Trends in
Cognitive Sciences

CellPress

Letter Response

Feature binding in biological and artificial vision

Pieter R. Roelfsema [1,2,3,4,*], and Thomas Serre[5,6]

In a recent article [1], Scholte and de Haan challenge the prevailing view by claiming that the degree of specialization of visual brain areas is limited. They also suggest that neuronal codes for object features are so broadly distributed across brain regions that binding problems rarely occur. Their claims partly stem from insights into deep neural networks. The underlying idea is that if neurons explicitly represent feature conjunctions ('base-groupings' [2]), binding issues rarely occur. Here, we argue that Scholte and de Haan underestimate both the degree of specialization of brain regions and the necessity of binding mechanisms. We provide examples of strong specialization of brain areas and illustrate how deep neural networks struggle with binding.

First, the evidence for the specialization of visual brain regions is stronger than Scholte and de Haan acknowledge. Their conclusions are partly based on a stroke lesion study that found little association between mid-level visual area damage and perceptual deficits [3]. But stroke lesions are imprecise and variable, fail to respect anatomical boundaries, and they rarely affect some of the mid-level areas. Controlled experiments paint a clearer picture. For example, lesions in area MT specifically impair motion perception, and electrical stimulation predictably disturbs motion perception [4]. Face-processing areas show similar specialization: neurons are predominantly tuned to faces [5], modifying neuronal activity selectively affects face perception and lesions cause prosopagnosia [6]. Comparable functional specialization occurs throughout the visual cortex.

Second, binding mechanisms are indispensable for the flexibility of perception, far beyond what base-groupings can achieve. Imagine a crowd where one person makes a hand gesture (Figure 1A). To identify the gesturer, base-groupings would require neurons to encode every person-gesture combination – an infeasible solution. Scholte and de Haan suggest that base-groupings for frequently co-occurring features suffice, but this does not help when any individual might make any gesture. Hardwired feature conjunctions cannot support the dynamic grouping seen in human perception. Instead, binding is often a time-consuming process [7], where object-based attention incrementally labels the parts of an object, for one object at a time. In our example scenario, attention would start at the gesturing hand and spread via the arm, trunk, and neck to the face so that the gesturer can be recognized (Figure 1B). At the neuronal level, enhanced activity gradually spreads across the object's representation [7]. Studies show that attention enhances neuronal responses throughout the visual cortex, enabling the flexible grouping of the diverse features of a single object [2,6], specifying feature conjunctions beyond the hardwired base-groupings [2]. Similar processes operate continuously in daily life. Consider picking a specific piece of cutlery – say, a fork – from a crowded kitchen drawer: your fingers must target the fork's edges. It must first be visually grouped and segregated from nearby forks, knives, and spoons.

Third, artificial neural networks also depend on explicit binding mechanisms. Early networks explored neuronal synchrony as a binding mechanism, but

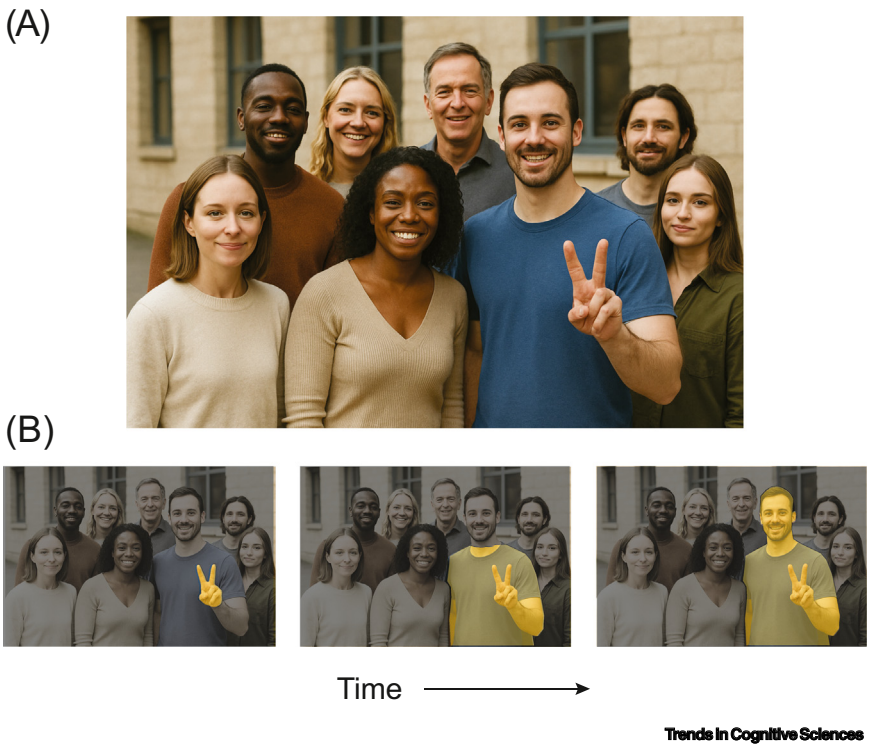

Figure 1. Binding with object-based attention. (A) Identifying the gesturer requires binding between the hand and face. (B) The formation of these incremental groups is associated with the spread of object-based attention (yellow).

