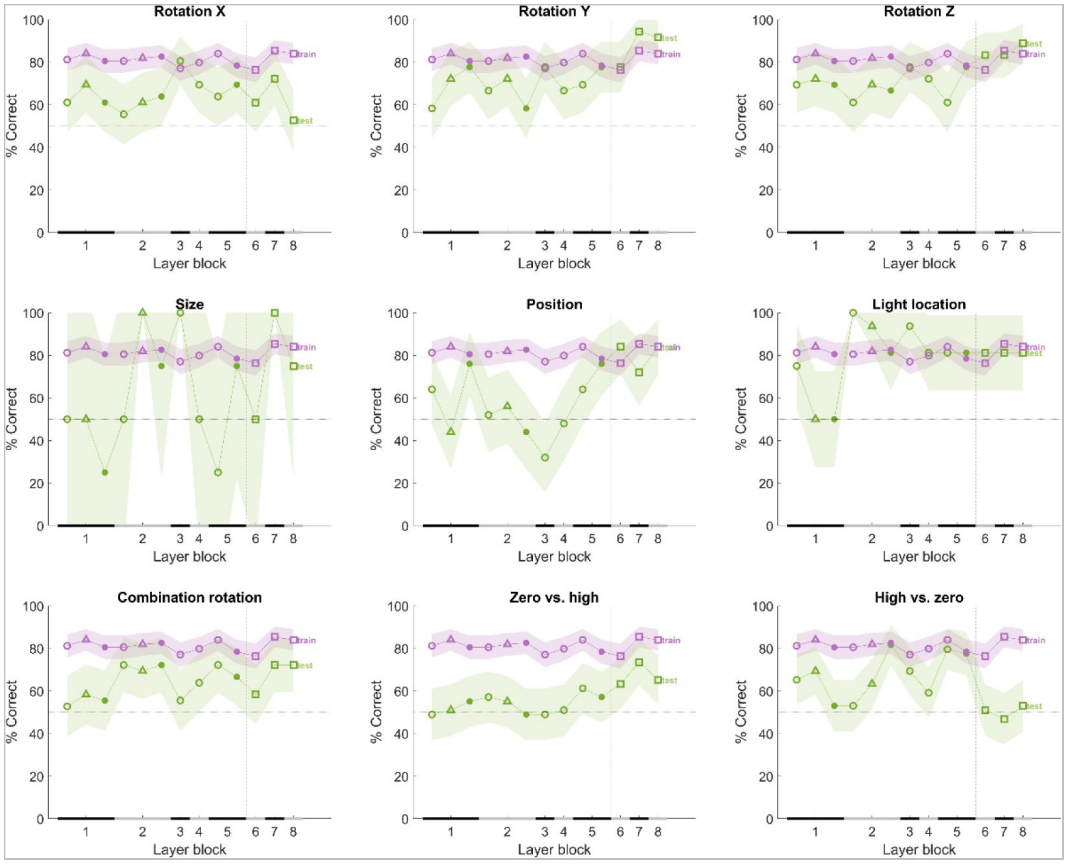


Figure 8

The performance of the CNN after training on our training stimuli, with noise added to its input.

The naming convention on the x axis corresponds to the layers of the network, identical as in **Figure 4**. The performance (y-axis) illustrates that each layer is challenged by at least part of the test protocols. The purple line indicates the training performance and the green line indicates the test performance of the neural network. The x axis on each subplot indicates the block of the layer: layer blocks 1-8 correspond to (convolutional layer 1, normalization layer 1, pool layer 1), (convolutional layer 2, normalization layer 2, pool layer 2), convolutional layer 3, convolutional layer 4, (convolutional layer 5, pool layer 5), fully connected layer 6, fully connected layer 7 and fully connected layer 8, respectively. The black and grey horizontal lines on the x-axis indicated the layer blocks (block 1 consisting of conv1, norm1, pool1; block 2 consisting of conv2, norm2, pool2; block 3-4 corresponding to conv3-4 (respectively); block 5 consisting of conv5, pool5; block 6-7-8 corresponding to fc6-7-8, respectively. The vertical grey dashed line indicates the division between convolutional and fully connected layer blocks. The horizontal dashed line indicates chance level. The shaded error bounds correspond to 95% confidence intervals calculated using Jackknife standard error estimates, as done previously in *(Vinken & Op de Beeck, 2021)*. The different markers indicate different sorts of layers: circle for convolutional layers, triangle for normalization layers, point for pool layers, and squares for fully connected layers.

We used this Classification Score to select image pairs for a CNN-informed stimulus set. To do so, we randomly chose one target and one distractor from a subset of the pool of all 4×4 stimuli, including all possible transformations on these stimuli. This resulted in a stimulus pool of 10.290 stimuli (5145 targets, 5145 distractors) to randomly choose two from, and 5145*5145 (26 471 025) possible resulting pairs of two stimuli. Once one random target and one random distractor was chosen, the DNN was tested in a similar manner as we did for the six test protocols. We performed a total of 10000 iterations of randomly choosing a target and distractor pair. For each iteration, we calculated the average Classification Score of layers 1-3 and of layers 11-13 as we wanted to compare those two levels of processing (earlier layers vs higher layers). After these 10000 iterations, we finetuned and filtered the results according to the profile of performance across earlier and higher layers (see **Supplemental Table 2**). This finetuning started by calculating the distribution and standard deviation for two profiles of interest, i.e. (i) where early layers show an average Classification Score close to zero but higher layers show high Classification Scores (Zero vs High), and (ii) where early layers show high Classification Scores but higher layers show close to zero Classification Scores (High vs Zero). The performance was expressed relative to the distribution of values across all pairs, summarized by de standard deviation of the average target-distractor difference in Classification Scores of the early layers and the higher layers. We found a total of 48 stimulus pairs for these two criteria, and we ended up choosing 14 pairs, 7 of each criterion, that we used for the final part of the animal and human study (see lower two rows in **Figure 1**).

Afterwards we also calculated the binary target-distractor CNN decision performance for the image pairs in the Zero vs High and High vs Zero tests, which is shown in **Figure 8** (bottom row). The image pairs in the Zero vs High protocol are more difficult than the other protocols, in particular for the first half of the CNN layers. In contrast, the *High vs Zero* protocol is the only protocol associated with chance performance in the last three layers. These analyses confirm that the CNN-based image pair selection resulted in protocols that are very different from protocols that zoom in on intuitively chosen transformations and their combinations.

Comparing the rat performances to the Classification Scores of the network was done by calculating the correlation across image pairs between these model scores and the rat performances averaged across animals. We concatenated the performance of the animals on all nine test protocols, as well as the distance to hyperplane of the network on all nine test protocols. Correlating these two arrays resulted in the correlations as visualized in **Figure 4**. To test whether these correlations are significant, we performed a permutation test. We permutated these arrays 1000 times, resulting in a normal distribution of permutated data per layer. We then calculated, per layer, how many of the permutated values are higher than or equal to the correlation that is presented in **Figure 4**, and divided this by the number of permutations.

4.3 Human study

4.3.1 Participants

Data was collected from 50 participants (average age 33.24 ± 12.23; 34 females) who participated in return for a gift voucher of 10 euro. Out of these 50 participants, 5 were excluded because of outlying behaviour during the quality check protocols (see **Section 4.3.3**). All participants had normal or corrected-to-normal vision. The experiment was approved by the ethical commission of KU Leuven (G-2020-1902-R3) and each participant digitally signed an informed consent before the start of the experiment.

4.3.2 Setup

For the human part of this study, we developed an online experiment using PsychoPy3 (v2020.1.3, Python version 3.8.10) and placed it on the online platform Pavlovia. All participants received the link and their individual participant number by e-mail with which they could participate in the experiment on their own computer. It took 30-45 minutes to complete the online study.

4.3.3 Stimuli and protocols

We used the same stimuli as in the animal study. The human experiment underwent the same phases as depicted in **Figure 1**, albeit with small changes. We dropped the 1/3rd old trials in the test protocols and included two additional *Dimension Learning* protocols in between the first counterbalanced tests as quality check (see **Supplemental Figure 7**). **Supplemental Table 3** provides an overview of the number of trials during the human experiment for each phase. **Supplemental Figure 7** shows an overview of all image pairs that were presented in the human study.

Similar as in Bossens & Op de Beeck (2016), we presented the targets and distractors briefly to the left and right side of a white fixation cross on a grey background. Each stimulus was presented for three frames, followed by a mask (a noise image with 1/f frequency spectrum for three frames). We used this fast and eccentric stimulus presentation with a mask to resemble the stimulus perception more closely to that of rats. Vermaercke & Op de Beeck (2012) have found that human visual acuity in these fast and eccentric presentations is not significantly better than the reported visual acuity of rats. By using this approach we avoid that differences in strategies between humans and rats would be explained by such a difference in acuity. Participants could then answer using the 'f' and 'j' keys to indicate which position they thought was the correct position. If they thought the target was on the left side of the fixation cross, they had to press 'f', and 'j' if they thought the target was on the right side. Participants received feedback during the shaping and the three training phases. This happened by colouring the fixation cross green if they answered correctly, and red if they answered incorrectly. Each trial was followed by an intertrial interval (ITI) of 0.5s. During the *Shaping* and *Training* phase, we kept a running average of the past 20 (*Shaping*) and 40 (*Training*) trials and participants continued to the next phase when they reached a performance of 80% or higher on the last 20 or 40 trials, similar as in Bossens & Op de Beeck (2016). There was no time limit for the participants for providing a response. The order of the first six test protocols (*Rotation X, Rotation Y, Rotation Z, Size, Light Location* and *Position)* was counterbalanced between the participants based on the participant number, as well as the order of the last two test protocols (*Zero vs. high* and *High vs. zero*), similar as the approach in the rat study. **Supplemental Table 1** indicates the average number of trials per test protocol for all human participants together.

In terms of instructions, we explained to participants that they would see two figures appearing at the same time very quickly next to a fixation cross, and they would have to make a decision of which figure is the correct one. We mentioned that during training, the fixation cross would turn green if they answered correctly, and red if they answered incorrectly. Participants were informed that during testing, they would not get feedback (changing colour of the fixation cross) anymore and that they would have to use the knowledge they gained throughout training to make their decision in the testing.

We performed a similar correlation analysis as with rat performance to investigate the correlation between pixel similarity and brightness with human performance. We followed the exact same steps as we did for rat performance.

Data availability

The data has been made publicly available via the Open Science Framework and can be accessed at *https://osf.io/9eqyz/*.

