Supplementary Material
Putting visual object recognition in context

Mengmi Zhang[1,2], Claire Tseng[3], and Gabriel Kreiman[1,2]

{mengmi.zhang@childrens, ctseng@college, kreiman.gabriel@childrens}.harvard.edu
[1]Children's Hospital, Harvard Medical School
[2]Center for Brains, Minds and Machines
[3]Harvard College, Harvard University

S1. List of supplementary figures

S2. List of supplementary tables

1. Table 1 Correlation between different computational methods and humans — Table S1

S3. Recurrent LSTM module

The CATNet architecture is described in Section 4 in the main text (Fig. 3). One of the components is the recurrent LSTM module. We use a long short-term memory (LSTM) network to output a predicted class label y_t based on the previous hidden state $\mathbf{h_{t-1}}$ and the gist vector $\widehat{\mathbf{z}}_\mathbf{t}$ for the target object I^o and the contextual information I^c (Fig. S14). Our implementation of LSTM closely follows [12] where $T_{s,t} : \mathbb{R}^s \rightarrow \mathbb{R}^t$ defines a linear transformation with learnable parameters. The variables $\mathbf{i_t}, \mathbf{f_t}, \mathbf{c_t}, \mathbf{o_t}, \mathbf{h_t}$ represent the input, forget, memory, output and hidden state of the LSTM respectively.

$$\begin{pmatrix} \mathbf{i_t} \\ \mathbf{f_t} \\ \mathbf{o_t} \\ \mathbf{g_t} \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} T_{D+n,n} \left(\widehat{\mathbf{z}}_\mathbf{t}, \mathbf{h_{t-1}} \right) \tag{1}$$

$$\mathbf{c_t} = \mathbf{f_t} \bigodot \mathbf{c_{t-1}} + \mathbf{i_t} \bigodot \mathbf{g_t}, \quad \mathbf{h_t} = \mathbf{o_t} \bigodot \tanh(\mathbf{c_t}) \tag{2}$$

where n is the dimensionality of the LSTM, σ is the logistic sigmoid activation, and $\bigodot$ indicates element-wise multiplication. To cue CATNet about the location of the target object, we initialize the memory state $\mathbf{c_0}$ and hidden state $\mathbf{h_0}$ of the LSTM based on a binary mask that contains zeros everywhere and ones in the target object location. Specifically, $\mathbf{c_0}$ and $\mathbf{h_0}$ are computed by an average of all feature vectors a_0 over all L locations with two separate linear transformations $W_{c0} \in R^{n \times D}$ and $W_{h0} \in R^{n \times D}$:

$$\mathbf{c_0} = W_{c0}(\frac{1}{L}\sum_i^L \mathbf{a_{0i}}), \quad \mathbf{h_0} = W_{h0}(\frac{1}{L}\sum_i^L \mathbf{a_{0i}}) \tag{3}$$

CATNet uses attention-modulated features to reduce the input dimension to a fully connected LSTM. We conducted another ablation study with a convolutional LSTM after the feature extraction encoder. The average accuracy was > 5 % lower than CATNet.

S4. Implementation Details

The dimension of the LSTM module was $n = 512$. For both I^c and I^o, the feature maps extracted from the last convolution layer was of size $2048 \times 28 \times 28$, and the total number of locations was $L = 28 \times 28 = 784$. The Adam optimizer [4] was used with a learning rate of 10^{-4} to fine-tune the VGG16 network, and a learning rate of 4×10^{-4} to train the attentional and the LSTM module. The network was developed in Pytorch, based on [11].

S5. Comparative Methods and Ablated Models

DeepLab-Conditional Random Field (CRF). An interesting solution to reason about the target object based on context is to run state-of-the-art semantic segmentation algorithms and use majority voting on the predicted labels over all pixels in the bounding box. We used the instantiation in DeepLab-CRF [1], which semantically segments regions using deep nets and uses conditional random field (CRF) for refinement. Figure S16 - Figure S22 compare CATNet and DeepLab results.

YOLO3. Similar to semantic segmentation, we adapt the object detection algorithms in context-aware object recognition tasks by majority voting on the predicted labels over all pixels in the bounding box of the target object. We used the instantiation in YOLO3 [6, 7], which is s highly competitive object detection method in computer vision. Figure S23 - Figure S29 compare CATNet and YOLO3 results.

Since most of these comparative models lack a sense of time, we cannot test them on Experiment C, which requires different exposure times. Therefore, comparisons are restricted to Experiments A and B.

VGG16 on cropped objects. To assess recognition on objects alone, we fine-tuned VGG16 [8] pre-trained on ImageNet using the training set containing 55 categories of MSCOCO and tested it in all the experiments. This includes stimulus sets with zero context (context object ratio CO = 0) in Exp A2. This comparison helps us gauge a lower bound of in-context object recognition in typical feed-forward networks relying exclusively on the target object information. Figure S30 - Figure S36 compare CATNet and VGG16 results.

VGG16 + binary mask. An intuitive way of solving context-aware object recognition problems is to use a feed-forward object recognition network pre-trained on ImageNet, e.g., VGG16 [8], and fine-tune it to classify the target object on MSCOCO dataset. During training, the input to the network was an image where one object on the image was randomly selected as the target object. To indicate the target object location, we concatenate the natural image of RGB channels with a binary mask as an additional channel to the network with 1 denoting the target location and 0 otherwise. Figure S37 - Figure S43 compare CATNet and VGG16+BinaryMask results.

Two-stream VGG16. Following the work in reference [10], we consider a two-stream VGG16 network that takes two inputs, the target object I^o and the context I^c, concatenates their activations from the second last fully connected layers of two separate pre-trained VGG16 networks and outputs a predicted class label. The two-stream network is fine-tuned using the same training set as in the other models. Figure S51 - Figure S57 compare CATNet and Two-stream VGG16 results.

VGG16 + attention. Previous work has demonstrated the efficiency of attention in many computer vision tasks [5], such as question answering and image captioning [11]. To study the effect of attention in object recognition, we added an attention module to the end of the two-stream VGG16 network above and produced attention-modulated features maps on I^o and I^c. To make the complexity of the architecture comparable with CATNet, we added the same number of fully connected layers as in the LSTM module. Figure S44 - Figure S50 compare CATNet and VGG16+Attention results.

VGG16 + attention + LSTM. Recent studies have argued that recurrent connections are useful for object recognition (e.g., [9]). We extended the **VGG16 + attention** model with an LSTM module (Section S3) which integrates attention-modulated feature maps over time and outputs predicted class label at every time step. Making predictions at every time step allows us to evaluate the temporal dynamics of contextual modulation in the three variations of Exp C and compare the results against human performance. These experiments evaluate the duration of object exposure, context exposure, the degree of synchrony between context and object presentation, and the effects of backward masking. Figure S58 - Figure S64 compare CATNet and VGG16+Attention+LSTM results.

S6. Implementation Details of Human Psychophysics Experiment

S6.1. Ground Truth Responses

Many visual recognition experiments are based on forced N-way categorization (e.g., [9]), where subjects have to choose among N possible answers. Here we introduced a different probing mechanism whereby there were no constraints on what words subjects could use to describe the target object in each trial (Fig. 2h). This evaluation procedure was introduced for two reasons: (i) it is difficult for humans to memorize 55 object classes and such a large number of classes could lead to non-uniform memory effects (certain labels remembered better than others); (ii) we were concerned that presenting specific choices could induce priming effects and introduce biases in the inference and decision processes.

We could not simply rely on the 55 category labels to evaluate behavioral performance because subjects could use other similar words or synonyms. We were interested in the context reasoning process rather than the subjects' language abilities. Therefore, to evaluate performance, we separately collected a distribution of ground truth answers for each target object. This ground truth distribution was created by presenting the same images with full context with the target object highlighted by a bounding box, without any time constraints, to 10 *other* subjects who did not participate in any of the 10 main experiments. In the 10 main experiments, a response was considered to be correct if it matched *any* of the ground truth labels, allowing for plurals and single letter misspellings.

We evaluated computational models (Sec. 4) on the same stimuli used in the behavioral experiments. Throughout the main text and supplementary figures, we report **top-1 classification accuracy**. Because of the behavioral probing mechanism described above, the computational results and human performance are not directly comparable. Chance level is $1/55$ for the computational models, whereas chance levels are not clearly defined for the behavioral experiments. One could try to define chance for the human observers by drawing nouns randomly according to their typical English usage, which would lead to levels well below $1/55$. We still find it instructive to plot computational results alongside human behavior for comparison purposes. Furthermore, relative changes and trends in humans can be directly compared to computational results. In Table 1 and Table S1, we report the correlation between human and model results, which is independent of the potential differences in absolute accuracy values due to the different evaluation methods.

S6.2. Evaluation Metrics and Statistics

To compare two sets of results (e.g. minimal context vs. full context, or humans vs. CATNet), we used a nonparametric Wilcoxon ranksum test [2]. We report sample size n_1, second sample size n_2, and p-value p. The null hypothesis is that the two population medians are equal.

When considering multiple groups (e.g. different object sizes, we used either a one-way or two-way ANOVA test [3]. The ANOVA test compares the variation in accuracy *between* versus *within* conditions (called the *F-ratio*). We report $F(a, b)$ where a and b are the degrees of freedom in the numerator and denominator of the F ratio distribution, and we also report the corresponding p-value p.

Correlation	expA1	expA2	expB1	expB2	expB3	expB4	expB5
CATNET	0.89	**0.89**	**0.95**	0.87	0.89	**0.92**	**0.93**
DeepLab	**0.90**	0.83	0.86	**0.88**	**0.90**	0.81	0.91
YOLO3	0.75	0.78	0.74	0.78	0.75	0.66	0.87
VGG16	0.76	0.47	0.63	0.73	0.74	0.71	0.71
VGG16+BinaryMask	0.69	0.69	0.78	0.75	0.58	0.74	0.75
VGG16+Attention	0.88	0.81	0.90	0.88	0.88	0.87	0.93
Two-stream VGG16	0.84	0.77	0.87	0.73	0.81	0.84	0.84
VGG16+Attention+LSTM	0.84	0.84	0.87	0.74	0.84	0.80	0.86

Table S1. Correlations between humans and models for Exp A and B. See Section 3 for definitions of evaluation metrics and Section S5 for a description of the different models. Best is in bold.

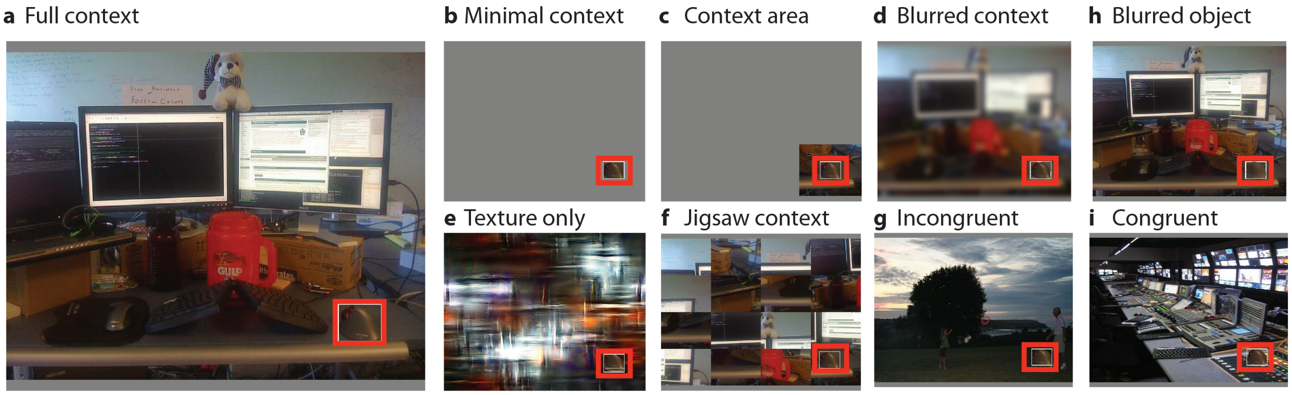

Figure S1. **Fundamental properties of context**. Example image with full context (**a**) and image modifications used in experiments. The target location (red box) is always the same across conditions for a given image (but subjects only see one of these versions). The full context (**a**) and minimal context (**b**) conditions are included in all the experiments for comparison purposes. Exp A2 titrates the amount of context (**c**). Exp B1 and B2 examine the impact of blurring the context (**d**) or the object (**h**). Exp B3 evaluates whether low-level texture properties of the context are sufficient (**e**). Exp B4 studies the geometrical properties of context (**f**). Exp B5 considers congruent (**i**) and incongruent (**g**) contextual information.

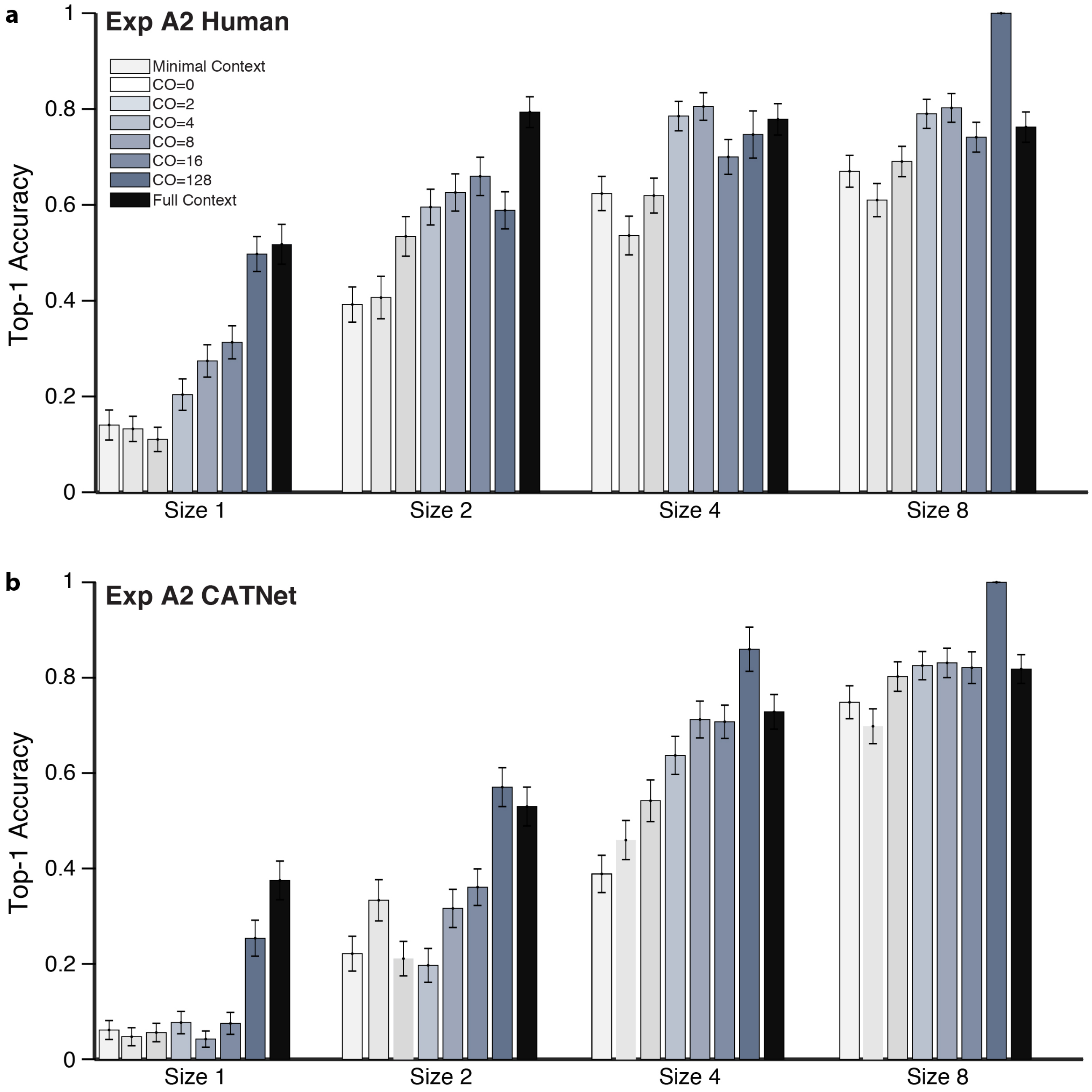

Figure S2. **Recognition performance improves with the amount of context both for humans (a) and CATANet (b) (Exp A2)**. This is an expansion of the discussion in Section 5.1 in the main text. The shade of gray denotes the context-object ratio (CO). The larger the CO ratio, the more contextual information. The error bar in this figure and subsequent figures denote SEM.

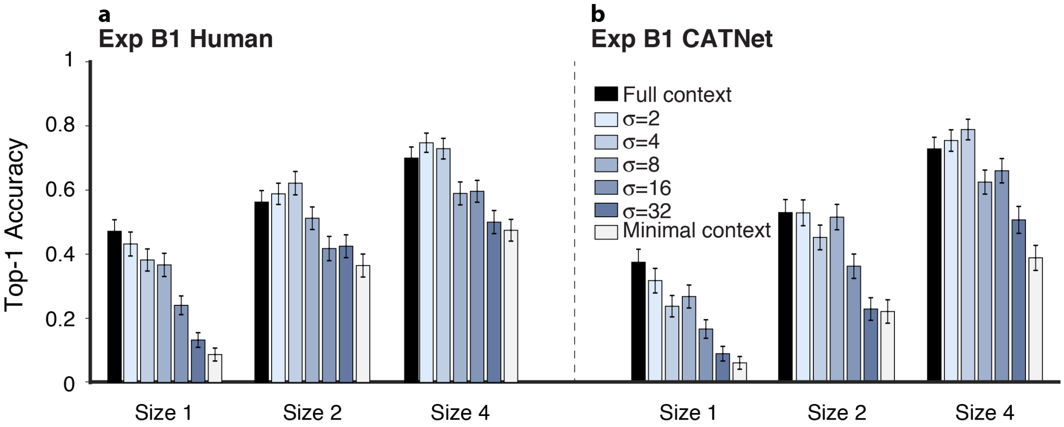

Figure S3. **Contextual facilitation persists even after small amounts of blurring (Exp B1)**. The shade of gray indicates the standard deviation of the gaussian blur applied to the context (the larger σ, the more blurring). A large amount of context blurring (Fig. 2d) is required to disrupt the recognition enhancement for humans and CATNet. This is an expansion of the discussion in Section 5.2.1 and Fig. 5 in the main text.

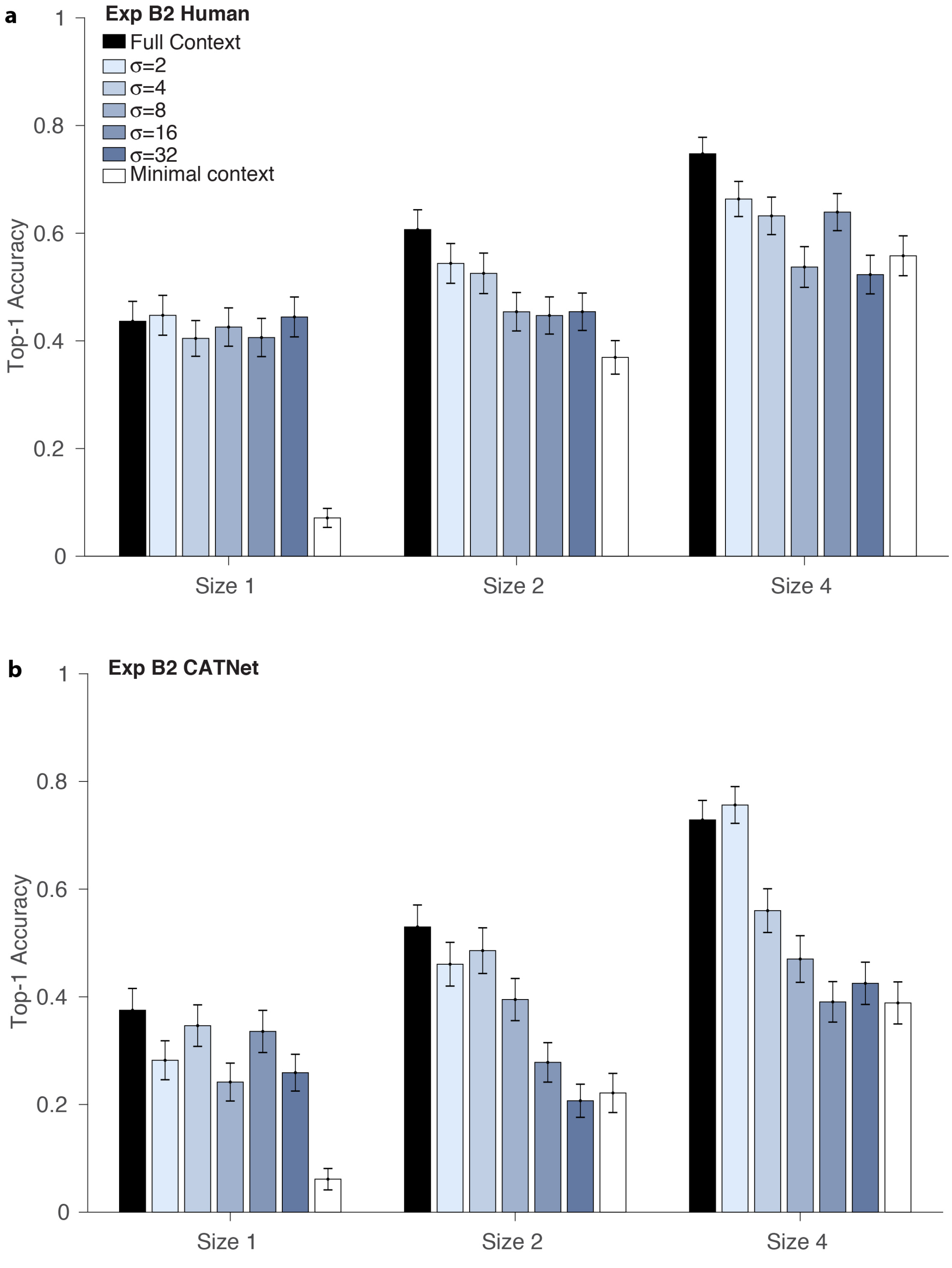

Figure S4. **Results for humans (a) and CATNet (b) when objects are blurred (Exp B2)**. The shade of gray indicates the standard deviation of the gaussian blur applied to the objects but *not* to the context (in contrast to Fig. S3). This is an expansion of the discussion in Section 5.2.1 in the main text.

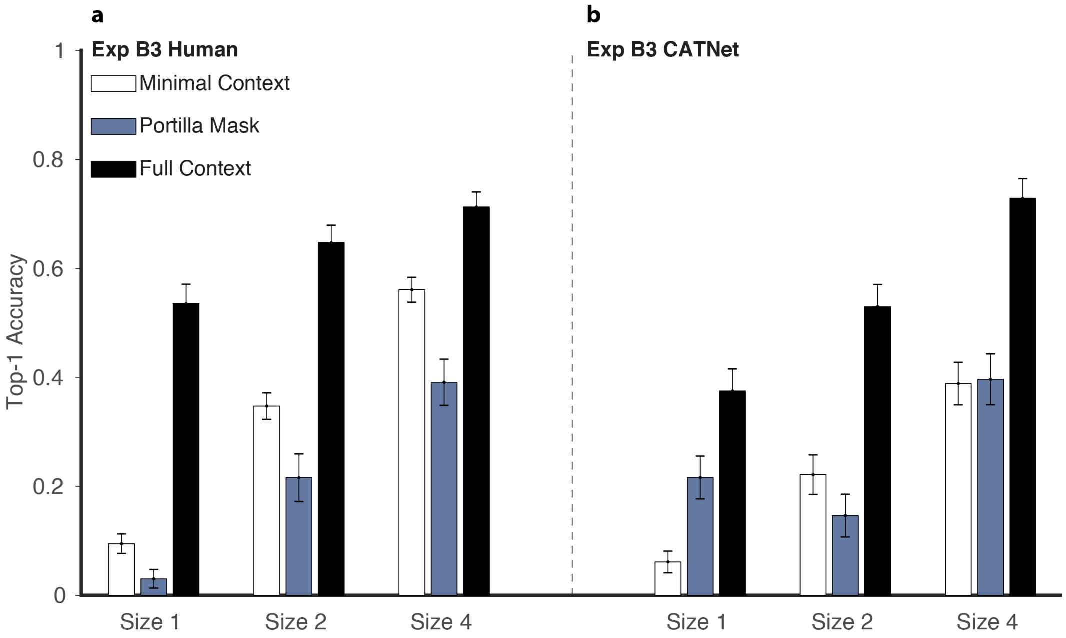

Figure S5. **Low-level contextual information matching the contextual texture does not yield contextual enhancement (Exp B3)**. In most cases, adding the texture-only context actually impaired recognition with respect to the minimal context condition. This is an expansion of the discussion in Section 5.2.2 in the main text.

