Generating object-based attention through reconstruction

Seoyoung Ahn[1*], Hossein Adeli[1,2] and Gregory J. Zelinsky[1,3]

[1]Department of Psychology, Stony Brook University, NY, USA.
[2]Zuckerman Mind Brain Behavior Institute, Columbia University, NY, USA.
[3]Department of Computer Science, Stony Brook University, NY, USA.

*Corresponding author(s). E-mail(s): ahnseoyoung@gmail.com;
Contributing authors: ha2366@columbia.edu;
gregory.zelinsky@stonybrook.edu;

Abstract

Behavioral and neural evidence suggests that the fundamental unit of attention selection for humans is a complete object representation. However, the mechanism underlying object-based attention remains poorly understood. One open question about object-based attention is how high-level object representations can effectively guide bottom-up processing of low-level visual features. Here, we propose that the visual system attempts to continuously generate reconstructions of the objects in an input and that these reconstructions serve as a mechanism for object-based attention. Building on auto-encoder neural networks, we show that reconstructing the object's appearance and location from its abstract representation enables the visual system to better bind features and locations to create hypothesized objects, thus providing a mechanism for generating top-down attentional biases and selectively routing low-level features of specific objects. We evaluated our model using the MNIST-C (handwritten digits under corruptions) and ImageNet-C (real-world objects under corruptions) datasets, which are challenging recognition tasks requiring top-down attention to effectively filter out noise and occlusion. Our model not only demonstrated superior performance on these tasks, but also better accounted for human behavioral reaction times and error patterns than a standard feedforward Convolutional

Neural Network. Our work suggests that object reconstruction is a biologically plausible and effective object-based attention mechanism that can endow humans with robust object perception and recognition abilities.

Keywords: Generative Vision, Object-based Attention, Robustness

1 Introduction

The human visual system evolved to interact with objects, and toward this end it uses a complete object representation as its fundamental unit of processing. This is known as object-based attention (OBA; Cavanagh et al., 2023; Z. Chen, 2012; Duncan, 1984; Scholl, 2001; Vecera, 2000). Theories and models aimed at understanding OBA have focused on the low-level grouping mechanisms responsible for combining visual pieces into object parts and complete objects. In one of the first theories of OBA (Logan, 1996), attention was proposed to select perceptual groups based on the spatial proximity between feature signals in the visual input. More recently but similar in principle, the incremental grouping model (Roelfsema & Houtkamp, 2011) proposed that stimulus-driven grouping cues, such as the Gestalt laws (Wagemans et al., 2012), are responsible for attention seeming to spread throughout an object. However, how these early bottom-up processes interact with higher-level object representations remains an open and poorly understood question.

Because object representations by higher cortical areas are largely invariant to changes in an object's location or features (DiCarlo, Zoccolan, & Rust, 2012), a model of OBA is needed to bridge these higher-level object representations to lower-level representations closer to the visual input (Cavanagh et al., 2023). Top-down feedback is believed to be crucial in building this bridge. Previous studies highlight the importance of top-down feedback in robust object perception, particularly in noisy and ambiguous visual environments where bottom-up processes alone are insufficient (Kietzmann et al., 2019; Spoerer, McClure, & Kriegeskorte, 2017; Wyatte, Jilk, & O'Reilly, 2014). Our study focuses on understanding the role of top-down feedback in OBA, particularly how top-down feedback contributes to the binding of specific features and locations to objects and how top-down object reconstructions can be generated and used to control spatial attention. Our work also provides insight into a long-standing open question in the literature on attention mechanisms (Kravitz & Behrmann, 2011), which is how spatial (Posner, 1980; Treisman, 1996) and feature-based (Carrasco, 2011; Wolfe, 1994) attention mechanisms relate to what we know as object-based attention.

Here, we propose a novel generative model of object-based attention called Object Reconstruction-guided Attention (ORA). The fundamental premise of ORA is that the process of object reconstruction serves as the underlying mechanism by which high-level object representations are used to recover specific location and feature information about an object from the visual input. Under

this framework, the object reconstruction becomes an attentional template for introducing top-down biases attributed to attention (Beck & Kastner, 2009; Bundesen, Habekost, & Kyllingsbæk, 2011; Desimone & Duncan, 1995). This biasing occurs by prioritizing the object features that offer the most accurate interpretation of the retinal input, that is, minimize the reconstruction error (similar to "explain-away", Clark, 2013). The idea that reconstruction plays a role in perception is not new in cognitive neuroscience theory and has taken various forms such as "prediction" in predictive coding theory (Clark, 2013; Friston, 2010) or "synthesis" in analysis-by-synthesis theory (Yuille & Kersten, 2006). A promising direction in recent years has been to model human visual cortical processing using a neural network equipped with a decoder to reconstruct the input (Adeli, Ahn, & Zelinsky, 2023; Al-Tahan & Mohsenzadeh, 2021; Boutin, Zerroug, Jung, & Serre, 2020; Csikor, Meszéna, Szabó, & Orbán, 2022; Fleming & Storrs, 2019; Hedayati, O'Donnell, & Wyble, 2022; Xing, Chrastil, Nitz, & Krichmar, 2022; Yildirim, Belledonne, Freiwald, & Tenenbaum, 2020). However, there is currently no research addressing the role that top-down reconstruction feedback plays in human object-based attention and robust object perception.

ORA fills this gap by providing a computational model of the OBA mechanism that leverages object reconstructions to generate top-down attentional biases. As part of its novel neural network architecture, the model directs more attention to the areas and features of the input that align with the current best object reconstruction, which ultimately reduces the reconstruction error. In extensive model evaluation, we found that ORA exhibits remarkable robustness when applied to challenging recognition tasks such as MNIST-C (handwritten digits under corruptions, Mu & Gilmer, 2019) and ImageNet-C (real-world objects under corruptions, Hendrycks & Dietterich, 2019), effectively managing noise and occlusion present in the input image. ORA also accounts for behavioral reaction time and error results that we observed from human participants performing the same tasks, better than a baseline feedforward network (e.g., Resnet; He, Zhang, Ren, & Sun, 2016). We are hopeful that ORA's use of object reconstruction will provide new insights into the underlying neural mechanism for robust human object-based perception and attention.

2 The ORA model of object-based attention

Prior computational and empirical work suggest that the human brain constantly generates predictions about the visual world and adjusts our neural responses accordingly to align our perception with a single best prediction (de Lange, Heilbron, & Kok, 2018; Panichello, Cheung, & Bar, 2013). We hypothesize that an object reconstruction serves as a generated prediction about the possible appearance of an object, and that the reconstruction as a whole serves to bind features to object locations and mediate what we know as object-based attention biases (Beck & Kastner, 2009; Bundesen et al., 2011; Desimone &

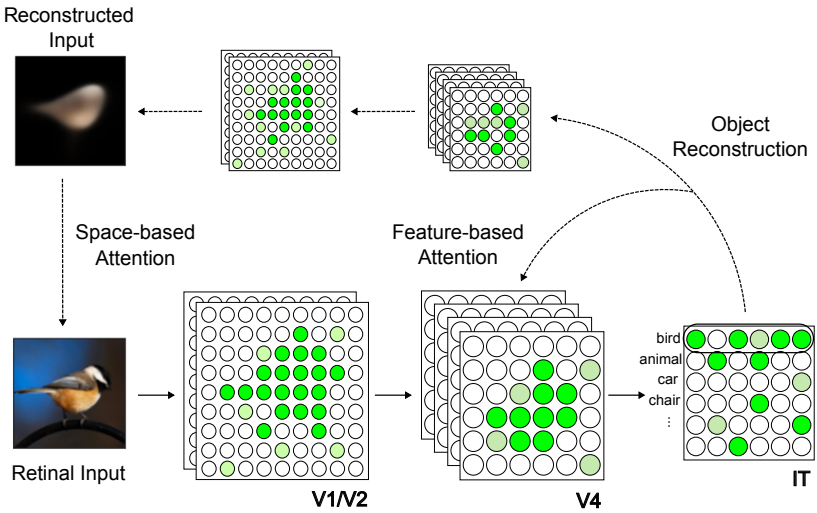


Fig. 1 A potential implementation of ORA in the brain. ORA uses object reconstruction as top-down attentional feedback and iteratively routes the most relevant visual information in successive steps of feedforward object recognition processes. This reconstruction-guided attention operates at two levels, 1) Spatial attention: a long-range projection that spatially constrains attention to the most likely object location and 2) Feature-based attention: more local feedback that biases feature binding strengths to favor the formation of a better object reconstruction. See main text for more details.

Duncan, 1995). We refer to our hypothesis as *Object Reconstruction-guided Attention* or ORA. According to this hypothesis, object reconstruction is achieved through top-down feedback originating from the most activated object category representation in the Inferior Temporal (IT) cortex (DiCarlo et al., 2012; Rajalingham & DiCarlo, 2019), and this serves as an attentional template by biasing subsequent bottom-up processing in the ventral stream.

Figure 1 shows how an ORA mechanism might be implemented in the brain. The retinal image is first processed through the ventral stream, which consists of multiple layers of linear and non-linear filters (e.g., V1, V2, and V4) specialized in extracting the visual features of objects (DiCarlo et al., 2012). As information flows through these layers, neuronal receptive field sizes progressively increase and responses grow more selective and abstract (depicted by the greenish circles). For our implementation, we model the ventral stream using a Convolutional Neural Network (CNN) incorporating skip connections between retinotopic feature maps (He et al., 2016). At the top of this hierarchy is the inferotemporal cortex (IT), where the low-dimensional latent features are encapsulated into separate category-selective slots (Rajalingham & DiCarlo, 2019)[1] Each slot in ORA comprises a set of neurons that represent a vector

[1]In computer vision, this encapsulation of information is sometimes referred to as a "capsule" (Doerig, Schmittwilken, Sayim, Manassi, & Herzog, 2020; Sabour, Frosst, & Hinton, 2017), although the term that we adopt is the more recently used "slot" (Greff, van Steenkiste, & Schmidhuber, 2020; Locatello et al., 2020). Note that in cognitive psychology, a slot corresponds to one "file" of information about an object as theorized by object-file theory (Kahneman, Treisman, & Gibbs, 1992; see also the idea of category-consistent features in Yu, Maxfield, & Zelinsky, 2016).

of category-specific information. For example, the neurons located in the top row of IT cortex in Figure 1 collectively represent the bird category, and their activity indicates the presence of that category. These object slots, despite being abstracted away from the visual input, maintain a coarse temporary binding between an object's features and its location (Peters & Kriegeskorte, 2021) and thus would be available to enable and mediate behaviors reflecting object-based attention (e.g., object tracking, Scholl, 2001).

The objects that ORA encodes into its IT slots are then decoded to generate top-down reconstructions of objects in the visual input. This encoding-decoding process is similar to an autoencoder architecture (Hinton & Salakhutdinov, 2006), where a model first encodes a retinal input to a highly compressed representation and then expands this latent representation through a decoder to reconstruct the original input. ORA's decoding process is different in that it iteratively generates reconstructions (rather than generating only one) and uses these preliminary reconstructions to recover a representation of retinotopy and facilitate the binding of object features to specific locations. Solving the feature binding problem (Greff et al., 2020) is essential in order for abstract object representations in higher-visual areas to influence lower-level processing in early visual cortex. ORA decodes an object reconstruction using dedicated feedback connections from an inversely mirrored CNN (Dumoulin & Visin, 2018). This CNN decoder takes the most activated object slot in ORA's IT cortex (measured by the sum of squares of all neural activation within that slot, i.e., vector magnitude), and sequentially reconstructs the input features using a series of convolutional and unpooling layers. More details about ORA's architecture can be found in the Methods.

We trained ORA to perform two tasks simultaneously: accurately classify an input and generate its reconstruction. Margin loss (Sabour et al., 2017) was used for classification training. By minimizing this loss, the activation of the target object slot (measured by vector magnitude) is pushed toward the desired value of 1, while the activation of non-target object slots is suppressed to near 0 (see Methods). For reconstruction loss, we simply compute the mean squared differences in pixel intensities between the input image and the model's reconstruction. During model evaluation (i.e., inference), we use ORA's object reconstructions as top-down attention feedback, which iterates through steps of ventral processing until a classification score reaches a threshold (see Methods). We propose that ORA's reconstruction-guided feedback plays two significant roles with respect to attention, indicated by the downward arrows in Figure 1. One is the control of spatial attention, a function potentially mediated by the dorsal pathway encompassing posterior parietal cortex (PPC) (Behrmann, Geng, & Shomstein, 2004; Xu, 2018). Another role is to facilitate the grouping and binding of object features, a function likely served by feature-based attention. These processes are hypothesized to occur in mid-level visual areas such as V4 (Pasupathy, Popovkina, & Kim, 2020), motivating ORA's separate top-down feedback connection to this brain structure. In the following sections, detailed methods are provided to implement

these connections and demonstrate how they capture the functions of spatial and feature-based attention in the brain.

