

# Towards Global Recurrent Models of Visual Processing: Capsule Networks

**Adrien Doerig (adrien.doerig@gmail.com)**

Brain Mind Institute, EPFL,  
Lausanne, Switzerland

**Lynn Schmittwilken (schmittwilken.lynn@gmail.com)**

Brain Mind Institute, EPFL,  
Lausanne, Switzerland

**Michael H. Herzog (michael.herzog@epfl.ch)**

Brain Mind Institute, EPFL,  
Lausanne, Switzerland

## Abstract:

Classically, visual processing is described as a cascade of local feedforward computations. Convolutional Neural Networks (CNNs) have shown how powerful such models can be. However, CNNs only roughly mimic human vision. For example, CNNs do not take the global spatial configuration of visual elements into account but mainly rely on local features such as textures. For CNNs, a face is not different from a scrambled version of it. For this reason, CNNs fail to explain many visual paradigms, such as crowding, where configuration strongly matters. In crowding, the perception of a target deteriorates in the presence of neighboring elements. Classically, adding flanking elements was thought to always *decrease* performance. However, *adding* flankers even far away from the target can *improve* performance, depending on the global configuration (an effect called *uncrowding*). We showed previously that no classic model of crowding, including CNNs, can explain uncrowding (Doerig et al., 2019). Here, we show that Capsule Networks (CapsNets; Sabour, Frosst, & Hinton, 2017), combining CNNs learning algorithms with recurrent object segmentation, explain both crowding and uncrowding. We also conducted psychophysical experiments investigating how time-consuming recurrent computations shape object formation. The results of these experiments about complex object-level effects are also well captured. These powerful recurrent networks offer a promising general framework to model global object shape recurrent processing.

**Keywords:** Vision, Neural Networks, Capsule Networks, Crowding, Global Processing, Recurrent Processing

## Introduction

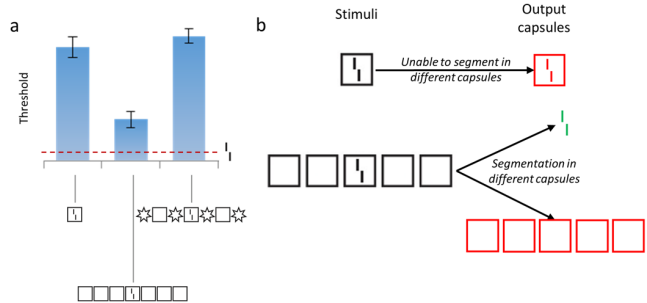
The visual system is often seen as a hierarchy of local, feedforward computations (DiCarlo, Zoccolan, & Rust, 2012). Low-level neurons detect simple elements such as edges, and higher-level neurons pool this information to detect higher-level features such as corners, shapes, and

ultimately complex objects. CNNs have shown that these architectures can indeed excel in object detection. Despite the amazing range of tasks accomplished by CNNs, they only roughly mimic human vision. For example, they lack the abundant recurrent processing found in humans (Kietzmann et al., 2019; Lamme & Roelfsema, 2000), perform differently from humans in many psychophysical tasks (Doerig et al., 2019; Funke et al., 2018), and are easily fooled by simple tricks (Geirhos et al., 2018; Su, Vargas, & Sakurai, 2019; Szegedy et al., 2013). The above results all suggest that CNNs are biased towards local, texture-like features, while the brain relies more on global object shape (Baker, Lu, Erlichman, & Kellman, 2018).

Here we show that CapsNets (Sabour et al., 2017), a recent type of recurrent deep network combining CNNs and recurrent object segmentation, can address several of the shortcomings of CNNs and provide a promising alternative framework for vision. In particular, CapsNets are good candidates to capture and use global object shape. As a probe to investigate these object-level computations, we focus on surprising aspects of crowding, a well-known and ubiquitous phenomenon. In crowding, perception of a target deteriorates in the presence of nearby flankers (review: Levi, 2008). Crowding is crucial for vision in general, since elements are rarely seen in isolation. For example, a vernier target (i.e., two vertical lines separated by a horizontal offset; Figure 1) is presented. When the vernier is displayed alone, observers easily discriminate the offset direction. When a single flanking square is added, performance drops, i.e., crowding occurs. Surprisingly, *adding* more flankers can *reduce* crowding, depending on the configuration (Figure 1a; Manassi, Lonchampt, Clarke, & Herzog, 2016). This configurational *uncrowding* effect is well known, powerful, and occurs for a wide range of stimuli in vision, audition and haptics (review: Doerig et al., 2019). Hence, uncrowding

seems to be an essential byproduct of the brain’s general strategy for vision. We showed previously that these very strong configurational effects cannot be explained by models based on the classic framework of vision, including CNNs, because of local feedforward processing (Doerig et al., 2019). A recurrent, flexible grouping and segmentation process seems crucial. Here, we show that CapsNets can naturally explain these complex configurational results.

