Using Generated Object Reconstructions to Study Object-based Attention

Seoyoung Ahn[1] **(ahnseoyoung@gmail.com)**, **Hossein Adeli**[1], **Gregory J. Zelinsky**[1,2]
[1]Stony Brook University, Department of Psychology, Stony Brook, NY,
[2]Stony Brook University, Department of Computer Science, Stony Brook, NY

Abstract

Object-based attention operates by selecting complete object representations as a fundamental unit, but existing theories tend to focus on bottom-up cues, such as Gestalt principles, and have given relatively little attention to the influence of top-down signals on this process, as well as how bottom-up and top-down factors may interact. Here we propose that Object Reconstruction-guided Attention (ORA) may provide a useful framework to study the interplay between bottom-up and top-down factors in object-based attention. The ORA model encodes object-based representations to reconstruct object location and appearance, and utilizes this reconstructed information to further bias the bottom-up signal in the feedforward pass. The objective of the model is to best explain the input by selecting and reconstructing target objects with the lowest reconstruction error, creating an object-selection bias. Our results demonstrate that this simple reconstruction-based selection principle can support various visual tasks, providing new insights into the brain mechanisms underlying robust object-based attention and visual perception. Future work will extend this work to more naturalistic images and examine the model's correspondence with human behavior.

Keywords: Generative networks; Reconstruction; Top-down feedback; Object-based attention; Robustness

Introduction

A fundamental role of attention is to select the information that is most relevant to our current goals and needs, in order to navigate our daily interactions with the world around us. Empirical evidence suggests that our attention may select a complete object representation as a basic unit, known as object-based attention (O'Craven et al., 1999; Egly et al., 1994; Einhäuser et al., 2008, for review Chen, 2012; Scholl, 2001). However, the few theories of object-based attention that exist primarily rely on the bottom-up cues to form object representations, and do not fully address the role of top-down processes (Logan, 1996; Roelfsema and Houtkamp, 2011; Jeurissen et al., 2016). Recent research suggests that object reconstruction, a process that recovers specific object locations and features through an autoencoder-like architecture, may explain the integration of top-down and bottom-up processes in object-based attention (Cavanagh et al., 2022). There is increasing evidence that generative or reconstructive processes play a significant role in creating top-down biasing signals in visual perception (Clark et al., 2019; Breedlove et al., 2020; Bi, 2021). However, few studies have investigated the specific mechanisms by which the visual system uses reconstruction for attentional modulation (Adeli et al., 2022; Ahn et al., 2022).

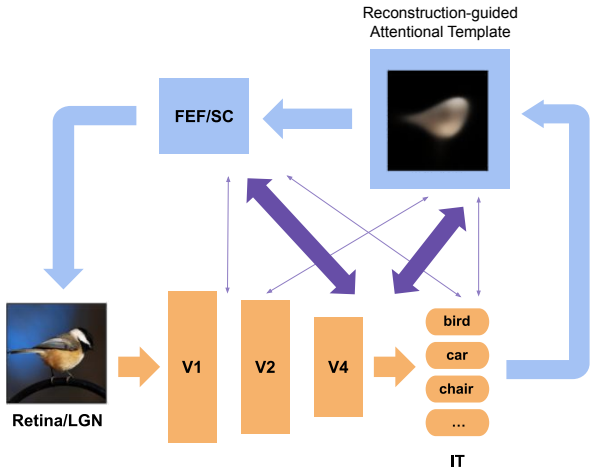

Figure 1: Object reconstruction-guided attention (ORA)

This work aims to investigate the potential brain mechanisms that underlie object-based attention facilitated by object reconstruction, using computational modeling. We implement a generative network that encodes visual features into compressed and abstract object representations (similar to the IT cortex), which are then used for Object Reconstruction-guided Attention (ORA, see Figure 1). Specifically, the ORA model uses reconstruction error as a top-down attentional bias to choose the object hypotheses that best explain how bottom-up features can be grouped to best match the visual input (known as "explain-away" behavior). We tested the proposed model across a wide range of visual tasks, including object recognition, grouping, and search. These tasks were selected to investigate how the model's object reconstructions might serve several different attention mechanisms, such as feature binding in mid-level visual areas (Pasupathy et al., 2020) and spatial attention control in the Frontal Eye Fields (FEF) and Superior Colliculus (SC) (Thompson, 2005; Schall, 2002; Krauzlis et al., 2013). In doing this, we aim to provide new insights into how the brain might use object reconstructions to mediate object-based attention control mechanisms in visual perception (Gershman, 2019; Yildirim et al., 2020; DiCarlo et al., 2021).