2.1 Reconstruction-based spatial masking

Spatial attention refers to the prioritization of information at a spatial location by filtering or masking out information from all other locations (Posner, 1980), with the neural expression of this being enhanced activity at the attended location and inhibition elsewhere (Carrasco, 2011). We hypothesize that object reconstructions, generated by the network of parietal brain areas implicated in representing priority in spatial attention (Behrmann et al., 2004; Xu, 2018), can serve this spatial masking function by feeding back to early ventral areas to modulate feedforward ventral processing. The spatial mask imposed by ORA constrains early feedforward neural resources to represent the shape of the reconstructed object, as supported by multiple empirical studies on human selective attention (Y. Chen & Zelinsky, 2019; Duncan, 1984; Müller & Hübner, 2002). Our approach can be contrasted with an attention modeling approach known as "hard attention" (Adeli, Ahn, & Zelinsky, 2023; Cordonnier et al., 2021), which selects a uniformly weighted patch from an image for attention. In this study we focus on ORA's covert spatial masking function and its role in facilitating object recognition, however, ORA can also be applied to overt attention given that the primate fovea, working with an oculomotor system, imposes another overt form of spatial masking via the spatial locations that we

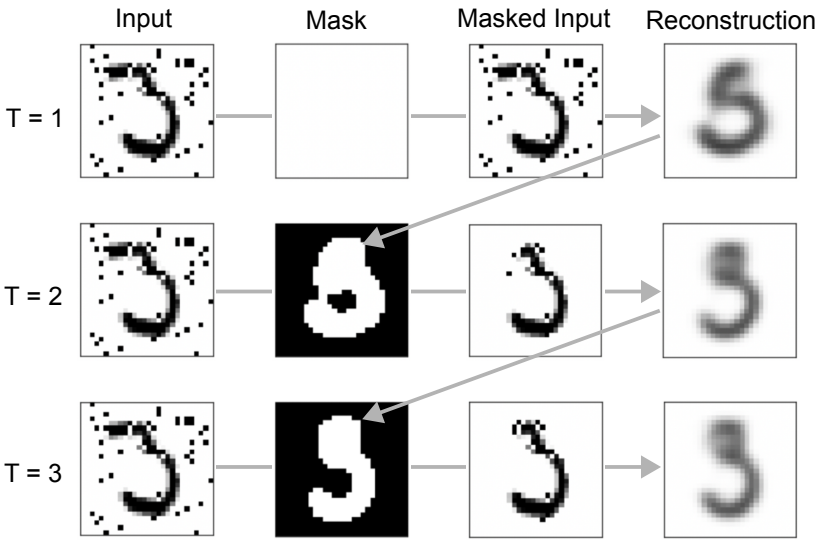


Fig. 2 Object reconstructions and generated spatial masks for three foward processing steps (out of a possible five in total). The spatial mask serves to limit ORA's attention to the shape of the most likely object. In this example, the model's most likely digit hypothesis changed from a 5 to a 3. The groundtruth class is 3.

select for eye-movement fixation. This application of ORA to fixation-location prediction is beyond the scope of this study.

By constraining its spatial attention to the shape of the most likely object, ORA is able to filter out background noise and improve its classification accuracy. Figure 2 illustrates how this accuracy benefit occurs. Initially, ORA hypothesized that the input was the digit 5, but after filtering out noise using its reconstruction mask the model correctly changed its hypothesis to a 3, matching the ground truth. On average, ORA's iterative spatial masking enabled it to successfully recover the object's correct identity on over 90% of the trials when the model started with an incorrect hypothesis (see Extended Data A). Previous research has similarly argued for benefits from a reconstruction-based masking approach, but in the context of augmenting datasets for improving the learning of robust visual representations for use in downstream tasks (M. Chen et al., 2020; He et al., 2022). ORA is different in that it leverages the object-based reconstructions that it learns during training to generate at inference spatial attention masks to focus processing on the hypothesized object shape. ORA updates its reconstruction-guided mask at the end of each forward pass, and globally repeats this process until the model's classification confidence surpasses a predetermined threshold or until the limit of five forward processing steps is reached, whichever comes first (see Methods for additional details).

2.2 Reconstruction-based feature binding

How the varied features of an object become spatially integrated into a coherent object representation is known as the binding problem (Humphreys, 2016; Treisman, 1996). The binding problem remains an open question despite a half century of active research. Object-based attention requires solving the binding problem, and many model solutions have been proposed (Greff et al., 2020). In a recent model, capsule networks (CapsNets, Sabour et al., 2017) used representational similarity to adjust binding strengths between objects and features (i.e., the color red would would bind more strongly to a strawberry than a banana). ORA also engages the binding problem. In addition to its long-range feedback serving a spatial masking function, ORA sends reconstruction-guided feedback to mid-level visual areas (V4 in Figure 1), which are believed to be involved in the grouping and binding of features into objects (Pasupathy et al., 2020). ORA builds on this approach by iteratively incorporating its incrementally-generated reconstructions into its feature-object binding process. Specifically, ORA modulates the weights between the V4 and IT layers, suppressing the binding strength between a feature and an object if it leads to a poorer reconstruction match to the input (see Methods for pseudocode). Because this algorithm operates solely by suppressing information inconsistent with ORA's object hypothesis, it also predicts a narrowing of the population response and is therefore consistent with the information tuning model of feature-based attention (Carrasco, 2011; Martinez-Trujillo & Treue,

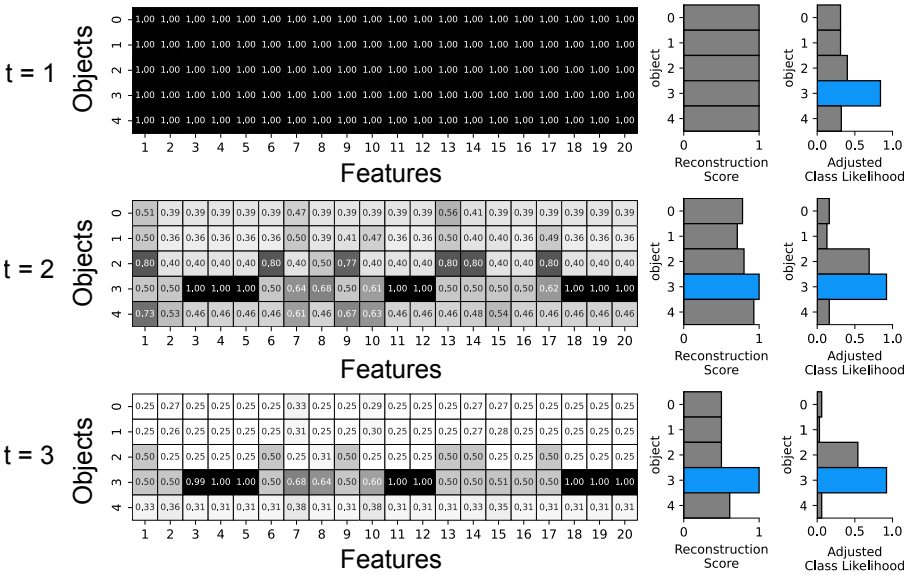


Fig. 3 Step-wise visualizations of reconstruction-based feature binding. The matrices show binding coefficients between object slots (rows) and feature slots (columns) for three time steps. The binding coefficients were initialized to one at t = 1, resulting in the dark matrix on the top, but in the subsequent time steps there is a rapid selective suppression of coefficients (matrix cells becoming lighter) as the model learns the object features leading to higher reconstruction accuracy. In the illustrated example, the coefficient matrix becomes sparser with each iteration as ORA focuses its attention on the features of the digit 3 (darker line forming along the fourth row). The middle column of bar graphs show the hypothesis yielding the highest reconstruction score (in blue), and the rightmost column shows the classification score (class likelihood) adjusted by the reconstruction score. For clarity, only 5 object slots and 20 feature slots are illustrated, but the full figure can be found in Extended Data B.

2004). Figure 3 shows ORA's binding coefficients (which modulate the original weights between the V4 and IT layers) sharpening over iterations and becoming more focused on the specific object features that best match the input. Binding coefficients were updated over three iterations during inference (testing); coefficients were not updated during model training. Available computing resources required us to limit ORA's reconstruction-based feedback to only three iterations, however we observed during piloting that three steps was usually sufficient for ORA to converge on a classification.

3 Results

We evaluated ORA in challenging object recognition tasks. First, we compared ORA's recognition performance to that of a standard feedforward CNN using a version of the MNIST dataset designed to test for effects of occlusions and specific forms of visual noise (MNIST-C; Mu & Gilmer, 2019). Both ORA and the CNN were trained on "clean" handwritten digits and then systematically

tested on each form of corrupted digit. This evaluation revealed that ORA used its reconstruction-based feedback to filter out irrelevant information and this led to greater recognition robustness to the local changes in texture introduced by the corruptions. Second, we ablated ORA's spatial masking and feature binding mechanisms and found the former to be most important for the accuracy of ORA's predictions and the latter to be most important for the speed (i.e., number of iterations) that ORA took to make its recognition judgement. Third, we collected reaction time (RT) data from humans performing the same recognition tasks and found that, on trials having the longest RTs, ORA also needed more reconstruction-based feedback before recognizing the corrupted digit. In a fourth analysis we found that this agreement between human and model extended to them making a shared pattern of errors (i.e., the same digit confusions). Lastly, we conducted a limited evaluation of ORA's generalization to natural images by using the ImageNet-C dataset (Hendrycks & Dietterich, 2019), which consists of various corruptions but now applied to the ImageNet objects. Details about the datasets and experimental procedures can be found in the Methods.

3.1 ORA improves recognition performance under visual corruptions

We used the MNIST-C dataset (Mu & Gilmer, 2019) to evaluate the effectiveness of ORA's reconstruction-based feedback in improving recognition robustness under visual corruptions. This dataset includes 15 types of corruptions: affine transformation, noise, blur, occlusion, and others (See Extended Data C for example images of each). We also report results separately on a subset of the dataset referred to as **MNIST-C-shape**, which includes those corruptions relating mainly to object shape. Given ORA's use of object-centric feedback, we hypothesized that robustness benefits would be greatest for these shape-specific corruptions. As shown in Table 1, this hypothesis was largely confirmed. We report results for two types of CNN model backbones used by ORA to approximate the features extracted by ventral visual brain areas: a shallow 2-layer CNN (2 Conv) and a deep 18-layer CNN with skip connections (Resnet-18; He et al., 2016). We also implemented three baseline models for

Table 1 Model comparison results using MNIST-C and MNIST-C-shape datasets. Recognition accuracy (means and standard deviations from 5 trained models, hereafter referred to as model "runs") from **ORA** and a **CNN** baseline, both of which were trained using identical CNN encoders (one a 2-layer CNN and the other a Resnet-18), and a **CapsNet** model following the implementation in Sabour et al. (2017).

Dataset	**ORA**		**CNN**		**CapsNet**
	Resnet-18	2 Conv	Resnet-18	2 Conv	
MNIST-C	**91.84 (0.67)**	88.32 (0.34)	81.94 (0.43)	88.94 (0.64)	75.14 (1.00)
MNIST-C-shape	**96.24 (0.74)**	92.82 (0.90)	86.43 (0.44)	91.80 (0.88)	81.87 (0.55)

comparative evaluation: two are CNNs having the same backbone encoders as ORA but followed by fully-connected layers for digit classification, and the third is a CapsNet model (Sabour et al., 2017, see Methods for details).

The best recognition accuracies on both MNIST-C and MNIST-C-shape were from ORA with a Resnet-18 encoder. Moreover, during piloting we observed that using only the low-spatial frequencies in the image to generate the object reconstruction was sufficient to produce recognition robustness comparable to what is achieved with full-resolution object reconstruction (Extended Data D). This shows that effective attention biases can be mediated by low-frequency neural feedback (Bar et al., 2006). ORA outperformed a CNN by a significant margin ($t(8) = 7.026$, $p < 0.001$, *Cohen's* $d = 4.444$), particularly excelling on the MNIST-C-shape dataset where it showed an improvement of over 4.5% ($t(8) = 8.617$, $p < 0.001$, *Cohen's* $d = 5.450$). Despite incorporating object-centric representations for reconstruction (but lacking top-down attentional feedback), the CapsNet model exhibited poor performance compared to ORA or a CNN, indicating that the addition of reconstruction alone does not translate into improved accuracy and robustness. ORA was even more accurate than models specifically designed for generalization, such as CNNs trained using PGD/GAN adversarial noise (Extended Data E). Lastly, we observed that increasing the capacity of the encoder from a 2-layer to an 18-layer CNN led to over-fitting by the CNN baseline and a consequent degradation of its accuracy from 88.94% to 81.94% on MNIST-C. ORA suffered no such degradation and achieved higher accuracy with the higher capacity encoder.