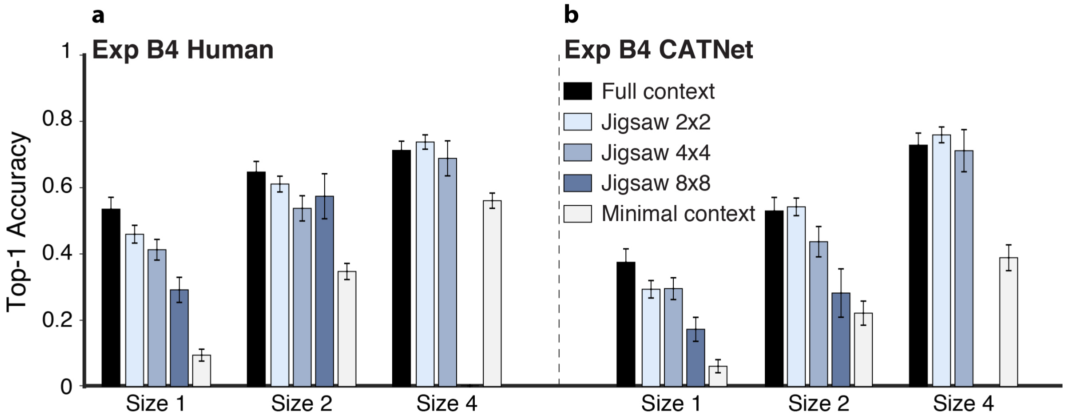

Figure S6. **Large geometrical context re-arrangements disrupt contextual enhancement (Exp B4)**. Scrambling context pieces (Fig. 2f) reduces the contextual enhancement only when many small context pieces are changed, both for humans (**a**) and CATNet (**b**).

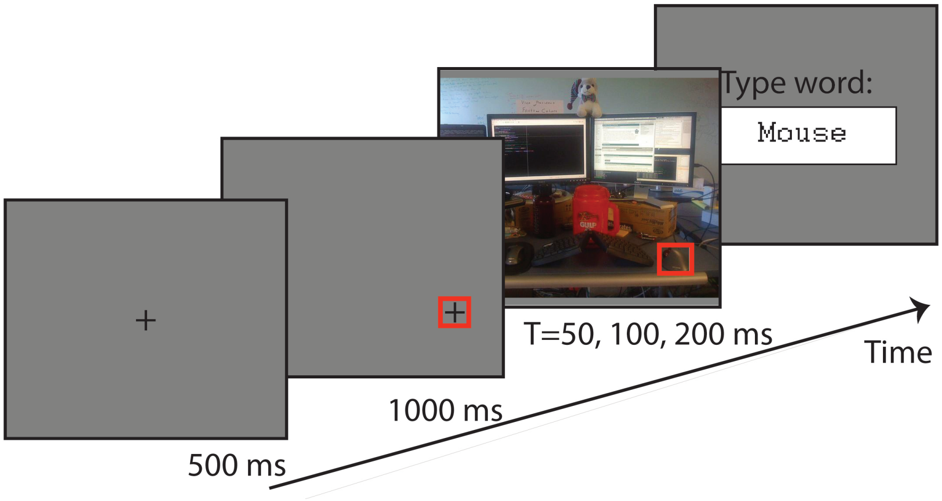

Figure S7. **Task schematic for Exp C1 (context exposure time)**. Subjects were presented with a fixation cross (500 ms), followed by a bounding box indicating the target object location (1000 ms). The target location (red box) was always the same across conditions for a given image but subjects only saw one condition for a given image. Expanding on Fig. 2h, in Exp C1, the image was shown for T ms where T could vary from 50, 100, 200 ms ($T = 200$ ms was the exposure time in Exp A and Exp B). After image offset, subjects typed one word to identify the target object. The correct answer (in this case: “mouse”) was not shown in the actual experiment.

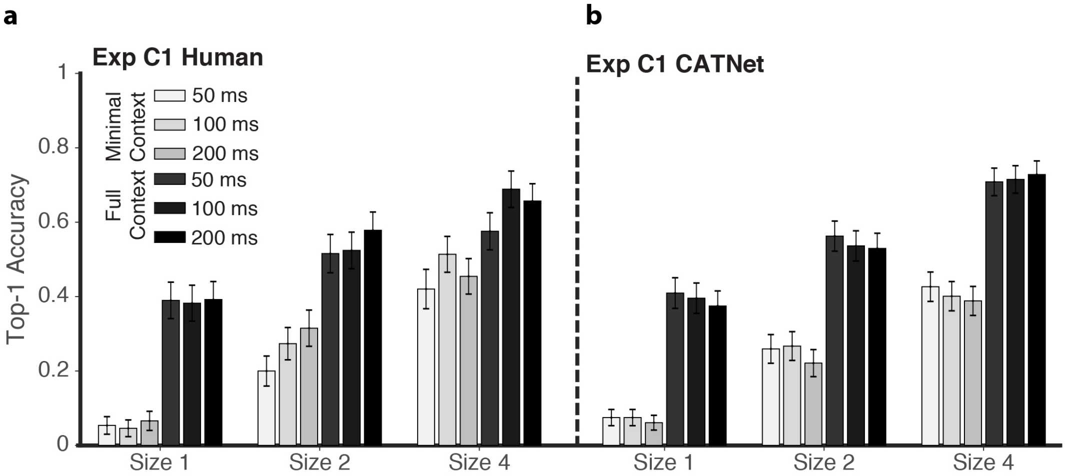

Figure S8. **Stimulus exposure time has little effect on recognition (Exp C1).** Exposure time was varied from 50 to 200 ms denoted by different shades of gray (Fig. S7). Exposure time of 50 ms is sufficient to get the "gist" of context. See Section 5.3.1 in the main text for details.

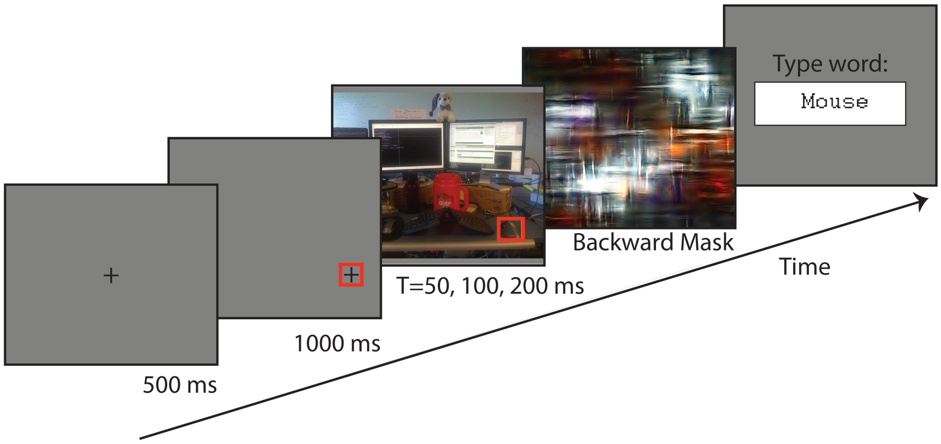

Figure S9. **Task schematic for Exp C2 (Backward masking)**. Subjects were presented with a fixation cross (500 ms), followed by a bounding box indicating the target object location (1000 ms). The target location (red box) was always the same across conditions for a given image but each subject only saw one of the conditions. Expanding on Fig. 2h, in Exp C2, the image was shown for T ms where T could be 50, 100, or 200 ms. In half of the trials (randomly), the image was followed by a backward mask for 500 ms.

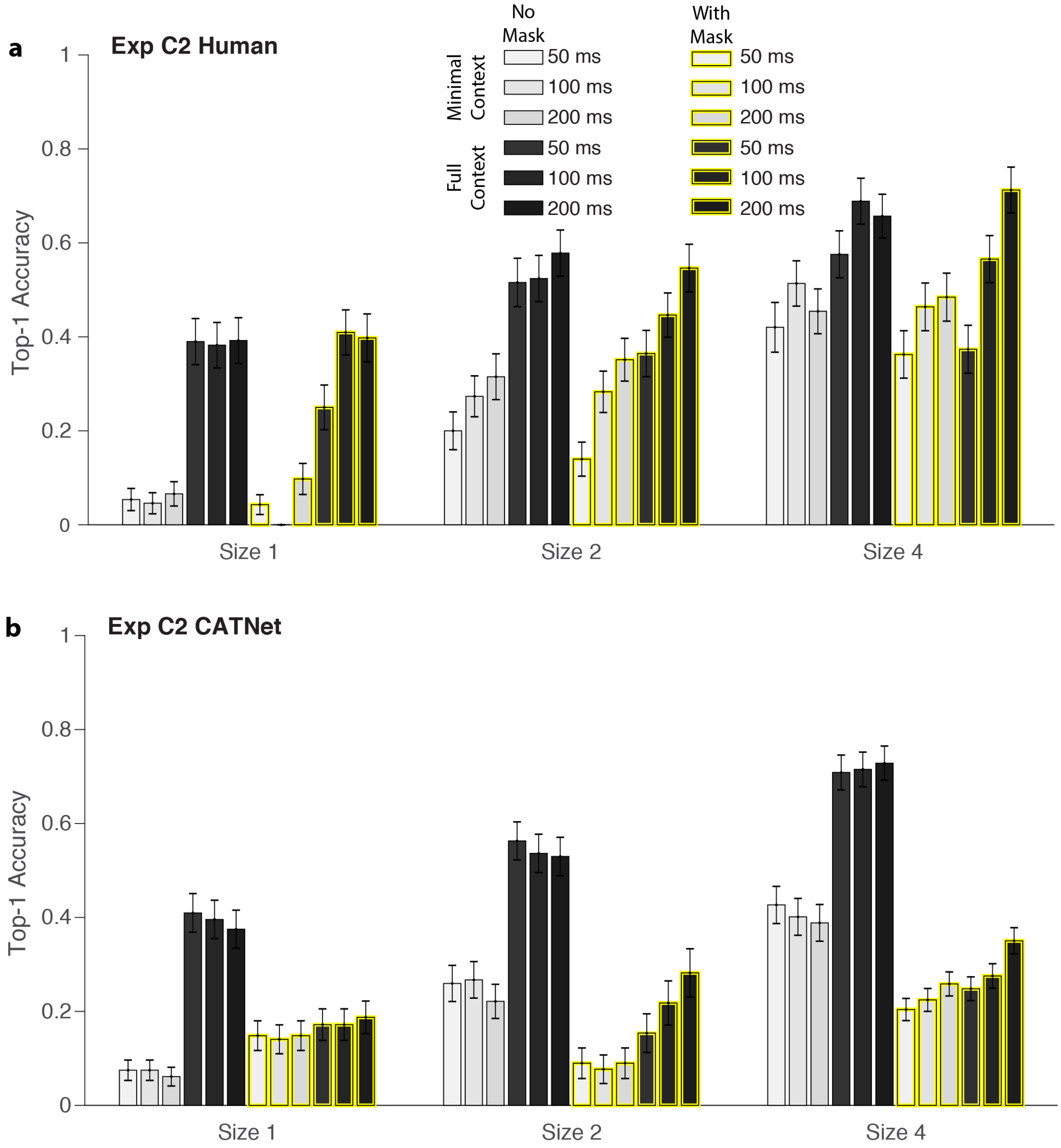

Figure S10. **Backward masking decreases contextual modulation effects (Exp C2)**. Conditions with backward masking (yellow borders) showed lower performance than the no mask counterparts. This is an expansion of the discussion in Section 5.3.2 in the main text.

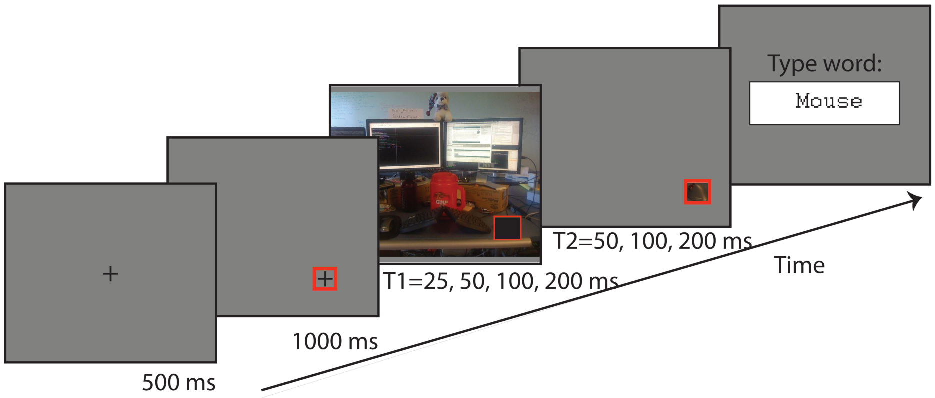

Figure S11. **Task schematic for Exp C3 (Asynchronous version of context)**. Subjects were presented with a fixation cross (500 ms), followed by a bounding box indicating the target object location (1000 ms). The target location (red box) is always the same across conditions for the same picture but each subject only saw one condition. Expanding on Fig. 2h, in Exp C3, the image containing *only full context without the object* was first shown for $T1$ ms, where $T1$ could be 25, 50, 100, or 200 ms. Next, the image containing *only the target object with minimal context* (Fig. S1b) was shown for $T2$ ms, where $T2$ could be 50, 100, or 200 ms.

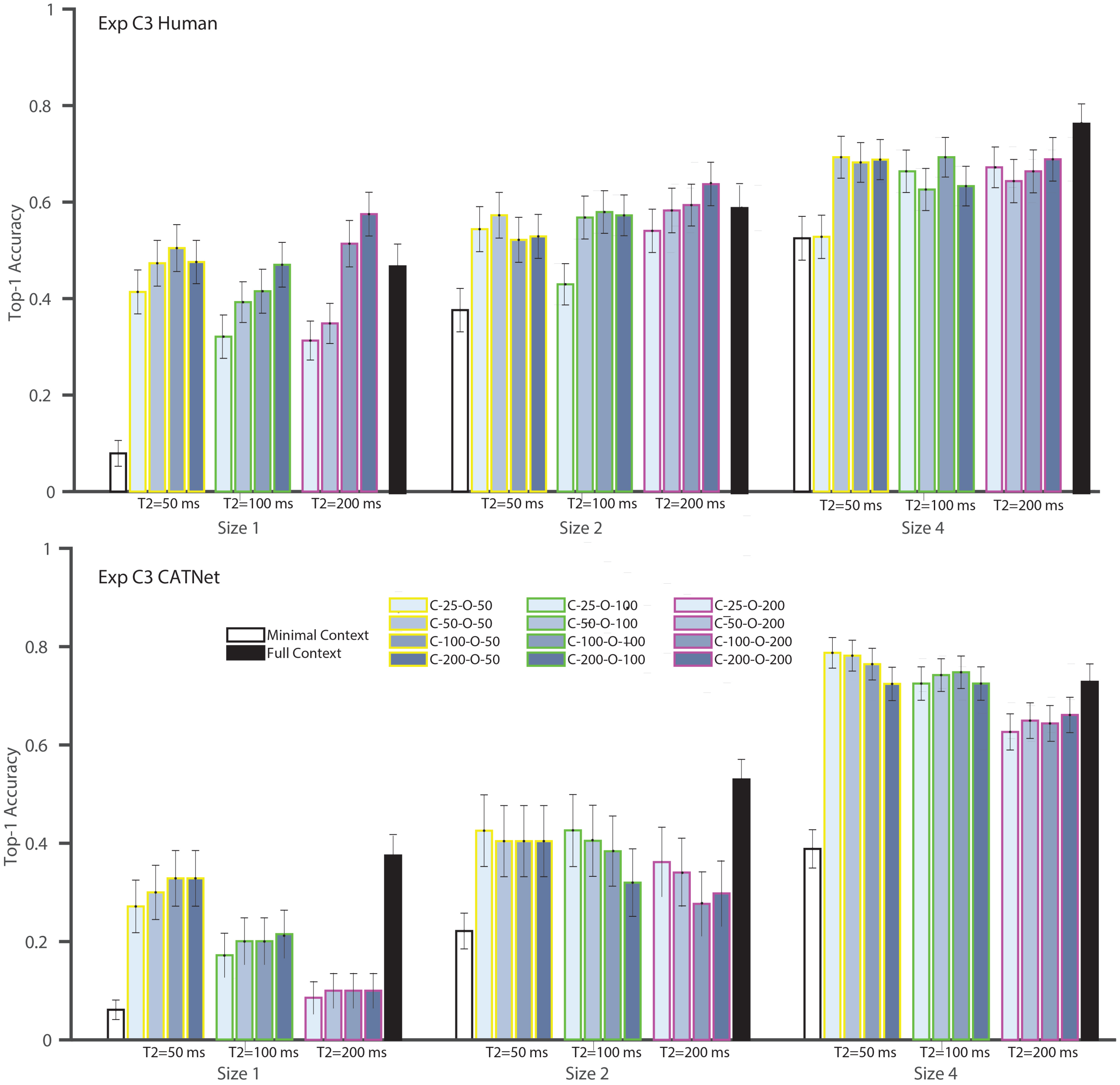

Figure S12. **Presenting the context before the object is sufficient to elicit facilitation.** Results for humans (top) and CATNet (bottom) on Exp C3 where the context is presented before the object. Bars with yellow (green — black) borders have a fixed object duration $T2$ of 50 ms (100 ms — 200 ms). The context duration $T1$ ranges from 25 ms to 200 ms. This is an expansion of the discussion in Section 5.3.3 in the main text.

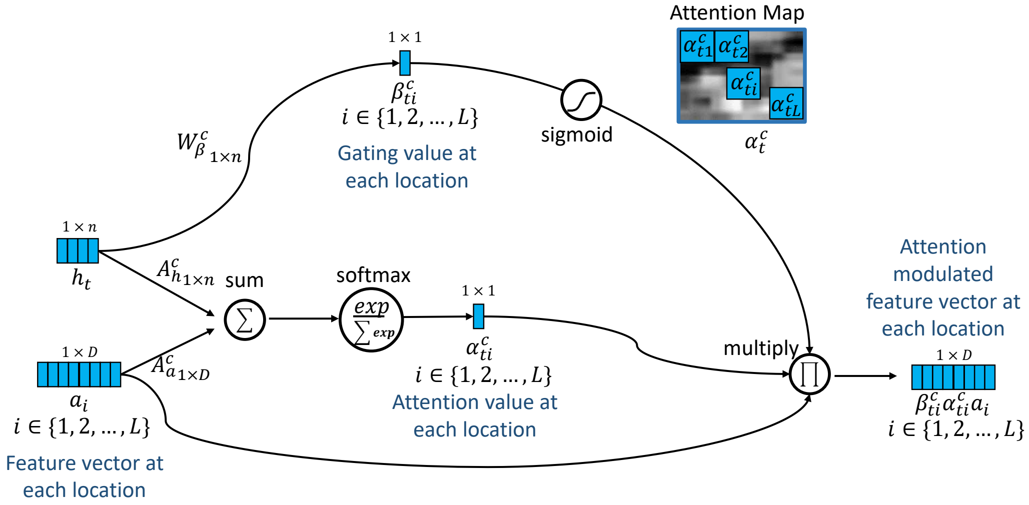
Attention Map
α_{t1}^{c}
α_{t2}^{c}
α_{ti}^{c}
α_{tL}^{c}
α_t^c
1×1
β_{ti}^{c}
$i \in \{1, 2, \dots, L\}$
Gating value at each location
sigmoid
$W_{\beta\,1\times n}^{c}$
$1 \times n$
h_t
$A_{h_{1\times n}}^{c}$
sum
$\sum$
softmax
$\frac{exp}{\sum exp}$
1×1
α_{ti}^{c}
$i \in \{1, 2, \dots, L\}$
Attention value at each location
multiply
$\prod$
Attention modulated feature vector at each location
$1 \times D$
$\beta_{ti}^{c}\alpha_{ti}^{c}a_i$
$i \in \{1, 2, \dots, L\}$
$1 \times D$
a_i
$i \in \{1, 2, \dots, L\}$
$A_{a_{1\times D}}^{c}$
Feature vector at each location

Figure S13. **Schematic illustration of the attention module implementation.** Expanding on the overall CATNet architecture shown in Fig. 3, here we zoom into the attention module (Section 4.2). The attention module takes as inputs the features at each location a_{ti} and the output of the LSTM module h_t, and selects the next attention location m_t and a map that modulates the features at each location (see Section 4 in the main text for a description of all the variables).

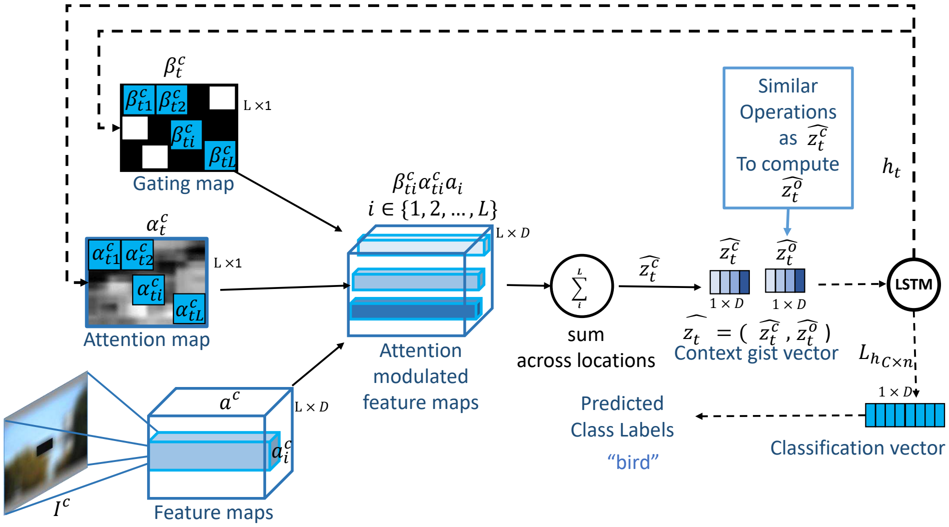

Figure S14. **Schematic illustration of the LSTM module implementation**. Expanding on the overall CATNet architecture shown in Fig. 3, here we zoom into the LSTM module (Section 4.3). The LSTM module takes as input the context gist vector $\widehat{\mathbf{z}}_\mathbf{t}$, and integrates the information with the previous state to inform the attention module in the next time step via h_t, and to predict a class label (see Section 4 for a description of all the variables).