Supplementary

Information about the pilot study

The goal of the pilot was to adjust the range of the two dimensions (concavity and alignment) in order to assure that the animals would use the two dimensions. Using the large stimulus set described in the previous paragraph, we tested a subset of them in a behavioural categorization experiment with two rats. This pilot study consisted of seven phases (see **Supplemental Figure 6** for an overview). We started by training rats in a base pair. This pair consists of a target and distractor that are maximally different in terms of concavity and alignment, and thus are placed at the very corners of the 4×11 stimulus grid. After *Training*, we tested them in a *Dimension learning* protocol, to investigate whether the animals use both dimensions (concavity and alignment) in this discrimination task. A total of two pairs were presented to the animals. One pair consisted of the original, base pair, whereas the other pair consisted of stimuli where the alignment dimension was the opposite as in the base pair. In the third phase, *Push towards concavity,* we pushed the animals to learn the concavity dimension. A total of three pairs were presented to the animals. One pair consisted of the original training pair, whereas for the other two stimulus pairs the target and distractor differed in only one of the two dimensions, that is either concavity or alignment. The fourth pilot phase (*Push towards concavity, only new)* consisted of only the two new pairs where the target and distractor differ in only one dimension. A next phase (*More pronounced dimensions)* included stimuli where the concavity dimension was more pronounced than the base pair we started with. We changed the concavity parameter, and again the target and distractor differed in only one dimension. The sixth phase was identical to the *Dimension learning* phase, i.e. checking whether the animals use both dimensions, but this time we used the stimuli where the concavity dimension was more pronounced. In the final phase, we decided to show all four possible combinations of the stimuli.

Tables

Figures

	Rats		Humans	
Protocol	Ø # trials per SP (SD)	Total # trials	Ø # trials per SP (SD)	Total # trials
Rotation X	87.50 (± 9.66)	3150	111.11 (± 2.55)	4000
Rotation Y	88.22 (± 8.90)	3176	111.11 (± 2.44)	4000
Rotation Z	88.25 (± 9.55)	3177	111.11 (± 3.01)	4000
Light location	94.00 (± 11.35)	1504	125 (± 3.72)	2000
Size	205.75 (± 194.86)	823	250 (± 3.16)	1000
Position	91.28 (± 9.06)	2282	120 (± 3.45)	3000
Combination rotation	122.81 (± 12.06)	4421	111.11 (± 3.45)	4000
Zero vs. high	95.53 (± 8.76)	4681	122.45 (± 4.07)	6000
High vs. zero	109.92 (± 10.85)	5386	122.45 (± 4.00)	6000

Supplemental Table 1

Average number of trials (SD) per test protocol and also per stimulus pair (SP).

In the second column, the average number of trials per stimulus pair, for all rats together, are indicated with the standard deviation between brackets. The third column shows the total number of trials per test protocol. The fourth and fifth columns indicate the same data but for the human participants.

Criterion	Early layers	Higher layers	# pairs found
1	Close to zero [-0.4*sd … 0.4*sd]	Very high > 3*sd	22
2	Very high > 3*sd	Close to zero [-0.4*sd … 0.4*sd]	26

Supplemental Table 2

The two criteria of choosing a CNN-informed stimulus set.

We calculated the standard deviation for the early layers and the higher layers. Based on their standard deviation, we then filtered the data and ensured that we had enough pairs per criterion.

Training protocol	# trials
Shaping	80% performance or higher in the last 20 trials
Training	80% performance or higher in the last 40 trials
Dimension learning	80
Transformations	100
Dimension learning (extra)	20

Testing protocol	# trials
Size	20
Position	60
Light location	40
Rotation X	80
Rotation Y	80
Rotation Z	80
Combination rotation	80
Zero vs. high	120
High vs. zero	120

Supplemental Table 3

Overview of the human experiment.

The left table indicates the number of trials for each training protocol. Note that for the first two phases the participants stayed in these protocols until they reached a performance of 80% or higher on the last 20 trials. The number of trials in the testing protocol is proportional to the number of pairs tested per protocol.

Test protocol	Planned # of sessions / animal (average # sessions / animal)	Mean performance on old stimuli (# old trials)	Mean performance on all new trials (# new trials)
Rotation (x-axis)	4 (3.50)	79.39% (1922)	56.49% (3083)
Rotation (y-axis)	4 (3.92)	75.86% (2138)	66.49% (3246)
Rotation (z-axis)	4 (3.83)	79.85% (2092)	63.33% (3315)
Light location	2 (1.92)	79.56% (1043)	72.95% (1645)
Size	1 (1.00)	82.73% (546)	63.49% (895)
Position	3 (2.83)	76.44% (1593)	64.65% (2423)

Supplemental Table 4

An overview of the performance of the animals on the first six test protocols.

The second column indicates the planned number of sessions per animal, and between brackets the average number of sessions per animal. We excluded some sessions if the performance on the old stimuli of that session was too low. The third and fourth column provides an overview of the average performance on the old and new pairs, respectively, with the total number of trials for all rats together between brackets. We excluded the correction trials in the analysis of the old pairs. For the new pairs, no correction trials were given and reward in 80% of the trials was given.

Test protocol	Mean % (old trials)	p-value (80%)	95% CI
Rotation (x-axis)	79.39%	0.50	[0.78;0.81]
Rotation (y-axis)	75.86%	< .0001	[0.74;0.78]
Rotation (z-axis)	79.85%	0.87	[0.78;0.82]
Light location	79.56%	0.74	[0.77;082]
Size	82.73%	0.13	[0.79;0.86]
Position	76.44%	< .0001	[0.74;0.79]

Test protocol	Mean % (new trials)	p-value (80%?)	p-value (50%?)	95% CI
Rotation (x-axis)	56.49%	< 0.0001	< 0.0001	[0.55;0.58]
Rotation (y-axis)	66.49%	< .0001	< 0.0001	[0.65;0.68]
Rotation (z-axis)	63.33%	< 0.0001	< 0.0001	[0.62;0.65]
Light location	72.95%	< 0.0001	< 0.0001	[0.71;0.75]
Size	63.49%	< .0001	< 0.0001	[0.57;0.64]
Position	64.65%	< .0001	< 0.0001	[0.63;0.67]

Supplemental Table 5

Results of binomial test on the six test protocols with the pooled data of all animals together, on the old trials (upper table) and new trials (lower table).

Here we test whether the average performance of all animals together differs significantly from chance level (indicated by *p-value (50%)*) or 80% (indicated by p-value (80%)) on both the old trials (upper table) as well as new trials (lower table).

Rotation X	Target	56.06 ± 8.54	56.60 ± 3.73	59.70 ± 5.01	55.66 ± 6.87	55.53 ± 9.54	55.37 ± 5.71
	Distractor	64.98 ± 2.12	59.21 ± 6.15	55.07 ± 5.64	52.05 ± 5.64	56.04 ± 5.81	51.55 ± 4.35
Rotation Z	Target	65.40 ± 9.54	63.42 ± 6.03	62.88 ± 10.66	60.20 ± 8.48	62.58 ± 6.78	65.52 ± 6.91
	Distractor	72.69 ± 4.47	68.25 ± 3.97	65.31 ± 6.27	60.27 ± 5.65	57.83 ± 6.83	55.64 ± 4.38

Supplemental Table 6

Marginal means and standard deviation of the *Rotation X* and *Rotation Z* protocols.

Layers	Correlation
1-2	0.11
2-3	-0.13
3-4	0.05
4-5	0.09
5-6	0.12
6-7	0.04
7-8	0.17
8-9	0.06
9-10	0.01
10-11	0.003
11-12	0.11
12-13	0.12

Supplemental Table 7

Correlation between neighbouring layers of the deep neural network.

Predictor	T statistic	p-value
(Intercept)	82.15	< 0.0001
Layer 1	1.19	0.23
Layer 2	0.56	0.57
Layer 3	0.91	0.37
Layer 4	0.52	0.60
Layer 5	-0.28	0.78
Layer 6	1.38	0.17
Layer 7	0.59	0.55
Layer 8	2.26	0.025
Layer 9	2.00	0.047
Layer 10	2.35	0.019
Layer 11	-0.46	0.64
Layer 12	-0.82	0.41
Layer 13	1.38	0.17

Supplemental Table 8

Results of the linear regression model with rat performances.

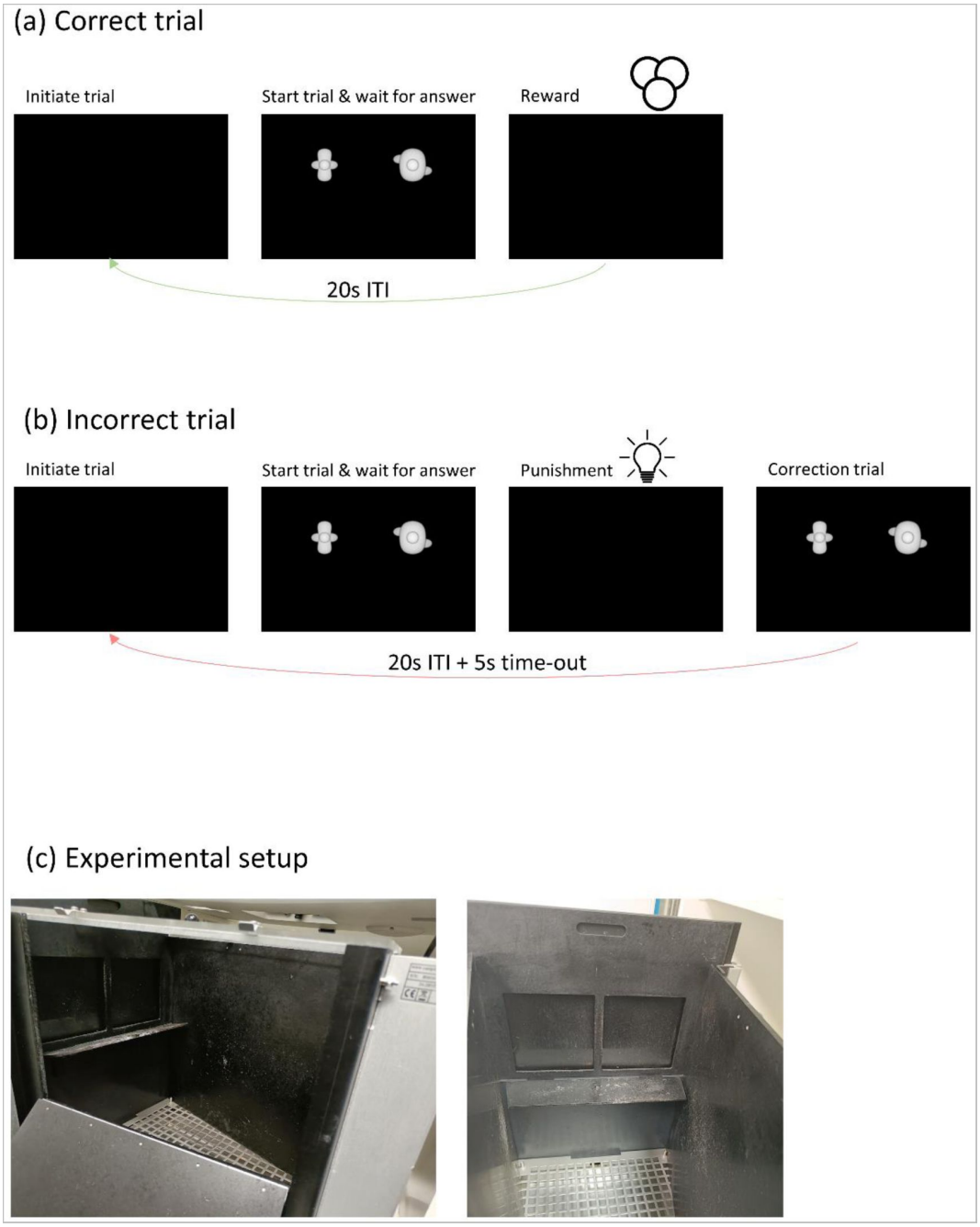
A linear regression was calculated with the Classification Scores of the 13 layers as predictors and the rat performances as response vector. Each row indicates one predictor, i.e. one layer, the middle and right column indicate the t-statistic and p-value of the associated predictor, respectively.