Modeling Method

Model Architecture: To implement ORA, we employ an object-centric generative network that was originally proposed by Monnier et al. (2021). Figure 2 provides an overview of the model's processing pipeline. The model encodes the in-

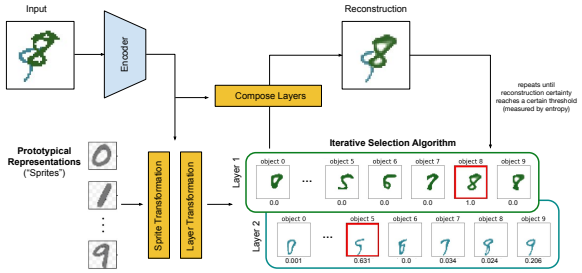

Figure 2: Model architecture. For details, see the text or zoom into the pdf of the figure.

put object(s) into abstract object representations, referred to as "sprites". These sprites are learnable category embeddings that are randomly initialized at the start of the training process and updated through training to form prototypical representations. The decoder consists of three stages: spatial transformations, sprite selection, and image composition. The spatial transformations include rotation, scaling, and translation (Jaderberg et al., 2016), while the sprite selection stage involves choosing the sprites that best explain the input. The image composition stage dictates the order in which the sprites should be placed and identifies which areas are occluded or occluding. We implemented baseline models having a Resnet50 architecture, followed by multiple dense layers to predict the target object class (for the recognition and grouping tasks) or the target object location (for the search task).

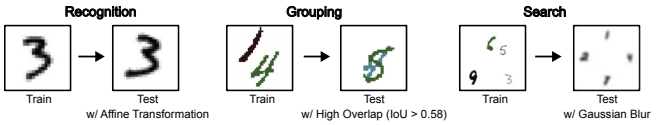

Figure 3: Training and testing dataset. Zoom in pdf file for a better view.

Model Training and Testing: To evaluate the performance of ORA, we conducted experiments on three visual tasks: object recognition, grouping, and search. For each task, we employed an out-of-distribution evaluation scheme to assess ORA's ability to generalize and maintain robustness to transformations (See Figure. 3 for example images for each task). This approach helped us determine whether the model had learned meaningful features applicable to complex scenarios and not just relying on trivial shortcuts. In the recognition task, we trained the model to reconstruct a single centrally-located digit and tested it with digits that had affine transformations applied. In the grouping task, we trained the model using two moderately overlapping MNIST digits and tested it under high overlap. For both tasks, we measured accuracy based on the class prototype selected for reconstruction relative to ground truth. In the search task, we trained the model with four digits randomly placed on the image and presented it with four digits, each Gaussian blurred, at fixed locations during inference. The model generates a guidance signal based on the reconstruction error of the target map from a specific target object hypothesis. Note that ORA was trained purely in an unsupervised manner, no ground truth target object labels were provided during training. This distinguishes our approach from the baseline models tested in our study, which relied on such labels.

Results and Discussion

Table 1 provides a quantitative assessment of the ORA model's performance in all three tasks. Since model training was unsupervised, it does not directly provide classification predictions for the recognition and grouping tasks. Instead, we estimated its recognition performance based on the class of sprite selected for reconstruction. The model exhibits high generalizability across different tasks, despite not being specifically trained for them, and achieves superior or comparable performance to feedforward baselines having a ResNet encoder devoted to each task. This performance is particularly noteworthy because the ORA did not rely on external supervision or category labels to perform the task, thereby better simulating the natural learning that occurs in the brain and more accurately capturing the neural mechanisms underlying object-based attention.

Table 1: Model performance and comparison

Task	ORA	ResNet (trained on each task)
Recognition	0.94	0.90
Grouping	0.73	0.42
Search	0.90	0.94

Figure 4 shows examples of ORA, not only making accurate attentional selections, but also making interpretable human-like errors. For instance, the model assigns an overlapping stroke between a 9 and 1 mainly to 1, thereby making an erroneous grouping and interpreting the remaining shape as a 5 (Fig. 4A). In the search task, the model reconstructs target-like objects from the peripheral input, which we call *target reconstruction*. This process involves "hallucinating" targets from the image, which is how ORA accounts for effects of target-distractor similarity on search guidance (e.g., mistaking a 7 for a 2 in Fig.4B). Future work will extend ORA to more naturalistic images and examine its ability to predict human recognition, grouping and search behavior, such as the reaction time, as well as patterns of human error.