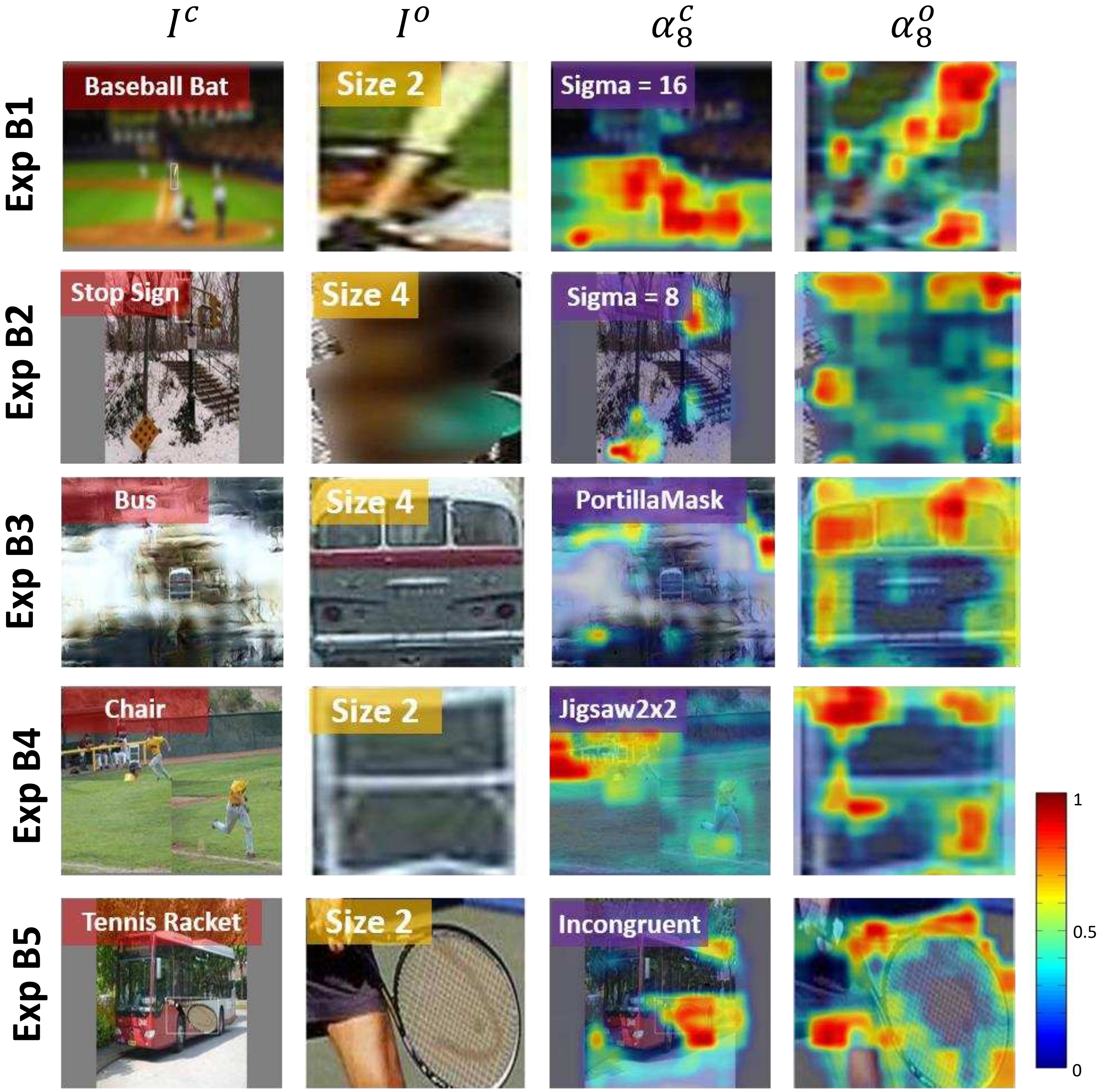

Figure S15. **Example visualization for CATNet at time step $T = 8$ (approximating 200 ms exposure time) under different context conditions for Exp B**. Example trials (one row per experiment) for Exp B1 (1st row), B2 (2nd row), B3 (3rd row), B4 (4th row), and B5 (5th row), In each trial, the ground truth label of the target object is highlighted in red (column 1), the object size is highlighted in yellow (column 2) and different experimental conditions are highlighted in purple (column 3). The first column of each trial shows the stimulus I^c. The second column of each trial shows the target object I^o. Predicted attention maps α^c and α^o by two-stream architecture are overlaid on I^c and I^o respectively. The colors on the attention map denote different levels of attentional values (see colorbar on the right).

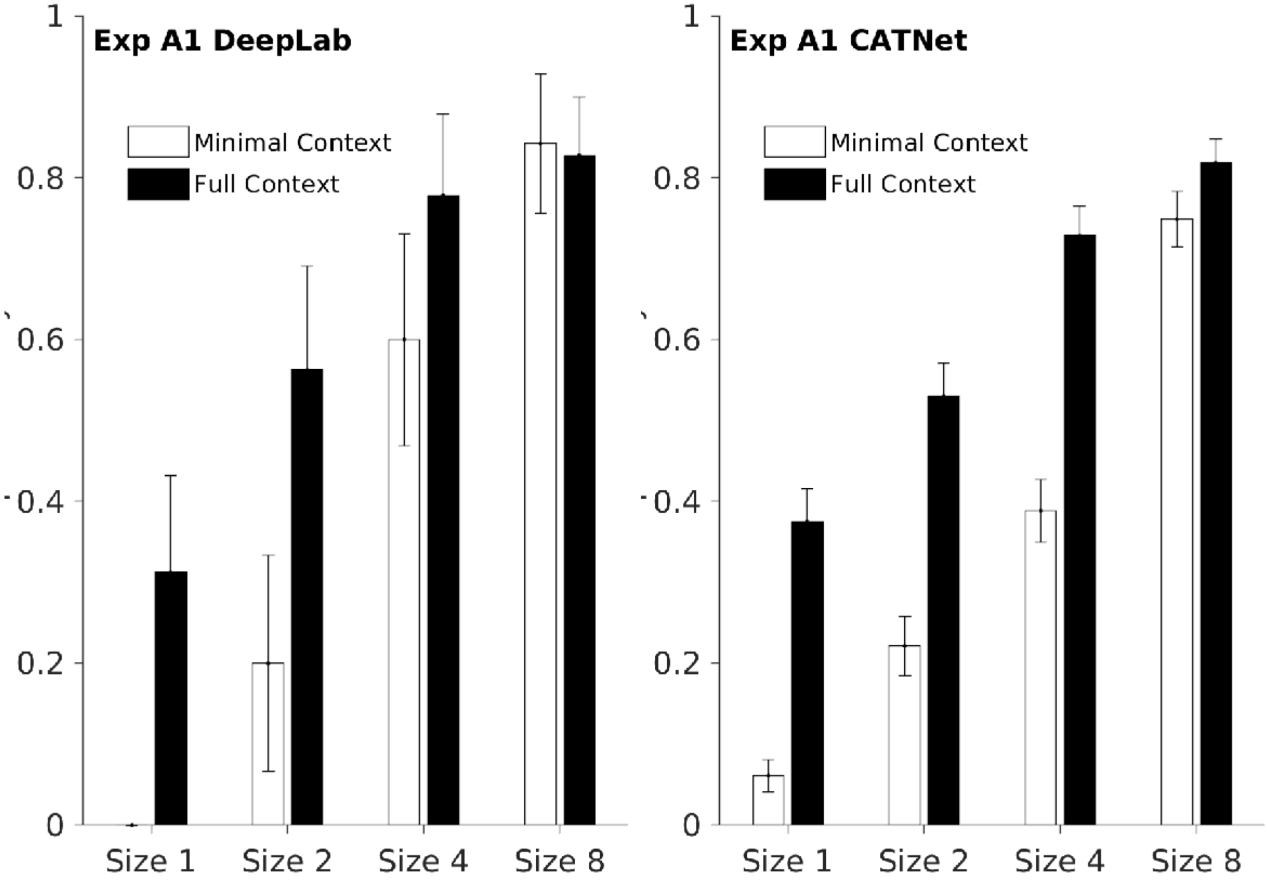

Figure S16. **Results on object size in Exp A1 between DeepLab and CATNet**. Expanding on the discussion in Sec 5.1 and Fig. 4, contextual modulation is stronger for smaller target objects for DeepLab and CATNet. CATNet outperforms DeepLab for smaller target objects.

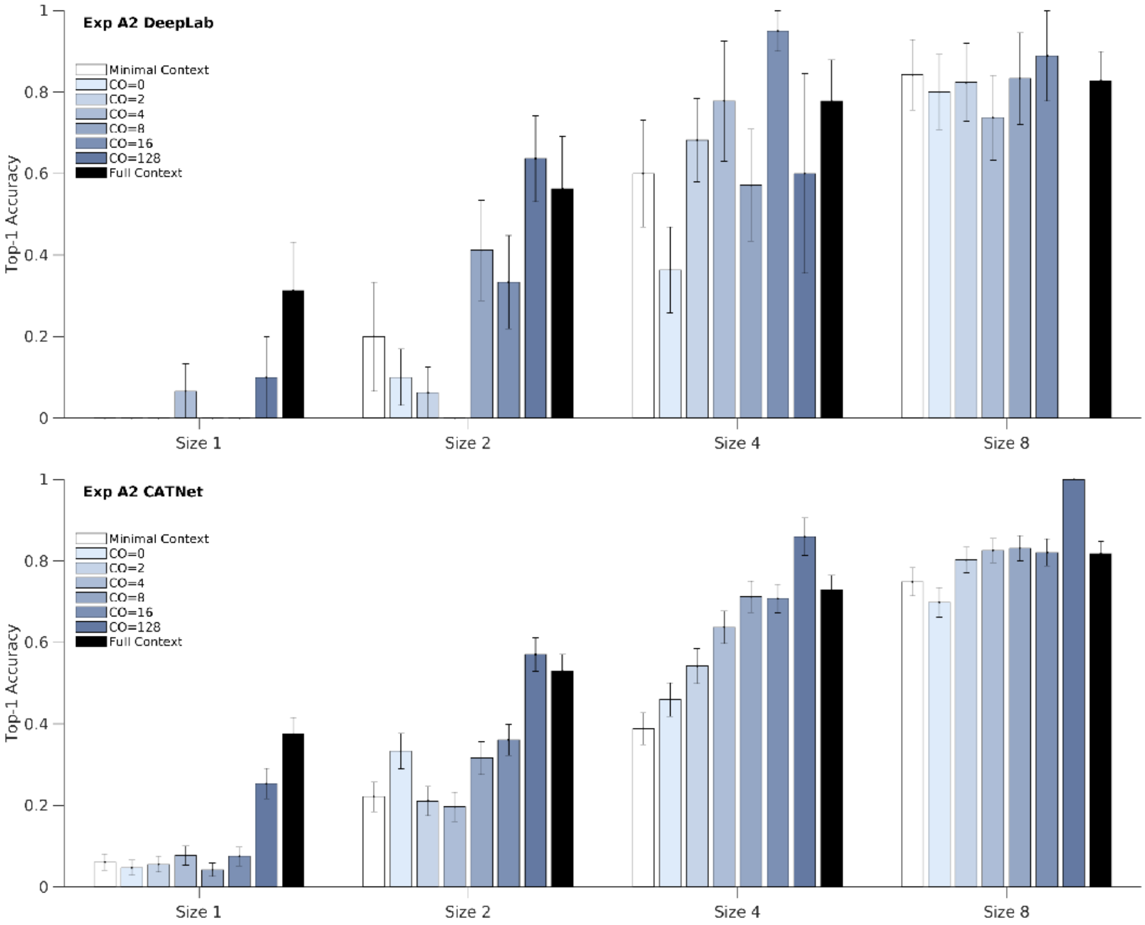

Figure S17. **Results on amount of context in Exp A2 between DeepLab and CATNet**. Expanding on the discussion in Sec 5.1 and Fig S2, CATNet outperforms DeepLab for smaller objects.

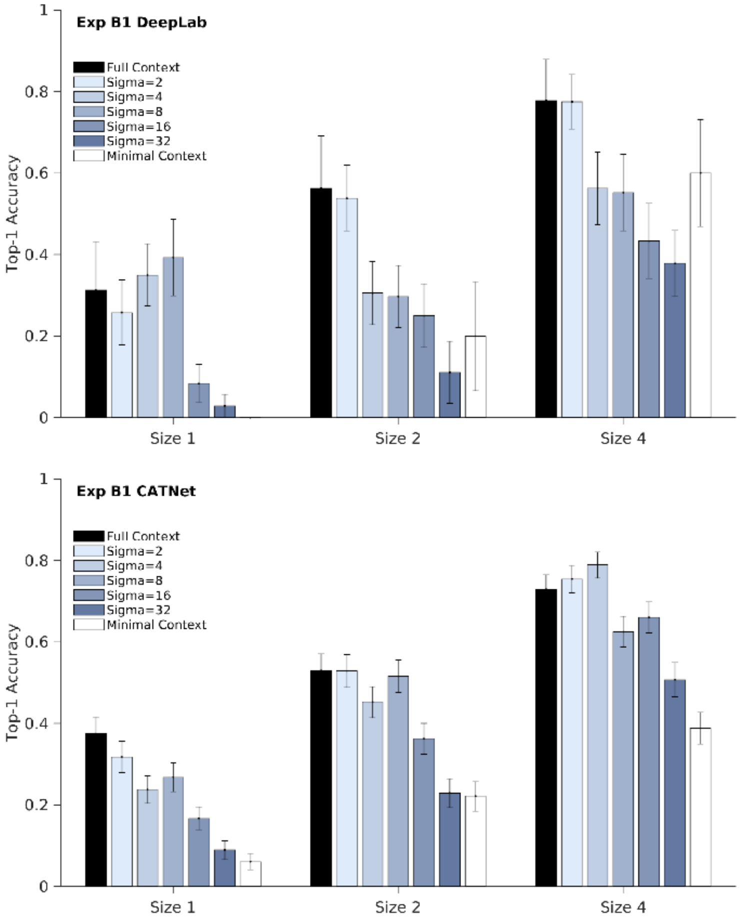

Figure S18. **Results on blurred context in Exp B1 between DeepLab and CATNet.** Expanding on the discussion in Sec 5.2 and Fig 5, contextual facilitation persists even after small amounts of blurring (Exp B1). CATNet outperforms DeepLab for smaller objects.

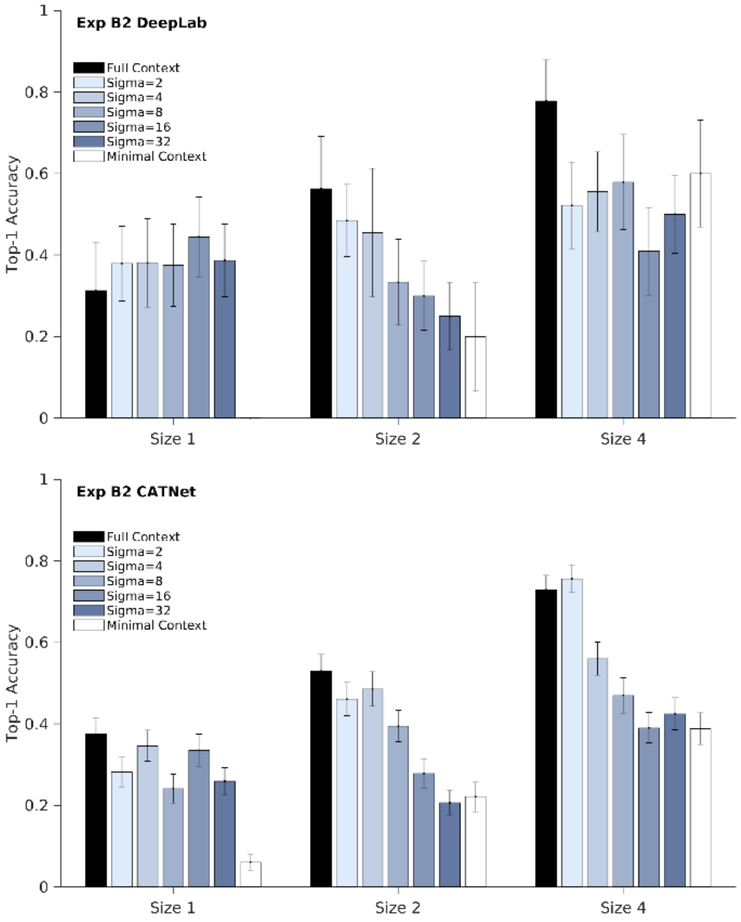

Figure S19. **Results on blurred objects in Exp B2 between DeepLab and CATNet.** Expanding on the discussion in Sec 5.2 and Fig S4, compared with context blurring, modifying the object led to larger accuracy drops. CATNet performs equivalently well as DeepLab.

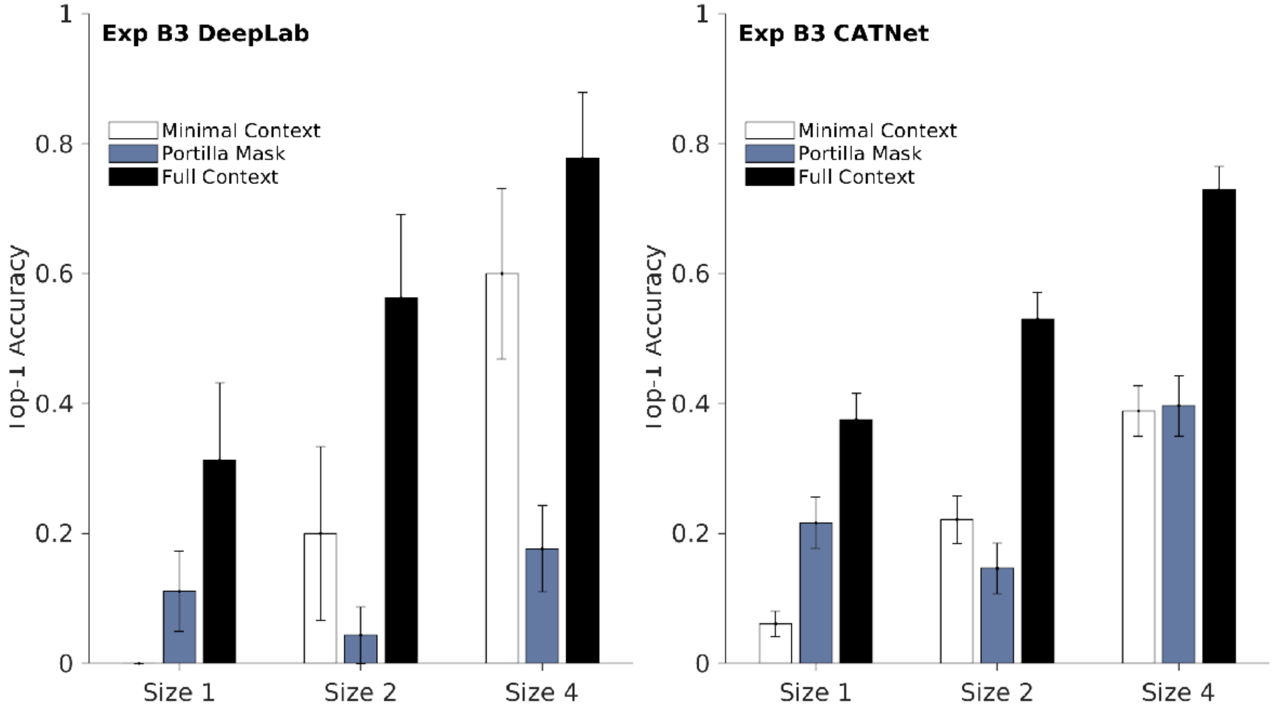

Figure S20. **Results on texture only in Exp B3 between DeepLab and CATNet.** Expanding on the discussion in Sec 5.2 and Fig S5, low-level features did not facilitate recognition. CATNet outperforms DeepLab on small objects.

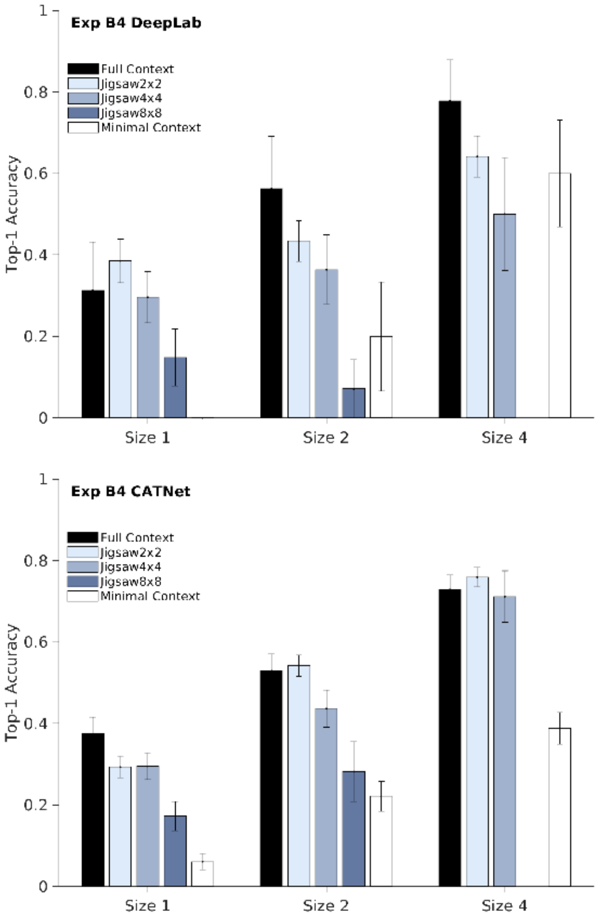

Figure S21. **Results on spatial configurations in Exp B4 between DeepLab and CATNet**. Expanding on the discussion in Sec 5.2 and Fig 6, large geometrical context re-arrangements disrupts contextual enhancement. CATNet performs equivalently well as DeepLab.

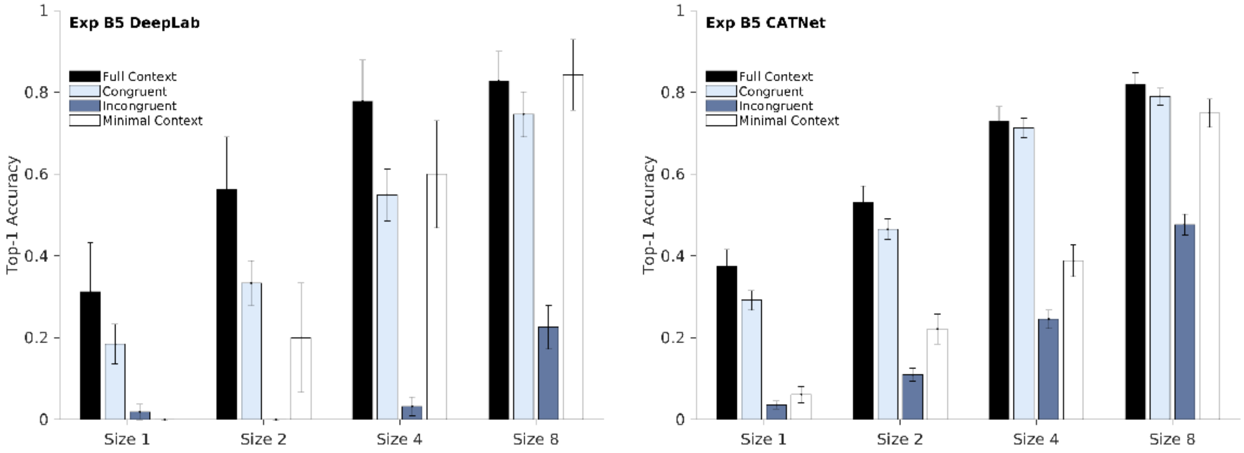

Figure S22. **Results on congruent versus incongruent context in Exp B5 between DeepLab and CATNet**. Expanding on the discussion in Sec 5.2 and Fig 7, incongruent context impairs object recognition. CATNet outperforms DeepLab for all object sizes.