Predictor	T statistic	p-value
(Intercept)	145.94	< 0.0001
Layer 1	0.63	0.53
Layer 2	-1.82	0.07
Layer 3	0.61	0.54
Layer 4	1.40	0.16
Layer 5	1.61	0.11
Layer 6	-1.23	0.22
Layer 7	0.14	0.89
Layer 8	0.62	0.54
Layer 9	0.80	0.43
Layer 10	1.18	0.24
Layer 11	2.90	< 0.01
Layer 12	5.08	< 0.0001
Layer 13	4.72	< 0.0001

Supplemental Table 9

Results of the linear regression model with human performances.

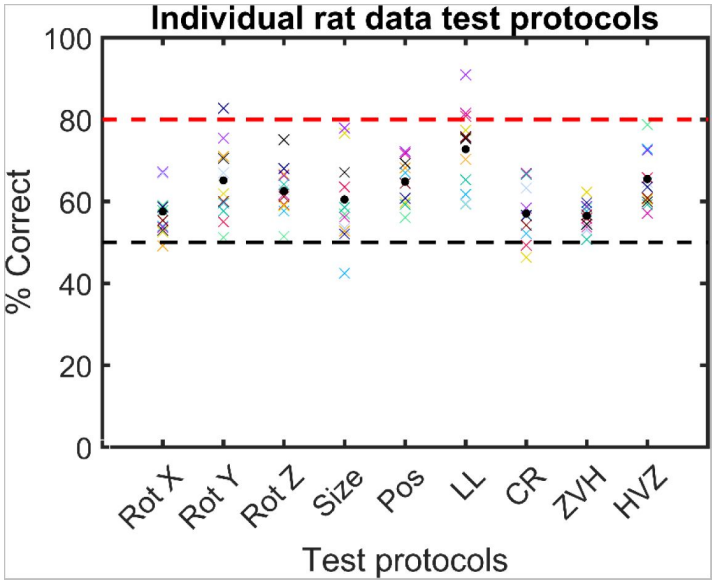
A linear regression was calculated with the Classification Scores of the 13 layers as predictors and the human performances as response vector. Each row indicates one predictor, i.e. one layer, the middle and right column indicate the t-statistic and p-value of the associated predictor, respectively.



Supplemental Figure 1

Timeline of a correct trial (a) and an incorrect trial (b).

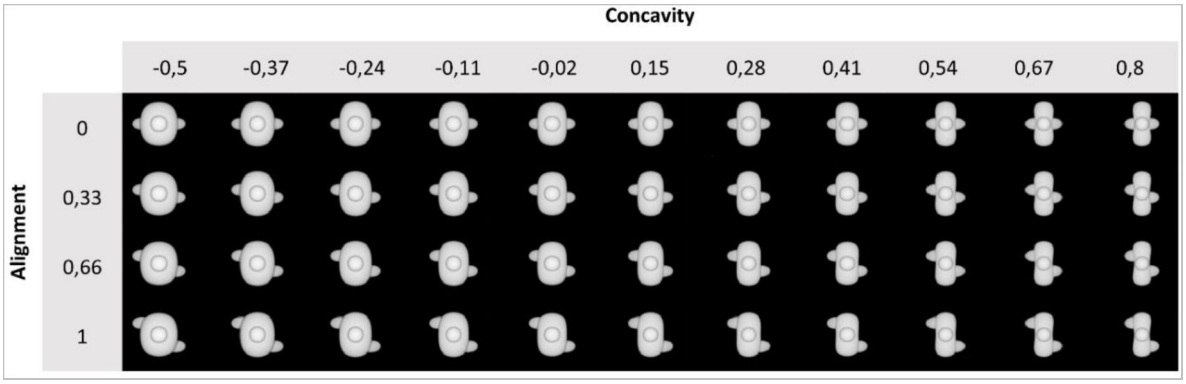
In case of a correct trial, animals receive a reward in the form of sugar pellets for touching the target. If they answer incorrectly, they do not receive sugar pellets but instead the house light is turned on. For training protocols, a correction trial is added.(c) shows images of the experimental setup.



Supplemental Figure 2

Individual rat accuracy for each testing protocol.

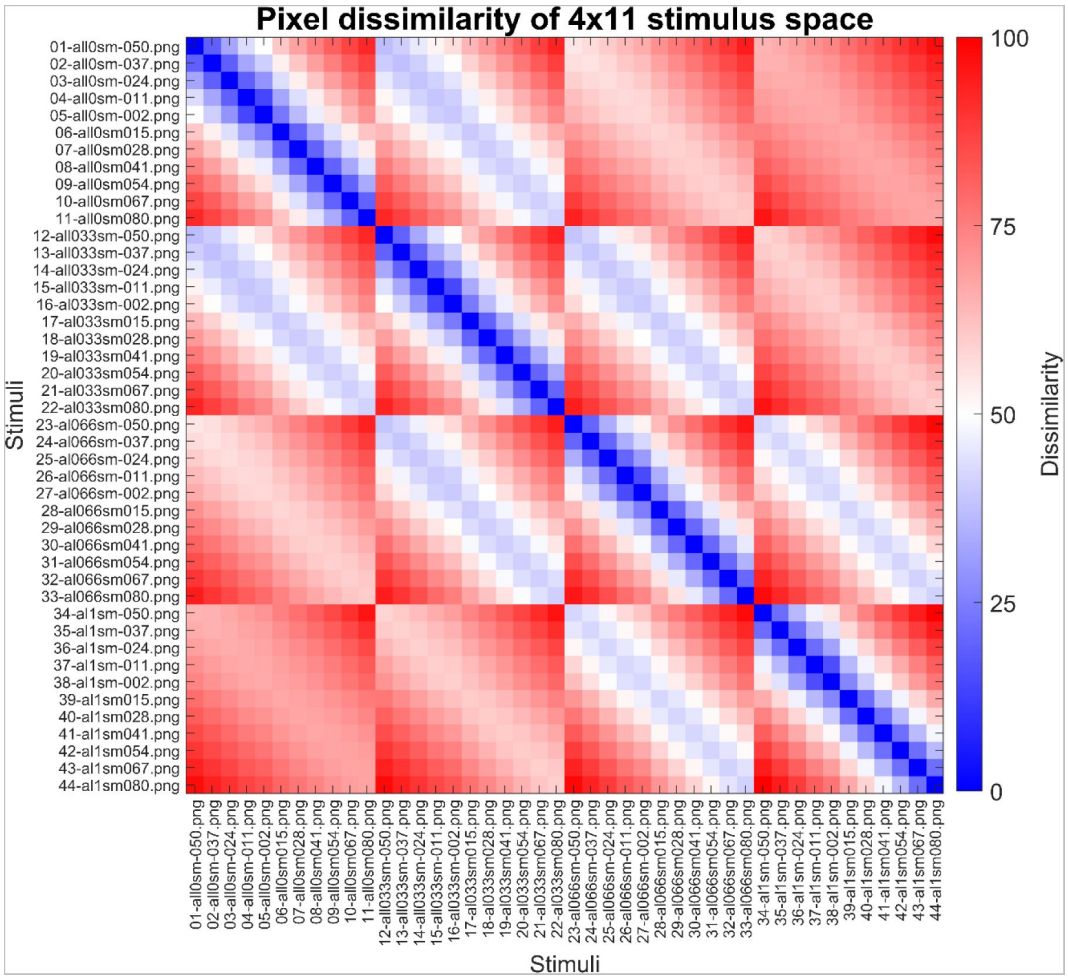
The black circle indicates the average accuracy across all rats together. Each coloured cross represents the accuracy of one rat.



Supplemental Figure 3

Illustration of the 4×11 stimulus grid.

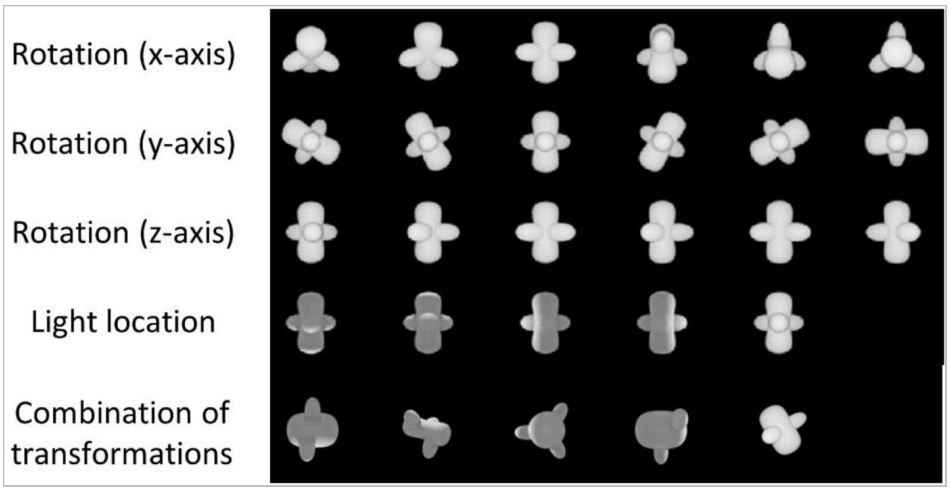
We chose eleven values for concavity, ranging between −0.5 and 0.8, and four values for alignment, ranging between 0 and 1.