neurophysiological evidence rather supports feature grouping via enhanced activity [2], which has been modeled in later neural networks [8]. Modern convolutional neural networks (CNNs) like Mask R-CNN[i] rely on region proposal mechanisms that explicitly group features across space, and they perform even better when they are enhanced with object-based attention [9]. If these binding mechanisms are omitted, neural networks, such as CNNs, struggle with simple visual reasoning tasks [10]. Even transformer-based large language models use attention to bind features, and transformer-based models for vision, such as 'Segment Anything', owe their performance to comparable flexible binding mechanisms. The accuracy of the advanced vision models deteriorates even more than that of humans when feature binding becomes challenging [11]. Finally, state-of-the-art vision-language models still make binding errors in basic reasoning tasks [12], underscoring the persistence of binding problems in neural networks while contradicting Scholte and de Haan's analysis.

In conclusion, the visual cortex shows clear specialization, and binding mechanisms are essential for flexible, object-based perception. Rather than eliminating binding problems, distributed representations in both biological and artificial vision systems make dedicated grouping mechanisms indispensable. Deep learning advances reinforce, rather than refute, the centrality of binding in perception.

Acknowledgments

The work was supported by NWO-grants (Crossover grant 17619 INTENSE and DBI2, a Gravitation program of the Dutch Ministry of Science), the European Union Horizon 2020 Framework Program under ERC grant (101052963 NUMEROUS) to P.R.R. and by ONR grants (N00014-24-1-2026 and REPRISM MURI N00014-24-1-2603), NSF grant (IIS-2402875 and EAR-1925481), and the ANR-3IA Artificial and Natural Intelligence Toulouse Institute (ANR-19-PI3A-0004) to T.S.

Declaration of interests

No interests are declared.

Resources

[i]https://github.com/facebookresearch/detectron2

[1]Department of Vision & Cognition, Netherlands Institute for Neuroscience (KNAW), 1105, BA, Amsterdam, The Netherlands
[2]Department of Integrative Neurophysiology, VU University, De Boelelaan 1085, 1081, HV, Amsterdam, The Netherlands
[3]Department of Neurosurgery, Academic Medical Centre, Postbus 22660, 1100 DD Amsterdam, The Netherlands
[4]Laboratory of Visual Brain Therapy, Sorbonne Universite´, INSERM, CNRS, Institut de la Vision, 17 rue Moreau, 75012 Paris, France
[5]Department of Cognitive and Psychological Science, Brown University, Thayer Street, Providence, RI 02906, USA
[6]Robert J. and Nancy D. Carney Institute for Brain Sciences, Brown University, Angell Street, Providence, RI 02906, USA

*Correspondence:
p.roelfsema@nin.knaw.nl (P.R. Roelfsema).

https://doi.org/10.1016/j.tics.2025.08.007

© 2025 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

References

1. Scholte, H.S. and de Haan, E.H.F. (2025) Beyond binding: from modular to natural vision. *Trends Cogn. Sci.* 29, 505–515
2. Roelfsema, P.R. (2006) Cortical algorithms for perceptual grouping. *Annu. Rev. Neurosci.* 29, 203–227
3. Lugtmeijer, S. *et al.* (2025) Visual feature processing in a large stroke cohort: evidence against modular organization. *Brain* 148, 1144–1154
4. Maunsell, R. and Newsome, W.T. (1987) Visual processing in monkey extrastriate cortex. *Annu. Rev. Neurosci.* 10, 363–401
5. Tsao, D.Y. *et al.* (2006) A cortical region consisting entirely of face-selective cells. *Nature* 311, 670–674
6. Moeller, S. *et al.* (2017) The effect of face patch microstimulation on perception of faces and objects. *Nat. Neurosci.* 20
7. Roelfsema, P.R. (2023) Solving the binding problem: assemblies form when neurons enhance their firing rate - they don't need to oscillate or synchronize. *Neuron* 111, 1003–1019
8. Schmid, D. and Neumann, H. (2025) A model of thalamocortical interaction for incremental binding in mental contour-tracing. *PLoS Comput. Biol.* 21, 1–51
9. Linsley, D. *et al.* (2020) Stable and expressive recurrent vision models. In *Advances in Neural Information Processing Systems* (33) (Larochelle, H. *et al.*, eds), NeurIPS
10. Ricci, M. *et al.* (2021) Same-different conceptualization: a machine vision perspective. *Curr. Opin. Behav. Sci.* 37, 47–55
11. Fan, J. and Zeng, Y. (2023) Challenging deep learning models with image distortion based on the abutting grating illusion. *Patterns* 4, 1–20
12. Campbell, D. *et al.* (2024) Understanding the limits of vision language models through the lens of the binding problem. In *Advances in Neural Information Processing Systems* (37) (Globerson, D. *et al.*, eds), NeurIPS