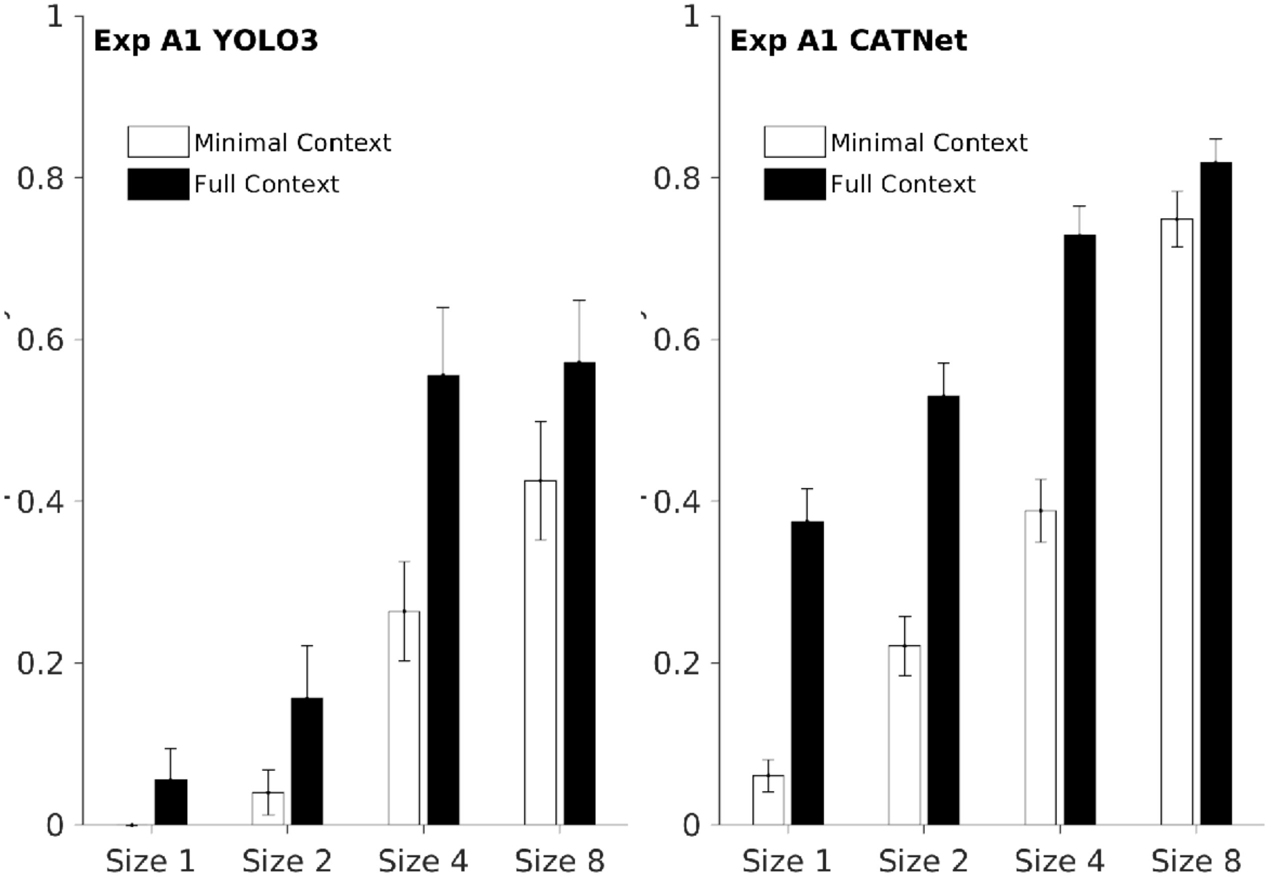

Figure S23. **Results on object size in Exp A1 between YOLO3 and CATNet**. Expanding on the discussion in Sec 5.1 and Fig 4, contextual modulation is stronger for smaller target objects for YOLO3 and CATNet. CATNet outperforms YOLO3 for all objects sizes.

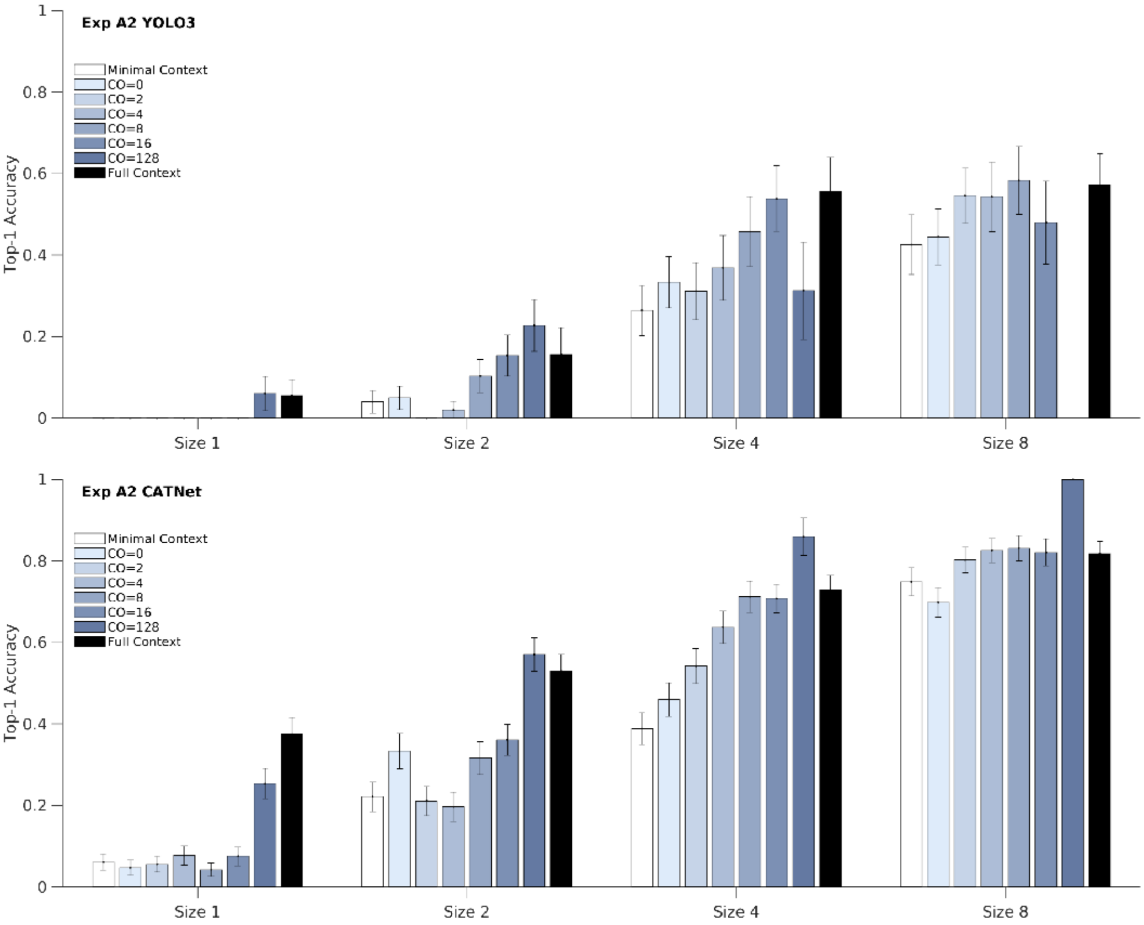

Figure S24. **Results on amount of context in Exp A2 between YOLO3 and CATNet**. Expanding on the discussion in Sec 5.1 and Fig S2, CATNet outperforms YOLO3 for all objects sizes.

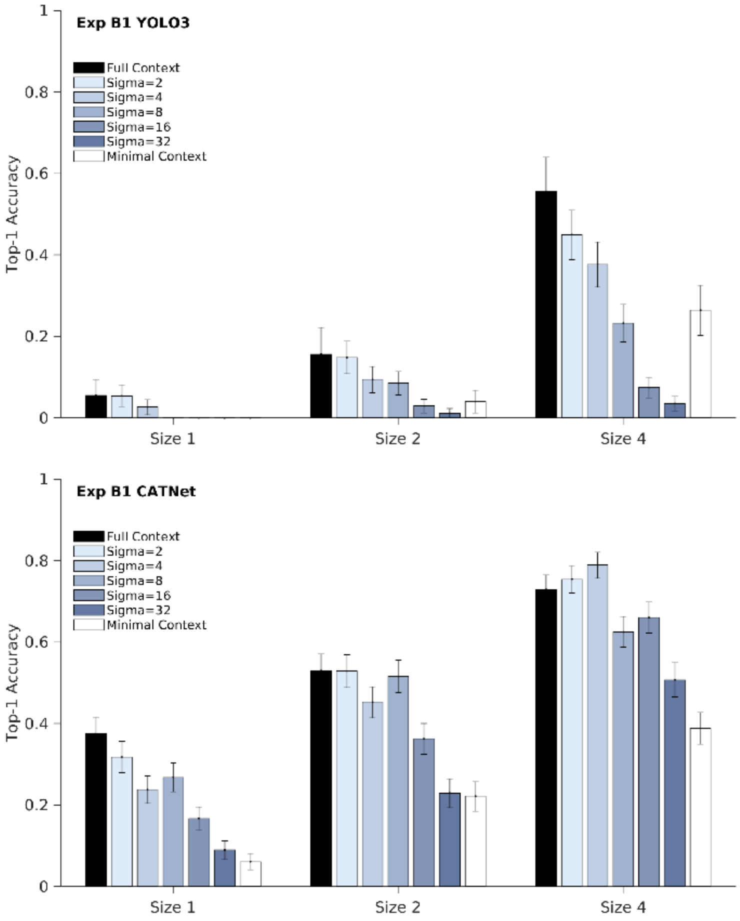

Figure S25. **Results on blurred context in Exp B1 between YOLO3 and CATNet.** Expanding on the discussion in Sec 5.2 and Fig 5, contextual facilitation persists even after small amounts of blurring (Exp B1). CATNet outperforms YOLO3 for all objects sizes.

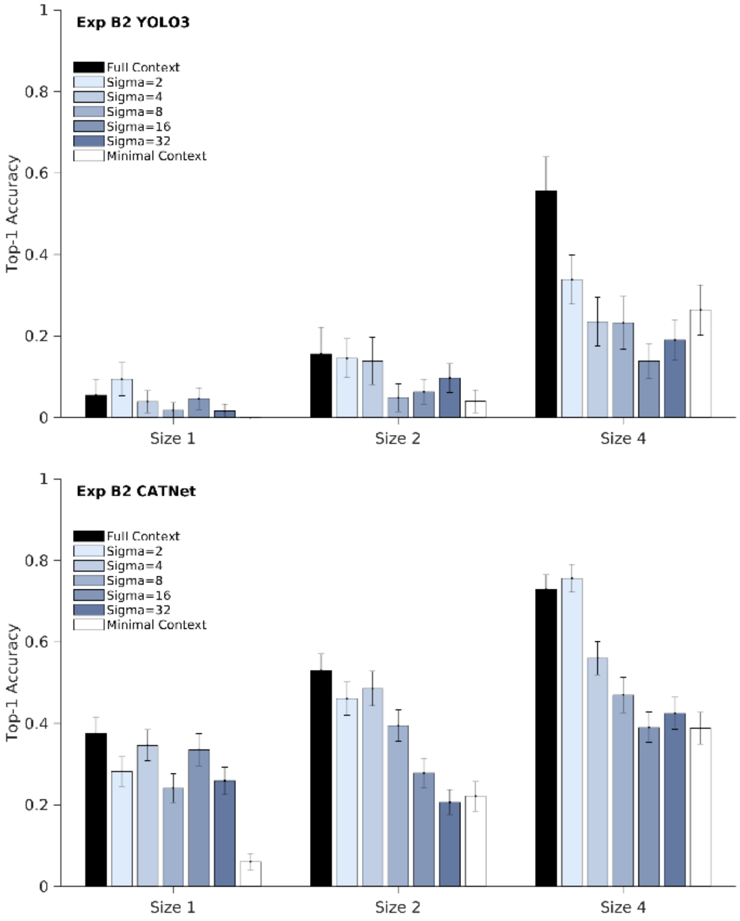

Figure S26. **Results on blurred objects in Exp B2 between YOLO3 and CATNet.** Expanding on the discussion in Sec 5.2 and Fig S4, compared with context blurring, modifying the object led to larger accuracy drops. CATNet outperforms YOLO3 for all objects sizes.

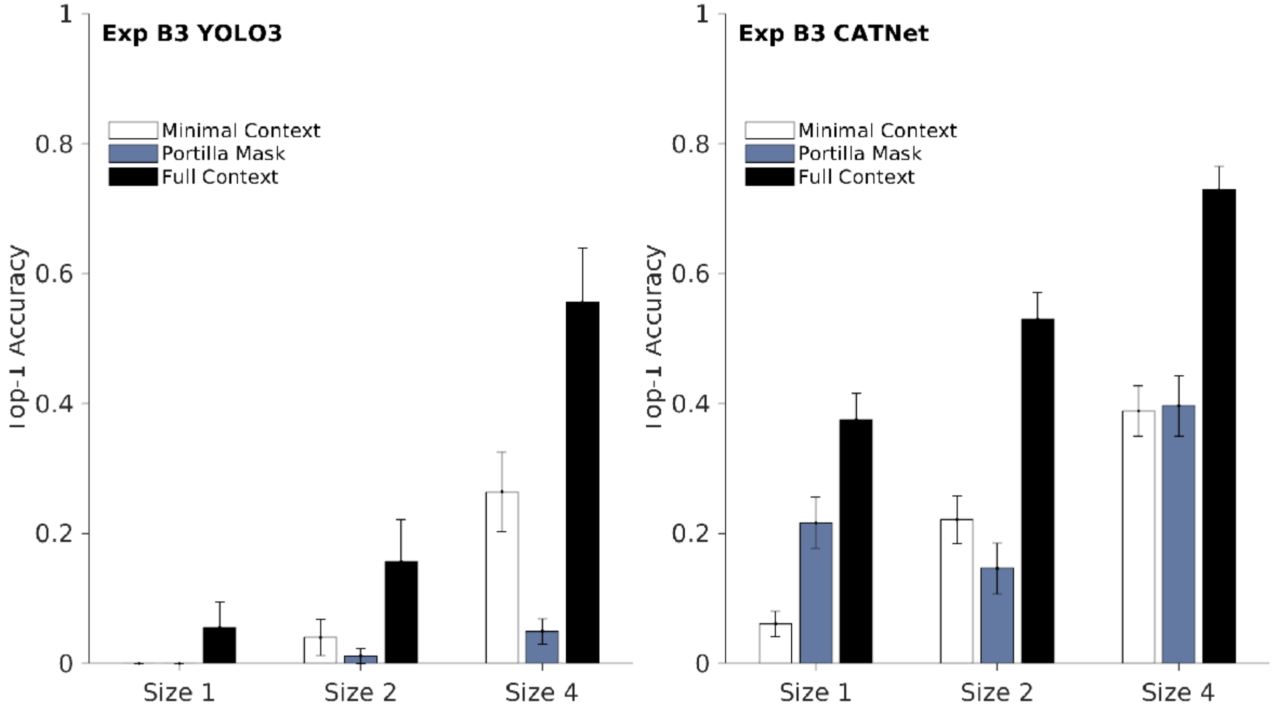

Figure S27. **Results on texture only in Exp B3 between YOLO3 and CATNet.** Expanding on the discussion in Sec 5.2 and Fig S5, low-level features did not facilitate recognition. CATNet outperforms YOLO3 for all objects sizes.

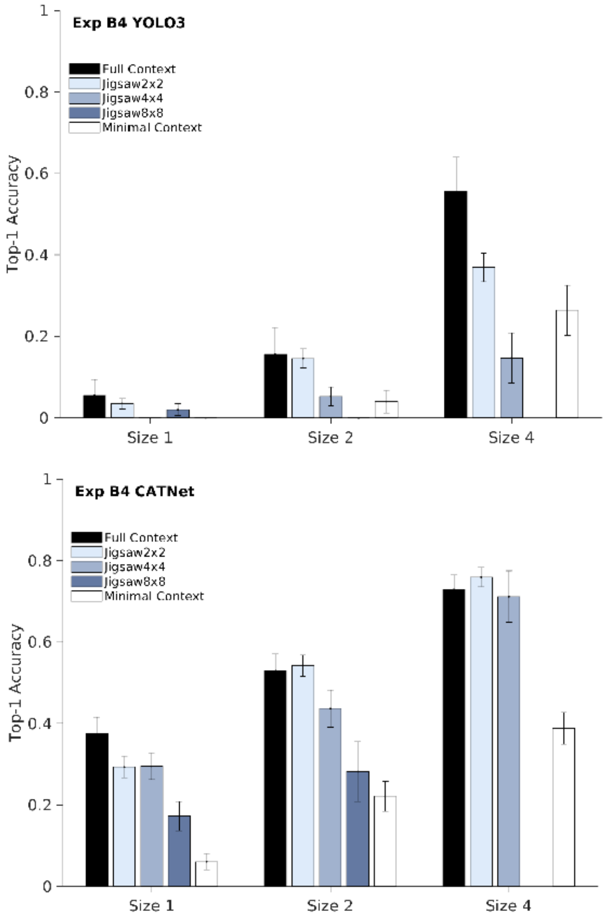

Figure S28. **Results on spatial configurations in Exp B4 between YOLO3 and CATNet**. Expanding on the discussion in Sec 5.2 and Fig 6, large geometrical context re-arrangements disrupts contextual enhancement. CATNet outperforms YOLO3 for all objects sizes.

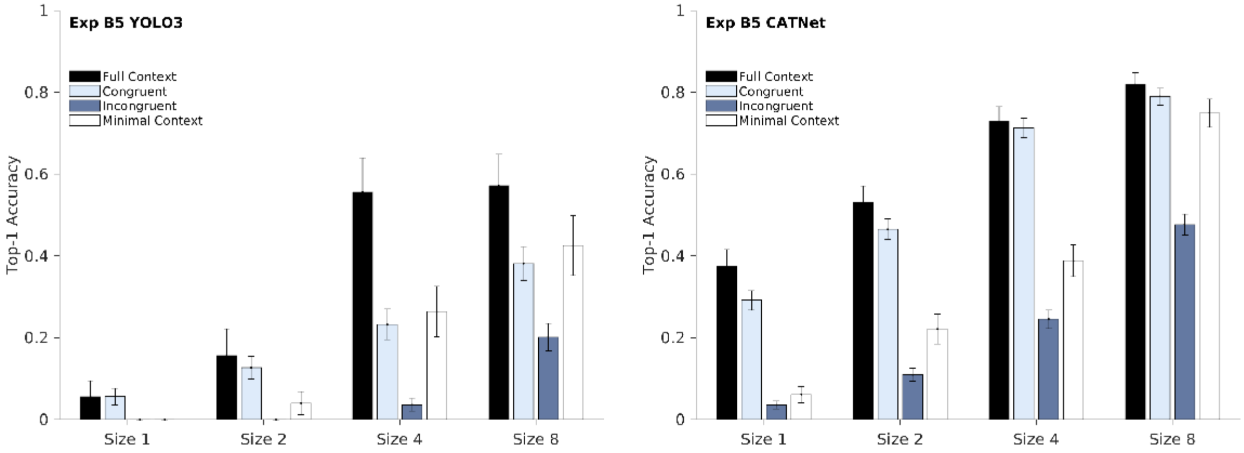

Figure S29. **Results on congruent versus incongruent context in Exp B5 between YOLO3 and CATNet**. Expanding on the discussion in Sec 5.2 and Fig 7, incongruent context impairs object recognition. CATNet outperforms YOLO3 for all objects sizes.

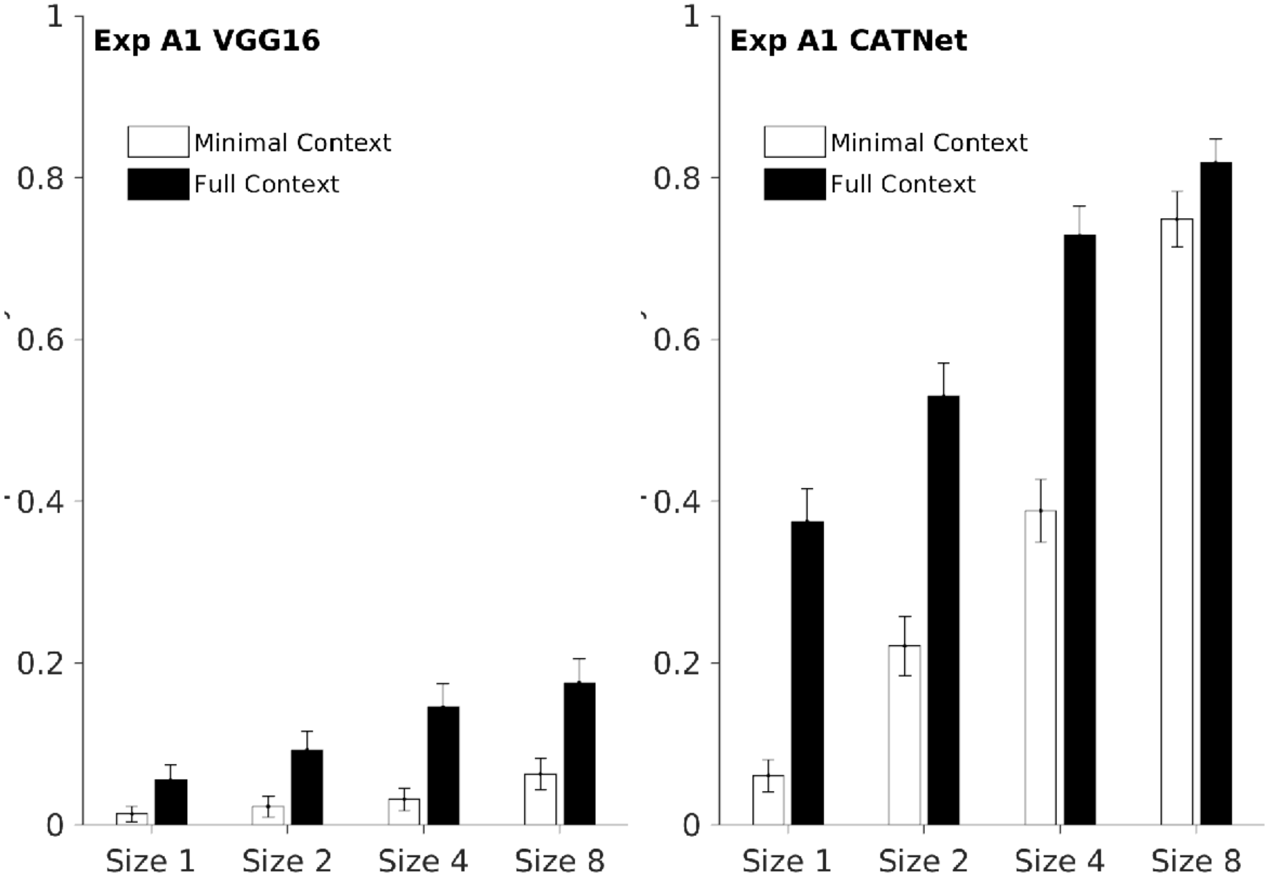

Figure S30. **Results on object size in Exp A1 between VGG16 and CATNet**. Expanding on the discussion in Sec 5.1 and Fig 4, contextual modulation is stronger for smaller target objects for VGG16 and CATNet. CATNet outperforms VGG16 for all objects sizes.