In CapsNets, early convolutional layers extract basic visual features. Recurrent processing then combines these features to group and segment objects from each other by a process called *routing by agreement*. *Capsules* are groups of neurons representing visual features crucial for this routing by agreement process. Low-level capsules iteratively predict the activity of high-level capsules in a recurrent loop. If the predictions agree, the corresponding high-level capsule is activated. For example, if a triangle capsule above a rectangle capsule are both active, they agree that the higher-level object should be a house and, therefore, the corresponding high-level capsule is activated. Through this process, CapsNets are able to recognize overlapping digits (Sabour et al., 2017) and, as we show, to explain (un)crowding (Figure 1b). Crowding occurs when the target and flankers are represented in the same capsule. In this case, a single capsule represents them both and so they interfere, because a single capsule cannot represent well two objects simultaneously due to limited neural resources. This mechanism is similar to pooling: information about the target gets pooled with information about the flankers, leading to poorer representations. However, if the flankers are segmented away and represented in a different capsule, the target is released from the flankers’ deleterious effect and *uncrowding* occurs. This segmentation can only happen if the network has learnt to group the flankers into a single higher-level object represented in a different capsule than the vernier target. Segmentation is facilitated when more flankers are added because more low-level capsules agree about the presence of the flanker group.



**Figure 1: a. (Un)crowding:** A vernier (two vertical bars with a horizontal offset) is presented in the visual periphery. The offset direction is easily reported (dotted red line; the y-axis shows the threshold, i.e., the minimal offset size at which observers can report the offset direction with 75% accuracy). When a square flanker surrounds the vernier, performance deteriorates, a classic crowding effect. When more squares are added, performance recovers (uncrowding). Critically,

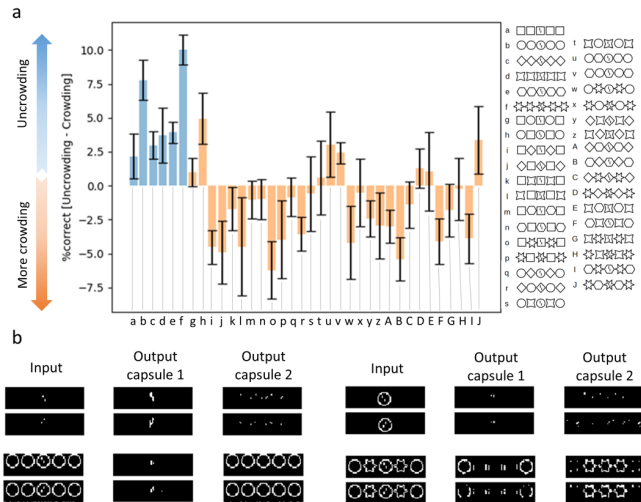
the uncrowding effect depends on the global stimulus configuration. For example, if some squares are replaced by stars, performance deteriorates again. It was shown empirically that, in a such displays with single lines of flankers, adding identical flankers usually leads to uncrowding, but adding different flankers does not. In more complex 2D displays, even arrays of different flankers can lead to uncrowding, depending on the configuration (Manassi et al., 2016). **b. Segmentation and (un)crowding in CapsNets:** If CapsNets can segment the vernier target away from the flankers during the recurrent routing by agreement process, uncrowding can occur. This is difficult when a single flanker surrounds the target because capsules disagree about what is shown at this location. But in the case of configurations that the network has learned to group, many primary capsules agree about the presence of a large shape group, which can therefore easily be segmented away from the vernier target.

## Methods & Results

### Experiment 1: Crowding And Uncrowding Naturally Occur In CapsNets

We trained a CapsNet to recognize greyscale images of vernier targets and groups of identical shapes. During training, either a vernier or a group of identical shapes was presented, and the network had to classify which shape type was present, as well as the number of shapes in the group, and the vernier offset. Hence, after training, the network knew about verniers and groups of identical shapes, but had never seen (un)crowding stimuli.

When we tested the trained network’s vernier offset discrimination performance on crowding and uncrowding stimuli, both crowding and uncrowding occurred (Figure 2a). This result was not affected by small changes in network hyperparameters or stimulus characteristics. Reconstructing the input image based on the network’s output capsules shows that (un)crowding occurs for the reasons described earlier: there is crowding when the target cannot be segmented from flankers, and uncrowding when the target is successfully segmented in its own capsule (Figure 2b). As we suggested, this segmentation becomes easier when the network recognizes a large group of shapes.