Figure 4A shows specific comparisons between our ORA model and the CNN baseline (the second best model) for each of the corruption types from

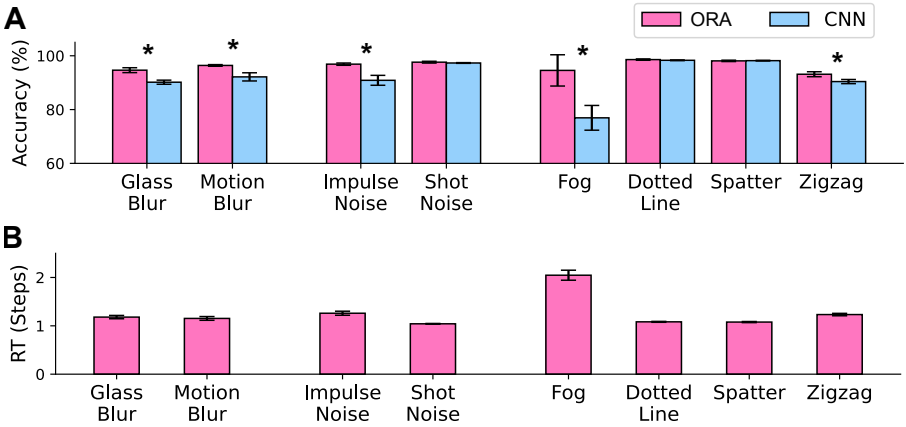


Fig. 4 Model accuracy and reaction time for specific corruption types from MNIST-C-shape. **A:** ORA vs CNN using each model's best performing encoder (Resnet-18 and a 2-layer CNN, respectively). **B:** ORA's reaction time (RT), estimated as the number of forward processing steps required to reach a confidence threshold. Error bars indicate standard deviations from 5 different model runs. Asterisks (*) indicate statistical significance at a p-value of < 0.05.

MNIST-C-shape. For comprehensive results, refer to Extended Data F. ORA consistently outperforms the CNN baselines and exhibits particularly strong performance under the fog corruption ($t(8) = 5.318$, $p < 0.001$, *Cohen's d* = 3.363). As shown in Figure 4B, on average ORA confidently recognizes corrupted digits in 1.4 steps (measured as the number of forward processing steps required to reach a confidence threshold; see Methods). However, it takes significantly longer (more than 2 steps) under the fog corruption, suggesting that top-down attentional feedback is particularly helpful in cases of visual corruptions that remove high-spatial frequencies.

3.2 Revealing distinct roles of spatial and feature-based attention

ORA uses two types of reconstruction-based feedback, roughly corresponding to spatial and feature-based attention, and we evaluated their unique contributions by ablating each from the best performing ORA model (i.e., Resnet-18 encoder) and observing the effect on performance. The left panel of Figure 5 shows that ORA's recognition accuracy stemmed from its ability to use a generated reconstruction to perform spatial masking. Removing spatial masking led to a significant drop in performance of over 6% (-Spatial / +Feature; $t(8) = 4.682$, *p-adjust* = 0.009, *Hedges' g* = 2.674). Ablating the feature binding component did not significantly affect prediction accuracy (+Spatial / -Feature; $t(8) = 0.294$, *p-adjust* = 1.0, *Hedges' g* = 0.168), but this does not mean that it had no effect on performance. As shown in the right panel of Figure 5, feature-binding ablation instead lead to an increase in ORA's RT ($t(8) = 10.207$, $p < 0.001$, *Cohen's d* = 6.455). Note that we cannot report the corresponding spatial ablation because ORA's RT is based on the number of spatial-mask iterations and would therefore always be zero. These findings suggest that ORA uses spatial and feature-based attention to perform distinct functional roles. Spatial attention improves recognition accuracy and enhances the neural signal-to-noise ratio by masking out noise from irrelevant

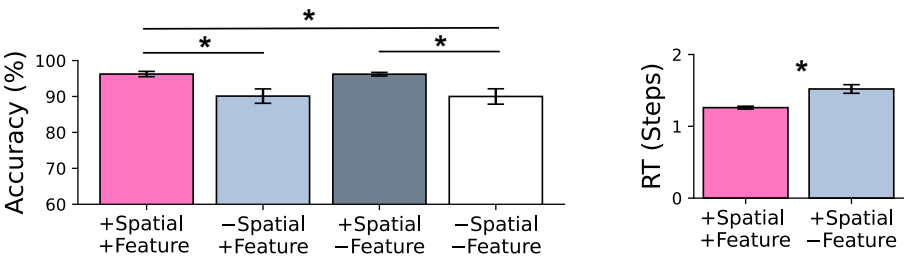


Fig. 5 Results from ablating (-) ORA's reconstruction-based spatial masking and feature binding components. Leftmost bars in pink indicate the complete (no ablation) ORA model. Left: Recognition accuracy. Right: reaction time, approximated by the number of forward processing steps taken to recognize digits with confidence. Error bars are standard deviations from 5 different model runs. Asterisks (*) indicate statistical significance at a p-value of < 0.05 with Bonferroni corrections.

spatial locations. On the other hand, feature-based attention enhances the overall discriminability of the encoder by suppressing features associated with competing object hypotheses. However, when the ventral encoder has a lower capacity for discriminability (i.e., the shallower 2-layer CNN), we observed that reconstruction-based feature binding also significantly influences both ORA's RT and accuracy (Extended Data G). The distinct roles played by feature and spatial attention in ORA's function is further supported by the differential impact that each had on the model's ability to recover from an initially incorrect hypothesis (Extended Data A). On average, spatial attention led to an incorrect hypothesis changing to a correct prediction on approximately 6% of these trials, with a maximum of 40% incorrect-to-correct changes observed for the fog corruption. In contrast, iterations of feature-binding resulted in hypothesis changes on less than 1% of trials. This suggests that ORA uses feature-binding primarily to increase the efficiency of its recognition of MNIST-C digits whereas spatial masking is used to modify digit hypotheses and improve recognition accuracy.

3.3 Comparison with human RT

As shown in Figure 4B, ORA's RT behavior changes depending on the difficulty of the recognition discrimination. The ORA model iteratively generates the object reconstruction that it hypothesizes is the most plausible explanation of the input and uses it as a spatial mask for the subsequent feedforward step. This process repeats until the model achieves a predetermined level of classification confidence measured by the entropy of the classification scores (softmax output). In the majority of cases, ORA reached the confidence threshold within a single feedforward sweep (Figure 4B). However, on trials where the digit is inherently ambiguous or becomes ambiguous due to noise, ORA alternated between two object hypotheses, resulting in a longer time required for one to reach the confidence threshold (Extended Data H).

To assess how this RT variability compares to human RTs, we conducted a psychophysical experiment on humans who were asked to perform the same challenging MNIST-C recognition tasks (Figure 6A; see Methods for details). Participants (N=146) viewed 480 images of digits and were instructed to press a keyboard spacebar as soon as they recognized the object in the image. This RT judgment was followed by a second accuracy check in which participants used the keypad to enter the recognized digit. Only correct trials were used for the RT analysis. To encourage a distribution of human RTs, we selected MNIST-C digits for use as experimental stimuli based on the level of difficulty predicted by ORA. Specifically, we created easy, medium, and difficult conditions consisting of digits taking ORA either 1 step, 2-3 steps, or 4-5 steps to recognize, respectively (Figure 6B, left panel). Figure 6B (right panel) shows evidence for an alignment between human recognition accuracy and the three difficulty groups predicted by ORA. Human accuracy was by far the highest in the easy condition (i.e., only a single feedforward step was needed by ORA

for accurate model recognition), followed by the medium (2-3 steps) and difficult (4-5 steps) conditions, with the latter showing the most errors. The mean human RT was 1130ms (SD = 530ms), but there was a significant effect of condition difficulty, $F(2, 284) = 105.266$, $p < 0.001$, $\eta^2 = 0.426$. Mean RTs in the difficult condition were 1265ms (SD = 649ms), but these decreased to 1170ms (SD = 517ms) in the medium condition and 985ms (SD = 473ms) in the easy condition. Lastly, Figure 6C shows a statistically significant positive Pearson correlation between human RTs and ORA's RTs ($r = 0.33$, $p < 0.001$), suggesting that ORA and humans found the same trials to be difficult.

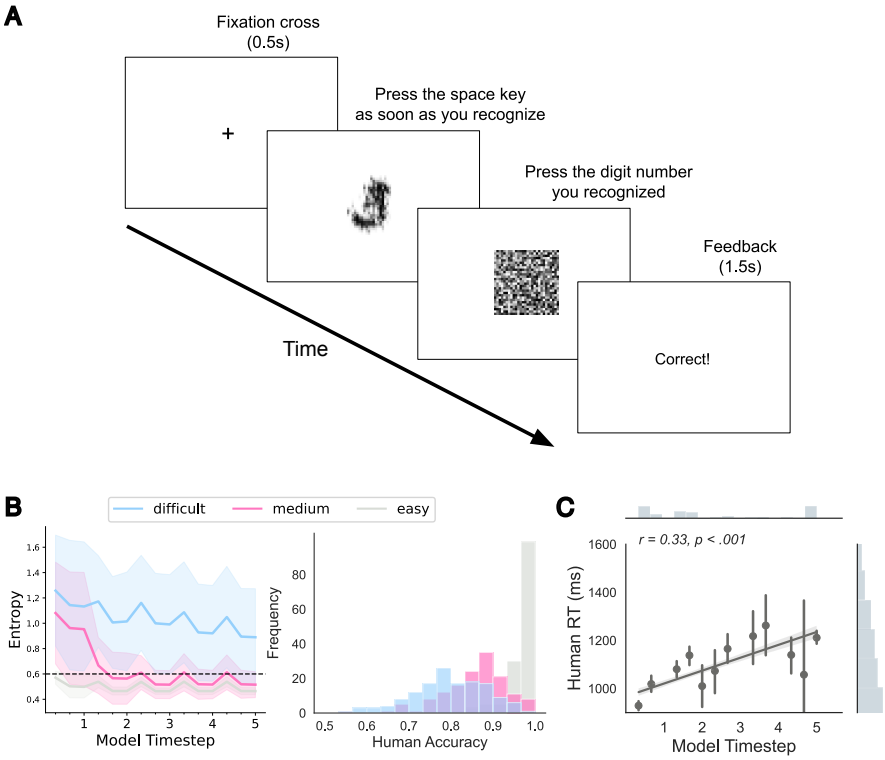


Fig. 6 Human recognition experiment and results. **A**: Overview of the behavioral paradigm for measuring recognition reaction time (first button press) and accuracy (second button press). **B**: Stimuli were grouped into easy, medium, and difficult levels of recognition difficulty based on the number of forward processing steps needed for ORA to reach a predetermined recognition confidence threshold (dashed line; left panel). Colored regions indicate the corresponding standard deviations (SDs) across images. The unnumbered ticks on the x-axis indicate three local iterations of feature-based attention occurring within each forward processing (i.e., global spatial-masking iteration). The right panel shows distributions of human accuracy (averaged over participants) for the three difficulty conditions. **C**: Correlation between human RT and ORA's RT, with the normalized marginal distribution of each variable shown on the top and right panels. Error bars represent standard error estimates bootstrapped from 1000 samples. Correlation plots for all MNIST-C corruption types can be found in Extended Data I.