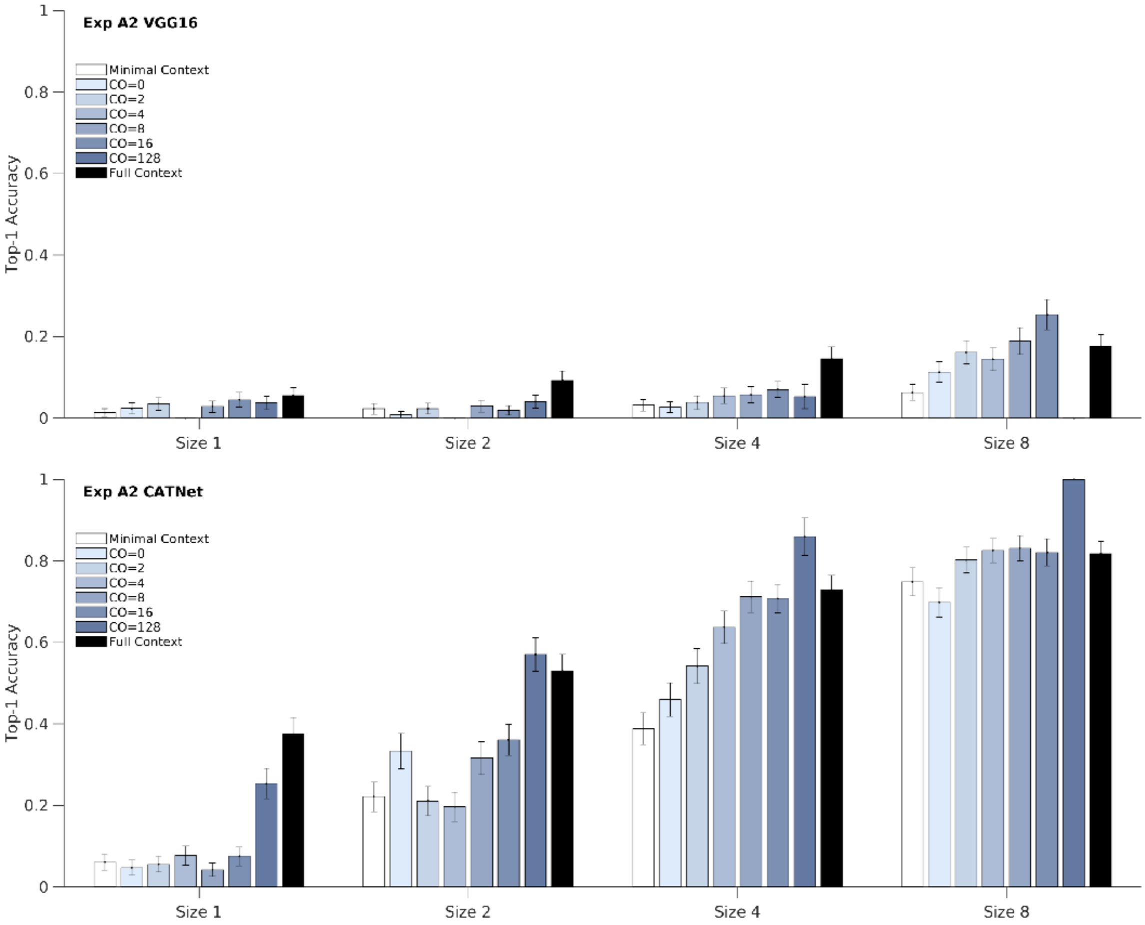

Figure S31. **Results on amount of context in Exp A2 between VGG16 and CATNet**. Expanding on the discussion in Sec 5.1 and Fig S2, CATNet outperforms VGG16 for all objects sizes.

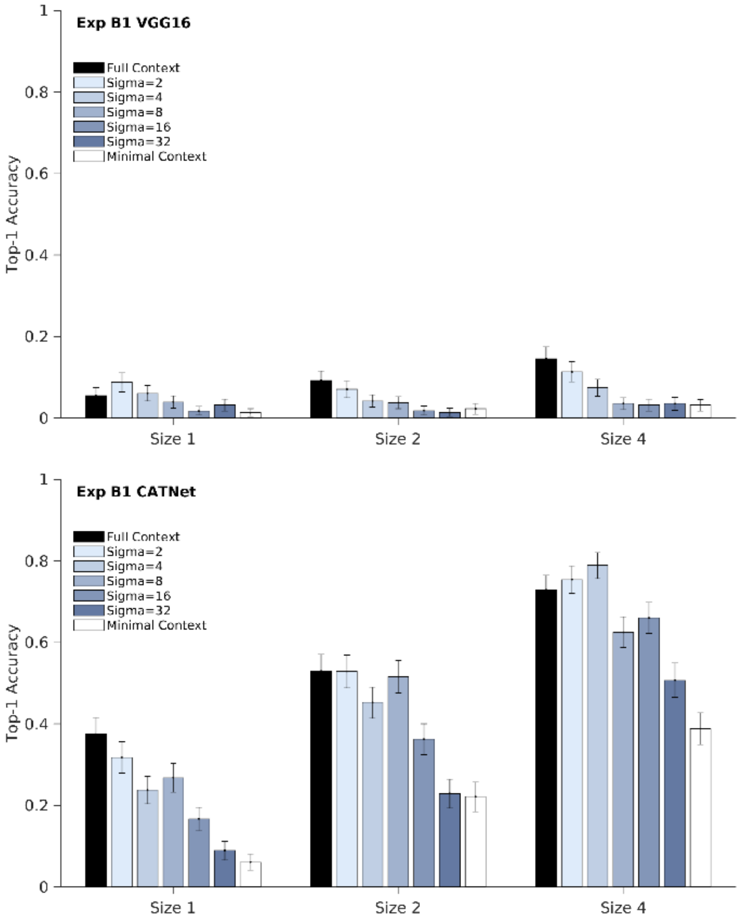

Figure S32. **Results on blurred context in Exp B1 between VGG16 and CATNet.** Expanding on the discussion in Sec 5.2 and Fig 5, contextual facilitation persists even after small amounts of blurring (Exp B1). CATNet outperforms VGG16 for all objects sizes.

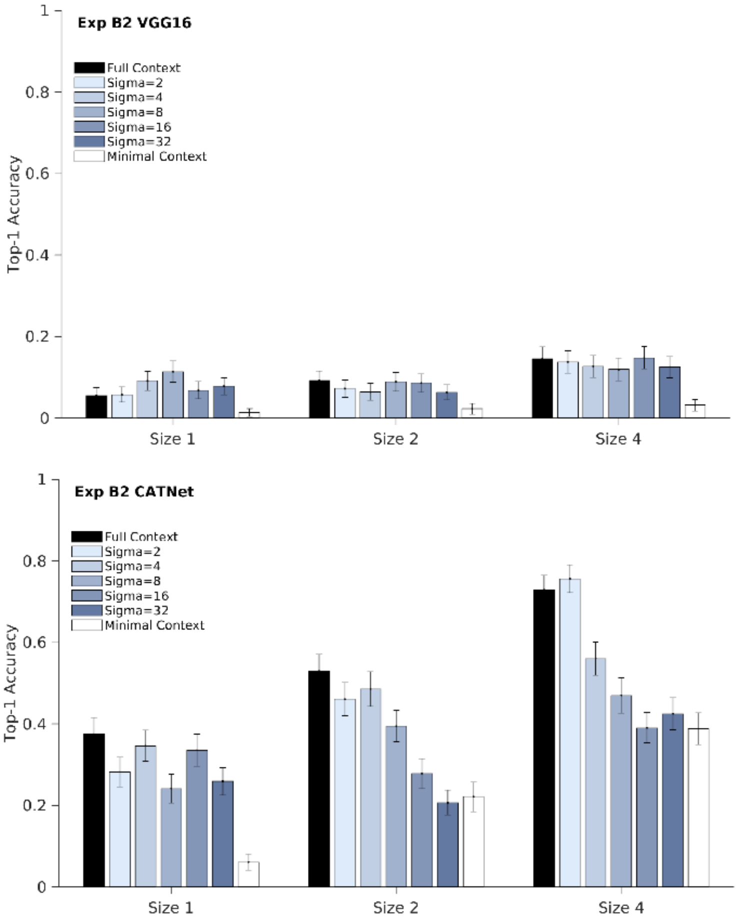

Figure S33. **Results on blurred objects in Exp B2 between VGG16 and CATNet.** Expanding on the discussion in Sec 5.2 and Fig S4, compared with context blurring, modifying the object led to larger accuracy drops. CATNet outperforms VGG16 for all objects sizes.

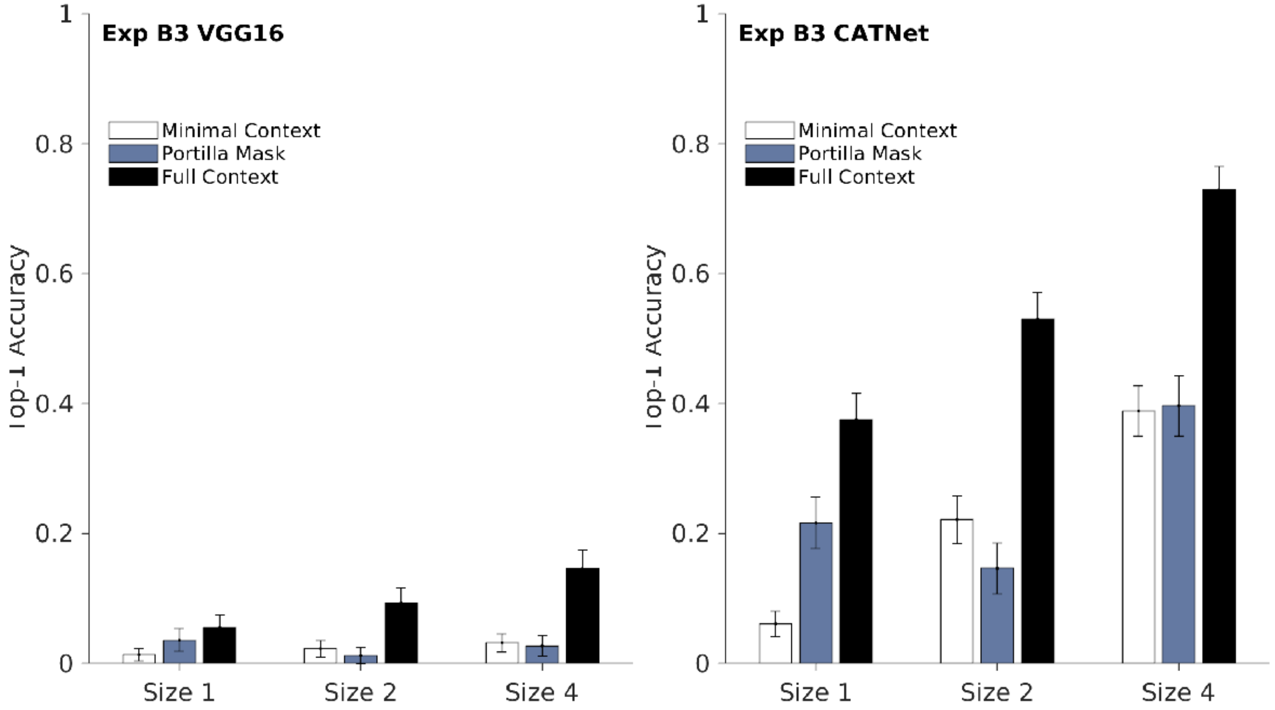

Figure S34. **Results on texture only in Exp B3 between VGG16 and CATNet.** Expanding on the discussion in Sec 5.2 and Fig S5, low-level features did not facilitate recognition. CATNet outperforms VGG16 for all objects sizes.

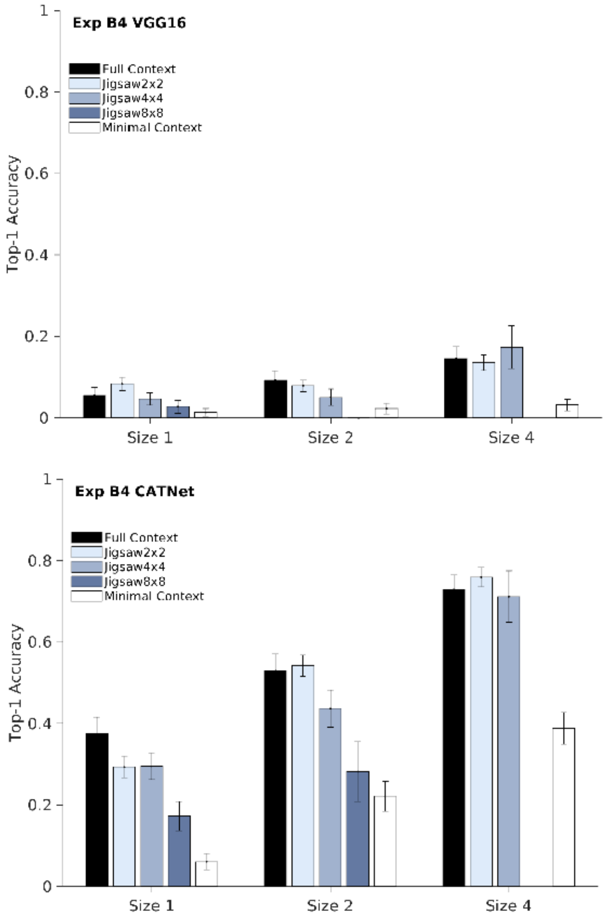

Figure S35. **Results on spatial configurations in Exp B4 between VGG16 and CATNet**. Expanding on the discussion in Sec 5.2 and Fig 6, large geometrical context re-arrangements disrupts contextual enhancement. CATNet outperforms VGG16 for all objects sizes.

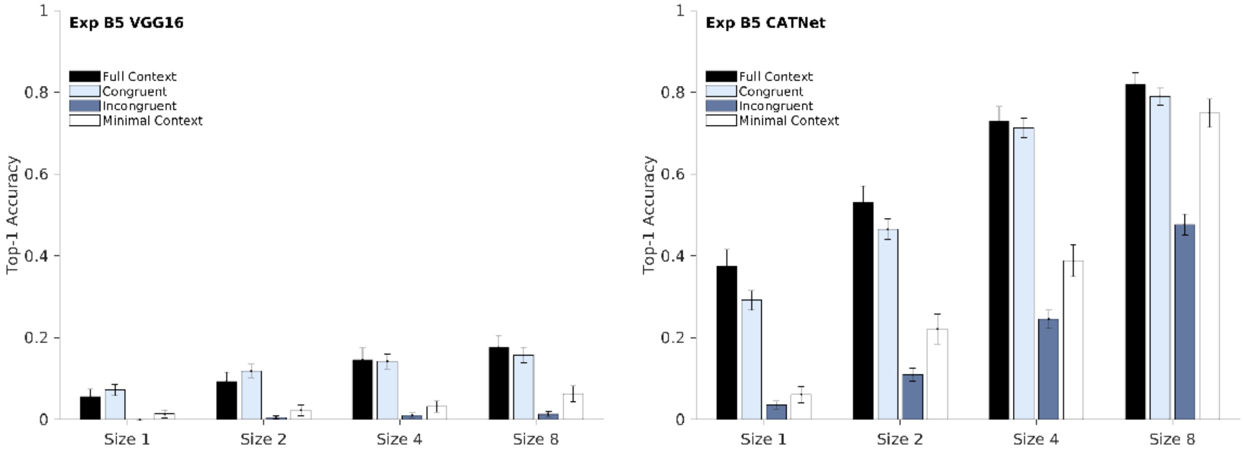

Figure S36. **Results on congruent versus incongruent context in Exp B5 between VGG16 and CATNet**. Expanding on the discussion in Sec 5.2 and Fig 7, incongruent context impairs object recognition. CATNet outperforms VGG16 for all objects sizes.

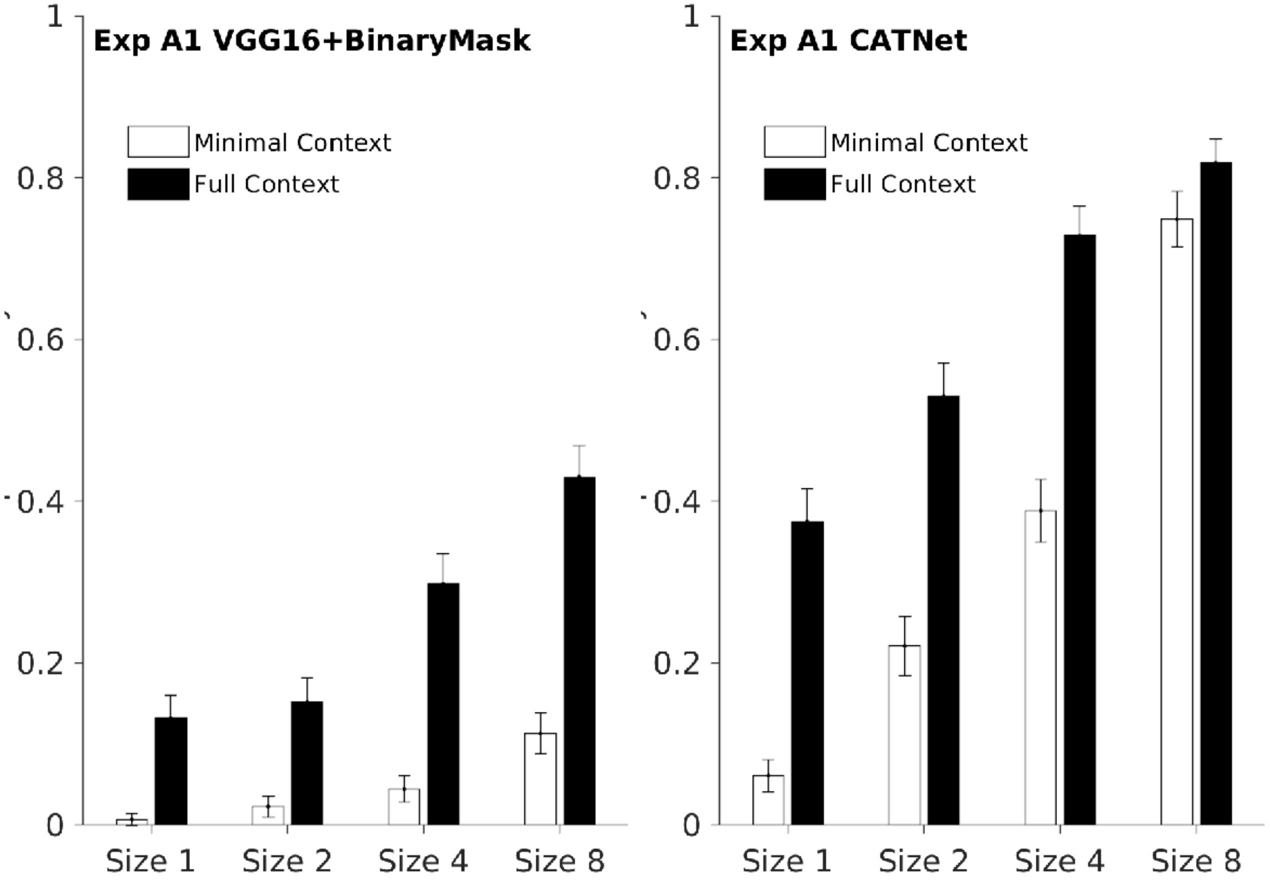

Figure S37. **Results on object size in Exp A1 between VGG16+BinaryMask and CATNet**. Expanding on the discussion in Sec 5.1 and Fig 4, contextual modulation is stronger for smaller target objects for VGG16+BinaryMask and CATNet. CATNet outperforms VGG16+BinaryMask for all objects sizes.

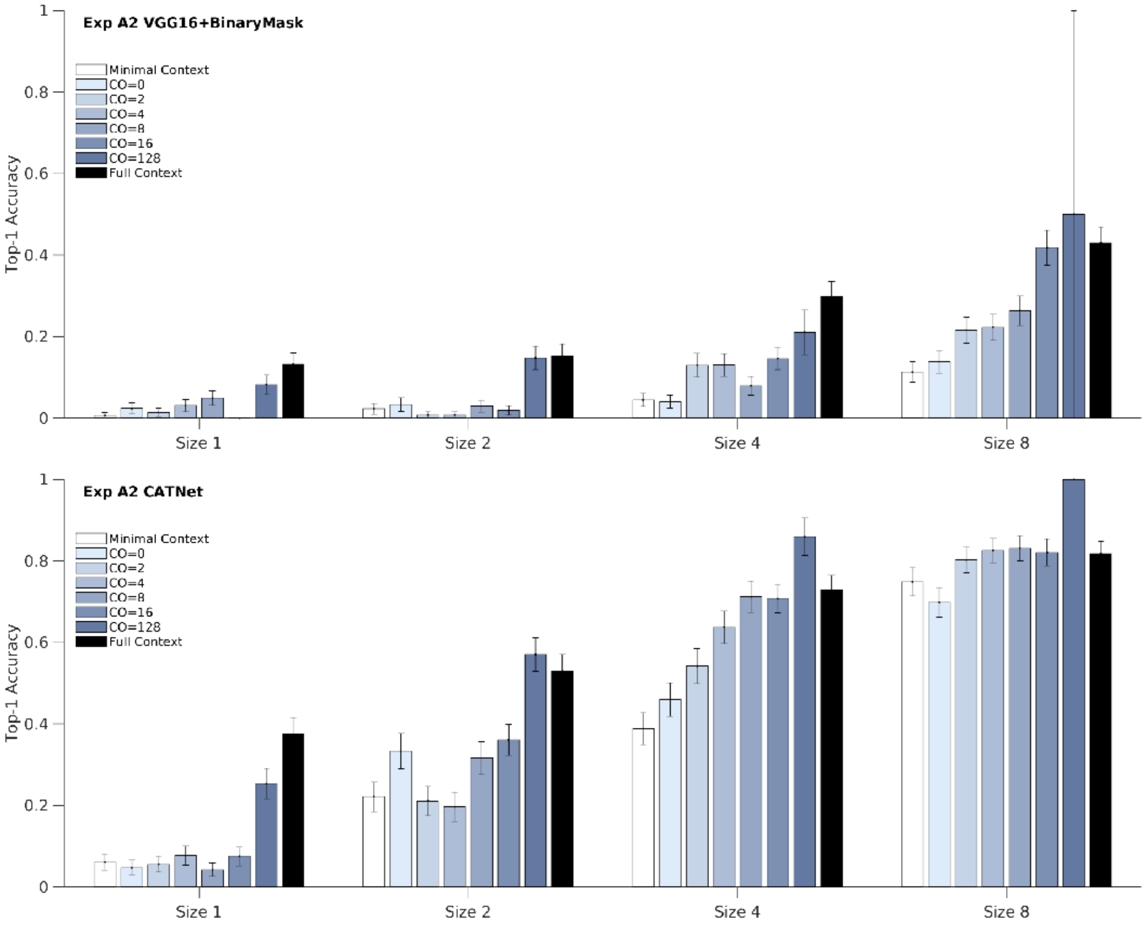

Figure S38. **Results on amount of context in Exp A2 between VGG16+BinaryMask and CATNet**. Expanding on the discussion in Sec 5.1 and Fig S2, CATNet outperforms VGG16+BinaryMask for all objects sizes.

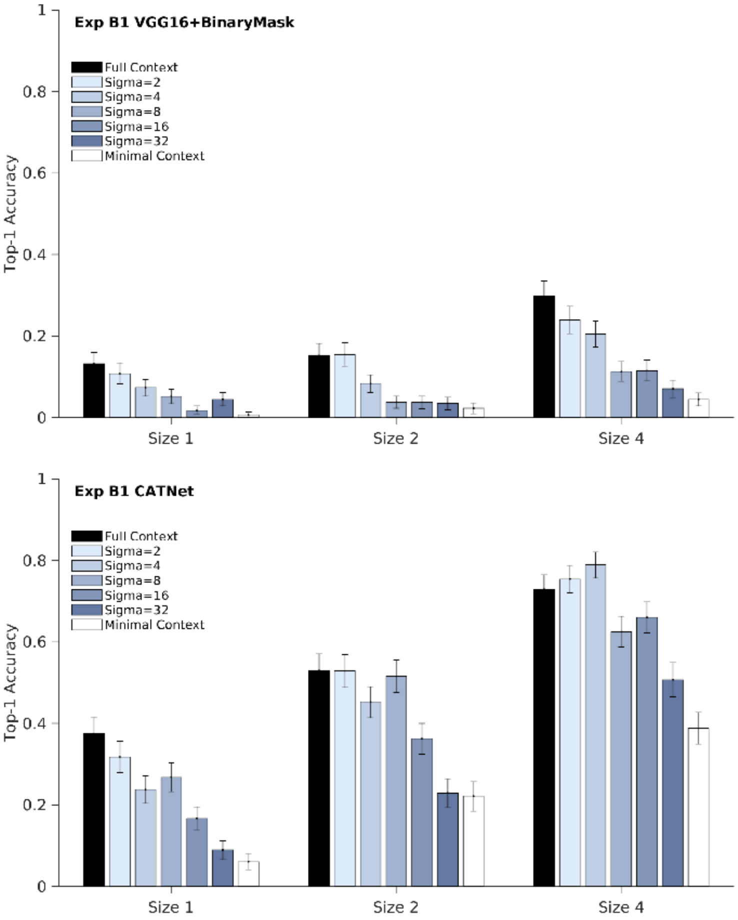

Figure S39. **Results on blurred context in Exp B1 between VGG16+BinaryMask and CATNet.** Expanding on the discussion in Sec 5.2 and Fig 5, contextual facilitation persists even after small amounts of blurring (Exp B1). CATNet outperforms VGG16+BinaryMask for all objects sizes.