Supplemental Figure 4

The pixel dissimilarity matrix of the 4×11 stimulus grid.

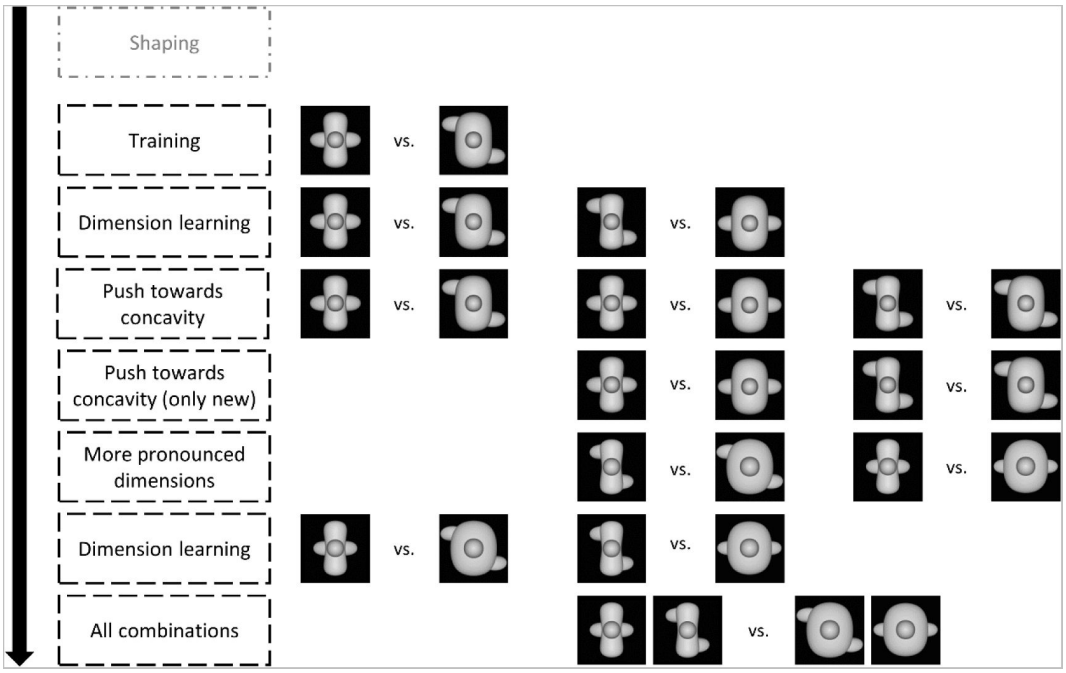
This matrix is calculated in the exact same way as (Schnell, et al., 2023). The colourbar indicates pixel dissimilarity values, normalized between 0 and 100. The higher this value, and thus the more red a cell, the more dissimilar two stimuli are. The naming convention of the stimuli is as follows: XX-alYYsmZZZ.png where XX is an ordering number, YY indicates the alignment level and ZZZ indicates the concavity (smoothness) level.



Supplemental Figure 5

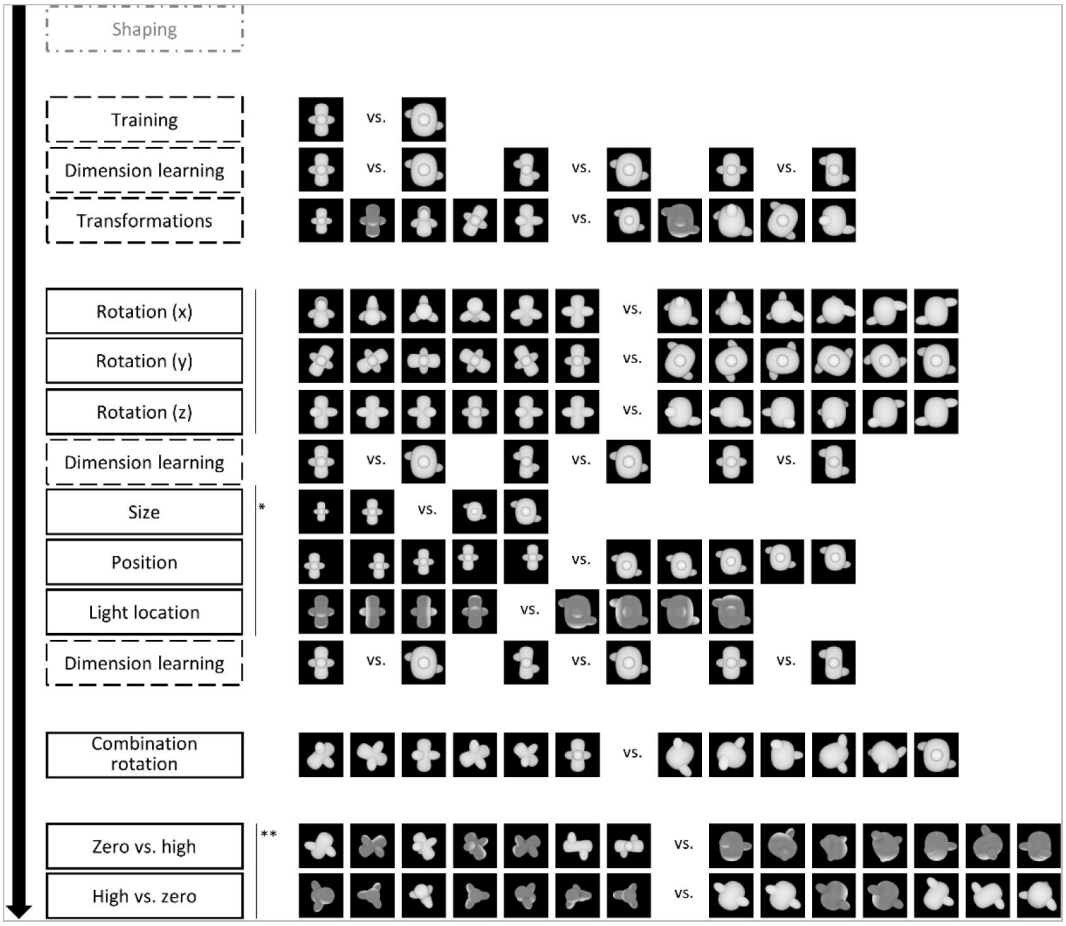
Identity-preserving transformations on one of the basic stimuli.

(a), (b) and (c) show the rotation along the x-axis, y-axis and z-axis respectively, always in angles of 30°, 60°, 90°, 120°, 150°, and 180°. (d) illustrates the light location transformation. (e) shows some combinations of transformations, with different values for the alignment and concavity parameter.



Supplemental Figure 6

Design of the pilot study.



Supplemental Figure 7

Design of the online human study.

The design of the human study resembles the design of the animal study, with minor adaptions. We removed the 1/3rd old trials in the test protocols, but included two additional *Dimension Learning* protocols.

References

Alemi-Neissi A., Rosselli F. B., Zoccolan D. (2013) **Multifeatural Shape Processing in Rats Engaged in Invariant Visual Object Recognition** *Journal of Neuroscience* **33**:5939–5956 https://doi.org/10.1523/JNEUROSCI.3629-12.2013

Avberšek L. K., Zeman A., Op de Beeck H. (2021) **Training for object recognition with increasing spatial frequency: A comparison of deep learning with human vision** *Journal of Vision* **21** https://doi.org/10.1167/jov.21.10.14

Bossens C., Op de Beeck H. P. (2016) **Linear and Non-Linear Visual Feature Learning in Rat and Humans** *Frontiers in Behavioral Neuroscience* **10** https://doi.org/10.3389/fnbeh.2016.00235

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** https://doi.org/10.1371/journal.pcbi.1003963

Callahan T. L., Petry H. M. (2000) **Psychophysical measurement of temporal modulation sensitivity in the tree shrew (Tupaia belangeri)** *Vision Research* **40**:455–458 https://doi.org/10.1016/S0042-6989(99)00194-7

Crijns E., Op de Beeck H. (2019) **The Visual Acuity of Rats in Touchscreen Setups** *Vision* **4** https://doi.org/10.3390/vision4010004

De Keyser R., Bossens C., Kubilius J., Op de Beeck H. P. (2015) **Cue-invariant shape recognition in rats as tested with second-order contours** *Journal of Vision* **15** https://doi.org/10.1167/15.15.14

Djurdjevic V., Ansuini A., Bertolini D., Macke J. H., Zoccolan D. (2018) **Accuracy of Rats in Discriminating Visual Objects Is Explained by the Complexity of Their Perceptual Strategy** *Current Biology* **28**:1005–1015 https://doi.org/10.1016/j.cub.2018.02.037

Duyck S., Martens F., Chen C.-Y., Op de Beeck H. (2021) **How Visual Expertise Changes Representational Geometry: A Behavioral and Neural Perspective** *Journal of Cognitive Neuroscience* **33**:2461–2476 https://doi.org/10.1162/jocn_a_01778

Groen I. I., Greene M. R., Baldassano C., Fei-Fei L., Beck D. M., Baker C. I. (2018) **Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior** *ELife* **7** https://doi.org/10.7554/eLife.32962

Güçlü U., Gerven M. A. J. van (2015) **Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream** *Journal of Neuroscience* **35**:10005–10014 https://doi.org/10.1523/JNEUROSCI.5023-14.2015

Kalfas I., Vinken K., Vogels R. (2018) **Representations of regular and irregular shapes by deep Convolutional Neural Networks, monkey inferotemporal neurons and human judgments** *PLOS Computational Biology* **14** https://doi.org/10.1371/journal.pcbi.1006557

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** *Nature Neuroscience* **22** https://doi.org/10.1038/s41593-019-0392-5

Kell A. J. E., Bokor S. L., Jeon Y.-N., Toosi T., Issa E. B. (2020) **Conserved core visual object recognition across simian primates: Marmoset image-by-image behavior mirrors that of humans and macaques**

Kell A. J. E., Bokor S. L., Jeon Y.-N., Toosi T., Issa E. B. (2021) **Brain organization, not size alone, as key to high-level vision: Evidence from marmoset monkeys** *bioRxiv* https://doi.org/10.1101/2020.10.19.345561

Kubilius J., Bracci S., Op de Beeck H. (2016) **Deep Neural Networks as a Computational Model for Human Shape Sensitivity** *PLOS Computational Biology* **12** https://doi.org/10.1371/journal.pcbi.1004896

Logothetis N. K., Sheinberg D. L. (1996) **Logothetis, N. K., & Sheinberg, D. L. (1996). Visual Object Recognition. 45.** *Visual Object Recognition* **45**