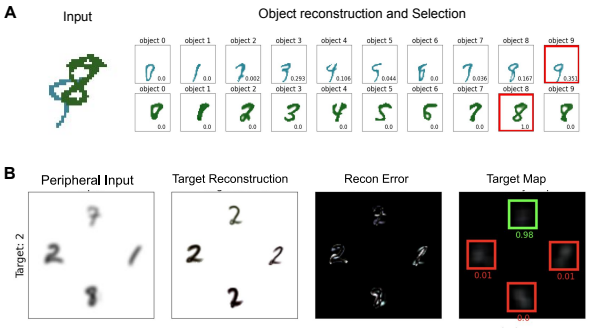

Figure 4: Visualizations of the model's incorrect predictions in the grouping task (A, top) and search task (B, bottom)

Acknowledgments

This work was supported in part by NSF IIS awards 1763981 and 2123920 to G.Z.

References

Adeli, H., Ahn, S., and Zelinsky, G. J. (2022). A brain-inspired object-based attention network for multi-object recognition and visual reasoning.

Ahn, S., Adeli, H., and Zelinsky, G. (2022). Reconstruction-guided attention improves the robustness and shape processing of neural networks. In *SVRHM 2022 Workshop@ NeurIPS*.

Bi, Z. (2021). Top-down generation of low-precision representations improves the perception and imagination of fine-scale visual information. preprint, Neuroscience.

Breedlove, J. L., St-Yves, G., Olman, C. A., and Naselaris, T. (2020). Generative Feedback Explains Distinct Brain Activity Codes for Seen and Mental Images. *Current Biology*, 30(12):2211–2224.e6.

Cavanagh, P., Caplovitz, G. P., Lytchenko, T. K., Maechler, M., Peter, U. T., and Sheinberg, D. (2022). Object-based attention.

Chen, Z. (2012). Object-based attention: A tutorial review. *Attention, Perception, & Psychophysics*, 74(5):784–802.

Clark, K., Khandelwal, U., Levy, O., and Manning, C. D. (2019). What Does BERT Look at? An Analysis of BERT's Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

DiCarlo, J. J., Haefner, R., Isik, L., Konkle, T., Kriegeskorte, N., Peters, B., Rust, N., Stachenfeld, K., Tenenbaum, J. B., Tsao, D., and others (2021). How does the brain combine generative models and direct discriminative computations in high-level vision?

Egly, R., Driver, J., and Rafal, R. D. (1994). Shifting visual attention between objects and locations: evidence from normal and parietal lesion subjects. *Journal of Experimental Psychology: General*, 123(2):161.

Einhäuser, W., Spain, M., and Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision*, 8(14):18–18.

Gershman, S. J. (2019). The Generative Adversarial Brain. *Frontiers in Artificial Intelligence*, 2:18.

Jaderberg, M., Simonyan, K., Zisserman, A., and Kavukcuoglu, K. (2016). Spatial Transformer Networks.

Jeurissen, D., Self, M. W., and Roelfsema, P. R. (2016). Serial grouping of 2D-image regions with object-based attention in humans. *Elife*, 5:e14320.

Krauzlis, R. J., Lovejoy, L. P., and Zénon, A. (2013). Superior colliculus and visual spatial attention. *Annual review of neuroscience*, 36.

Logan, G. D. (1996). The CODE theory of visual attention: an integration of space-based and object-based attention. *Psychological review*, 103(4):603.

Monnier, T., Vincent, E., Ponce, J., and Aubry, M. (2021). Unsupervised layered image decomposition into object prototypes. In *International Conference on Computer Vision*.

O'Craven, K. M., Downing, P. E., and Kanwisher, N. (1999). fMRI evidence for objects as the units of attentional selection. *Nature*, 401(6753):584–587.

Pasupathy, A., Popovkina, D. V., and Kim, T. (2020). Visual Functions of Primate Area V4. *Annual Review of Vision Science*, 6(1):363–385.

Roelfsema, P. R. and Houtkamp, R. (2011). Incremental grouping of image elements in vision. *Attention, Perception, & Psychophysics*, 73(8):2542–2572.

Schall, J. D. (2002). The neural selection and control of saccades by the frontal eye field. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 357(1424):1073–1082.

Scholl, B. J. (2001). Objects and attention: The state of the art. *Cognition*, 80(1-2):1–46.

Thompson, K. G. (2005). Neuronal Basis of Covert Spatial Attention in the Frontal Eye Field. *Journal of Neuroscience*, 25(41):9479–9487.

Yildirim, I., Belledonne, M., Freiwald, W., and Tenenbaum, J. (2020). Efficient inverse graphics in biological face processing. *Science advances*, 6(10):eaax5979.