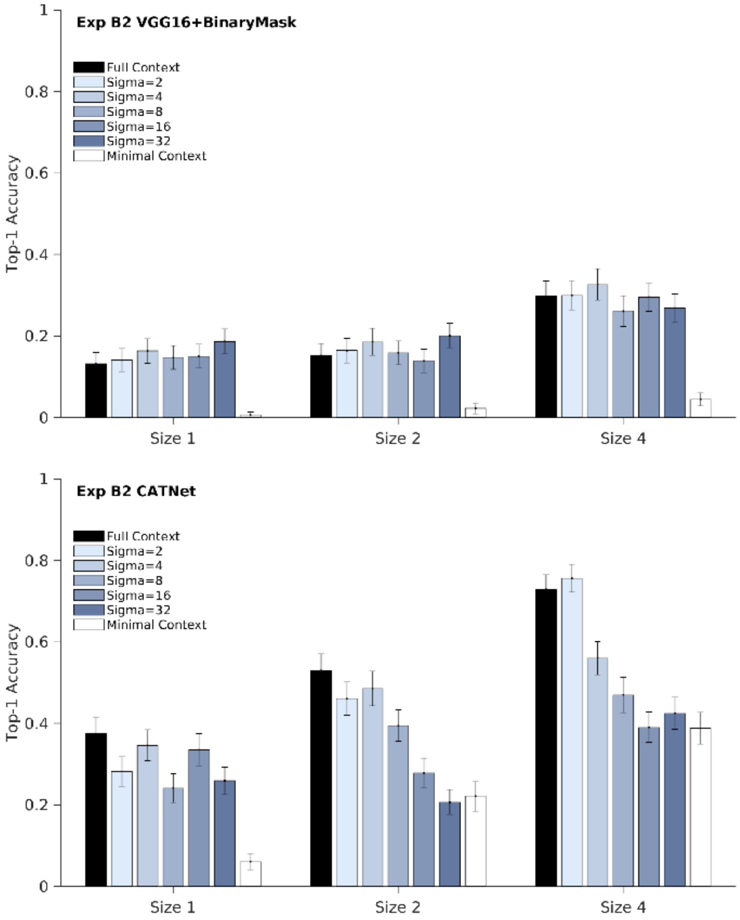

Figure S40. **Results on blurred objects in Exp B2 between VGG16+BinaryMask and CATNet.** Expanding on the discussion in Sec 5.2 and Fig S4, compared with context blurring, modifying the object led to larger accuracy drops. CATNet outperforms VGG16+BinaryMask for all objects sizes.

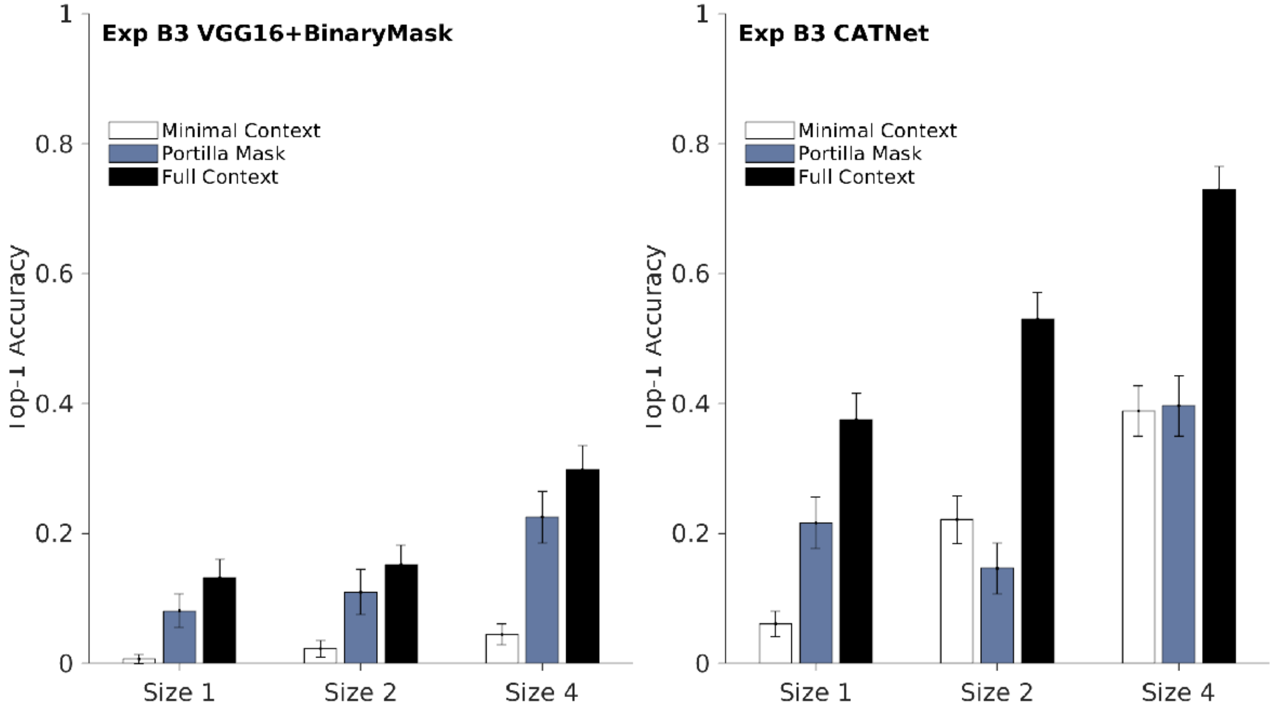

Figure S41. **Results on texture only in Exp B3 between VGG16+BinaryMask and CATNet.** Expanding on the discussion in Sec 5.2 and Fig S5, low-level features did not facilitate recognition. CATNet outperforms VGG16+BinaryMask for all objects sizes.

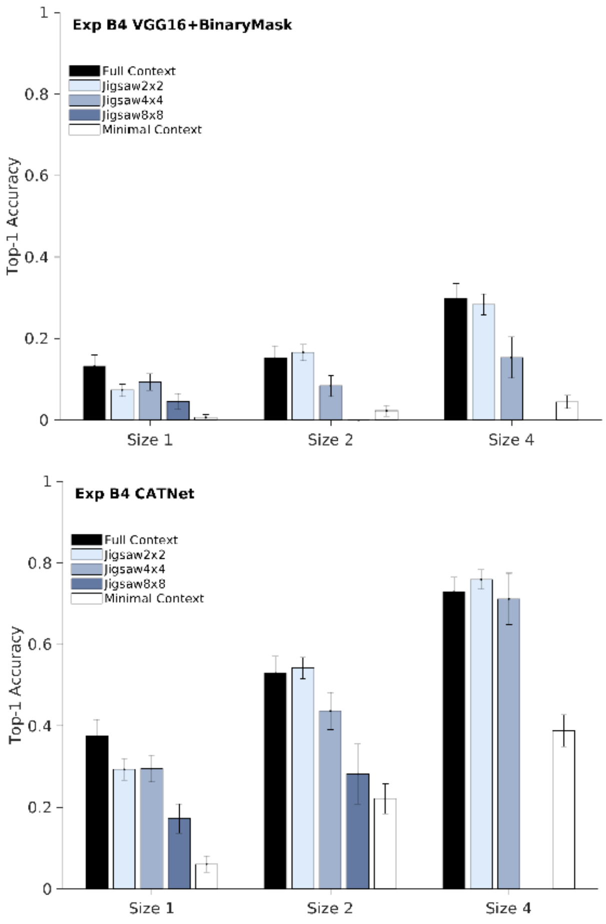

Figure S42. **Results on spatial configurations in Exp B4 between VGG16+BinaryMask and CATNet**. Expanding on the discussion in Sec 5.2 and Fig 6, large geometrical context re-arrangements disrupts contextual enhancement. CATNet outperforms VGG16+BinaryMask for all objects sizes.

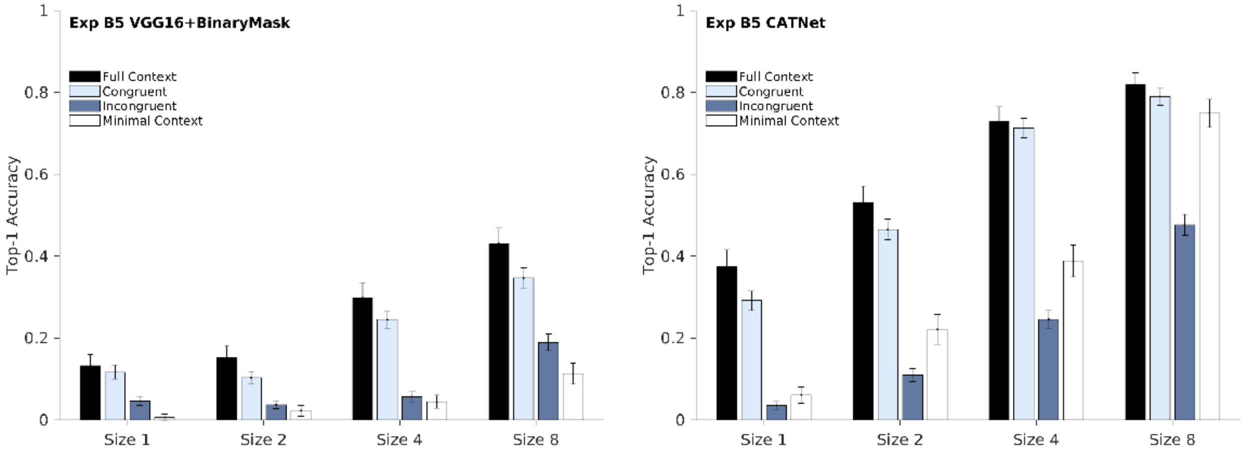

Figure S43. **Results on congruent versus incongruent context in Exp B5 between VGG16+BinaryMask and CATNet**. Expanding on the discussion in Sec 5.2 and Fig 7, incongruent context impairs object recognition. CATNet outperforms VGG16+BinaryMask for all objects sizes.

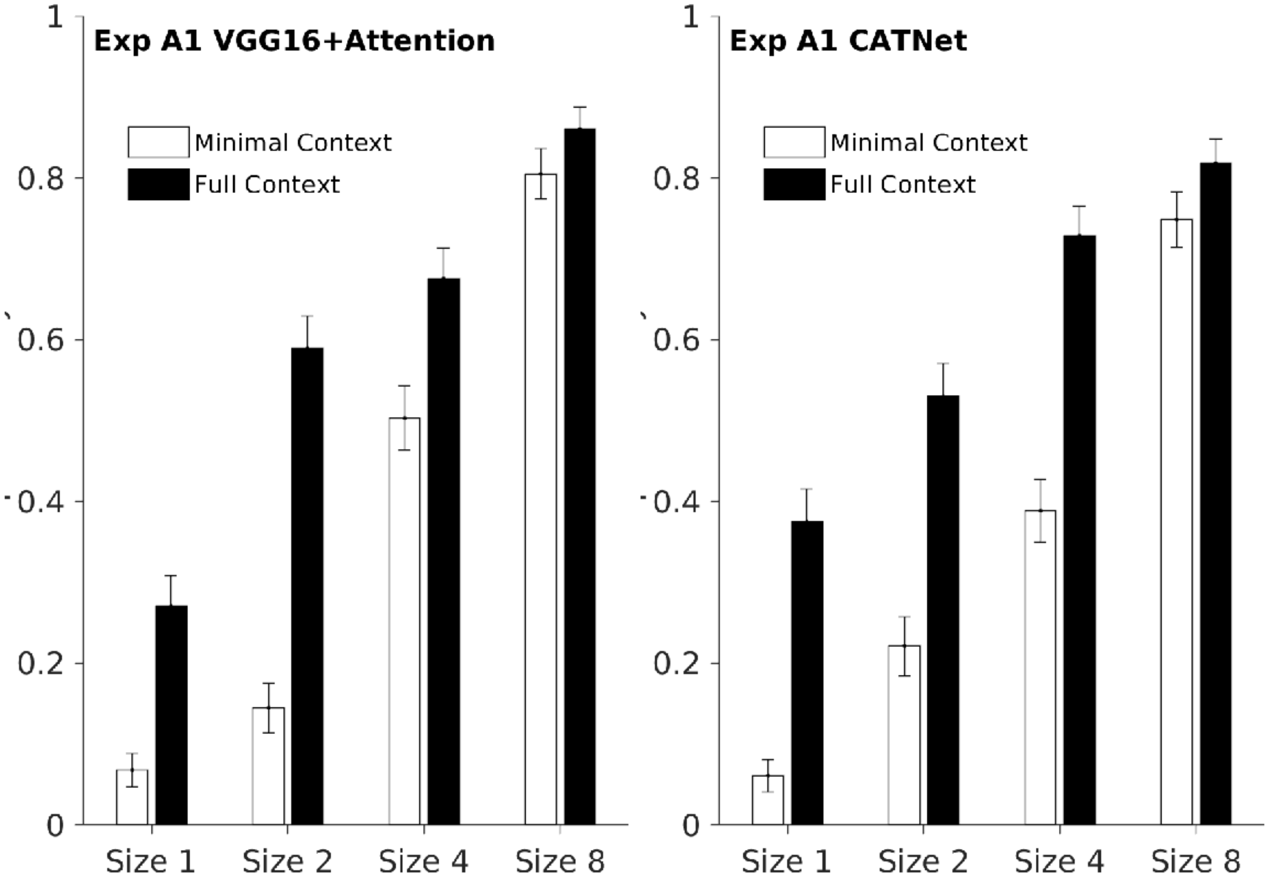

Figure S44. **Results on object size in Exp A1 between VGG16+Attention and CATNet**. Expanding on the discussion in Sec 5.1 and Fig 4, contextual modulation is stronger for smaller target objects for VGG16+Attention and CATNet. CATNet performs equally well as VGG16+Attention for all objects sizes.

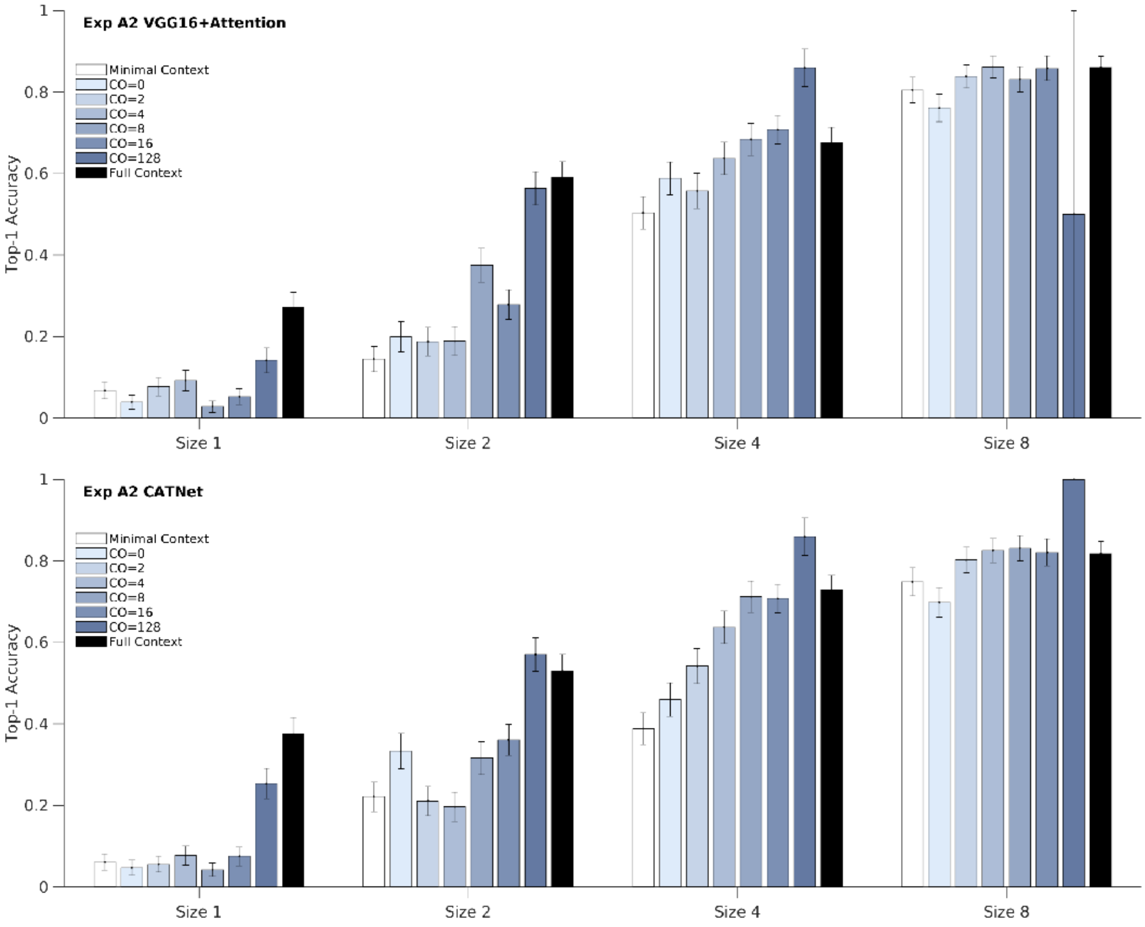

Figure S45. **Results on amount of context in Exp A2 between VGG16+Attention and CATNet**. Expanding on the discussion in Sec 5.1 and Fig S2, CATNet outperforms VGG16+Attention for smaller sizes.

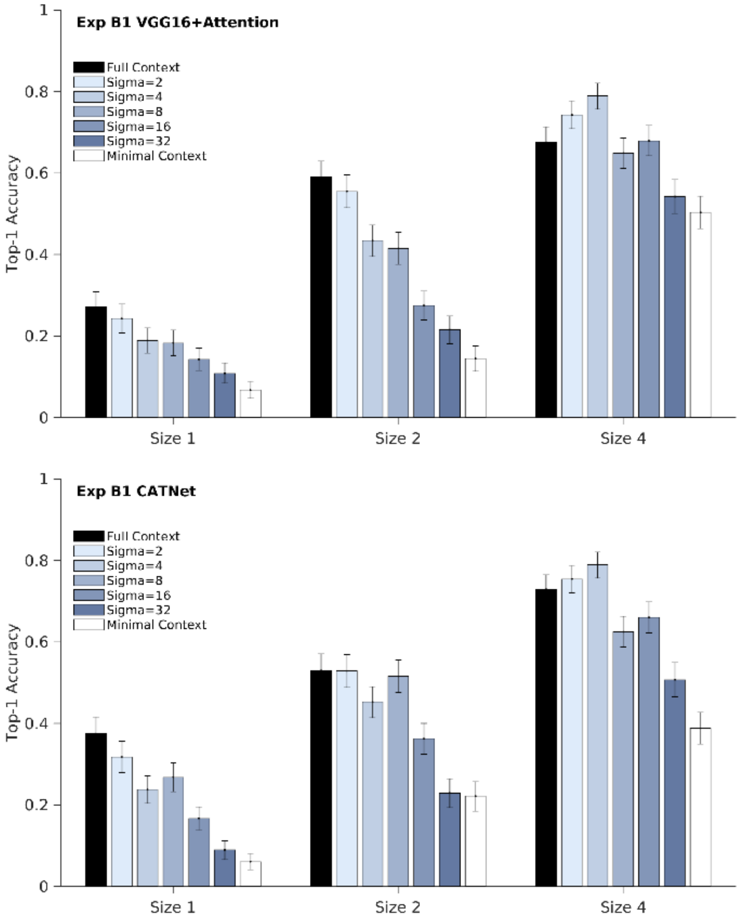

Figure S46. **Results on blurred context in Exp B1 between VGG16+Attention and CATNet.** Expanding on the discussion in Sec 5.2 and Fig 5, contextual facilitation persists even after small amounts of blurring (Exp B1). CATNet performs equally well as VGG16+Attention for all objects sizes.

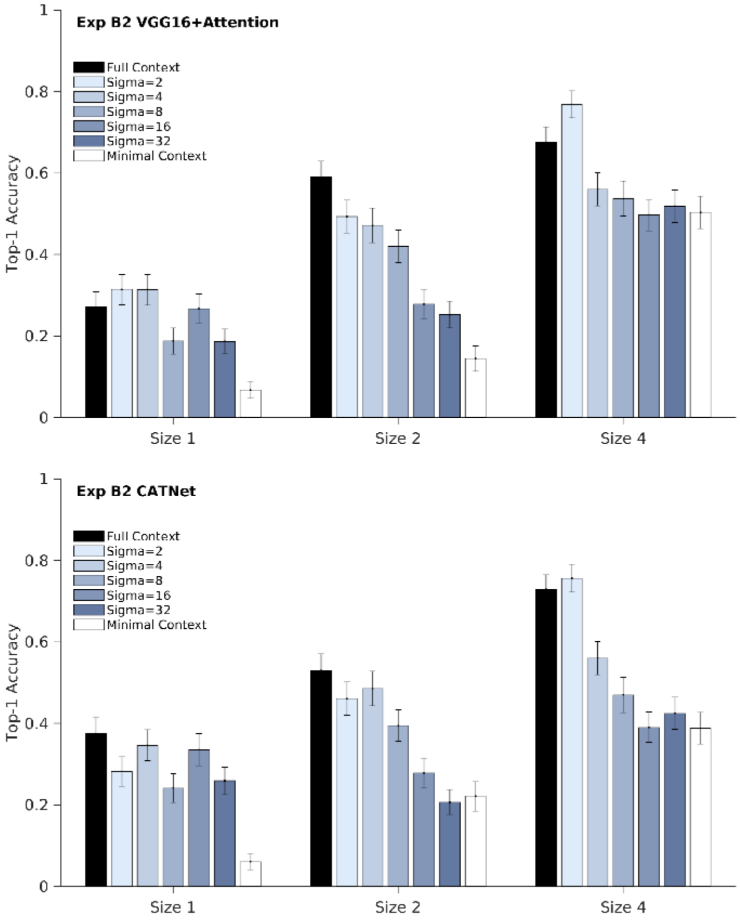

Figure S47. **Results on blurred objects in Exp B2 between VGG16+Attention and CATNet.** Expanding on the discussion in Sec 5.2 and Fig S4, compared with context blurring, modifying the object led to larger accuracy drops. CATNet performs equally well as VGG16+Attention for all objects sizes.