3.4 ORA's errors and explain-away behavior

To gain a clearer insight into underlying processing mechanisms, here we make a deeper comparison between the errors generated by ORA and those from the best-performing CNN from Table 1. We found that ORA exhibits interesting systematic errors as it tries to explain-away random noise in the image and reconstruct it into a specific pattern that it has learned. For example, the model often connects noise speckles to generate a line, as demonstrated in the first row of Figure 7A, thereby mistakenly reconstructing a digit 9 instead of the ground truth digit 4. Similarly, ORA occasionally misinterprets zig-zag noise as being part of the digit, thereby reconstructing a digit 7 instead of the

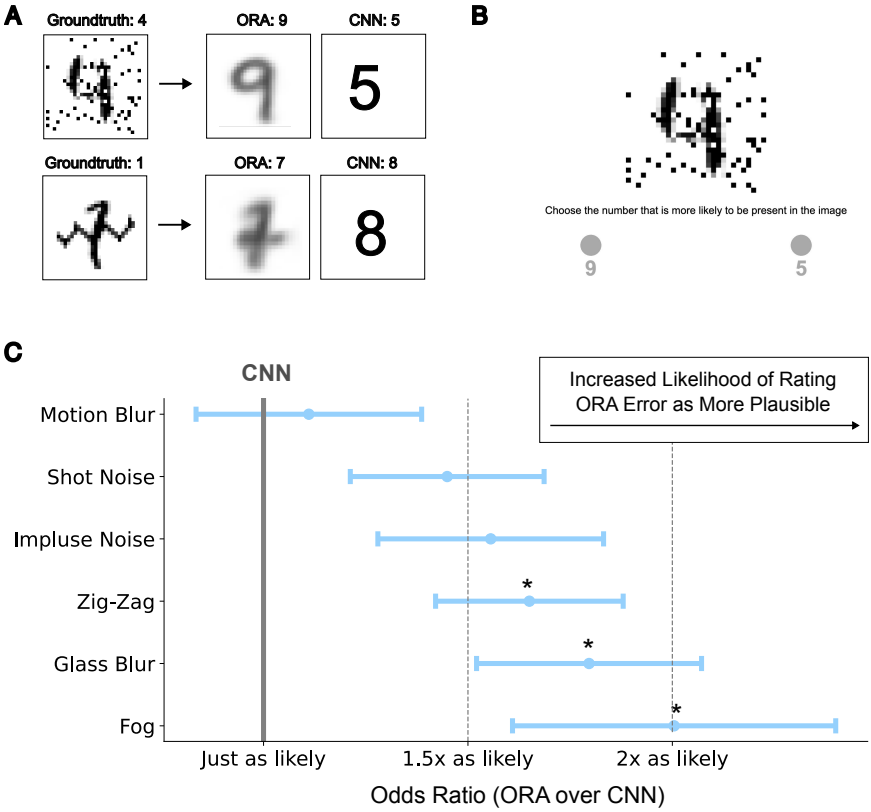


Fig. 7 ORA often hallucinates a non-existing pattern out of noise, and these errors are perceived as more likely by humans compared to errors made by a CNN baseline. **A**: Example of ORA's explain-way behavior and the resulting errors in classification. **B**: Two alternative forced choice experimental procedure requiring participants to choose which of two model predictions, one from ORA and the other from a CNN (but both errors), is the more plausible human interpretation of the corrupted digit. **C**: The relative likelihood of ORA's errors being perceived as more like those from a human, quantified as an odds ratio (a higher value indicates a more plausible ORA error). Error bars indicate standard errors, and asterisks (*) indicate odds ratios greater than 1 at a significance level of $p < 0.05$.

ground truth digit 1 (second row). These types of interpretable errors were not observed in the CNN baseline.

To determine whether ORA's recognition errors are more similar to human errors compared to those from a CNN, we collected data from 184 participants in a two-alternative forced-choice task. On each trial, participants were shown a corrupted digit together with two mistaken model predictions, one from ORA and the other from the CNN, and they had to indicate which they thought would be more plausibly generated by a human (Figure 7B; see Methods for details). Note that we selected MNIST-C images as stimuli to satisfy two constraints: (1) that both ORA and the CNN baseline made an error in predicting the ground truth digit, and (2) that these two model errors differed from each other. To evaluate our data we computed an odds ratio, which measures the relative perceived likelihood of ORA's errors compared to the CNN's errors. We found that this ratio was consistently greater than 1 (with 1 indicating equally likely) for every type of corruption (Figure 7C). This was particularly true for the fog corruption, where ORA's errors were rated twice as likely than those made by the CNN baseline ($t(19) = 2.61$, $p = 0.02$, *cohen's* $d = 0.58$, testing against being > 1). Interestingly, the fog corruption type is also where ORA exhibited the highest performance advantage over the CNN (Figure 4). These results suggest that humans find ORA's errors to be plausible, at least more so than errors from a CNN.

3.5 Generalization to natural images

We evalauted whether ORA's reconstruction-based feedback mechanism, shown to be beneficial for digit recognition, would generalize to more complex and natural images of objects. We did this using the ImageNet-C dataset (Hendrycks & Dietterich, 2019), which is similar to MNIST-C but visual corruptions are applied to ImageNet objects (Deng et al., 2009). A dataset called 16-class ImageNet (Geirhos, Temme, et al., 2018) was developed to teach models a more human-like semantic structure by selecting from ImageNet a more comparable set of 16 basic-level categories (see Methods for the list of categories), and we used this dataset for model training. To accommodate the complex backgrounds upon which objects appear in ImageNet images, we incorporated a recent segmentation model (Cheng, Schwing, & Kirillov, 2021) to estimate the ground-truth object boundaries, then trained the model to reconstruct only the pixels within the object segments. The resulting object reconstructions capture coarse representations of object shape, missing texture detail (see Extended Data J).

Despite the coarseness of the generated reconstructions, their use as feedback provided sufficient spatial masking to significantly improve recognition performance. Figure 8 reports model recognition accuracy (y-axes) for five increasing levels of corruption (x-axes) from the ImageNet-C dataset. We compared ORA's performance to a CNN baseline model, now a ResNet-50 trained to recognize ImageNet objects. ORA's feedforward ventral process was approximated by the identical ResNet-50 (see Methods). Due to limitations

in computational resources, we were able to conduct only a single training for each model (i.e., no error bars). Still, a clear pattern emerged. Although both models exhibit similarly accurate recognition at low corruption levels, at higher levels ORA is more robust on several types (on average > 7% greater than the CNN baseline without reconstruction feedback; Extended Data K). The most notable benefits are observed with noise corruptions (top row), where accuracy increased for Gaussian noise by up to 35%. For some types of blur, the CNN baseline performed as well as ORA or slightly better, with the CNN working best for a pixelation corruption (bottom-right). These effects, however are generally not as large as the improvements achieved by ORA (see Extended Data K for more results).

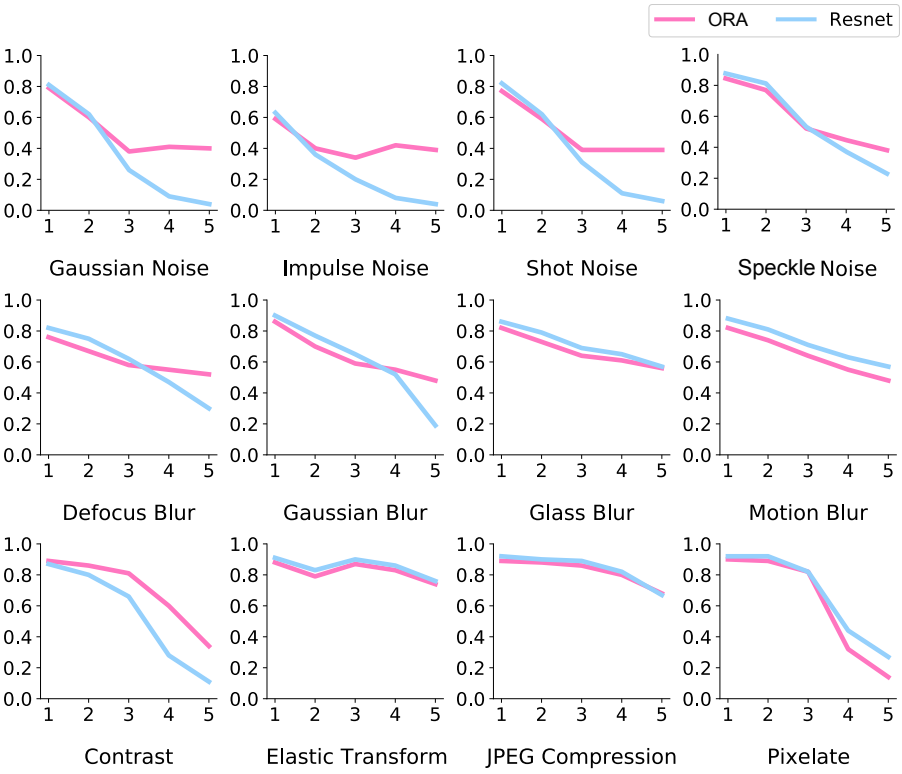


Fig. 8 Model performance on the ImageNet-C dataset, with classification accuracy on the y-axes and level of corruption on the x-axes. ORA often demonstrates a clear advantage over the CNN baseline (ResNet-50), particularly with high levels of noise corruption.

4 Discussion

By whatever name is preferred, our common visual experience suggests that top-down, generative or reconstructive processes play a significant role in creating our visual perceptions. This is evident in phenomena such as visual imagery, object completion, and pareidolia. However, there are still open questions regarding *why* we have this ability to reconstruct objects and *what* functional role it plays in shaping our percepts (Breedlove, St-Yves, Olman, & Naselaris, 2020; DiCarlo et al., 2021; Gershman, 2019). Through a combination of computational modeling and behavioral experiments, we demonstrate that object reconstruction serves as a mechanism for generating an effective object-based attentional template that can bias bottom-up processing to support challenging object recognition tasks. We implemented a neural network model of Object Reconstruction-guided Attention (ORA) using an autoencoder-like architecture (Hinton & Salakhutdinov, 2006; Sabour et al., 2017). ORA consists of an encoder that encodes the visual input into low-dimensional object-centric representations and a decoder that reconstructs objects from these representations. These object reconstructions are generated iteratively, thereby enabling their use to route the most relevant spatial and feature information to feedforward object recognition processes. We found that ORA is highly robust in recognizing digits under a wide range of corruptions, outperforming other feedforward model baselines. Ablation studies revealed distinct contributions of spatial and feature-based attention in enhancing the model's performance. In humans, spatial attention is believed to improve recognition accuracy by increasing the signal-to-noise ratio in neural responses, whereas feature-based attention is believed to improve recognition efficiency (i.e., speed) by sharpening the population of neural responses. ORA's attention mechanisms serve similar functions, enabling it to predict human RTs and error patterns in recognition tasks better than comparable CNN baselines. Lastly, we showed that the benefits of ORA's reconstruction-based attention mechanisms generalize to recognition tasks using natural objects, with larger benefits observed in more challenging recognition tasks.

One important open question in the study of object-based attention is how abstract object representations in high-level visual cortex feed back to bias specific locations and features represented in lower visual cortical areas (Cavanagh et al., 2023; Vecera, 2000). According to ORA, object-based attention is a generative process of reconstructing information about an object's location and features, and this attention mechanism is what creates the top-down object-based biases. ORA also suggests that object-based attention plays a role in mediating spatial and feature-based attention, as suggested by previous findings (Kravitz & Behrmann, 2011; Scolari, Ester, & Serences, 2014; Shomstein & Behrmann, 2006). We propose that spatial and feature-based attention operate through two independent reconstruction-based feedback pathways. Spatial attention is mediated by long-range feedback connections between parietal areas (Behrmann et al., 2004; Xu, 2018), which we believe to be involved in the computation of global shape representations (Ayzenberg & Behrmann, 2022),

and early visual areas, where bottom-up processing becomes constrained by the shape of the top-down hypothesized object. Feature-based attention uses relatively shorter-range feedback to control the feature binding process occurring in mid-level visual areas (Pasupathy et al., 2020). Our suggestion of parallel pathways of reconstruction-based attentional feedback might help to clarify mixed findings in the clinical cognitive neuroscience literature reporting intact spatial attention in patients who have impaired feature-based or object-based attention (de-Wit et al., 2009; Finsterwalder et al., 2017, as cited in Cavanagh et al., 2023).

With our ORA model, we propose that top-down generative feedback plays a role in the human brain's ability to robustly recognize objects. One hallmark of human visual recognition is its robustness against various types of noise and corruption (Avidan et al., 2002; Vogels & Biederman, 2002). However, the currently most popular model for the ventral recognition stream, Convolutional Neural Networks (CNNs), often fall alarmingly short in replicating this human-like robustness (Dodge & Karam, 2017; Szegedy et al., 2013). Although this gap is closing as models and datasets become larger (Geirhos et al., 2021), CNNs seem to rely on fundamentally different feature representations than those used by humans (e.g., using texture more than shape, Baker, Lu, Erlikhman, & Kellman, 2018; Geirhos, Rubisch, et al., 2018). While controlling for model size with our CNN baselines, we show that ORA's use of object reconstruction as top-down generative feedback significantly enhances its recognition robustness, enabling its performance to generalize to image corruptions that were not part of its training. ORA is particularly robust to noise corruptions in both MNIST-C and ImageNet-C that require the greatest global shape processing. Our findings suggest that the top-down generative feedback required for object reconstruction, together with ventral feedforwad processing, may contribute to building the human representation of global object shape and facilitating robust object recognition (Ayzenberg & Behrmann, 2022).