Matteucci G., Marotti R., Riggi M., Rosselli F. B., Zoccolan D. (2019) **Nonlinear Processing of Shape Information in Rat Lateral Extrastriate Cortex | Journal of Neuroscience** *Journal of Neuroscience* https://doi.org/10.1523/JNEUROSCI.1938-18.2018

Meyer E. E., Ong W. S., Balboa M., Arcaro M. J. (2022) **Assessing tree shrew high-level visual behavior using conventional and natural paradigms [Poster]** *Society for Neuroscience*

Minini L., Jeffery K. J. (2006) **Do rats use shape to solve "shape discriminations"?** *Learning & Memory* **13**:287–297 https://doi.org/10.1101/lm.84406

Nayebi A., Zhuang C., Norcia A. M., Kong N. C. L., Gardner J. L., Yamins D. L. K. (2023) **Mouse visual cortex as a limited resource system that self-learns an ecologically-general representation** *bioRxiv* https://doi.org/10.1101/2021.06.16.448730

Ohayon S., Freiwald W. A., Tsao D. Y. (2012) **What Makes a Cell Face Selective? The Importance of Contrast** *Neuron* **74**:567–581 https://doi.org/10.1016/j.neuron.2012.03.024

Op de Beeck H., Torfs K., Wagemans J. (2008) **Perceived Shape Similarity among Unfamiliar Objects and the Organization of the Human Object Vision Pathway** *Journal of Neuroscience* **28**:10111–10123 https://doi.org/10.1523/JNEUROSCI.2511-08.2008

Petry H. M., Bickford M. E. (2019) **The Second Visual System of The Tree Shrew** *Journal of Comparative Neurology* **527**:679–693 https://doi.org/10.1002/cne.24413

Petry H. M., Clark C., Day-Brown J., Bolin R. T., Bickford M. (2012) **Behavioral measurement of RDK velocity discrimination thresholds in the tree shrew** *Journal of Vision* **12** https://doi.org/10.1167/12.9.1223

Pinto N., Cox D. D., DiCarlo J. J. (2008) **Why is Real-World Visual Object Recognition Hard?** *PLOS Computational Biology* **4** https://doi.org/10.1371/journal.pcbi.0040027

Pospisil D. A., Pasupathy A., Bair W. (2018) **"Artiphysiology" reveals V4-like shape tuning in a deep network trained for image classification** *ELife* **7** https://doi.org/10.7554/eLife.38242

Schnell A. E., Bergh G. V. den, Vermaercke B., Gijbels K., Bossens C., Op de Beeck H. (2019) **Face categorization and behavioral templates in rats** *Journal of Vision* **19**:9–9 https://doi.org/10.1167/19.14.9

Schnell A. E., Vinken K., Op de Beeck H. (2023) **The importance of contrast features in rat vision** *Scientific Reports* **13** https://doi.org/10.1038/s41598-023-27533-3

Sinha P., Bülthoff H. H., Wallraven C., Lee S.-W., Poggio T. A. (2002) **Qualitative Representations for Recognition** *Biologically Motivated Computer Vision* :249–262 https://doi.org/10.1007/3-540-36181-2_25

Tafazoli S., Filippo A. D., Zoccolan D. (2012) **Transformation-Tolerant Object Recognition in Rats Revealed by Visual Priming** *Journal of Neuroscience* **32**:21–34 https://doi.org/10.1523/JNEUROSCI.3932-11.2012

Tafazoli S., Safaai H., De Franceschi G., Rosselli F. B., Vanzella W., Riggi M., Buffolo F., Panzeri S., Zoccolan D. (2017) **Emergence of transformation-tolerant representations of visual objects in rat lateral extrastriate cortex** *ELife* **6** https://doi.org/10.7554/eLife.22794

Vermaercke B., Gerich F. J., Ytebrouck E., Arckens L., Op de Beeck H. P., Van den Bergh G. (2014) **Functional specialization in rat occipital and temporal visual cortex** *Journal of Neurophysiology* **112**:1963–1983 https://doi.org/10.1152/jn.00737.2013

Vermaercke B., Op de Beeck H. P. (2012) **A Multivariate Approach Reveals the Behavioral Templates Underlying Visual Discrimination in Rats** *Current Biology* **22**:50–55 https://doi.org/10.1016/j.cub.2011.11.041

Vinken K., Op de Beeck H. (2021) **Using deep neural networks to evaluate object vision tasks in rats** *PLOS Computational Biology* **17** https://doi.org/10.1371/journal.pcbi.1008714

Vinken K., Van den Bergh G., Vermaercke B., Op de Beeck H. P. (2016) **Neural Representations of Natural and Scrambled Movies Progressively Change from Rat Striate to Temporal Cortex** *Cerebral Cortex* **26**:3310–3322 https://doi.org/10.1093/cercor/bhw111

Vinken K., Vermaercke B., Op de Beeck H. (2014) **Visual Categorization of Natural Movies by Rats** *Journal of Neuroscience* **34**:10645–10658 https://doi.org/10.1523/JNEUROSCI.3663-13.2014

Yue X., Cassidy B. S., Devaney K. J., Holt D. J., Tootell R. B. H. (2011) **Lower-Level Stimulus Features Strongly Influence Responses in the Fusiform Face Area** *Cerebral Cortex* **21**:35–47 https://doi.org/10.1093/cercor/bhq050

Zeman A. A., Ritchie J. B., Bracci S., Op de Beeck H. (2020) **Orthogonal Representations of Object Shape and Category in Deep Convolutional Neural Networks and Human Visual Cortex** *Scientific Reports* **10** https://doi.org/10.1038/s41598-020-59175-0

Zoccolan D. (2015) **Invariant visual object recognition and shape processing in rats** *Behavioural Brain Research* **285**:10–33 https://doi.org/10.1016/j.bbr.2014.12.053

Zoccolan D., Oertelt N., DiCarlo J. J., Cox D. D. (2009) **A rodent model for the study of invariant visual object recognition** *Proceedings of the National Academy of Sciences* **106**:8748–8753 https://doi.org/10.1073/pnas.0811583106

Article and author information

Anna Elisabeth Schnell
Department of Brain and Cognition & Leuven Brain Institute, KU Leuven, Leuven, Belgium
For correspondence: annaelisabeth.schnell@kuleuven.be

Maarten Leemans
Department of Brain and Cognition & Leuven Brain Institute, KU Leuven, Leuven, Belgium

Kasper Vinken
Department of Neurobiology, Harvard Medical School, Boston, Massachusetts, United States of America
ORCID iD: 0000-0001-7038-9638

Hans Op de Beeck
Department of Brain and Cognition & Leuven Brain Institute, KU Leuven, Leuven, Belgium

Copyright

Editors

Reviewing Editor
SP Arun
Indian Institute of Science Bangalore, India

Senior Editor
Joshua Gold
University of Pennsylvania, United States of America

Reviewer #1 (Public Review):

Schnell et al. performed two extensive behavioral experiments concerning the processing of objects in rats and humans. To this aim, they designed a set of objects parametrically varying along alignment and concavity and then they used activations from a pretrained deep convolutional neural network to select stimuli that would require one of two different discrimination strategies, i.e. relying on either low- or high-level processing exclusively. The results show that rodents rely more on low-level processing than humans.

Strengths:

1. The results are challenging and call for a different interpretation of previous evidence. Indeed, this work shows that common assumptions about task complexity and visual processing are probably biased by our personal intuitions and are not equivalent in rodents, which instead tend to rely more on low-level properties.
2. This is an innovative (and assumption-free) approach that will prove useful to many visual neuroscientists. Personally, I second the authors' excitement about the proposed approach,

and its potential to overcome the limits of experimenters' creativity and intuitions. In general, the claims seem well supported and the effects sufficiently clear.
3. This work provides an insightful link between rodent and human literature on object processing. Given the increasing number of studies on visual perception involving rodents, these kinds of comparisons are becoming crucial.
4. The paper raises several novel questions that will prompt more research in this direction.

Weaknesses:
1. The choice of alignment and concavity as baseline properties of the stimuli is not properly discussed.
2. From the low-correlations I got the feeling that AlexNet is not the best baseline model for rat visual processing.

Reviewer #2 (Public Review):

Schnell and colleagues trained rats on a two-alternative forced choice visual discrimination task. They used object pairs that differed in their concavity and the alignment of features. They found that rats could discriminate objects across various image transformations. Rat performance correlated best with late convolutional layers of an artificial neural network and was partially explained by factors of brightness and pixel-level similarity. In contrast, human performance showed the strongest correlation with higher, fully connected layers, indicating that rats employed simpler strategies to accomplish this task as compared to humans.

Strengths:
1. This is a methodologically rigorous study. The authors tested a substantial number of rats across a large variety of stimuli.
2. The innovative use of neural networks to generate stimuli with varying levels of complexity is a compelling approach that motivates principled experimental design.
3. The study provides important data points for cross-species comparisons of object discrimination behavior
4. The data strongly support the authors' conclusion that rats and humans rely on different visual features for discrimination tasks.
5. This is a valuable study that provides novel, important insights into the visual capabilities of rats.

Weaknesses:
1. The impact of rat visual acuity (~1cycle/degree) on the discriminability of stimuli could be more directly modeled and taken into consideration when comparing rat behavior to humans, who possess substantially higher acuity.
2. The distinction between low- and high-level visual behavior is coarse, and it remains uncertain which specific features rats utilized for discrimination. The correlations with brightness and pixel-level similarity do provide some insight.
3. The relatively weak correspondence between rat behavior and AlexNet raises the question of which network architecture, whether computational or biological, might better capture rat behavior, particularly to the level of cross-rat consistency.

Author response

The following is the authors’ response to the original reviews.

We would like to thank you for giving us the opportunity to submit a revised draft our manuscript. We appreciate the time and effort that you dedicated to providing insightful feedback on our manuscript and are grateful for the valuable comments and improvements on our paper. It helped us to improve our manuscript. We have carefully considered the

comments and tried our best to address every one of them. We have added clarifications in the Discussion concerning the type of neural network that we used, about which visual features might play a role in our results as well as clarified the experimental setup and protocol in the Methods section as these two sections were lacking key information points.

Below we provide a response to the public comments and concerns of the reviewers.

Several key points were addressed by at least two reviewers, and we will respond to them first.