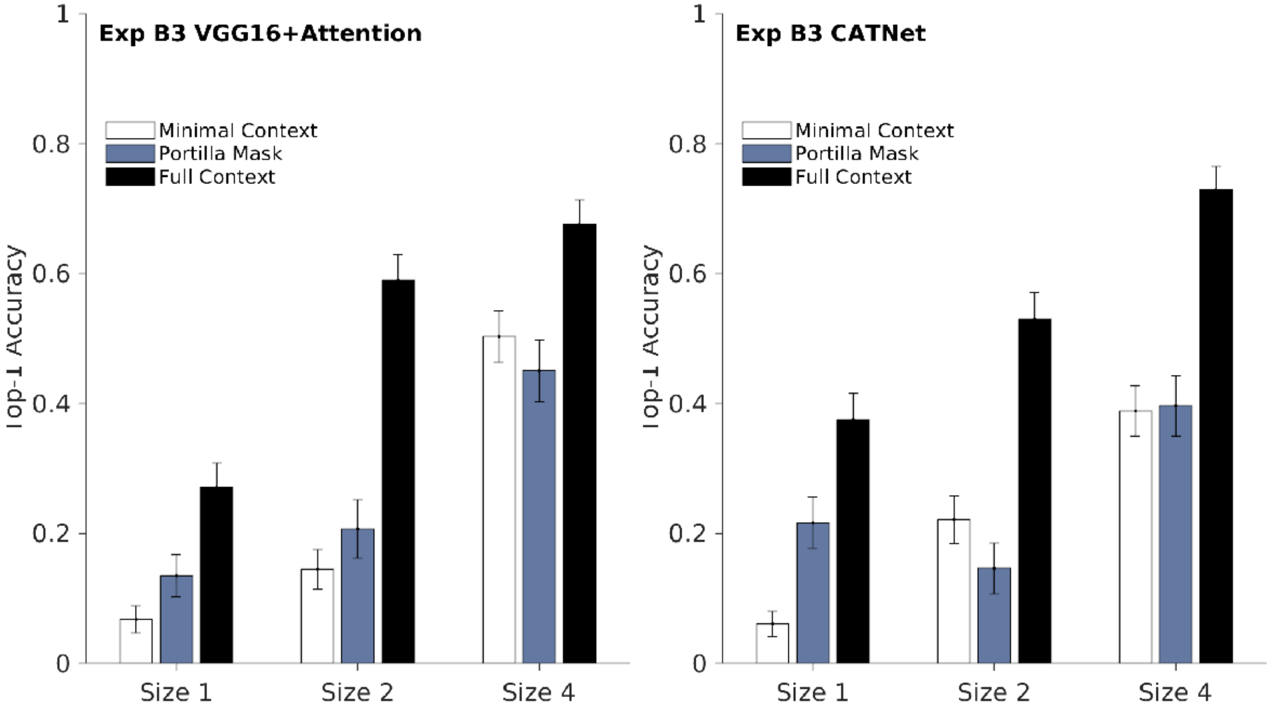

Figure S48. **Results on texture only in Exp B3 between VGG16+Attention and CATNet.** Expanding on the discussion in Sec 5.2 and Fig S5, low-level features did not facilitate recognition. CATNet outperforms VGG16+Attention for smaller objects.

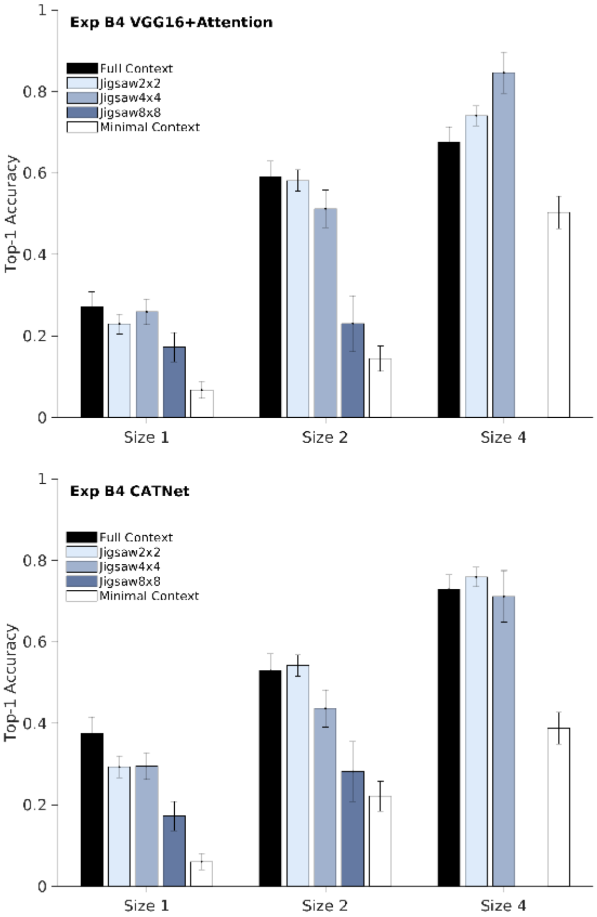

Figure S49. **Results on spatial configurations in Exp B4 between VGG16+Attention and CATNet**. Expanding on the discussion in Sec 5.2 and Fig 6, large geometrical context re-arrangements disrupts contextual enhancement. CATNet performs equally well as VGG16+Attention for all objects sizes.

References

[1] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 2

[2] Tammy Harris and James W Hardin. Exact wilcoxon signed-rank and wilcoxon mann–whitney ranksum tests. *The Stata Journal*, 13(2):337–343, 2013. 3

[3] PK Ito. 7 robustness of anova and manova test procedures. *Handbook of statistics*, 1:199–236, 1980. 4

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[5] Tam V Nguyen, Qi Zhao, and Shuicheng Yan. Attentive systems: A survey. *International Journal of Computer Vision*, 126(1):86–110, 2018. 3

[6] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 2

[7] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv*, 2018. 2

[8] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2, 3

[9] Hanlin Tang, Martin Schrimpf, William Lotter, Charlotte Moerman, Ana Paredes, Josue Ortega Caro, Walter Hardesty, David Cox, and Gabriel Kreiman. Recurrent computations for visual pattern completion. *Proceedings of the National Academy of Sciences*, 115(35):8835–8840, 2018. 3

[10] Kevin Wu, Eric Wu, and Gabriel Kreiman. Learning scene gist with convolutional neural networks to improve object recognition. In *Information Sciences and Systems (CISS), 2018 52nd Annual Conference on*, pages 1–6. IEEE, 2018. 3

[11] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015. 2, 3

[12] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*, 2014. 2

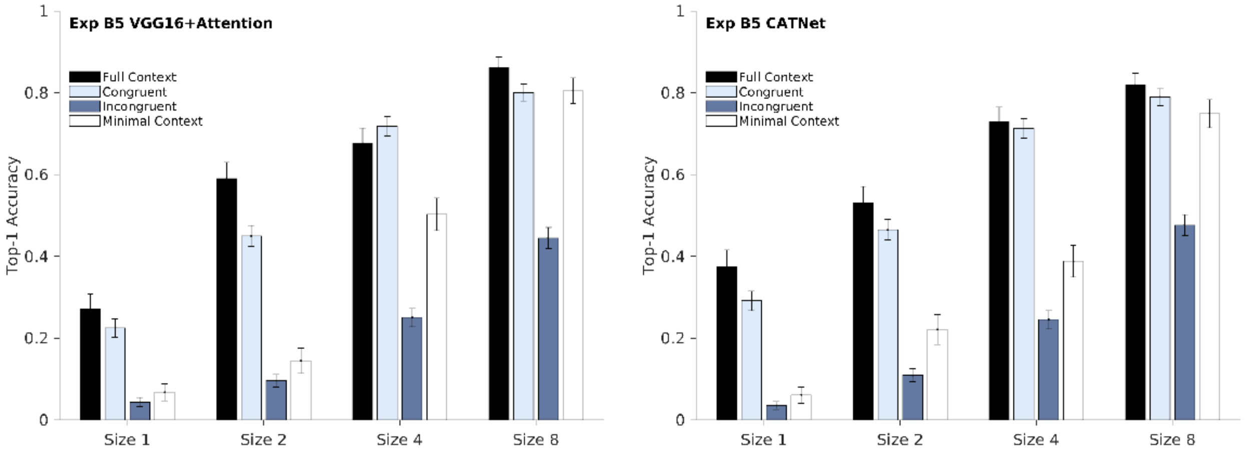

Figure S50. **Results on congruent versus incongruent context in Exp B5 between VGG16+Attention and CATNet**. Expanding on the discussion in Sec 5.2 and Fig 7, incongruent context impairs object recognition. CATNet performs equally well as VGG16+Attention for all objects sizes.

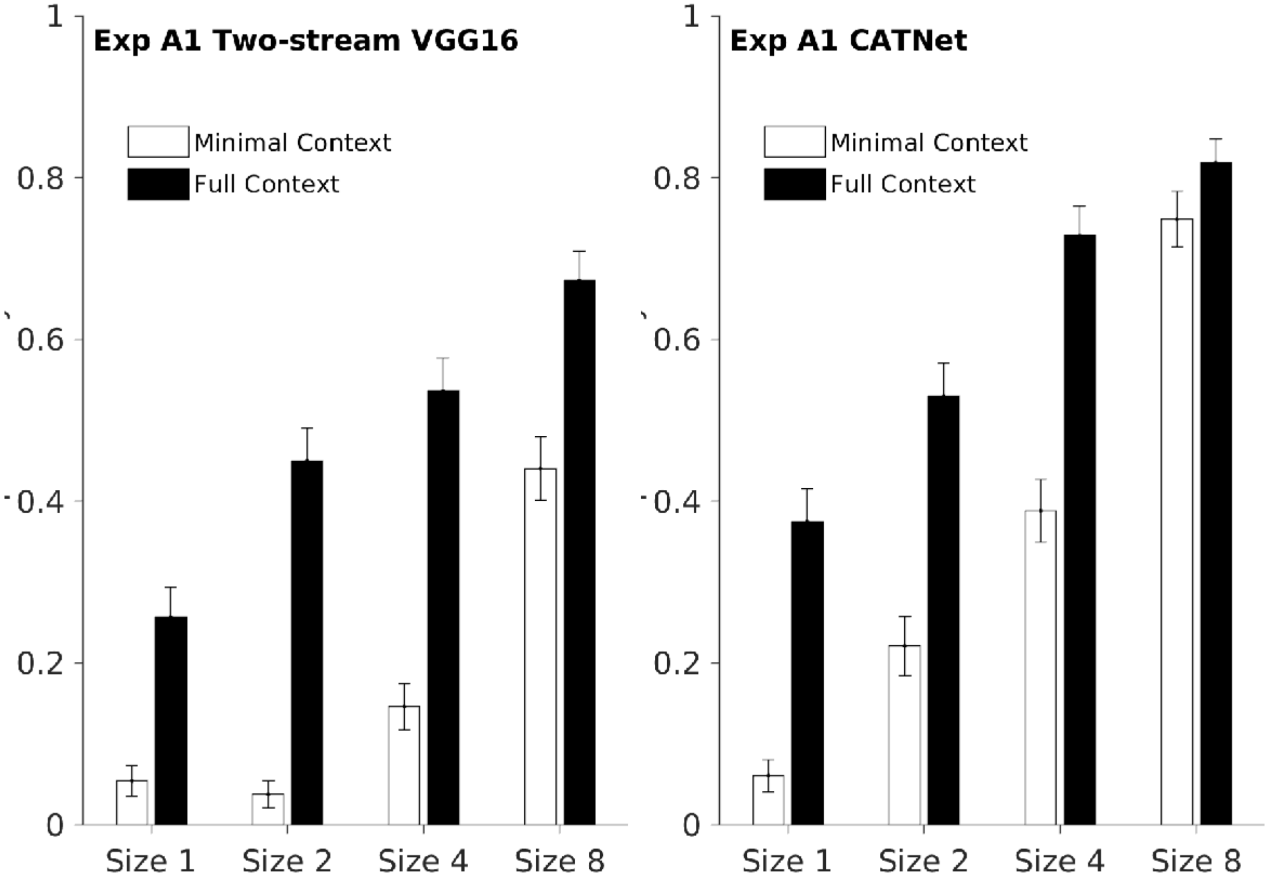

Figure S51. **Results on object size in Exp A1 between Two-stream VGG16 and CATNet**. Expanding on the discussion in Sec 5.1 and Fig 4, contextual modulation is stronger for smaller target objects for Two-stream VGG16 and CATNet. CATNet outperforms Two-stream VGG16 for all objects sizes.

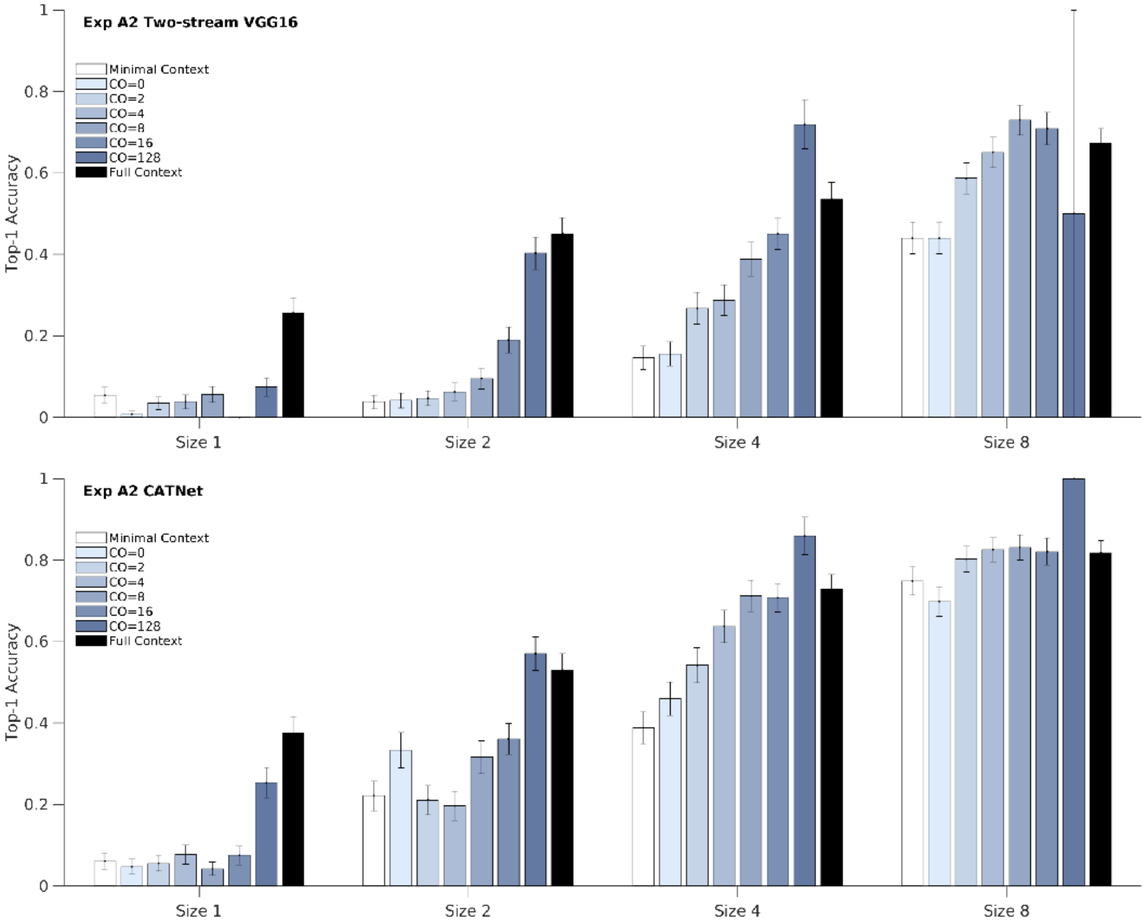

Figure S52. **Results on amount of context in Exp A2 between Two-stream VGG16 and CATNet**. Expanding on the discussion in Sec 5.1 and Fig S2, CATNet outperforms Two-stream VGG16 for all objects sizes.

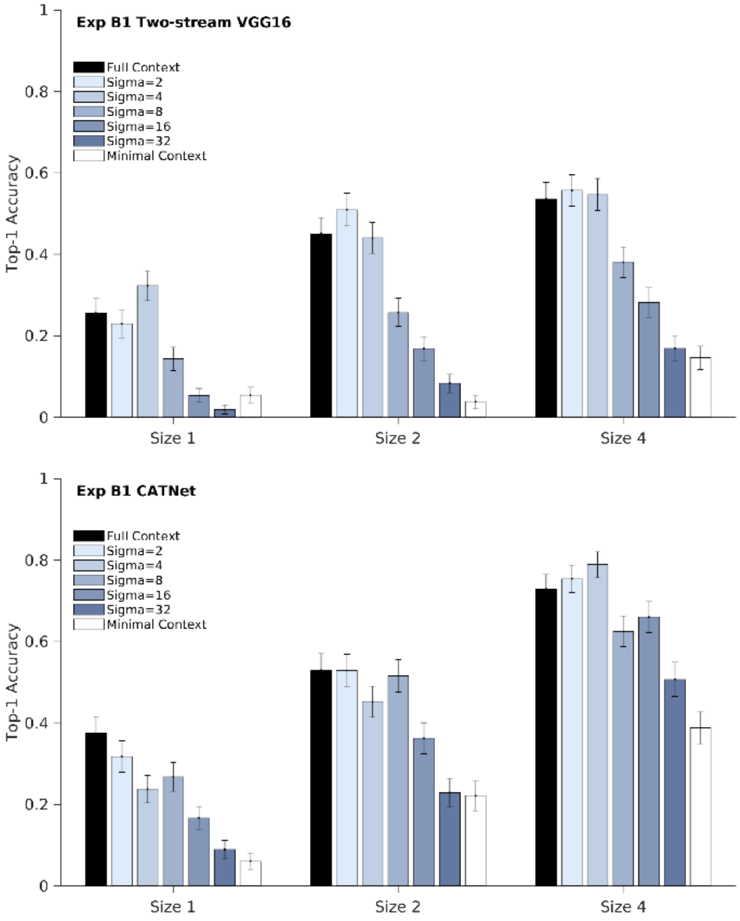

Figure S53. **Results on blurred context in Exp B1 between Two-stream VGG16 and CATNet.** Expanding on the discussion in Sec 5.2 and Fig 5, contextual facilitation persists even after small amounts of blurring (Exp B1). CATNet outperforms Two-stream VGG16 for all objects sizes.

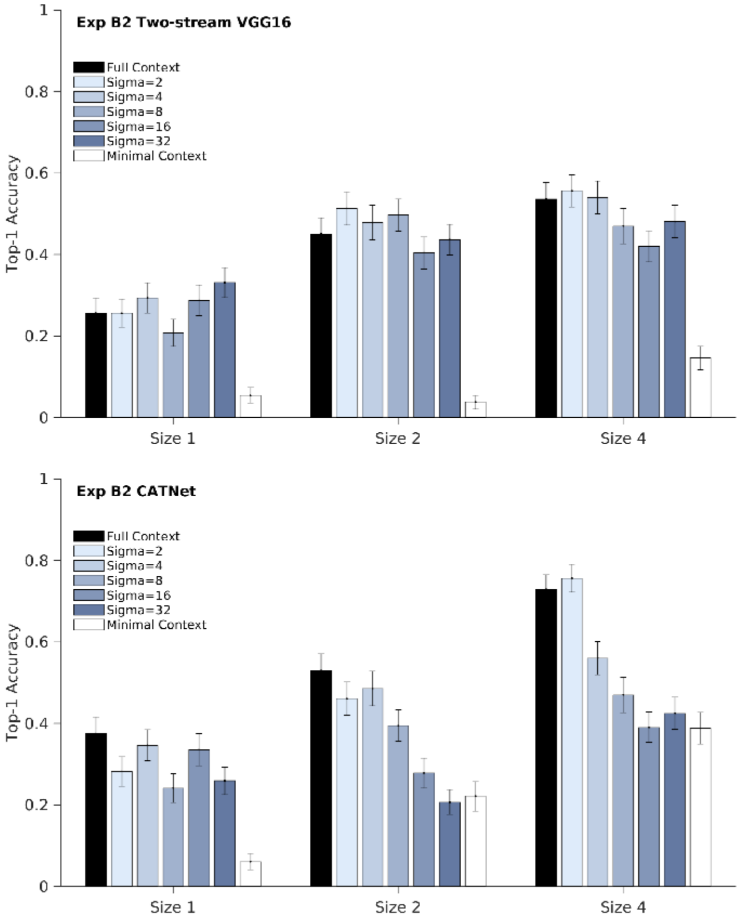

Figure S54. **Results on blurred objects in Exp B2 between Two-stream VGG16 and CATNet.** Expanding on the discussion in Sec 5.2 and Fig S4, compared with context blurring, modifying the object led to larger accuracy drops. CATNet outperforms Two-stream VGG16 for all objects sizes.

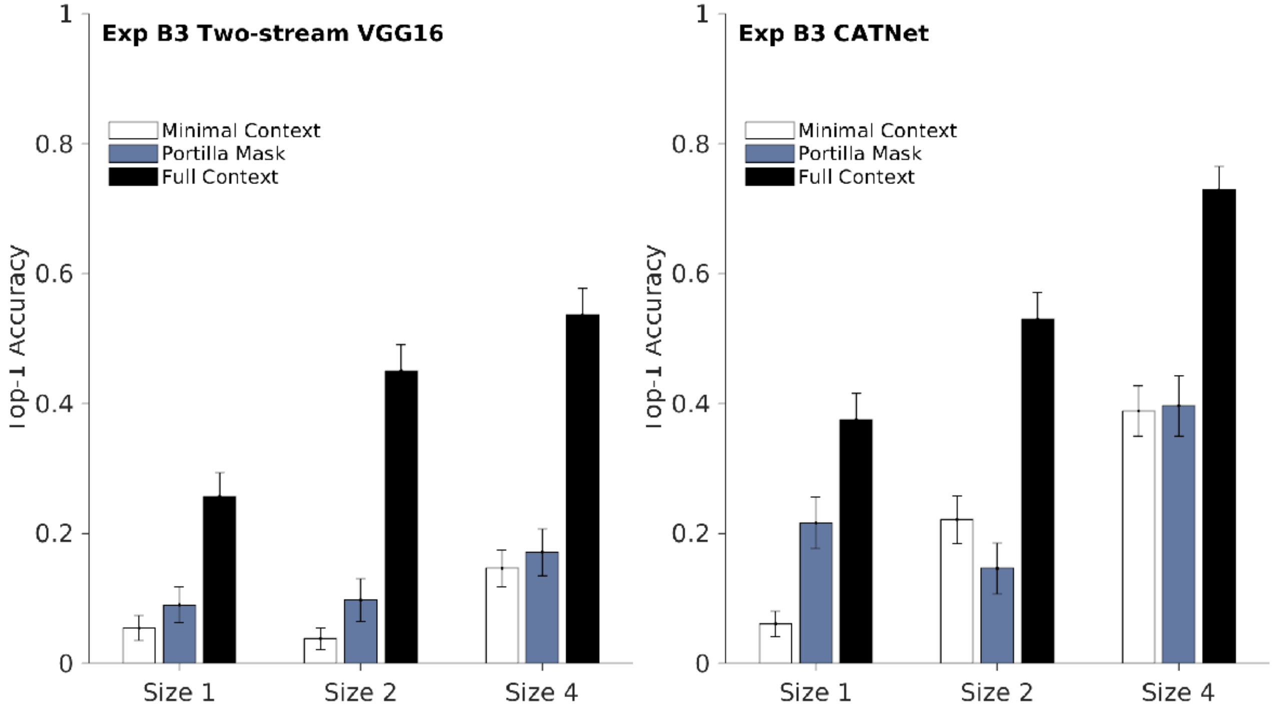

Figure S55. **Results on texture only in Exp B3 between Two-stream VGG16 and CATNet.** Expanding on the discussion in Sec 5.2 and Fig S5, low-level features did not facilitate recognition. CATNet outperforms Two-stream VGG16 for all objects sizes.