One question not answered in our current work is how completely an object must be reconstructed in order to serve as an effective attentional template. Our finding that even low-resolution object reconstructions are sufficient to generate attentional biases and performance benefits align with previous studies showing that the brain initially processes low spatial frequency information about objects for the purpose of generating top-down feedback signals for guiding the bottom-up processing of finer-grained details (Bar, 2003; Bar et al., 2006; Bi, 2021). However, effects of feature-based attention are also found in a range of specific low-level features (e.g., color, orientation, etc, Maunsell & Treue, 2006), making further computational experiments needed to determine whether optimization of the information and level of object reconstruction might yield even greater performance benefits. Relatedly, although ORA's current high compression rate limited its ability to reconstruct fine-grained object detail, it may be possible to address this limitation by exploiting recent advancements in decoder architecture and composition-based methods in the computer vision literature to generate reconstructions of varying kind and

quality (M. Chen et al., 2020; Singh, Deng, & Ahn, 2022). These questions, and broader questions regarding the biological plausibility of a reconstruction-based modeling approach, would benefit by additional study of the neural mechanisms underlying representation generation and object reconstruction in the brain. Studies on visual imagery have shown substantial overlap in neural processing between perception and mental imagery (Breedlove et al., 2020; Dijkstra, Bosch, & van Gerven, 2019), and more comparisons revealing similarities and differences between patterns of brain activity evoked by seen images and imagined images could provide valuable insights into the nature of the reconstruction process.

In our study we evaluated the effectiveness of ORA's reconstruction-based feedback using the relatively simple task of recognizing a single object, albeit under sometimes severe visual corruption. However, in the real world multiple objects typically compete for neural representation (Duncan, 1984; Luck, Chelazzi, Hillyard, & Desimone, 1997), making the role of ORA's top-down attention mechanism (e.g., spatially masking out background noise) even more critical. In future work, we aim to extend ORA and its evaluation to attentional phenomena involving multiple objects. One potential phenomenon is perceptual grouping, which involves separating an object of interest from other overlapping objects or a complex background (or both). In our application of ORA to the ImageNet-C images, we observed that the current implementation of the model found it difficult to differentiate the figure object from the complex naturalistic background, which is why we found it necessary to train ORA using supervision from an object segmentation method. We will address this limitation in future work by incorporating into ORA's attention mechanism an unsupervised (yet efficient) new method for parsing an object's features from a surrounding background or an overlapping object (Adeli, Ahn, Kriegeskorte, & Zelinsky, 2023; Greff et al., 2020). Another promising task is visual search, the orientation of attention to a specific "target" object in a visual input. Like object recognition and grouping, visual search is a common everyday activity that has deep theoretical connections to spatial and feature-based attention mechanisms (Treisman, 1996; Wolfe, 1994) and this makes the visual search task nicely aligned with ORA's reconstruction-based attention mechanisms. For example, the guidance of attention to a target has been modeled as a comparison between a visual input and an internal representation of the target object known as a "target template" (Zelinsky, 2008). Among the functions that we already identified, in future work we will explore whether ORA's generated object reconstructions can also serve as target templates to guide eye movements during visual search.

Our study has broad implications for the fields of behavioral vision science and computer vision. While there is still only limited empirical evidence for the use of generative processes in vision (Dijkstra & Fleming, 2023; Gershman, 2019), there is growing interest in capturing the robustness of human perception by modeling it as a reconstructive process (Adeli, Ahn, & Zelinsky, 2023; Al-Tahan & Mohsenzadeh, 2021; Boutin et al., 2020; Csikor et al., 2022;

Fleming & Storrs, 2019; Hedayati et al., 2022; Xing et al., 2022; Yildirim et al., 2020). A reconstruction framework offers the potential to unify our understanding of top-down cognitive processes, including attention biases (Cavanagh et al., 2023), memory retrieval (Hedayati et al., 2022), hypothesis testing (Bar et al., 2006), and mental imagery (Breedlove et al., 2020). We also hope that our work will inspire the development of novel computational architectures in computer vision. Methods that train models to reconstruct objects have been shown to enhance generalization performance (Dittadi et al., 2022; Locatello et al., 2020), but this approach has primarily been explored in the context of pretraining or incorporating auxiliary loss (He et al., 2022; Rasmus, Berglund, Honkala, Valpola, & Raiko, 2015; Sabour et al., 2017). Our work is the first to show that a reconstruction approach can effectively model attentional biases during inference, which had been an open research question (Shi, Darrell, & Wang, 2023).

In summary, our study demonstrates that object reconstruction can serve as a plausible neural mechanism for achieving robust object-based perception and attention. We model this mechanism as ORA, which uses object reconstruction as a top-down attention bias to focus bottom-up processing on information relevant to what it considers to be the most plausible (i.e., recognizable) object shape in the visual input. This reconstruction-based mechanism unifies object-based, space-based, and feature-based modes of attention into one computational attention system.

5 Methods

5.1 ORA Implementation

For ORA's encoder, which models the ventral stream, we used different pytorch implementations of feedforward CNNs: one a shallow 2-layer CNN[2] (2 Conv) and a deeper CNN[3] with skip connections (Resnet with 18 and 50 layers for the MNIST-C and ImageNet-C tasks, respectively; He et al., 2016). For ORA's decoder we used three fully-connected dense layers to reconstruct the MNIST image and an inversely mirrored Resnet to reconstruct the ImageNet images. As in the original CapsNet model (Sabour et al., 2017), we used a vector implementation of object and feature slots, where the length of each slot vector represents the presence of an object class in the image. To make this value range from 0 to 1 (i.e., 0 for absence and 1 for presence), we used the following nonlinear squash function. We also provide specification details for object slot representations (Sabour et al., 2017) and the total number of trainable parameters.

$$o_t^j = \frac{\|v_t^j\|^2}{1 + \|v_t^j\|^2} \frac{v_t^j}{\|v_t^j\|} \tag{1}$$

[2]https://github.com/pytorch/examples/blob/main/mnist/main.py
[3]https://github.com/pytorch/vision/blob/main/torchvision/models/resnet.py

	for MNIST-C task		**for ImageNet-C task**
Encoder type	2 Conv	Resnet-18	Resnet-50
Num of feature slots	1152	288	3136
Feature slot dimension	8	8	8
Num of object slots	10	10	16
Object slot dimension	16	16	49
Total number of parameters	2,904,720	4,581,296	67,116,160

5.2 Loss Function

ORA outputs both object slot activations and image reconstructions, and the losses for each output are computed and combined using a weighting parameter of $\lambda_{recon} = 0.392$ (Eq.2). For classification, we used margin loss (Sabour et al., 2017) to match the object slot activations to the ground truth (Eq. 3). The first term of the loss is active only when the target object is present ($T_j > 0$), and minimizing this term pushes the object slot activation (measured by the vector magnitude $\|o_j\|$) to be closer to 1, with a small margin (m). The second term applies only when the target object is absent, and minimizing this term pushes the predicted object slot activation below the small margin (m).

$$Total\,Loss = \sum_{j \in class} Class\,Loss_j + \lambda_{recon} Recon\,Loss \tag{2}$$

$$Class\,Loss_j = \begin{cases} T_j \cdot max(0, (T_j - m) - \|o_j\|)^2, & \text{if } T_j > 0 \\ max(0, \|o_j\| - m)^2, & \text{if } T_j = 0 \end{cases} \tag{3}$$

5.3 Baseline Implementation

The CNN baselines consist of the same encoders used for ORA (See Section 5.1) followed by two fully connected layers for classification readout. The readout layers are trained using the dropout technique for regularization (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014). Negative Log Likelihood loss is computed on the log softmax output for classification. The total number of model parameters are: 1,199,882 for the 2-layer CNN baseline, 2,803,882 for the ResNet-18 baseline in the MNIST-C experiment, and 23,540,816 for the ResNet-50 baseline in the ImageNet experiment. Performance for the other baselines reported in Extended Data are taken from the original benchmarking paper (Mu & Gilmer, 2019).

5.4 Training Details

All model parameters were updated using Adam optimization (Kingma & Ba, 2014) with a mini-batch size of 128 for the MNIST-C experiment and 32 for the ImageNet-C experiment. For the MNIST-C experiment, we used the exponentially decaying learning rate by a factor of 0.96 at the end of each training epoch, starting from 0.1. Similarly, for the ImageNet-C experiment,

the learning rate followed the same exponential decay pattern, but with a starting value of 0.0001. Model training was terminated based on the results from the validation dataset (10% of randomly selected training dataset) and was stopped early if there was no improvement in the validation accuracy after 20 epochs. Models were implemented in the PyTorch framework and trained on a single GPU using a NVIDIA GeForce RTX 3090 (24 GB) for the MNIST-C experiment and a RTX A6000 (48 GB) for the ImageNet-C experiment.

5.5 Measuring ORA's Reaction Time

ORA's processing is limited to five forward processing steps, each followed by reconstruction-based spatial attention feedback (i.e., global spatial masking iteration). However, the model typically did not need all five of these forward processing steps in order to confidently recognize an object, and the number of steps actually taken by the model to reach a classification threshold is ORA's RT. This threshold is measured by entropy over the softmax distribution of classification scores, similar to the method used in Spoerer, Kietzmann, Mehrer, Charest, and Kriegeskorte (2020), and we chose an entropy threshold of 0.6 based on a grid search. If this threshold is not reached within the five forward processing steps, we use the prediction from the final time step for classification.

5.6 Spatial Masking Methods

In each forward processing step, the model generates a reconstruction-guided spatial mask based on the most likely object reconstruction. To create this mask, we simply apply a threshold to the reconstructed pixel values. We used thresholds of 0.1 for MNIST images and 0.05 for ImageNet images, where the pixel intensities in both range from 0 to 1. This thresholding process creates a boolean mask where pixels with values below the threshold are assigned 0 and pixels with values above the threshold are assigned 1. The masked input for the next forward processing step is obtained by element-wise multiplication of the reconstruction with the boolean mask. The non-zero areas of the masked input are normalized using MaxMin normalization (Equation 4, when p corresponds to the reconstructed pixel value at image coordinates i and j). For ImageNet images, the RGB reconstructed images are first converted to grayscale before applying the thresholding and normalization processes.

$$p^{ij} = lb + (ub - lb)\frac{p^{ij} - min(p^{ij})}{max(p^{ij}) - min(p^{ij})} \tag{4}$$

5.7 Feature Binding Algorithm

ORA's reconstruction-guided feature binding process computes coefficients to dynamically modulate weights between features and objects. These coefficients, denoted as C_{ij}, are computed iteratively over three local iterations occurring within each forward processing step (i.e., global spatial masking

iteration). We update the original weights by multiplying them by the final coefficients, $C_{ij} \cdot W_{ij}$. In the original CapsNet model (Sabour et al., 2017), these binding coefficients are modulated based on representational similarity, which is calculated as the dot product between the features and objects (Line 6 in the Pseudocode). In our model, we also adjust these binding coefficients to suppress the features that are connected to objects having high reconstruction error. By suppressing the features leading to inaccurate object reconstructions, those features that are most selective to the hypothesized object become bound. All coefficients are normalized over objects using MaxMin normalization (like Equation 4).

Algorithm 1 Reconstruction-guided Feature Binding

1: **initialize** for all object j and feature i:
2: Binding coefficients $c^{ij} \leftarrow 1$
3: Feature similarity score $b^{ij} \leftarrow 0$
4: Reconstruction score coefficients $r^{ij} \leftarrow 0$
5: **while** feature binding iteration $t = 1, 2, 3$ **do**
6: For all object j: $\hat{f}^{j|i} \leftarrow W^{ij} f^{i}$
7: For all object j: $o_t^j \leftarrow squash(\sum_i c_t^{ij} \hat{f}^{j|i})$ ▷ Eq. 1
8: For all object j and feature i: $b_t^{ij} \leftarrow b_t^{ij} + \hat{f}^{j|i} \cdot o_t^j$
9: $b_t^{ij} \leftarrow max(maxmin(b_t^{ij}, dim = j), 0.5)$ ▷ Eq.4
10: For all object j: $reconstruction_t^j \leftarrow Decoder(o_t^j)$
11: For all object j: $r_t^j \leftarrow -MSE(image^j, reconstruction_t^j)$
12: For all feature i: $r_t^{ij} \leftarrow Repeat(r_t^j)$
13: $r_t^{ij} \leftarrow max(maxmin(r_t^{ij}, dim = j), 0.5)$ ▷ Eq.4
14: $c_t^{ij} \leftarrow b_t^{ij} \cdot r_t^{ij}$
15: **end while**
16: For all object j: $o^j \leftarrow squash(\sum_i c^{ij} \hat{f}^{j|i})$

5.8 16-Class-ImageNet Training Dataset

The original Imagenet dataset includes 1,000 subordinate-level labels. These include 120 different dog breeds, which results in a semantic structure that is highly atypical compared to that of an average person (Ahn, Zelinsky, & Lupyan, 2021) that may influence model performance. To address this problem, we also trained ORA using 16-class-ImageNet, a subset of ImageNet that groups its 1000 fine-grained classes into 16 basic-level categories: airplane, bicycle, boat, car, chair, dog, keyboard, oven, bear, bird, bottle, cat, clock, elephant, knife, and truck (Geirhos et al., 2021; Geirhos, Temme, et al., 2018).