A first point concerns the type of network we used. In our study, we used AlexNet to simulate the ventral visual stream and to further examine rat and human performance. While other, more complex neural networks might lead to other results, we chose to work with AlexNet because it has been used in many other vision studies that are published in high impact journals ((Cadieu et al., 2014; Groen et al., 2018; Kalfas et al., 2018; Nayebi et al., 2023; Zeman et al., 2020). We did not try to find a best DNN model for rat vision but instead, we were looking for an explanation of which visual features play a role in our complex discrimination task. We added a consideration to our Discussion addressing why we worked with AlexNet. Since our data will be published on OSF, we encourage to researchers to use our data with other, more complex neural networks and to further investigate this issue.

A second point that was addressed by multiple reviewers concerns the visual acuity of the animals and its impact on their performance. The position of the rat was not monitored in the setup. In a previous study in our lab (Crijns & Op de Beeck, 2019), we investigated the visual acuity of rats in the touchscreen setups by presenting gratings with different cycles per screen to see how it affects their performance in orientation discrimination. With the results from this study and general knowledge about rat visual acuity, we derived that the decision distance of rats lies around 12.5cm from the screen. We have added this paragraph to the Discussion.

A third key point that needs to be addressed as a general point involves which visual features could explain rat and human performance. We reported marked differences between rat and human data in how performance varied across image trials, and we concluded through our computationally informed tests and analyses that rat performance was explained better by lower levels of processing. Yet, we did not investigate which exact features might underlie rat performance. As a starter, we have focused on taking a closer look at pixel similarity and brightness and calculating the correlation between rat/human performance and these two visual features.

We calculated the correlation between the rat performances and image brightness of the transformations. We did this by calculating the difference in brightness of the base pair (brightness base target – brightness base distractor), and subtracting the difference in brightness of every test target-distractor pair for each test protocol (brightness test target – brightness test distractor for each test pair). We then correlated these 287 brightness values (1 for each test image pair) with the average rat performance for each test image pair. This resulted in a correlation of 0.39, suggesting that there is an influence of brightness in the test protocols. If we perform the same correlation with the human performances, we get a correlation of -0.12, suggesting a negative influence of brightness in the human study.

We calculated the correlation between pixel similarity of the test stimuli in relation to the base stimuli with the average performance of the animals on all nine test protocols. We did this by calculating the pixel similarity between the base target with every other testing distractor (A), the pixel similarity between the base target with every other testing target (B), the pixel similarity between the base distractor with every other testing distractor (C) and the pixel similarity between the base distractor with every other testing target (D). For each test image pair, we then calculated the average of (A) and (D), and subtracted the average of (C)

and (B) from it. We correlated these 287 values (one for each image pair) with the average rat performance on all test image pairs, which resulted in a correlation of 0.34, suggesting an influence of pixel similarity in rat behaviour. Performing the same correlation analysis with the human performances results in a correlation of 0.12.

We have also addressed this in the Discussion of the revised manuscript. Note that the reliability of the rat data was 0.58, clearly higher than the correlations with brightness and pixel similarity, thus these features capture only part of the strategies used by rats.

We have also responded to all other insightful suggestions and comments of the reviewers, and a point-by-point response to the more major comments will follow now.

> ### *Reviewer #1, general comments:*
>
> *The authors should also discuss the potential reason for the human-rat differences too, and importantly discuss whether these differences are coming from the rather unusual approach of training used in rats (i.e. to identify one item among a single pair of images), or perhaps due to the visual differences in the stimuli used (what were the image sizes used in rats and humans?). Can they address whether rats trained on more generic visual tasks (e.g. same-different, or category matching tasks) would show similar performance as humans?*

The task that we used is typically referred to as a two-alternative forced choice (2AFC). This is a simple task to learn. A same-different task is cognitively much more demanding, also for artificial neural networks (see e.g. Puebla & Bowers, 2022, J. Vision). A one-stimulus choice task (probably what the reviewer refers to with category matching) is known to be more difficult compared to 2AFC, with a sensitivity that is predicted to be Sqrt(2) lower according to signal detection theory (MacMillan & Creelman, 1991). We confirmed this prediction empirically in our lab (unpublished observations). Thus, we predict that rats perform less good in the suggested alternatives, potentially even (in case of same-different) resulting in a wider performance gap with humans.

> *I also found that a lot of essential information is not conveyed clearly in the manuscript. Perhaps it is there in earlier studies but it is very tedious for a reader to go back to some other studies to understand this one. For instance, the exact number of image pairs used for training and testing for rats and humans was either missing or hard to find out. The task used on rats was also extremely difficult to understand. An image of the experimental setup or a timeline graphic showing the entire trial with screenshots would have helped greatly.*

All the image pairs used for training and testing for rats and humans are depicted in Figure 1 (for rats) and Supplemental Figure 6 (for humans). For the first training protocol (Training), only one image pair was shown, with the target being the concave object with horizontal alignment of the spheres. For the second training protocol (Dimension learning), three image pairs were shown, consisting of the base pair, a pair which differs only in concavity, and a pair which differs only in alignment. For the third training protocol (Transformations) and all testing protocols, all combination of targets and distractors were presented. For example, in the Rotation X protocol, the stimuli consisted of 6 targets and 6 distractors, resulting in a total of 36 image pairs for this protocol. The task used on rats is exactly as shown in Figure 1. A trial started with two blank screens. Once the animal initiated a trial by sticking its head in the reward tray, one stimulus was presented on each screen. There was no time limit and so the stimuli remained on the screen until the animal made a decision. If the animal touched the target, it received a sugar pellet as reward and a ITI of 20s started. If the animal touched the distractor, it did not receive a sugar pellet and a time-out of 5s started in addition to the 20s ITI.

We have clarified this in the manuscript.

> *The authors state that the rats received random reward on 80% of the trials, but is that on 80% of the correctly responded trials or on 80% of trials regardless of the correctness of the response? If these are free choice experiments, then the task demands are quite different. This needs to be clarified. Similarly, the authors mention that 1/3 of the trials in a given test block contained the old base pair - are these included in the accuracy calculations?*

The animals receive random reward on 80% on all testing trials with new stimuli, regardless of the correctness of the response. This was done to ensure that we can measure true generalization based upon learning in the training phase, and that the animals do not learn/are not trained in these testing stimuli. For the trials with the old stimuli (base pair), the animals always received real reward (reward when correct; no reward in case of error).

The 1/3rd trials with old stimuli are not included in the accuracy calculations but were used as a quality check/control to investigate which sessions have to be excluded and to assure that the rats were still doing the task properly. We have added this in the manuscript.

> *The authors were injecting noise with stimuli to cDNN to match its accuracy to rat. However, that noise potentially can interacted with the signal in cDNN and further influence the results. That could generate hidden confound in the results. Can they acknowledge/discuss this possibility?*

Yes, adding noise can potentially interact with the signal and further influence the results. Without noise, the average training data of the network would lie around 100% which would be unrealistic, given the performances of the animals. To match the training performance of the neural networks with that of the rats, we added noise 100 times and averaged over these iterations (cfr. (Schnell et al., 2023; Vinken & Op de Beeck, 2021)).

> ***Reviewer #2, weaknesses:***
>
> 1. *There are a few inconsistencies in the number of subjects reported. Sometimes 45 humans are mentioned and sometimes 50. Probably they are just typos, but it's unclear.*

Thank you for your feedback. We have doublechecked this and changed the number of subjects where necessary. We collected data from 50 human participants, but had to exclude 5 of them due to low performance during the quality check (Dimension learning) protocols. Similarly, we collected data from 12 rats but had to exclude one animal because of health issues. All these data exclusion steps were mentioned in the Methods section of the original version of the manuscript, but the subject numbers were not always properly adjusted in the description in the Results section. This is now corrected.

> 1. *A few aspects mentioned in the introduction and results are only defined in the Methods thus making the manuscript a bit hard to follow (e.g. the alignment dimension), thus I had to jump often from the main text to the methods to get a sense of their meaning.*

Thank you for your feedback. We have clarified some aspects in the Introduction, such as the alignment dimension.

> 1. *Many important aspects of the task are not fully described in the Methods (e.g. size of the stimuli, reaction times and basic statistics on the responses).*

We have added the size of the stimuli to the Methods section and clarified that the stimuli remained on the screen until the animals made a choice. Reaction time in our task would not be interpretable given that stimuli come on the screen when the animal initiates a trial with its back to the screen. Therefore we do not have this kind of information.

> ***Reviewer #1***
>
> *• Can the authors show all the high vs zero and zero vs high stimulus pairs either in the main or supplementary figures? It would be instructive to know if some other simple property covaried between these two sets.*

In Figure 1, all images of all protocols are shown. For the High vs. Zero and Zero vs. High protocols, we used a deep neural network to select a total of 7 targets and 7 distractors. This results in 49 image pairs (every combination of target-distractor).

> *• Are there individual differences across animals? It would be useful for the authors to show individual accuracy for each animal where possible.*

We now added individual rat data for all test protocols – 1 colour per rat, black circle = average. We have added this picture to the Supplementary material (Supplementary Figure 1).

> *• Figure 1 - it was not truly clear to me how many image pairs were used in the actual experiment. Also, it was very confusing to me what was the target for the test trials. Additionally, authors reported their task as a categorisation task, but it is a discrimination task.*

Figure 1 shows all the images that were used in this study. Every combination of every target-distractor in each protocol (except for Dimension learning) was presented to the animals. For example in Rotation X, the test stimuli as shown in Fig. 1 consisted of 6 targets and 6 distractors, resulting in a total of 36 image pairs for this test protocol.

In each test protocol, the target corresponded to the concave object with horizontally attached spheres, or the object from the pair that in the stimulus space was closed to this object. We have added this clarification in the Introduction: “We started by training the animals in a base stimulus pair, with the target being the concave object with horizontally aligned spheres. Once the animals were trained in this base stimulus pair, we used the identity-preserving transformations to test for generalization.” as well as in the caption of Figure 1. We have changed the term “categorisation task” to “discrimination task” throughout the manuscript.

> *• Figure 2 - what are the red and black lines? How many new pairs are being tested here? Panel labels are missing (a/b/c etc)*

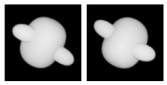
We have changed this figure by adding panel labels, and clarifying the missing information in the caption. All images that were shown to the animals are presented on this figure. For Dimension Learning, only three image pairs were shown (base pair, concavity pair, alignment pair) and for the Transformations protocol, every combination of every target and distractor were shown, i.e. 25 image pairs in total.

> *• Figure 3 - last panel: the 1st and 2nd distractor look identical.*

We understand your concern as these two distractors indeed look quite similar. They are different however in terms of how they are rotated along the x, y and z axes (see Author

response image 1 for a bigger image of these two distractors). The similarity is due to the existence of near-symmetry in the object shape which causes high self-similarity for some large rotations.

Author response image 1.