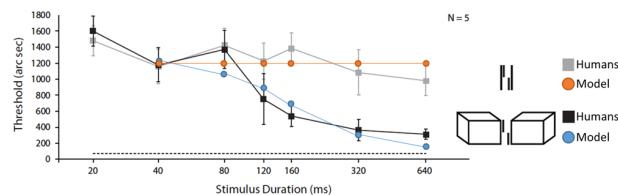
**a. Simulation results:** Both crowding and uncrowding occur in capsule networks. The x-axis shows different stimuli. Performance for these stimuli is shown on the y-axis as the trained network’s [%correct on stimuli with an entire line of flankers *minus* %correct with only the central flanker]. For example, in column a, vernier offset direction is easier to report with 5 square flankers than with 1 square flanker, as expected. Error bars are the standard error over 10 network trainings. The blue bars represent configurations for which uncrowding is expected (i.e., the blue bars must go up to explain the empirical data) and orange bars represent configurations for which crowding is expected (the orange bars need to go down to explain the empirical data). **b. Reconstructions:** We reconstruct the input image based on the network’s output capsules’ activities. The reconstructions based on the two first “winning” capsules are shown. When the vernier is alone (top left), the reconstructions are good. When a single flanker is added (top right), the vernier reconstruction deteriorates (crowding). The vernier is not well segmented from the flanker. When identical flankers are added (bottom left), the vernier reconstruction recovers, i.e., it is well segmented from the flankers (uncrowding). With different flankers (bottom right), the vernier is not represented at all in the two winning capsules (crowding).

## Experiment 2: Temporal Dynamics Of Uncrowding Naturally Occur In CapsNets

We experimentally investigated temporal dynamics of (un)crowding and modeled the results with CapsNets to study how time-consuming recurrent computations shape object formation. First, we performed a psychophysical crowding experiment with a vernier target flanked by either two simple lines or two more complex cuboids (Figure 2). The stimuli were displayed for varying durations from 20 to 640ms and five observers reported vernier offset direction. For short stimulus durations, crowding occurred for both flanker types. Crucially, uncrowding occurred for the complex cuboid flankers only when stimulus duration was long enough (Figure 2). We hypothesize that this reflects the time-

consuming recurrent computations necessary to segment the cuboid flankers away from the target. The line flankers cannot be segmented away from the target, so there is no uncrowding even for long stimulus durations.

CapsNets can explain this result by varying the number of iterations in the recurrent routing by agreement process (Figure 3). With more iterations of recurrent processing, the cuboids can be better segmented from the target, and uncrowding occurs. The simple lines, however, can never be segmented because they strongly group with the vernier. This result was not affected by small changes in network hyperparameters or stimulus characteristics.



**Figure 3: Temporal dynamics of uncrowding:** In humans, uncrowding occurs with cuboid flankers only after about 100ms of stimulus presentation (black). Uncrowding does not occur with single line flankers, even with long stimulus times (grey). We hypothesize that the cuboids are segmented from the vernier target through time-consuming recurrent processing (the line flankers are grouped with the target and cannot be segmented at all). CapsNets can explain these results by varying the number of recurrent routing by agreement iterations (blue and orange; the model’s %correct is converted to a threshold through a sigmoid psychometric function).

## Discussion

Powerful and flexible recurrent models are needed to go beyond current conceptions in vision science and AI. For example, flexible object segmentation is crucial for visual processing, but is absent from the architecture of CNNs (Doerig et al., 2019; Lamme & Roelfsema, 2000). Here, for the first time, we showed that CapsNets are able to explain complex, shape-level recurrent spatiotemporal processing in psychophysical experiments.

Uncrowding can be used as an experimental probe to investigate how the brain flexibly forms object representations based on grouping and segmentation. Our results show that CapsNets are a good model of this process. Although other segmentation networks exist (e.g. Francis, Manassi, & Herzog, 2017), CapsNets are much more flexible and can be trained to solve any task. We focused on vernier experiments in this contribution, but the exact same procedure can plausibly explain (un)crowding and other shape-level recurrent processing with different stimuli, across different modalities.