5.9 Behavioral Methods for the MNIST-C Recognition Experiment

Participants A total of 146 Stony Brook University undergraduates, self-identified as 94 women, 47 men, and 2 non-binary individuals, participated in the experiment for course credit. This sample size was determined based on an a priori power analysis using G*Power (Faul, Erdfelder, Lang, & Buchner, 2007), which indicated that a minimum of 28 participants would be sufficient for a repeated-measures ANOVA to detect a medium effect size with a power of .80 at a significance level of $\alpha = .05$. This estimate is based on each participant viewing a 96 image subset of the 480 total testing images (see Stimuli and Apparatus), a design choice motivated by a desire to avoid participant fatigue and to maintain their engagement in the task. Participants all had normal or corrected to normal vision, by self report. This study was approved by the Institutional Review Board responsible for overseeing research involving human participants.

Stimuli and Apparatus Stimuli consisted of 480 images from the MNIST-C testing dataset Mu and Gilmer (2019), 60 images for each of the 8 corruption types in MNIST-C-shape. Images were divided into 5 sets of 96 images, where each participant viewed a 96 image set. To experimentally test ORA's RTs we selected the images used for behavioral testing based on the RT predicted by the model. Specifically, for each corruption type we selected 20 images by randomly sampling from 3 different model RT conditions: short (1 step), medium (2-3 step), and long (4-5 step). Note that all images were correctly classified by the model, meaning that this manipulation is targeted to the number of forward processing steps needed by ORA to recognize an object. Selected images were presented on a 19-inch flat-screen CRT ViewSonic SVGA monitor at a screen resolution of 1024×768 pixels and a refresh rate of 100 Hz. Participants were seated approximately 70 cm from the monitor and at this viewing distance the image height subtended approximately 8°of a visual angle. Space bar responses were made with the participant's dominant hand using a standard computer keyboard.

Procedure Participants were instructed that they would see a series of centrally presented digits, ranging from 0 to 9, and that they should press the space bar as soon as they recognized the digit class. However, the task emphasized accuracy as well as speed. Participants were instructed to make their key press when they felt confident in their answer, although a reminder was provided if the response was too slow (> 5 sec). Upon pressing the space bar, the original stimuli was masked by a randomly generated pixel image and participants were asked as part of a secondary judgment to report the digit that they recognized by pressing the digit on the numeric keypad. At the start of each trial a central cross appeared immediately prior to stimuli presentation (500 ms), and feedback was provided at the end of each trial on the accuracy of

the digit recognition response (10 alternative forced choice). Each participant viewed 96 images, 4 images $\times$ 3 model RT conditions $\times$ 8 corruption types. Ten practice trials were provided before the experiment so that participants could familiarize themselves with the task. The order of presentation of the images was randomized over participants.

5.10 Behavioral Methods for the MNIST-C Error Likelihood Rating Experiment

Participants A total of 184 undergraduate students from Stony Brook University, self-identified as 120 women, 61 men, and 3 non-binary individuals, participated in the experiment for course credit. Sample size was determined based on an a priori power analysis using G*Power (Faul et al., 2007), which indicated that a minimum of 41 participants per stimulus set (4 stimulus sets in total) would be sufficient for a paired t-test to detect a medium effect size with a power of .80 at a $\alpha = .05$ level of significance. All participants had normal or corrected to normal vision, by self report. This study was approved by the appropriate Institutional Review Board.

Stimuli and Apparatus Stimuli consisted of 120 images selected from the MNIST-C testing dataset Mu and Gilmer (2019), 20 images per corruption type, to satisfy the experimental constraint of having competing models make different errors to a stimulus. These ambiguous images were divided into 4 separate sets, with each set containing 30 images. Only six corruption types from the eight MNIST-C-shape categories were used because the number of images meeting the experimental constraint (i.e., different errors) were too few for analysis in the dotted-line and spatter corruption conditions. The stimuli image spanned 40% of their screen size in height. All responses were made with a computer mouse or trackpad.

Procedure Participants accessed the rating experiment remotely using their desktop devices. On each trial, a single corrupted digit was presented at a size of 40% of the vertical screen dimension. Below this stimulus were the two model predictions, horizontally separated (the location was randomized), and participants were instructed to rate which of the two would be more plausibly made by a person (two alternative forced choice; see Fig. 7B). A response was required on each trial and participants were not allowed to skip images, although they could change their answers anytime before proceeding to the next trial (which they initiated by clicking a "Next" button). Three practice trials were provided before the experiment so that participants would familiarize themselves with the task and interface. Each participant viewed 30 images. The order of presentation of the images was randomized over participants.

Code availability. The code necessary to reproduce our results is available on the github repository at https://github.com/ahnchive/recon-mnist

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

References

Adeli, H., Ahn, S., Kriegeskorte, N., Zelinsky, G. (2023, May). *Affinity-based Attention in Self-supervised Transformers Predicts Dynamics of Object Grouping in Humans* (No. arXiv:2306.00294). arXiv.

Adeli, H., Ahn, S., Zelinsky, G.J. (2023, May). A brain-inspired object-based attention network for multiobject recognition and visual reasoning. *Journal of Vision*, *23*(5), 16.

10.1167/jov.23.5.16

Ahn, S., Zelinsky, G.J., Lupyan, G. (2021, December). Use of superordinate labels yields more robust and human-like visual representations in convolutional neural networks. *Journal of Vision*, *21*(13), 13.

10.1167/jov.21.13.13

Al-Tahan, H., & Mohsenzadeh, Y. (2021, March). Reconstructing feedback representations in the ventral visual pathway with a generative adversarial autoencoder. *PLOS Computational Biology*, *17*(3), e1008775.

10.1371/journal.pcbi.1008775

Avidan, G., Harel, M., Hendler, T., Ben-Bashat, D., Zohary, E., Malach, R. (2002). Contrast sensitivity in human visual areas and its relationship to object recognition. *Journal of neurophysiology*, *87*(6), 3102–3116.

10.1152/jn.2002.87.6.3102

Ayzenberg, V., & Behrmann, M. (2022, October). Does the brain's ventral visual pathway compute object shape? *Trends in Cognitive Sciences*, S1364661322002406.

10.1016/j.tics.2022.09.019

Baker, N., Lu, H., Erlikhman, G., Kellman, P.J. (2018, December). Deep convolutional networks do not classify based on global object shape. *PLOS Computational Biology*, *14*(12), e1006613.

10.1371/journal.pcbi.1006613

Bar, M. (2003, May). A cortical mechanism for triggering top-down facilitation in visual object recognition. *Journal of Cognitive Neuroscience*, *15*(4), 600–609.

10.1162/089892903321662976

Bar, M., Kassam, K.S., Ghuman, A.S., Boshyan, J., Schmid, A.M., Dale, A.M., ... Halgren, E. (2006, January). Top-down facilitation of visual recognition. *Proceedings of the National Academy of Sciences*, *103*(2), 449–454.

10.1073/pnas.0507062103

Beck, D.M., & Kastner, S. (2009, June). Top-down and bottom-up mechanisms in biasing competition in the human brain. *Vision Research*, *49*(10), 1154–1165.

10.1016/j.visres.2008.07.012

Behrmann, M., Geng, J.J., Shomstein, S. (2004, April). Parietal cortex and attention. *Current Opinion in Neurobiology*, *14*(2), 212–217.

10.1016/j.conb.2004.03.012

Bi, Z. (2021, January). Top-down generation of low-precision representations improves the perception and imagination of fine-scale visual information. *bioRxiv*, 2021.05.07.443208.

10.1101/2021.05.07.443208

Boutin, V., Zerroug, A., Jung, M., Serre, T. (2020). Iterative VAE as a predictive brain model for out-of-distribution generalization. *NeurIPS 2020 workshop SVRHM.*

Breedlove, J.L., St-Yves, G., Olman, C.A., Naselaris, T. (2020, June). Generative feedback explains distinct brain activity codes for seen and mental images. *Current Biology*, *30*(12), 2211-2224.e6.

10.1016/j.cub.2020.04.014

Bundesen, C., Habekost, T., Kyllingsbæk, S. (2011, May). A neural theory of visual attention and short-term memory (NTVA). *Neuropsychologia*, *49*(6), 1446–1457.

10.1016/j.neuropsychologia.2010.12.006

Carrasco, M. (2011). Visual attention: The past 25 years. *Vision research*, *51*(13), 1484–1525.

10.1016/j.visres.2011.04.012

Cavanagh, P., Caplovitz, G.P., Lytchenko, T.K., Maechler, M.R., Tse, P.U., Sheinberg, D.L. (2023, April). The architecture of object-based

attention. *Psychonomic Bulletin & Review*.

10.3758/s13423-023-02281-7

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., Sutskever, I. (2020). Generative pretraining from pixels. *International Conference on Machine Learning* (pp. 1691–1703). PMLR.

Chen, Y., & Zelinsky, G.J. (2019, January). Is there a shape to the attention spotlight? Computing saliency over proto-objects predicts fixations during scene viewing. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(1), 139–154.

10.1037/xhp0000593

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

10.3758/s13414-012-0322-z

Cheng, B., Schwing, A., Kirillov, A. (2021). Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems* (Vol. 34, pp. 17864–17875). Curran Associates, Inc.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, *36*(3), 181–204.

10.1017/S0140525X12000477

Cordonnier, J.-B., Mahendran, A., Dosovitskiy, A., Weissenborn, D., Uszkoreit, J., Unterthiner, T. (2021). Differentiable patch selection for image recognition. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2351–2360).

Csikor, F., Meszéna, B., Szabó, B., Orbán, G. (2022, June). *Top-down inference in an early visual cortex inspired hierarchical Variational Autoencoder* (No. arXiv:2206.00436). arXiv.

de Lange, F.P., Heilbron, M., Kok, P. (2018, September). How do expectations shape perception? *Trends in Cognitive Sciences*, *22*(9), 764–779.

10.1016/j.tics.2018.06.002

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). Ieee. 10.1109/

CVPR.2009.5206848

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual review of neuroscience*, *18*(1), 193–222.

10.1146/annurev.ne.18.030195.001205

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

DiCarlo, J.J., Zoccolan, D., Rust, N.C. (2012, February). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434.

10.1016/j.neuron.2012.01.010

Dijkstra, N., Bosch, S.E., van Gerven, M.A. (2019). Shared neural mechanisms of visual perception and imagery. *Trends in cognitive sciences*.

10.1016/j.tics.2019.02.004

Dijkstra, N., & Fleming, S.M. (2023, March). Subjective signal strength distinguishes reality from imagination. *Nature Communications*, *14*(1), 1627.

10.1038/s41467-023-37322-1

Dittadi, A., Papa, S.S., Vita, M.D., Schölkopf, B., Winther, O., Locatello, F. (2022). Generalization and robustness implications in object-centric learning. K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, & S. Sabato (Eds.), *International conference on machine learning, ICML 2022, 17-23 july 2022, baltimore, maryland, USA* (Vol. 162, pp. 5221–5285). PMLR.

Dodge, S., & Karam, L. (2017). A study and comparison of human and deep learning recognition performance under visual distortions. *2017 26th international conference on computer communication and networks (ICCCN)* (pp. 1–7). 10.1109/ICCCN.2017.8038465

Doerig, A., Schmittwilken, L., Sayim, B., Manassi, M., Herzog, M.H. (2020). Capsule networks as recurrent models of grouping and segmentation. *PLoS computational biology*, *16*(7), e1008017.

10.1371/journal.pcbi.1008017

Dumoulin, V., & Visin, F. (2018, January). *A guide to convolution arithmetic for deep learning* (No. arXiv:1603.07285). arXiv.

Duncan, J. (1984). Selective attention and the organization of visual information. *Journal of experimental psychology: General*, *113*(4), 501.

10.1037/0096-3445.113.4.501

Faul, F., Erdfelder, E., Lang, A.-G., Buchner, A. (2007, May). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191.

10.3758/BF03193146

Fleming, R.W., & Storrs, K.R. (2019, December). Learning to see stuff. *Current Opinion in Behavioral Sciences*, *30*, 100–108.

10.1016/j.cobeha.2019.07.004

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature reviews neuroscience*, *11*(2), 127–138.

10.1038/nrn2787

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. *Advances in Neural Information Processing Systems*, *34*, 23885–23899.

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. *International conference on learning representations.*

Geirhos, R., Temme, C.R., Rauber, J., Schütt, H.H., Bethge, M., Wichmann, F.A. (2018). Generalisation in humans and deep neural networks. *Advances in neural information processing systems*, *31*.

Gershman, S.J. (2019, September). The generative adversarial brain. *Frontiers in Artificial Intelligence*, *2*, 18.

10.3389/frai.2019.00018

Greff, K., van Steenkiste, S., Schmidhuber, J. (2020, December). On the Binding Problem in Artificial Neural Networks. *arXiv:2012.05208 [cs]*. https://arxiv.org/abs/2012.05208 [cs]

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R. (2022, June). Masked autoencoders are scalable vision learners. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 16000–16009).