> *• Line 542 – authors say they have 'concatenated' the performance of the animals, but do they mean they are taking the average across animals?*

It is both. In this specific analysis we calculated the performance of the animals, which was indeed averaged across animals, per test protocol, per stimulus pair. This resulted in 9 arrays (one for each test protocol) of several performances (1 for each stimulus pair). These 9 arrays were concatenated by linking them together in one big array (i.e. placing them one after the other). We did the same concatenation with the distance to hyperplane of the network on all nine test protocols. These two concatenated arrays with 287 values each (one with the animal performance and one with the DNN performance) were correlated.

> *• Line 164 - What are these 287 image pairs - this is not clear.*

The 287 image pairs correspond to all image pairs of all 9 test protocols: 36 (Rotation X) + 36 (Rotation Y) + 36 (Rotation Z) + 4 (Size) + 25 (Position) + 16 (Light location) + 36 (Combination Rotation) + 49 (Zero vs. high) + 49 (High vs. zero) = 287 image pairs in total. We have clarified this in the manuscript.

> *• Line 215 - Human rat correlation (0.18) was comparable to the best cDNN layer correlation. What does this mean?*

The human rat correlation (0.18) was closest to the best cDNN layer - rat correlation (about 0.15). In the manuscript we emphasize that rat performance is not well captured by individual cDNN layers.

> ### *Reviewer #2*
>
> *Major comments*
>
> *• In l.23 (and in the methods) the authors mention 50 humans, but in l.87 they are 45. Also, both in l.95 and in the Methods the authors mention "twelve animals" but they wrote 11 elsewhere (e.g. abstract and first paragraph of the results).*

In our human study design, we introduced several Dimension learning protocols. These were later used as a quality check to indicate which participants were outliers, using outlier detection in R. This resulted in 5 outlying human participants, and thus we ended with a pool of 45 human participants that were included in the analyses. This information was given in the Methods section of the original manuscript, but we did not mention the correct numbers everywhere. We have corrected this in the manuscript. We also changed the number of participants (humans and rats) to the correct one throughout the entire manuscript.

> *• At l.95 when I first met the "4x4 stimulus grid" I had to guess its meaning. It would be really useful to see the stimulus grid as a panel in Figure 1 (in general Figures S1 and S4*

> *could be integrated as panels of Figure 1). Also, even if the description of the stimulus generation in the Methods is probably clear enough, the authors might want to consider adding a simple schematic in Figure 1 as well (e.g. show the base, either concave or convex, and then how the 3 spheres are added to control alignment).*

We have added the 4x4 stimulus grid in the main text.

> *• There is also another important point related to the choice of the network. As I wrote, I find the overall approach very interesting and powerful, but I'm actually worried that AlexNet might not be a good choice. I have experience trying to model neuronal responses from IT in monkeys, and there even the higher layers of AlexNet aren't that helpful. I need to use much deeper networks (e.g. ResNet or GoogleNet) to get decent fits. So I'm afraid that what is deemed as "high" in AlexNet might not be as high as the authors think. It would be helpful, as a sanity check, to see if the authors get the same sort of stimulus categories when using a different, deeper network.*

We added a consideration to the manuscript about which network to use (see the Discussion): "We chose to work with Alexnet, as this is a network that has been used as a benchmark in many previous studies (e.g. (Cadieu et al., 2014; Groen et al., 2018; Kalfas et al., 2018; Nayebi et al., 2023; Zeman et al., 2020)), including studies that used more complex stimuli than the stimulus space in our current study. [...] . It is in line with the literature that a typical deep neural network, AlexNet and also more complex ones, can explain human and animal behaviour to a certain extent but not fully. The explained variance might differ among DNNs, and there might be DNNs that can explain a higher proportion of rat or human behaviour. Most relevant for our current study is that DNNs tend to agree in terms of how representations change from lower to higher hierarchical layers, because this is the transformation that we have targeted in the Zero vs. high and High vs. zero testing protocols. (Pinto et al., 2008) already revealed that a simple V1-like model can sometimes result in surprisingly good object recognition performance. This aspect of our findings is also in line with the observation of Vinken & Op de Beeck (2021) that the performance of rats in many previous tasks might not be indicative of highly complex representations. Nevertheless, there is still a relative difference in complexity between lower and higher levels in the hierarchy. That is what we capitalize upon with the Zero vs. high and High vs. zero protocols. Thus, it might be more fruitful to explicitly contrast different levels of processing in a relative way rather than trying to pinpoint behaviour to specific levels of processing."

> *• The task description needs way more detail. For how long were the stimuli presented? What was their size? Were the positions of the stimuli randomized? Was it a reaction time task? Was the time-out used as a negative feedback? In case, when (e.g. mistakes or slow responses)? Also, it is important to report some statistics about the basic responses. What was the average response time, what was the performance of individual animals (over days)? Did they show any bias for a particular dimension (either the 2 baseline dimensions or the identity preserving ones) or side of response? Was there a correlation within animals between performance on the baseline task and performance on the more complex tasks?*

Thank you for your feedback. We have added more details to the task description in the manuscript.

The stimuli were presented on the screens until the animals reacted to one of the two screens. The size of the stimuli was 100 x 100 pixel. The position of the stimuli was always centred/full screen on the touchscreens. It was not a reaction time task and we also did not measure reaction time.

> *• Related to my previous comment, I wonder if the relative size/position of the stimulus with respect to the position of the animal in the setup might have had an impact on the performance, also given the impact of size shown in Figure 2. Was the position of the rat in the setup monitored (e.g. with DeepLabCut)? I guess that on average any effect of the animal position might be averaged away, but was this actually checked and/or controlled for?*

The position of the rat was not monitored in the setup. In a previous study from our lab (Crijns & Op de Beeck, 2019), we investigated the visual acuity of rats in the touchscreen setups by presenting gratings with different cycles per screen to see how it affects their performance in orientation discrimination. With the results from this study and general knowledge about rat visual acuity, we derived that the decision distance of rats lies around 12.5cm from the screen. We have added this to the discussion.

> *Minor comments*
>
> *• l.33 The sentence mentions humans, but the references are about monkeys. I believe that this concept is universal enough not to require any citation to support it.*

Thank you for your feedback. We have removed the citations.

> *• This is very minor and totally negligible. The acronymous cDNN is not that common for convents (and it's kind of similar to cuDNN), it might help clarity to stick to a more popular acronymous, e.g. CNN or ANN. Also, given that the "high" layers used for stimulus selection where not convolutional layers after all (if I'm not mistaken).*

Thank you for your feedback. We have changed the acronym to ‘CNN’ in the entire manuscript.

> *• In l.107-109 the authors identified a few potential biases in their stimuli, and they claim these biases cannot explain the results. However, the explanation is given only in the next pages. It might help to mention that before or to move that paragraph later, as I was just wondering about it until I finally got to the part on the brightness bias.*

We expanded the analysis of these dimensions (e.g. brightness) throughout the manuscript.

> *• It would help a lot the readability to put also a label close to each dimension in Figures 2 and 3. I had to go and look at Figure S4 to figure that out.*

Figures 2 and 3 have been updated, also including changes related to other comments.

> *• In Figure 2A, please specify what the red dashed line means.*

We have edited the caption of Figure 2: “Figure 2 (a) Results of the Dimension learning training protocol. The black dashed horizontal line indicates chance level performance and the red dashed line represents the 80% performance threshold. The blue circles on top of each bar represent individual rat performances. The three bars represent the average performance of all animals on the old pair (Old), the pair that differs only in concavity (Conc) and on the pair that differs only in alignment (Align). (b) Results of the Transformations training protocol. Each cell of the matrix indicates the average performance per stimulus pair, pooled over all animals. The columns represent the distractors, whereas the rows separate the targets. The colour bar indicates the performance correct. ”

> *• Related to that, why performing a binomial test on 80%? It sounds arbitrary.*

We performed the binomial test on 80% as 80% is our performance threshold for the animals

> *• The way the cDNN methods are introduced makes it sound like the authors actually fine-tuned the weights of AlexNet, while (if I'm not mistaken), they trained a classifier on the activations of a pre-trained AlexNet with frozen weights. It might be a bit confusing to readers. The rest of the paragraph instead is very clear and easy to follow.*

We think the most confusing sentence was " Figure 7 shows the performance of the network after training the network on our training stimuli for all test protocols. " We changed this sentence to " Figure 8 shows the performance of the network for each of the test protocols after training classifiers on the training stimuli using the different DNN layers."

Reviewer #3

Main recommendations:

Although it may not fully explain the entire pattern of visual behavior, it is important to discuss rat visual acuity and its impact on the perception of visual features in the stimulus set.

We have added a paragraph to the Discussion that discusses the visual acuity of rats and its impact on perceiving the visual features of the stimuli.

> *The authors observed a potential influence of image brightness on behavior during the dimension learning protocol. Was there a correlation between image brightness and the subsequent image transformations?*

We have added this to the Discussion: "To further investigate to which visual features the rat performance and human performance correlates best with, we calculated the correlation between rat performance and pixel similarity of the test image pairs, as well as the correlation between rat performance and brightness in the test image pairs. Here we found a correlation of 0.34 for pixel similarity and 0.39 for brightness, suggesting that these two visual features partly explain our results when compared to the full-set reliability of rat performance (0.58). If we perform the same correlation with the human performances, we get a correlation of 0.12 for pixel similarity and -0.12 for brightness. With the full-set reliability of 0.58 (rats) and 0.63 (humans) in mind, this suggests that even pixel similarity and brightness only partly explain the performances of rats and humans."

> *Did the rats rely on consistent visual features to perform the tasks? I assume the split-half analysis was on data pooled across rats. What was the average correlation between rats? Were rats more internally consistent (split-half within rat) than consistent with other rats?*

The split-half analysis was indeed performed on data pooled across rats. We checked whether rats are more internally consistent by comparing the split-half within correlations with the split-half between correlations. For the split-half within correlations, we split the data for each rat in two subsets and calculated the performance vectors (performance across all image pairs). We then calculated the correlation between these two vectors for each animal. To get the split-half between correlation, we calculated the correlation between the performance vector of every subset data of every rat with every other subset data from the other rats. Finally, we compared for each animal its split-half within correlation with the

split-half between correlations involving that animal. The result of this paired t-test (p = 0.93, 95%CI [-0.09; 0.08]) suggests that rats were not internally more consistent.

> *Discussion of the cDNN performance and its relation to rat behavior could be expanded and clarified in several ways:*
>
> *• The paper would benefit from further discussion regarding the low correlations between rat behavior and cDNN layers. Is the main message that cDNNs are not a suitable model for rat vision? Or can we conclude that the peak in mid layers indicates that rat behavior reflects mid-level visual processing? It would be valuable to explore what we currently know about the organization of the rat visual cortex and how applicable these models are to their visual system in terms of architecture and hierarchy.*

We added a consideration to the manuscript about which network to use (see Discussion).

