Current DNNs are Unable to Integrate Visual Information Across Object Discontinuities...

Ben Lonnqvist, Elsa Scialom, Abdulkadir Gökce, Zehra Merchant, Michael H. Herzog, Martin Schrimpf

SCAN ME!

benlonnqvist.github.io/

Question

Humans perform visual grouping to group object parts across the entire visual field[1].

Are current DNNs able to recognize objects that require grouping?

Which variables predict their alignment to humans?

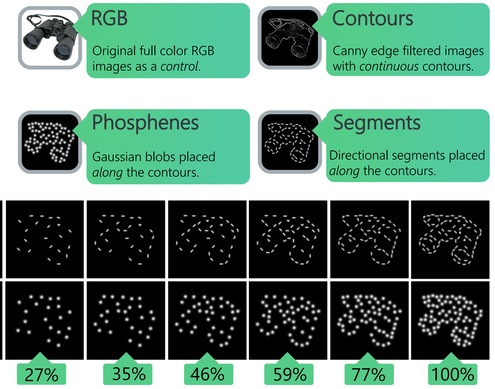
Approach

Using a new task, we map out the *range* of human performance from *near-chance* to *ceiling*.

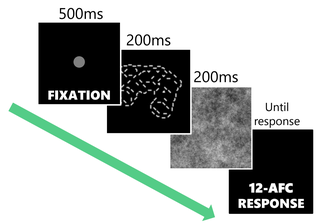
We *contrast* this with the ***entire range*** of extant **DNN** models, from tiny (**mini-ResNet**) to huge (**GPT-4o**).

Dataset

- 4 image types: RGB, Contours, Phosphenes, Segments[2].
- A *wide range* of element densities.

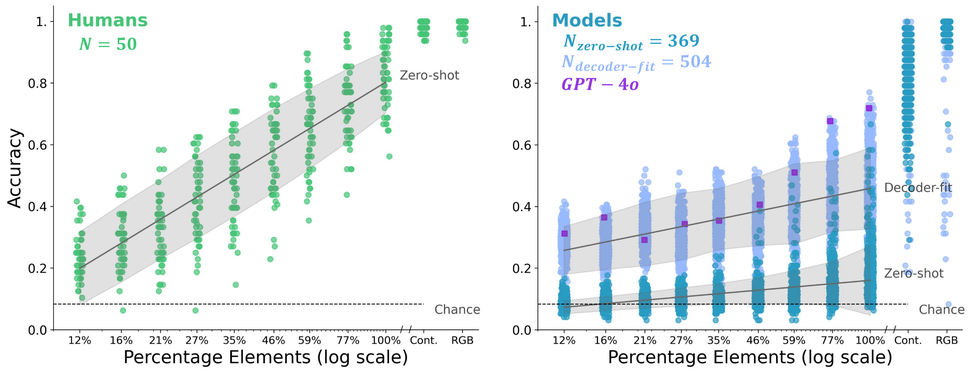


Experimental Methodology



- $N = 50$ subjects total.
- Subjects were split into two groups of 25 subjects, each performing either the phosphene task or the segment task.
- All subjects saw Contour and RGB images.

Humans outperform models on average



TWO VIEWS ON RESULTS

View 1:

DNNs fail on average.

When prompted zero-shot, DNNs fail to capture the human slope ($p < 10^{-100}$).

This demonstrates a DNN **failure** to integrate visual elements *across* object discontinuities.

View 2:

Some DNNs approach humans.

The best open-source DNN (*ConvNeXt*) surpasses the worst human on average. GPT-4o is **nearly indistinguishable** from the human average, though its model and training details are unknown.

...BUT MODEL AND TRAINING DATASET SIZE STRONGLY PREDICT PERFORMANCE

Financial support

*This work was supported by the Swiss National Science Foundation grant n. 176153 "Basics of visual processing : from elements to figures".

References

1. Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., & von der Heydt, R. (2012). A Century of Gestalt Psychology in Visual Perception I. Perceptual Grouping and Figure-Ground Organization. Psychological bulletin.
2. Rotermund, D., Scialom, E., Repnow, M., Herzog, M., &Ernst, U. (2024). davrot/percept simulator 2023:V1.0.0 (neuroprosthesis). Zenodo. doi: 10.5281/zenodo.10978899
3. Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T.,Bethge, M., Wichmann, F. A., & Brendel, W. (2021). Partial success in closing the gap between human and machine vision. In Advances in Neural Information Processing Systems.

Modeling methodology

- A total of **985 unique** *model architecture-training dataset* pairs.
- Model responses were obtained in *two ways*: **zero-shot**, as well as by **fitting a decoder**.

Brain-Score

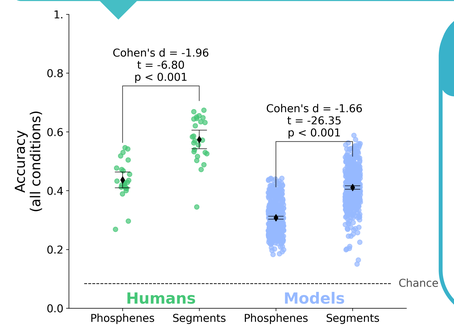
ZERO-SHOT

If a model is capable of outputting ImageNet labels, we *map* them to our 12-AFC task labels using a *WordNet SynSet* mapping[3].

DECODER-FIT

We extract the *penultimate layer activations* of the model and use **120 novel samples** (10 per class) to fit a **linear decoder** on the model activations.

GLASS HALF FULL?

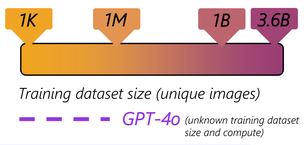


DNNs perform worse...

but exhibit the same segment preference.

*A **t-test** for* the difference of means showed that both models and humans exhibit a preference for segments over phosphenes, with a similar effect size (d=1.96; d=1.66).

MODEL AND DATA SCALING PREDICT ALIGNMENT



ZERO-SHOT MODELS

*Log-regression from FLOPs and training dataset size to accuracy $R^2 = 0.366$**

DECODER-FIT MODELS

*Log-regression from FLOPs and training dataset size to accuracy $R^2 = 0.654$**