He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hedayati, S., O'Donnell, R.E., Wyble, B. (2022, May). A model of working memory for latent representations. *Nature Human Behaviour*, *6*(5), 709–719.

10.1038/s41562-021-01264-9

Hendrycks, D., & Dietterich, T.G. (2019). Benchmarking neural network robustness to common corruptions and perturbations. *International Conference On Learning Representations*.

Hinton, G.E., & Salakhutdinov, R.R. (2006, July). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507.

10.1126/science.1127647

Humphreys, G.W. (2016, October). Feature confirmation in object perception: Feature integration theory 26 years on from the Treisman Bartlett lecture. *Quarterly Journal of Experimental Psychology*, *69*(10), 1910–1940.

10.1080/17470218.2014.988736

Kahneman, D., Treisman, A., Gibbs, B.J. (1992, April). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology*, *24*(2), 175–219.

10.1016/0010-0285(92)90007-O

Kietzmann, T.C., Spoerer, C.J., Sörensen, L.K., Cichy, R.M., Hauk, O., Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, *116*(43), 21854–21863.

10.1073/pnas.1905544116

Kingma, D.P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. https://arxiv.org/abs/1412.6980

Kravitz, D.J., & Behrmann, M. (2011). Space-, object-, and feature-based attention interact to organize visual scenes. *Attention, Perception, & Psychophysics*, *73*, 2434–2447.

10.3758/s13414-011-0201-z

Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., ... Kipf, T. (2020). Object-centric learning with slot attention. *Advances in Neural Information Processing Systems*, *33*, 11525–11538.

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

10.1037/0033-295X.103.4.603

Luck, S.J., Chelazzi, L., Hillyard, S.A., Desimone, R. (1997). Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. *Journal of neurophysiology*, *77*(1), 24–42.

10.1152/jn.1997.77.1.24

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. *International conference on learning representations.*

Martinez-Trujillo, J.C., & Treue, S. (2004, May). Feature-based attention increases the selectivity of population responses in primate visual cortex. *Current Biology*, *14*(9), 744–751.

10.1016/j.cub.2004.04.028

Maunsell, J.H., & Treue, S. (2006). Feature-based attention in visual cortex. *Trends in neurosciences*, *29*(6), 317–322.

10.1016/j.tins.2006.04.001

Mu, N., & Gilmer, J. (2019, June). MNIST-C: A Robustness Benchmark for Computer Vision. *arXiv:1906.02337 [cs]*. https://arxiv.org/abs/1906.02337 [cs]

Müller, M.M., & Hübner, R. (2002, March). Can the spotlight of attention be shaped like a doughnut? Evidence from steady-state visual evoked potentials. *Psychological Science*, *13*(2), 119–124.

10.1111/1467-9280.00422

Panichello, M.F., Cheung, O.S., Bar, M. (2013). Predictive feedback and conscious visual experience. *Frontiers in Psychology*, *3*.

10.3389/fpsyg.2012.00620

Pasupathy, A., Popovkina, D.V., Kim, T. (2020, September). Visual functions of primate area V4. *Annual Review of Vision Science*, *6*(1), 363–385.

10.1146/annurev-vision-030320-041306

Peters, B., & Kriegeskorte, N. (2021, September). Capturing the objects of vision with neural networks. *Nature Human Behaviour*, *5*(9), 1127–1144.

10.1038/s41562-021-01194-6

Posner, M.I. (1980). Orienting of attention. *Quarterly journal of experimental psychology*, *32*(1), 3–25.

10.1080/00335558008248231

Rajalingham, R., & DiCarlo, J.J. (2019, April). Reversible inactivation of different millimeter-scale regions of primate IT results in different patterns of core object recognition deficits. *Neuron*, *102*(2), 493-505.e5.

10.1016/j.neuron.2019.02.001

Rasmus, A., Berglund, M., Honkala, M., Valpola, H., Raiko, T. (2015). Semi-supervised learning with ladder networks. *Advances in neural information processing systems*, *28*.

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

10.3758/s13414-011-0200-0

Sabour, S., Frosst, N., Hinton, G.E. (2017). Dynamic routing between capsules. *Advances in neural information processing systems* (pp. 3856–3866).

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

10.1016/S0010-0277(00)00152-9

Schott, L., Rauber, J., Bethge, M., Brendel, W. (2018). Towards the first adversarially robust neural network model on MNIST. *International conference on learning representations.*

Scolari, M., Ester, E.F., Serences, J.T. (2014, January). Feature- and Object-Based Attentional Modulation in the Human Visual System. A.C.K. Nobre & S. Kastner (Eds.), *The Oxford Handbook of Attention* (p. 0). Oxford University Press. 10.1093/oxfordhb/9780199675111.013.009

Shi, B., Darrell, T., Wang, X. (2023). Top-down visual attention from analysis by synthesis. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2102–2112).

Shomstein, S., & Behrmann, M. (2006, July). Cortical systems mediating visual attention to both objects and spatial locations. *Proceedings of the National Academy of Sciences*, *103*(30), 11387–11392.

10.1073/pnas.0601813103

Singh, G., Deng, F., Ahn, S. (2022). Illiterate DALL-E learns to compose. *International conference on learning representations.*

Spoerer, C.J., Kietzmann, T.C., Mehrer, J., Charest, I., Kriegeskorte, N. (2020, October). Recurrent neural networks can explain flexible trading of speed and accuracy in biological vision. *PLOS Computational Biology*, *16*(10), e1008215.

10.1371/journal.pcbi.1008215

Spoerer, C.J., McClure, P., Kriegeskorte, N. (2017, September). Recurrent convolutional neural networks: A better model of biological object recognition. *Frontiers in Psychology*, *8*, 1551.

10.3389/fpsyg.2017.01551

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929–1958.

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*. https://arxiv.org/abs/1312.6199

Treisman, A. (1996). The binding problem. *Current opinion in neurobiology*, *6*(2), 171–178.

10.1016/S0959-4388(96)80070-5

Vecera, S.P. (2000). Toward a biased competition account of object-based segregation and attention. *Brain and Mind*, *1*(3), 353–384.

10.1023/A:1011565623996

Vogels, R., & Biederman, I. (2002). Effects of illumination intensity and direction on object coding in macaque inferior temporal cortex. *Cerebral Cortex*, *12*(7), 756–766.

10.1093/cercor/12.7.756

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*, *138*(6), 1172.

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., ... Change Loy, C. (2018). Esrgan: Enhanced super-resolution generative adversarial networks. *Proceedings of the European conference on computer vision (ECCV) workshops* (pp. 0–0).

Wolfe, J.M. (1994). Guided search 2.0 a revised model of visual search. *Psychonomic bulletin & review*, *1*(2), 202–238.

10.3758/BF03200774

Wyatte, D., Jilk, D.J., O'Reilly, R.C. (2014). Early recurrent feedback facilitates visual object recognition under challenging conditions. *Frontiers in psychology*, *5*, 674.

10.3389/fpsyg.2014.00674

Xing, J., Chrastil, E.R., Nitz, D.A., Krichmar, J.L. (2022, November). Linking global top-down views to first-person views in the brain. *Proceedings of the National Academy of Sciences*, *119*(45), e2202024119.

10.1073/pnas.2202024119

Xu, Y. (2018, November). The Posterior Parietal Cortex in Adaptive Visual Processing. *Trends in Neurosciences*, *41*(11), 806–822.

10.1016/j.tins.2018.07.012

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

10.1126/sciadv.aax5979

Yu, C.-P., Maxfield, J.T., Zelinsky, G.J. (2016). Searching for category-consistent features: A computational approach to understanding visual category representation. *Psychological science*, *27*(6), 870–884.

10.1177/0956797616640237

Yuille, A., & Kersten, D. (2006, July). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences*, *10*(7), 301–308.

10.1016/j.tics.2006.05.002

Zelinsky, G.J. (2008). A theory of eye movements during target acquisition. *Psychological review*, *115*(4), 787.

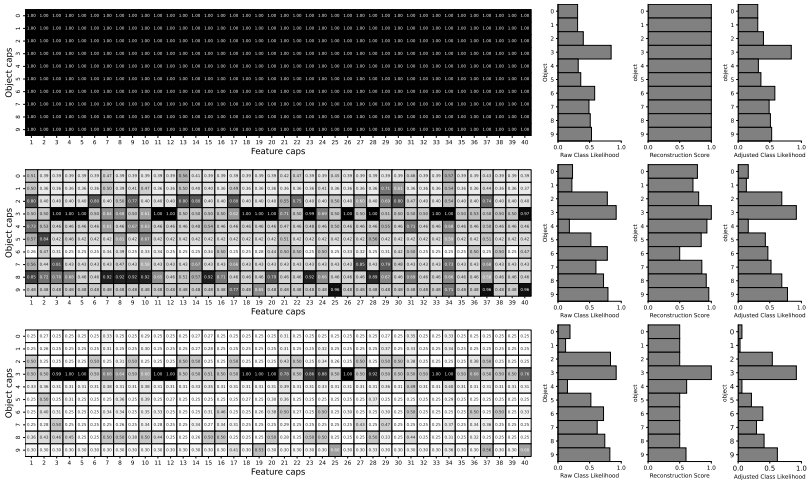
10.1037/a0013118

Extended Data A Percentage of changes in ORA's class prediction during spatial masking and feature binding iterations

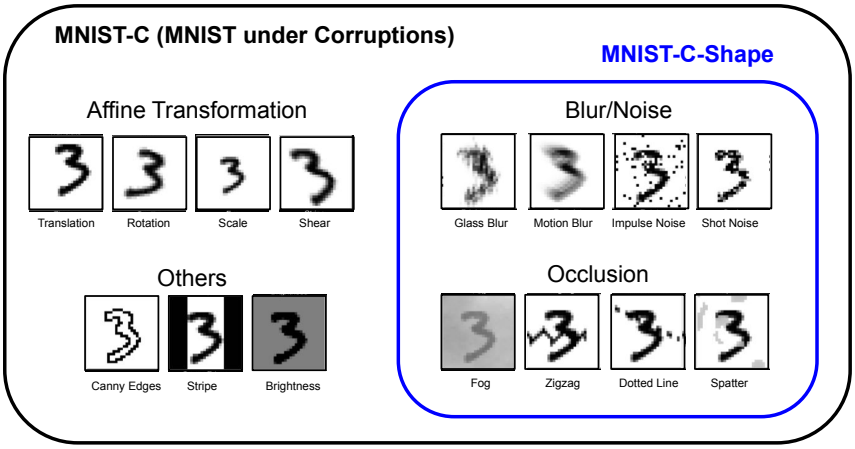
(%)	Spatial Masking Iteration			Feature Binding Iteration		
	False to True	False to False	% Recovery from Initial False Guess	False to True	False to False	% Recovery from Initial False Guess
Short Noise	0.20	1.40	12.50	0.12	0.05	71.15
Impulse Noise	3.40	0.70	82.93	0.57	0.34	62.77
Glass Blur	0.80	0.70	53.33	0.34	0.00	100.00
Motion Blur	0.60	0.70	46.15	0.13	0.05	72.72
Fog	40.10	0.20	99.50	0.73	1.19	38.01
Spatter	0.60	0.50	54.55	0.07	0.04	64.70
Dotted Line	0.80	0.40	66.67	0.17	0.02	89.65
Zigzag	3.30	0.40	89.19	0.44	0.08	84.56
Average	6.23	0.63	90.88	0.32	0.22	59.33

Extended Data B Step-wise visualization of reconstruction-guided feature binding process

The left column of the figure below shows binding coefficient matrices between all 10 of the possible digit categories (rows) and the first 40 feature slots (columns) at each of the 3 feature binding iterations. Binding coefficients are initially unweighted (step 1, top) but those object features leading to greater reconstruction error become suppressed (become lighter) over iterations, resulting in the feature binding coefficients focusing on the object features that best match the input (middle and bottom matrices). The nine bar plots in the right of the figure illustrate the effect of reconstruction-guided feature binding on ORA's digit classifications. The leftmost three plots show the raw likelihood of classification for each of the 10 object slots at the start of each time step (i.e., before the application of reconstruction-guided feature binding). The middle three plots show reconstruction scores (inverse of reconstruction error) at each step, and the rightmost three plots show the updated class likelihood scores after reconstruction-guided feature binding is applied (i.e., the two are multiplied). Best viewed with zoom in PDF.