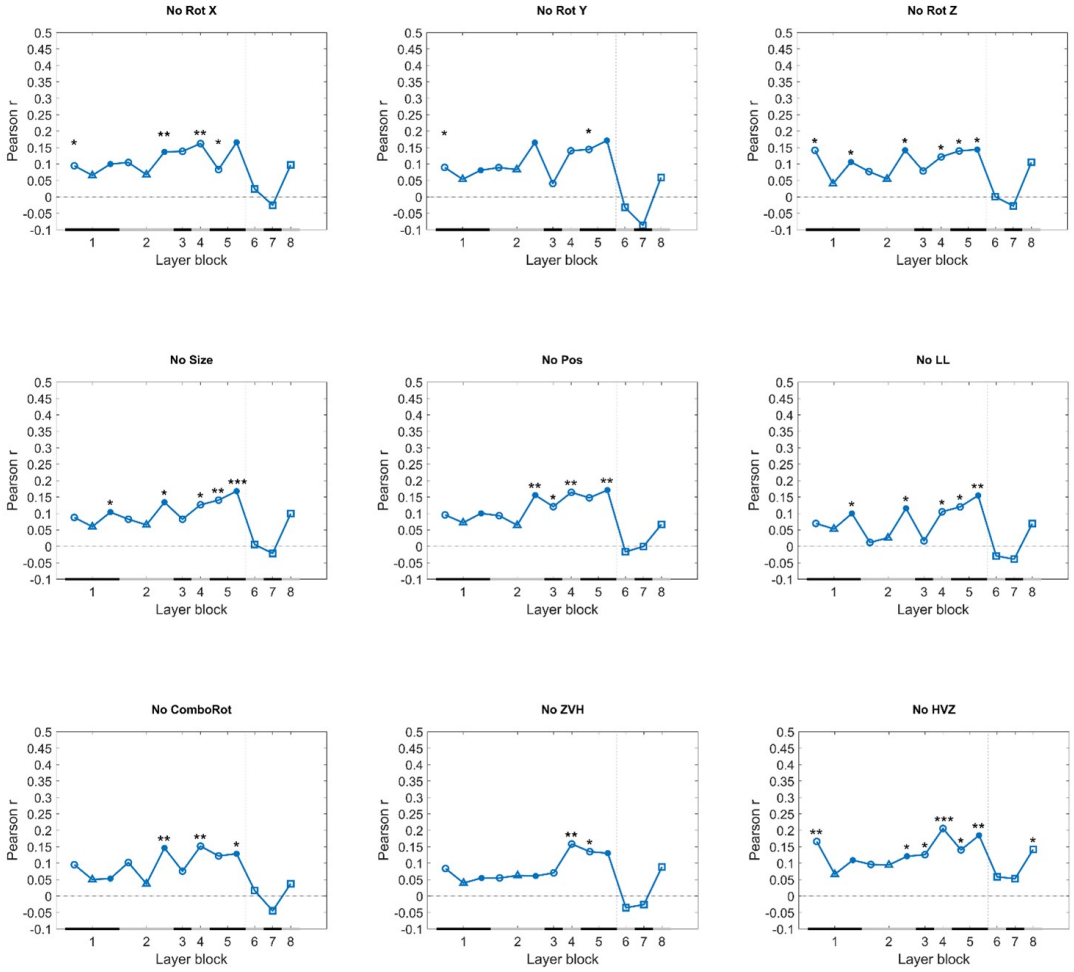
> *• The cDNN exhibited above chance performance in various early layers for several test protocols (e.g., rotations, light location, combination rotation). Does this limit the interpretation of the complexity of visual behavior required to perform these tasks?*

This is not uncommon to find. Pinto et al. (2008) already revealed that a simple V1-like model can sometimes result in surprisingly good object recognition performance. This aspect of our findings is also in line with the observation of Vinken & Op de Beeck (2021) that the performance of rats in many previous tasks might not be indicative of highly complex representations. Nevertheless, there is still a relative difference in complexity between lower and higher levels in the hierarchy. That is what we capitalize upon with the High vs zero and the Zero vs high protocols. Thus, it might be more fruitful to explicitly contrast different levels of processing in a relative way rather than trying to pinpoint behavior to specific levels of processing. This argumentation is added to the Discussion section.

> *• How representative is the correlation profile between cDNN layers and behavior across protocols? Pooling stimuli across protocols may be necessary to obtain stable correlations due to relatively modest sample numbers. However, the authors could address how much each individual protocol influences the overall correlations in leave-one-out analyses. Are there protocols where rat behavior correlates more strongly with higher layers (e.g., when excluding zero vs. high)?*

We prefer to base our conclusions mostly on the pooled analyses rather than individual protocols. As the reviewer also mentions, we can expect that the pooled analyses will provide the most stable results. For information, we included leave-one-out analyses in the supplemental material. Excluding the Zero vs. High protocol did not result in a stronger correlation with the higher layers. It was rare to see correlations with higher layers, and in the one case that we did (when excluding High versus zero) the correlations were still higher in several mid-level layers.

Author response image 2.



> • *The authors hypothesize that the cDNN results indicate that rats rely on visual features such as contrast. Can this link be established more firmly? e.g., what are the receptive fields in the layers that correlate with rat behavior sensitive to?*

This hypothesis was made based on previous in-lab research ((Schnell et al., 2023) where we found rats indeed rely on contrast features. In this study, we performed a face categorization task, parameterized on contrast features, and we investigated to what extent rats use contrast features to perform in a face categorization task. Similarly as in the current study, we used a DNN that as trained and tested on the same stimuli as the animals to investigate the representations of the animals. There, we found that the animals use contrast features to some extent and that this correlated best with the lower layers of the network. Hence, we would say that the lower layers correlate best with rat behaviour that is sensitive to contrast. Earlier layers of the network include local filters that simulate V1-like receptive fields. Higher layers of the network, on the other hand, are used for object selectivity.

> • *There seems to be a disconnect between rat behavior and the selection of stimuli for the high (zero) vs. zero (high) protocols. Specifically, rat behavior correlated best with mid layers, whereas the image selection process relied on earlier layers. What is the interpretation when rat behavior correlates with higher layers than those used to select the stimuli?*

We agree that it is difficult to pinpoint a particular level of processing, and it might be better to use relative terms: lower/higher than. This is addressed in the manuscript by the edit in response to three comments back.

> *• To what extent can we attribute the performance below the ceiling for many protocols to sensory/perceptual limitations as opposed to other factors such as task structure, motivation, or distractibility?*

We agree that these factors play a role in the overall performance difference. In Figure 5, the most right bar shows the percentage of all animals (light blue) vs all humans (dark blue) on the old pair that was presented during the testing protocol. Even here, the performance of the animals was lower than humans, and this pattern extended to the testing protocols as well. This was most likely due to motivation and/or distractibility which we know can happen in both humans and rats but affects the rat results more with our methodology.

> *Minor recommendations:*
>
> *• What was the trial-to-trial variability in the distance and position of the rat's head relative to the stimuli displayed on the screen? Can this variability be taken into account in the size and position protocols? How meaningful is the cDNN modelling of these protocols considering that the training and testing of the model does not incorporate this trial-to-trial variability?*

We have no information on this trial-to-trial variability. We have information though on what rats typically do overall from an earlier paper that was mentioned in response to an earlier comment (Crijns et al.).

We have added a disclaimer in the Discussion on our lack of information on trial-to-trial variability.

> *• Several of the protocols varied a visual feature dimension (e.g., concavity & alignment) relative to the base pair. Did rat performance correlate with these manipulations? How did rat behavior relate to pixel dissimilarity, either between target and distractor or in relation to the trained base pair?*

We have added this to the Discussion. See also our general comments in the Public responses.

> *• What could be the underlying factor(s) contributing to the difference in accuracy between the "small transformations" depicted in Figure 2 and some of the transformations displayed in Figure 3? In particular, it seems that the variability of targets and distractors is greater for the "small transformations" in Figure 2 compared to the rotation along the y-axis shown in Figure 3.*

There are several differences between these protocols. Before considering the stimulus properties, we should take into account other factors. The Transformations protocol was a training protocol, meaning that the animals underwent several sessions in this protocol, always receiving real reward during the trials, and only stopping once a high enough performance was reached. For the protocols in Figure 3, the animals were also placed in these protocols for multiple sessions in order to obtain enough trials, however, the difference here is that they did not receive real reward and testing was also stopped if performance was still low.

> *• In Figure 3, it is unclear which pairwise transformation accuracies were above chance. It would be helpful if the authors could indicate significant cells with an asterisk. The*

> *scale for percentage correct is cut off at 50%. Were there any instances where the behaviors were below 50%? Specifically, did the rats consistently choose the wrong option for any of the pairs? It would be helpful to add "old pair", "concavity" and "alignment" to x-axis labels in Fig 2A .*

We have added “old”, “conc” and “align” to the x-axis labels in Figure 2A.

> *• Considering the overall performance across protocols, it seems overstated to claim that the rats were able to "master the task."*

When talking about “mastering the task”, we talk about the training protocols where we aimed that the animals would perform at 80% and not significantly less. We checked this throughout the testing protocols as well, where we also presented the old pair as quality control, and their performance was never significantly lower than our 80% performance threshold on this pair, suggesting that they mastered the task in which they were trained. To avoid discussion on semantics, we also rephrased “master the task” into “learn the task”.

> *• What are the criteria for the claim that the "animal model of choice for vision studies has become the rodent model"? It is likely that researchers in primate vision may hold a different viewpoint, and data such as yearly total publication counts might not align with this claim.*

Primate vision is important for investigating complex visual aspects. With the advancements in experimental techniques for rodent vision, e.g. genetics and imaging techniques as well as behavioural tasks, the rodent model has become an important model as well. It is not necessarily an “either” or “or” question (primates or rodents), but more a complementary issue: using both primates and rodents to unravel the full picture of vision.

We have changed this part in the introduction to “Lately, the rodent model has become an important model in vision studies, motivated by the applicability of molecular and genetic tools rather than by the visual capabilities of rodents”.

> *• The correspondence between the list of layers in Supplementary Tables 8 and 9 and the layers shown in Figures 4 and 6 could be clarified.*

We have clarified this in the caption of Figure 7

> *• The titles in Figures 4 and 6 could be updated from "DNN" to "cDNN" to ensure consistency with the rest of the manuscript.*

Thank you for your feedback. We have changed the titles in Figures 4 and 6 such that they are consistent with the rest of the manuscript.