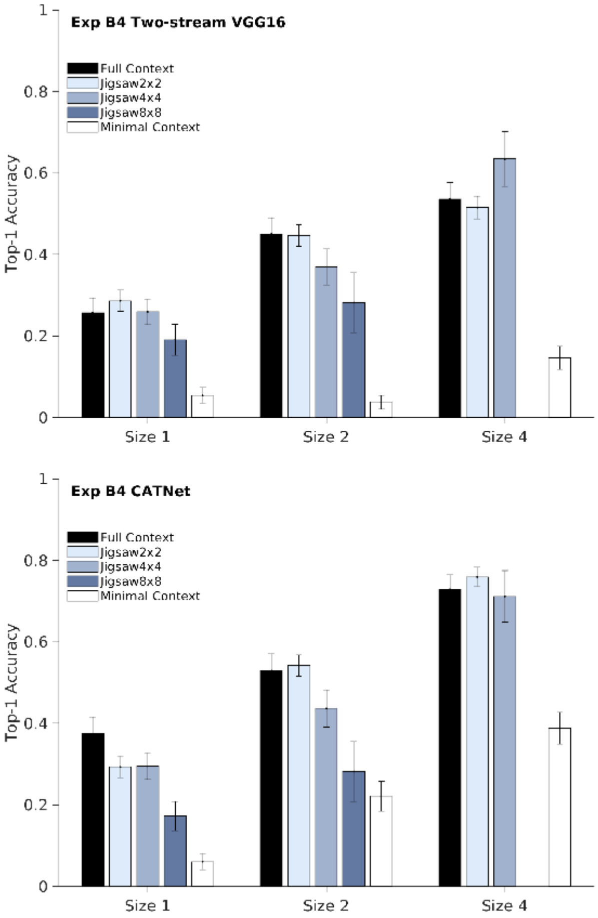

Figure S56. **Results on spatial configurations in Exp B4 between Two-stream VGG16 and CATNet**. Expanding on the discussion in Sec 5.2 and Fig 6, large geometrical context re-arrangements disrupts contextual enhancement. CATNet outperforms Two-stream VGG16 for all objects sizes.

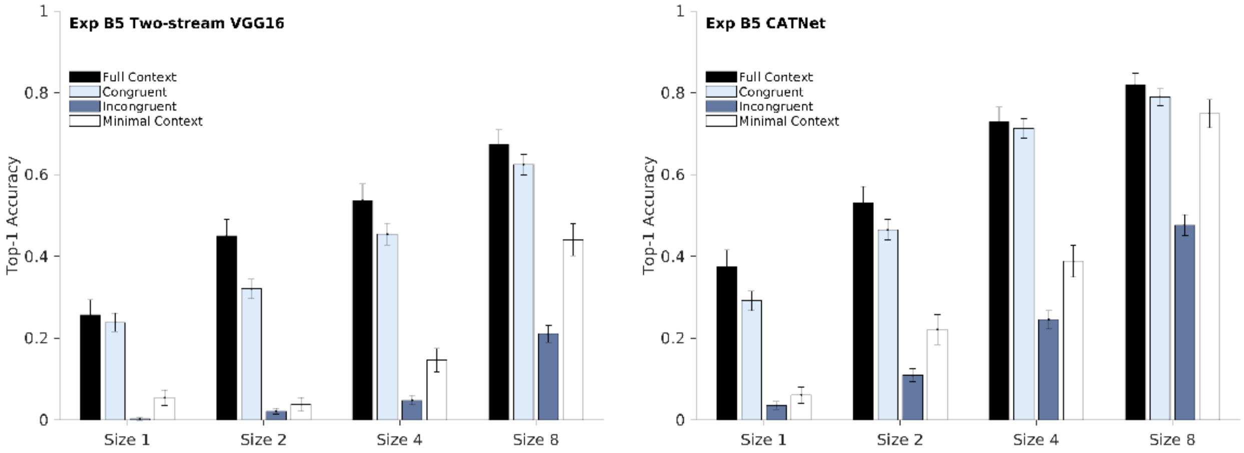

Figure S57. **Results on congruent versus incongruent context in Exp B5 between Two-stream VGG16 and CATNet**. Expanding on the discussion in Sec 5.2 and Fig 7, incongruent context impairs object recognition. CATNet outperforms Two-stream VGG16 for all objects sizes.

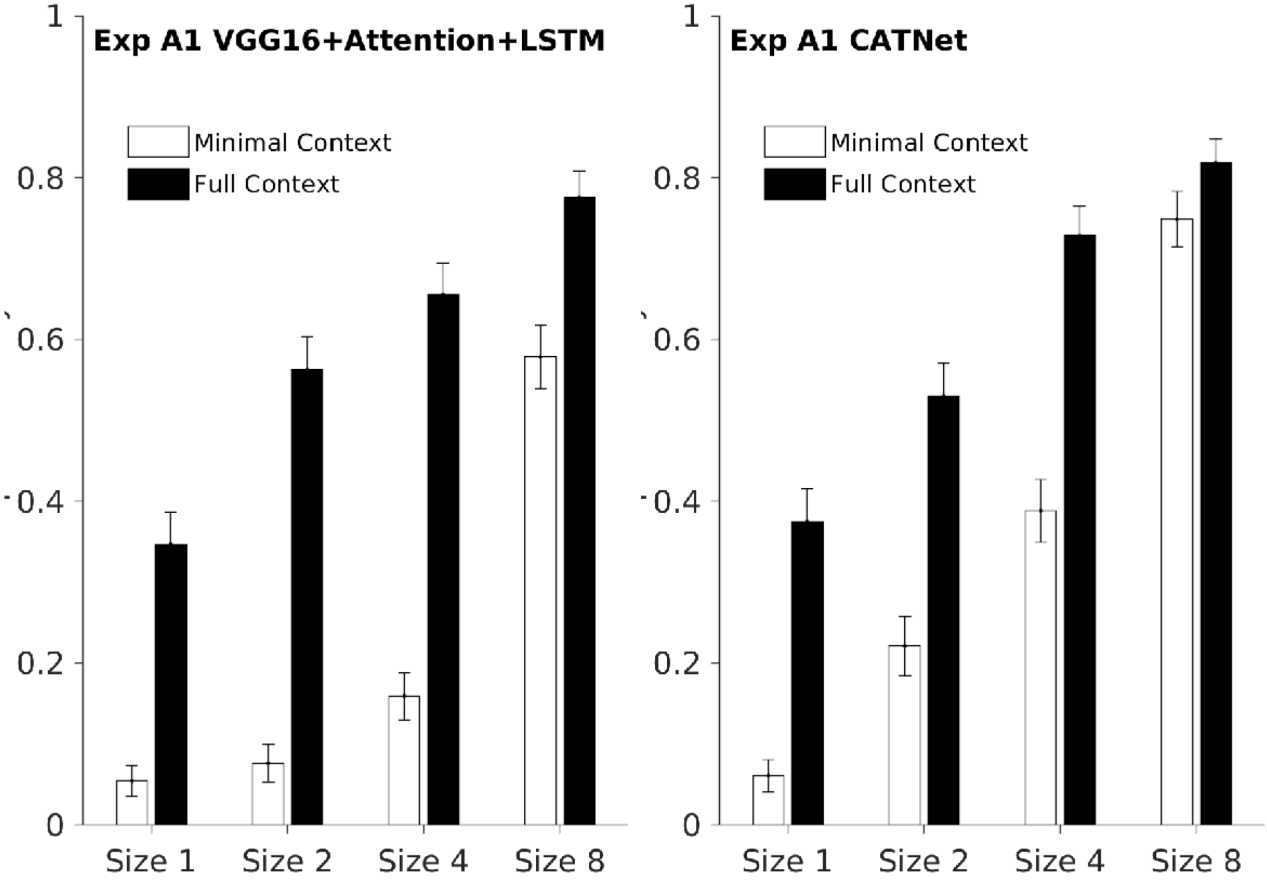

Figure S58. **Results on object size in Exp A1 between VGG+Attention+LSTM and CATNet**. Expanding on the discussion in Sec 5.1 and Fig 4, contextual modulation is stronger for smaller target objects for VGG+Attention+LSTM and CATNet. CATNet outperforms VGG+Attention+LSTM for all objects sizes.

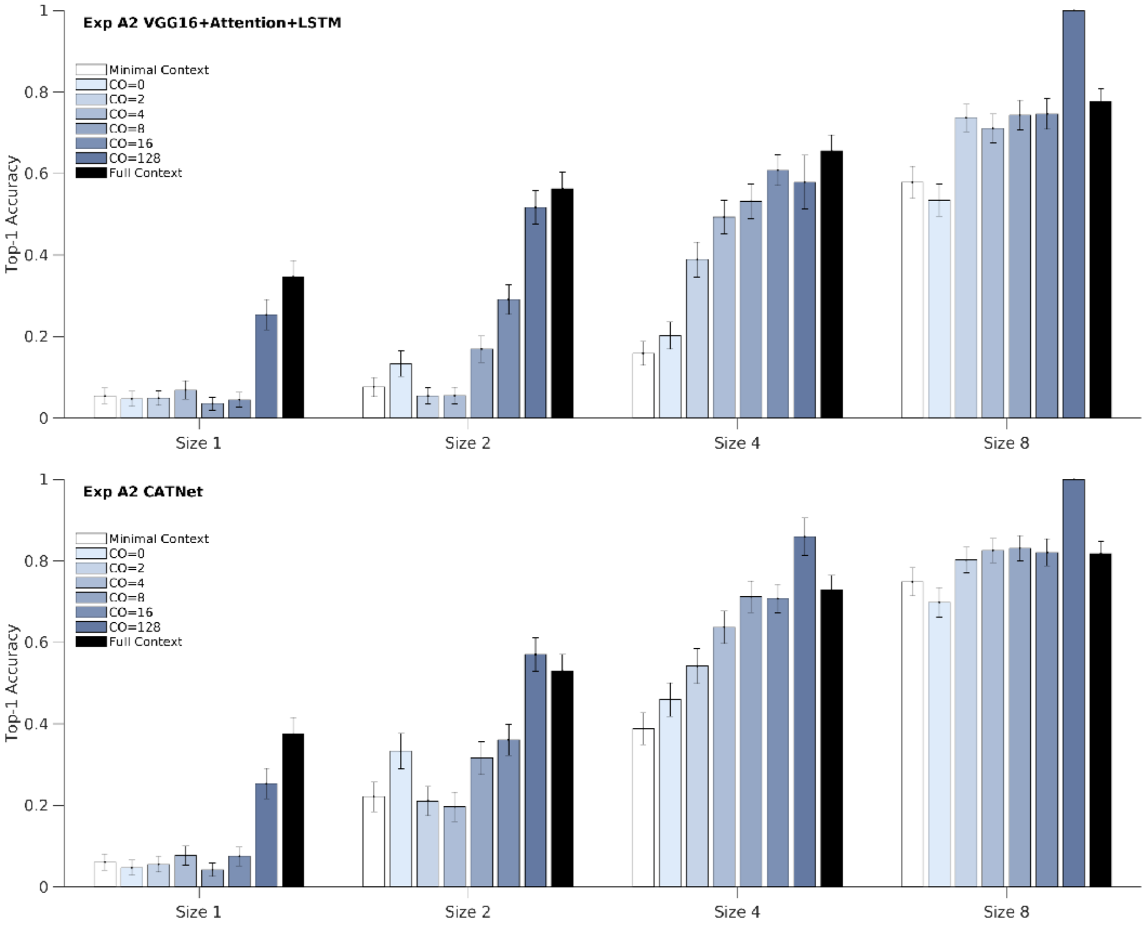

Figure S59. **Results on amount of context in Exp A2 between VGG+Attention+LSTM and CATNet**. Expanding on the discussion in Sec 5.1 and Fig S2, CATNet outperforms VGG+Attention+LSTM for smaller objects.

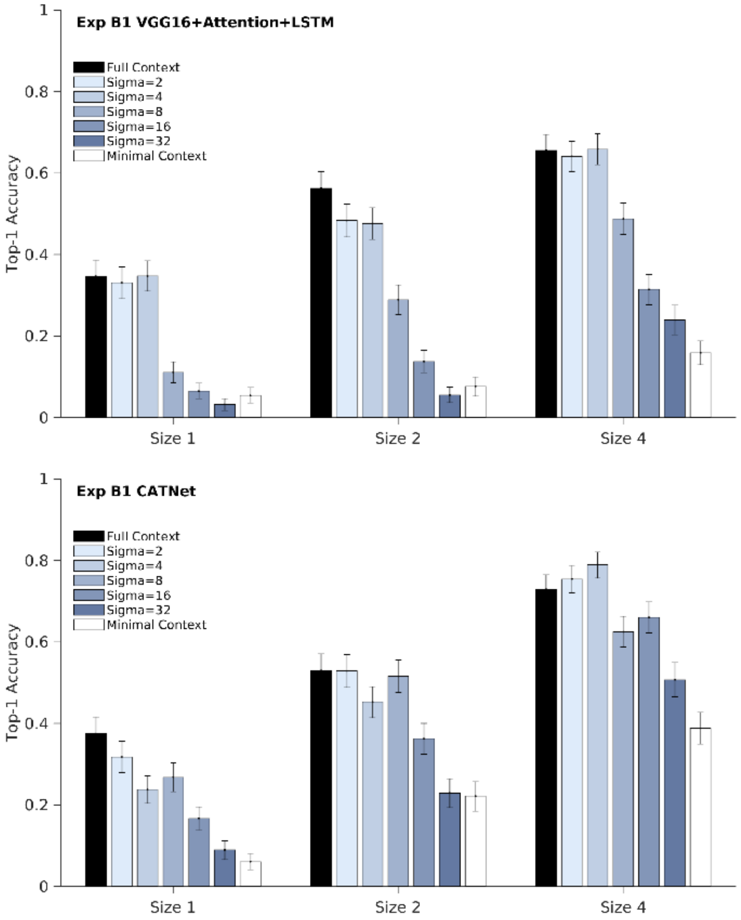

Figure S60. **Results on blurred context in Exp B1 between VGG+Attention+LSTM and CATNet.** Expanding on the discussion in Sec 5.2 and Fig 5, contextual facilitation persists even after small amounts of blurring (Exp B1). CATNet outperforms VGG+Attention+LSTM for smaller objects.

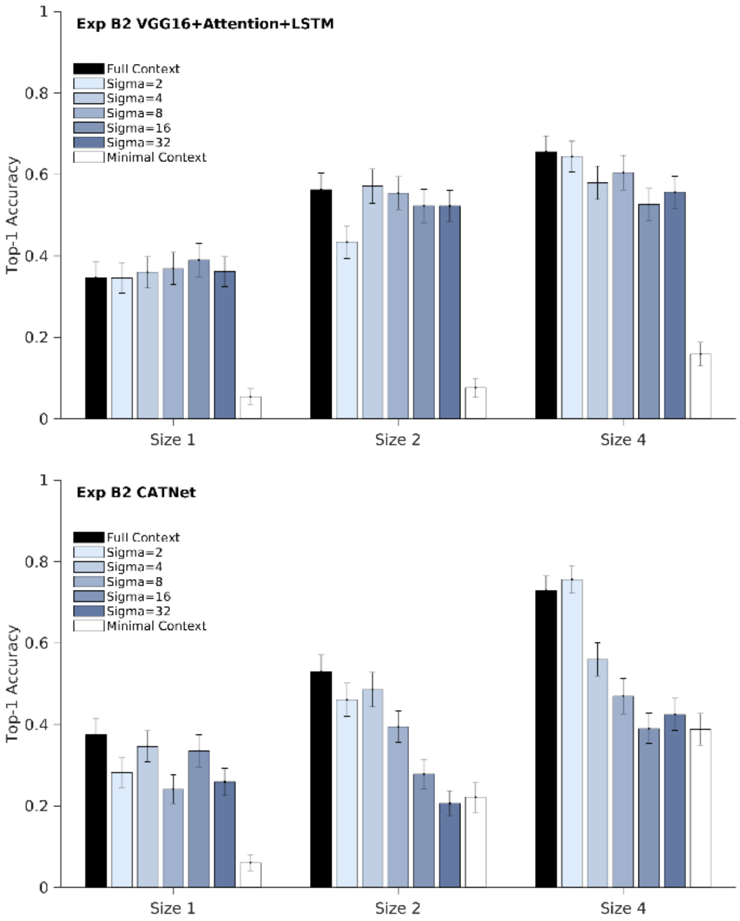

Figure S61. **Results on blurred objects in Exp B2 between VGG+Attention+LSTM and CATNet.** Expanding on the discussion in Sec 5.2 and Fig S4, compared with context blurring, modifying the object led to larger accuracy drops. CATNet performs equivalently well as VGG+Attention+LSTM for all objects sizes.

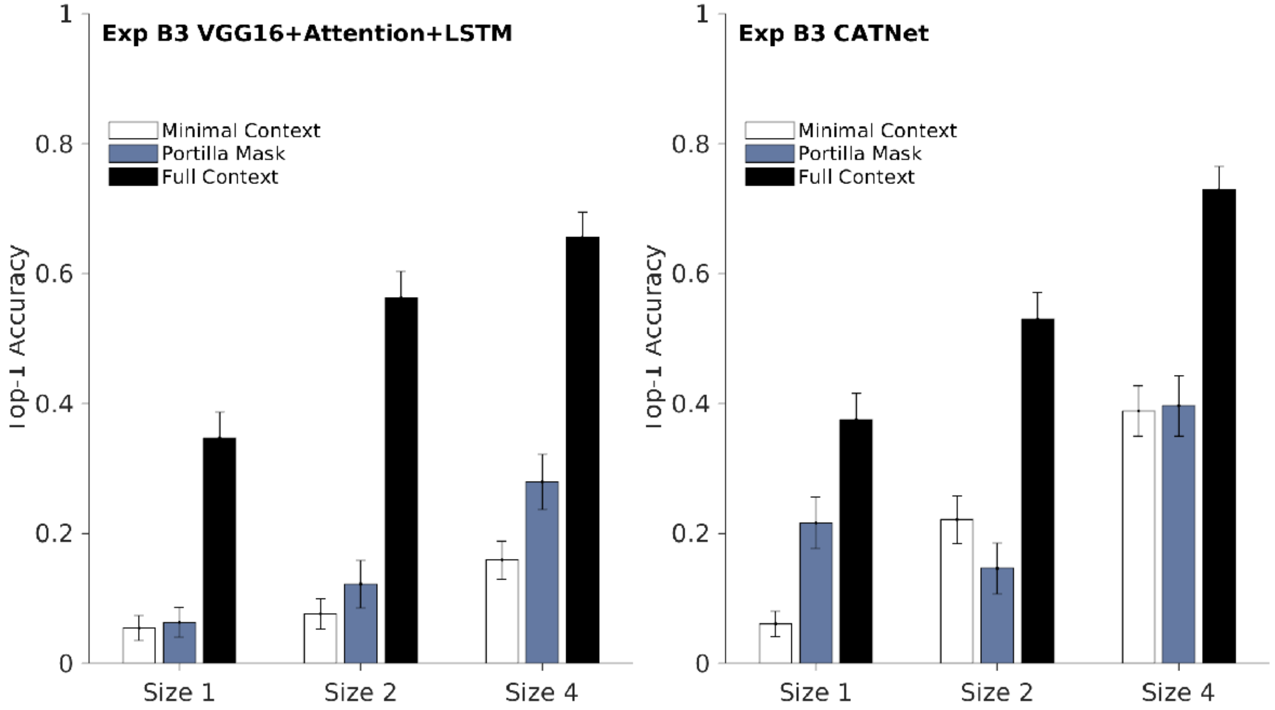

Figure S62. **Results on texture only in Exp B3 between VGG+Attention+LSTM and CATNet.** Expanding on the discussion in Sec 5.2 and Fig S5, low-level features did not facilitate recognition. CATNet outperforms VGG+Attention+LSTM for all objects sizes.

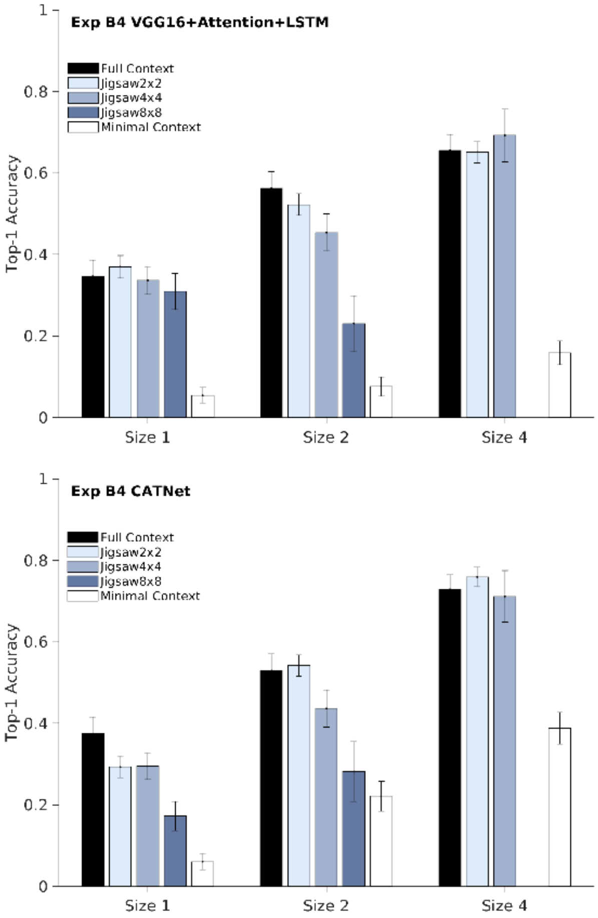

Figure S63. **Results on spatial configurations in Exp B4 between VGG+Attention+LSTM and CATNet**. Expanding on the discussion in Sec 5.2 and Fig 6, large geometrical context re-arrangements disrupts contextual enhancement. CATNet performs equivalently well as VGG+Attention+LSTM for all objects sizes.

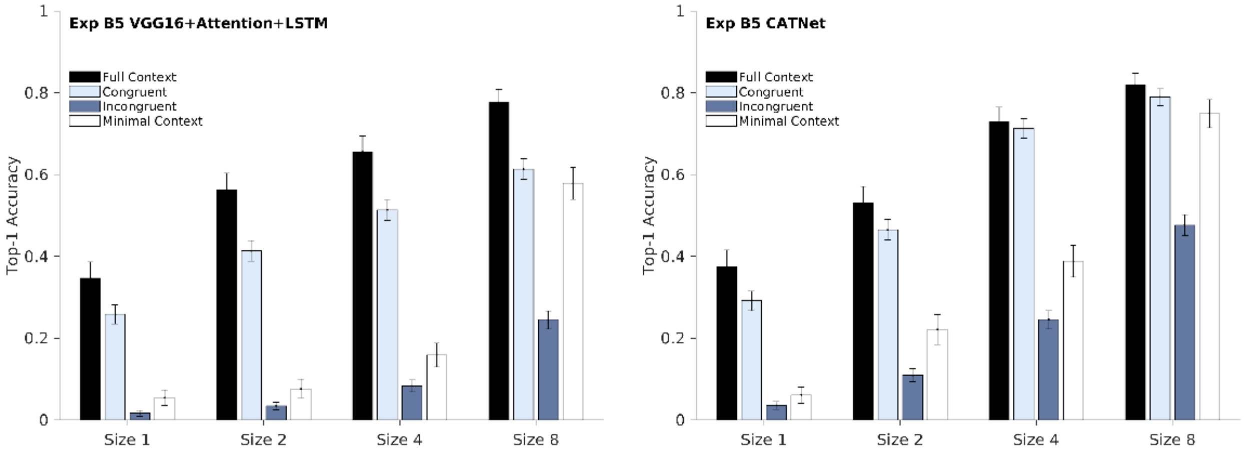

Figure S64. **Results on congruent versus incongruent context in Exp B5 between VGG+Attention+LSTM and CATNet**. Expanding on the discussion in Sec 5.2 and Fig 7, incongruent context impairs object recognition. CATNet outperforms VGG+Attention+LSTM for all objects sizes.