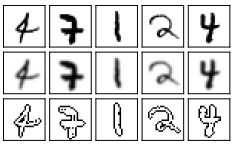


Extended Data C MNIST-C dataset examples



Extended Data D Results from manipulating the spatial frequencies used for image reconstruction

To test how reconstruction efficacy is effected by spatial frequency, we manipulated the spatial frequencies of the visual input that ORA used to generate its reconstructions. we trained different models to reconstruct different spatial frequency components of the input (e.g., generating a low spatial frequency reconstruction of the original full spectrum image) and then tested them to evaluate the effect of spatial frequency on model performance. We extracted different spatial frequency information from the input using a broadband Gaussian filter with cutoff frequencies below 6 cycles per image and above 30 cycles per image for generating low and high spatial frequency images, respectively. As shown below, we found that the use of a low-resolution object reconstruction yields comparable performance to a model trained with the full spectrum. The figure on the left shows samples of reconstructions from full-spectrum images (top) and images filtered for low (middle) and high (bottom) spatial frequencies. The table on the right shows the average performance of models trained to reconstruct either full-spectrum (FS), low-spatial frequency (LSF), or high-spatial frequency (HSF) images. Values in parentheses indicate standard deviations from 5 model runs.



	FS	LSF	HSF
MNIST-C	92.09 (0.28)	91.84 (0.67)	91.29 (0.49)
MNIST-C-shape	96.38 (0.45)	96.24 (0.74)	94.73 (0.67)

Extended Data E Model performance comparison

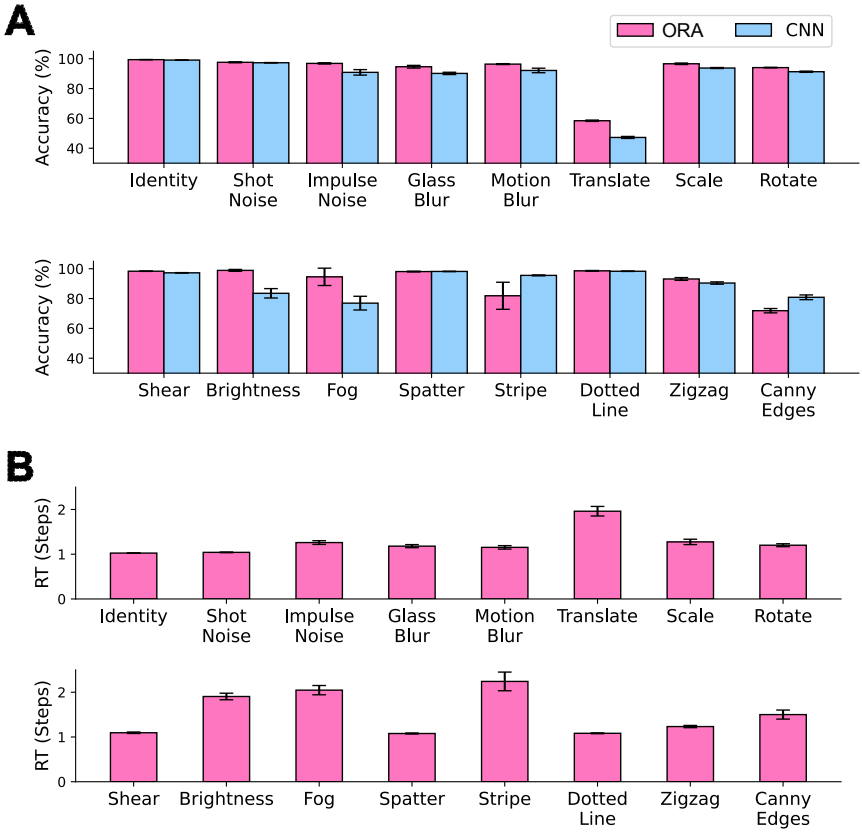
Table E1 Extended model comparison. Average performance and standard deviations from 5 model runs are reported for **ORA**, **CNN**, and **CapsNet** models that we trained for this study (see text for details). Best model performance for the other baselines are taken from Mu and Gilmer (2019).

Dataset	**ORA**		**CNN**		**CapsNet**	**ABS**	**PGD1**	**PGD2**	**GAN**
	Resnet-18	2 Conv	Resnet-18	2 Conv					
MNIST-C	**91.84 (0.67)**	88.32 (0.34)	81.94 (0.43)	88.94 (0.64)	75.14 (1.00)	82.46	80.06	78.86	81.14
MNIST-C-shape	**96.24 (0.74)**	92.82 (0.90)	86.43 (0.44)	91.80 (0.88)	81.87 (0.55)	88.69	83.43	81.38	83.53

Note: **ABS**: a generative model (Schott, Rauber, Bethge, & Brendel, 2018). **PGD1**: a CNN trained against PGD adversarial noise (Madry, Makelov, Schmidt, Tsipras, & Vladu, 2018). **PGD2/GAN**: another CNN trained against PGD/GAN adversarial noise (Wang et al., 2018).

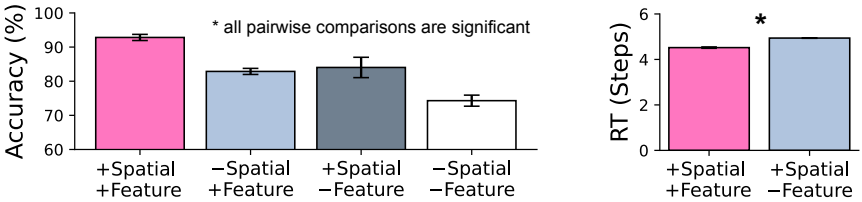
Extended Data F Model comparison on all MNIST-C corruption types

Average model accuracy and reaction time for all of the MNIST-C corruption types. A: ORA and CNN accuracy using each model's best performing encoder (Resnet-18 and a 2-layer CNN, respectively). B: ORA's reaction time (RT), estimated as the number of forward processing steps required to reach a confidence threshold. Error bars indicate the standard deviations from 5 different runs.



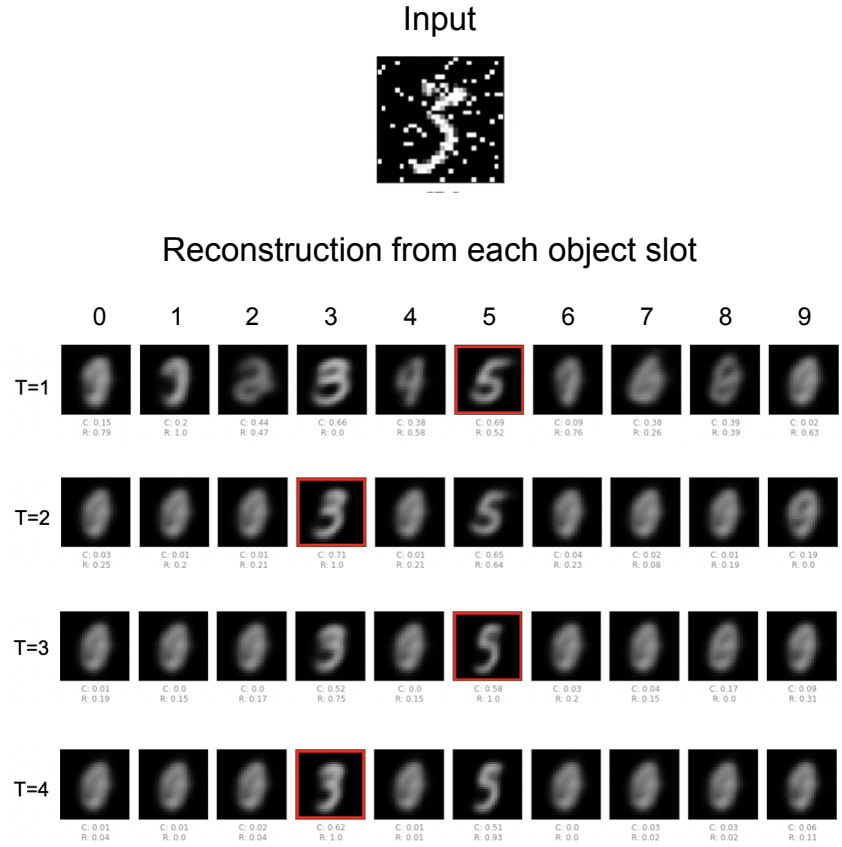
Extended Data G Ablation studies on ORA with 2 Convolutional encoder

Results from ablating ORA's reconstruction-based spatial masking and feature binding components. Ablated components are indicated by a - symbol, none ablated components are indicated by a +. Leftmost bars in pink indicate the complete (no ablation) ORA model. Left: Recognition accuracy. Right: reaction time, approximated by the number of forward processing steps taken to recognize digits with confidence. Error bars are standard deviations from 5 different runs.

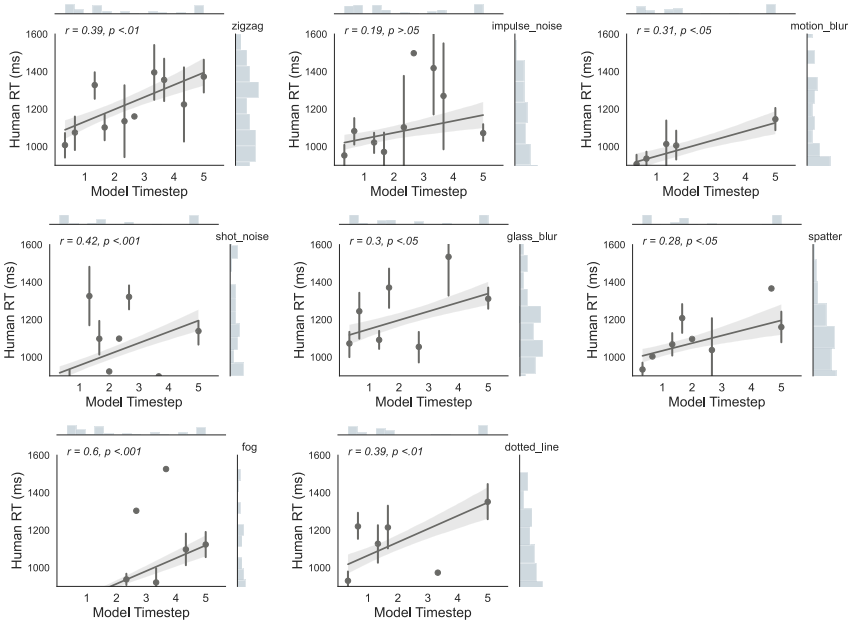


Extended Data H ORA's behavior on an ambiguous stimulus

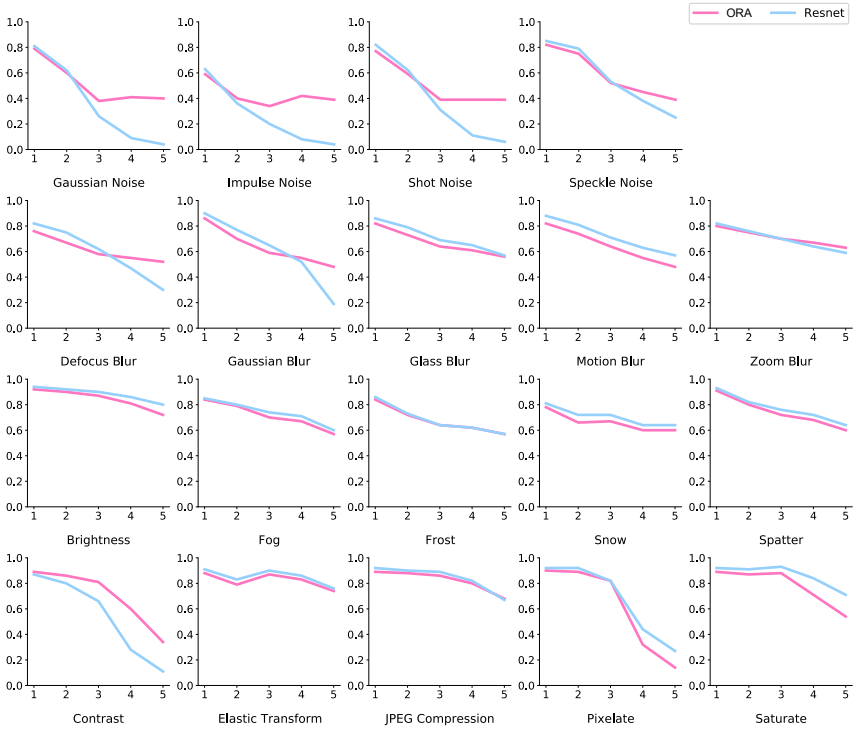
The figure below shows ORA taking longer to converge on a single digit hypothesis when the input is ambiguous. On the top is the input image, which could be plausibly interpreted as either the digit 3 or the digit 5. On the bottom are reconstructions hypothesized for all 10 digits, with the respective classification (C) and reconstruction (R) scores below each. ORA's most likely object hypothesis (with the highest classification score) at each step is indicated by the red square. Note that ORA's best guess about this object vacillates between the 3 and 5 hypotheses over time steps. The ground truth for this image is the digit 3.



Extended Data I Correlations between human and ORA reaction times for all MNIST-C corruption types



Extended Data J Examples of ORA reconstructions and spatial masks from Imagenet-C

Extended Data K Model comparison for all ImageNet-C corruption types

The x-axes show five level of noise severity and the y-axes show classification accuracy.