It is well known that humans can solve a number visual of tasks very quickly, presumably in a single feedforward pass of neural activity (such as analysing briefly viewed natural scenes; Thorpe, Fize, & Marlot, 1996). In this regime, CNNs

have been shown to be good models of visual processing (Khaligh-Razavi & Kriegeskorte, 2014; Kietzmann, McClure, & Kriegeskorte, 2018; Yamins et al., 2014). However, neural activities are not determined by the feedforward sweep alone: recurrent activity is also crucial and offers distinct modes of processing (Kietzmann et al., 2019; Lamme & Roelfsema, 2000). For this, new models are needed. CapsNets naturally include both fast feedforward and time-consuming recurrent regimes, depending on the time allowed for routing by agreement. We showed how these two regimes in CapsNets explain previously unexplained psychophysical results: object segmentation depends on the presence or absence of recurrent computations, and, again, (un)crowding can be used as a probe into this process.

In conclusion, CapsNets propose solutions to several shortcomings of CNNs: they are good candidates to capture and use global object shape, include a powerful and flexible segmentation process, and naturally link the feedforward and recurrent modes of visual processing. Although much work is needed to show the extent to which CapsNets match the human visual system, they constitute a promising alternative framework for vision.

### Acknowledgments

This work was supported by the SNF grant “Basics of visual processing: from elements to figures” (176153).

### References

- Baker, N., Lu, H., Erlichman, G., & Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLoS Computational Biology*, *14*(12), e1006613.
- DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How Does the Brain Solve Visual Object Recognition? *Neuron*, *73*(3), 415–434. <https://doi.org/10.1016/j.neuron.2012.01.010>
- Doerig, A., Bornet, A., Rosenholtz, R., Francis, G., Clarke, A. M., & Herzog, M. H. (2019). Beyond Bouma’s window: How to explain global aspects of crowding? *PLOS Computational Biology*, *15*(5), e1006580. <https://doi.org/10.1371/journal.pcbi.1006580>
- Francis, G., Manassi, M., & Herzog, M. H. (2017). Neural dynamics of grouping and segmentation explain properties of visual crowding. *Psychological Review*, *124*(4), 483.
- Funke, C. M., Borowski, J., Wallis, T. S. A., Brendel, W., Ecker, A. S., & Bethge, M. (2018). Comparing the ability of humans and DNNs to recognise closed contours in cluttered images. *18th Annual Meeting of the Vision Sciences Society (VSS 2018)*, 213.
- Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *ArXiv Preprint ArXiv:1811.12231*.
- Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, *10*(11), e1003915.
- Kietzmann, T. C., Spoerer, C. J., Sörensen, L., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence required to capture the dynamic computations of the human ventral visual stream. *ArXiv Preprint ArXiv:1903.05946*.
- Kietzmann, T. C., McClure, P., & Kriegeskorte, N. (2018). Deep neural networks in computational neuroscience. *BioRxiv*, 133504.
- Lamme, V. A., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, *23*(11), 571–579.
- Levi, D. M. (2008). Crowding—An essential bottleneck for object recognition: A mini-review. *Vision Research*, *48*(5), 635–654. <https://doi.org/10.1016/j.visres.2007.12.009>
- Manassi, M., Lonchamp, S., Clarke, A., & Herzog, M. H. (2016). What crowding can tell us about object representations. *Journal of Vision*, *16*(3), 35–35. <https://doi.org/10.1167/16.3.35>
- Sabour, S., Frosst, N., & Hinton, G. E. (2017). Dynamic routing between capsules. *Advances in Neural Information Processing Systems*, 3856–3866.
- Su, J., Vargas, D. V., & Sakurai, K. (2019). One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*.
- Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *ArXiv Preprint ArXiv:1312.6199*.
- Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*(6582), 520.
- Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.

**DO NOT READ THIS I JUST NEED TO  
KEEP IT HERE FOR REFERENCE  
FORMATTING ISSUES**

**References**

- Chalnick, A., & Billman, D. (1988). Unsupervised learning of correlational structure. *Proceedings of the tenth annual conference of the cognitive science society* (pp. 510-516). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Feigenbaum, E. A. (1963). The simulation of verbal learning behavior. In E. A. Feigenbaum & J. Feldman (Eds.), *Computers and thought*. New York: McGraw-Hill.
- Hill, J. A. C. (1983). A computational model of language acquisition in the two-year old. *Cognition and Brain Theory*, 6, 287-317.
- Matlock, T. (2001). *How real is fictive motion?* Doctoral dissertation, Psychology Department, University of California, Santa Cruz.
- Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.
- Ohlsson, S., & Langley, P. (1985). *Identifying solution paths in cognitive diagnosis* (Tech. Rep. CMU-RI-TR-85-2). Pittsburgh, PA: Carnegie Mellon University, The Robotics Institute.
- Shrager, J., & Langley, P. (Eds.) (1990). *Computational models of scientific discovery and theory formation*. San Mateo, CA: Morgan Kaufmann.