Putting visual object recognition in context

Mengmi Zhang[1] Claire Tseng[2] Gabriel Kreiman[1]

[1] Boston Children's Hospital, Harvard Medical School
[2] Harvard College
Address correspondence to gabriel.kreiman@tch.harvard.edu

Abstract

Context plays an important role in visual recognition. Recent studies have shown that visual recognition networks can be fooled by placing objects in inconsistent contexts (e.g. a cow in the ocean). To understand and model the role of contextual information in visual recognition, we systematically and quantitatively investigated ten critical properties of where, when, and how context modulates recognition including amount of context, context and object resolution, geometrical structure of context, context congruence, time required to incorporate contextual information, and temporal dynamics of contextual modulation. The tasks involve recognizing a target object surrounded with context in a natural image. As an essential benchmark, we first describe a series of psychophysics experiments, where we alter one aspect of context at a time, and quantify human recognition accuracy. To computationally assess performance on the same tasks, we propose a biologically inspired context aware object recognition model consisting of a two-stream architecture. The model processes visual information at the fovea and periphery in parallel, dynamically incorporates both object and contextual information, and sequentially reasons about the class label for the target object. Across a wide range of behavioral tasks, the model approximates human level performance without retraining for each task, captures the dependence of context enhancement on image properties, and provides initial steps towards integrating scene and object information for visual recognition.

1. Introduction

The tiny object on the table is probably a spoon, not an elephant. Objects do not appear in isolation. Instead, they co-vary with other objects and scene properties, their sizes and colors usually respect regularities relative to nearby elements, and objects tend to appear at stereotypical locations. The success in object recognition and detection

Figure 1. **Mis-classification of objects in unfamiliar contexts**. State-of-the-art deep visual recognition networks, such as InceptionV3 [42], ResNet50 [53] and VGG16 [40], make mistakes when the context is incongruent. The top-5 labels and confidence levels by each model are shown on the right.

tasks in natural images relies on *implicit* incorporation of contextual information. Deep convolutional neural networks jointly learn statistical associations between objects, image properties, and labels [12, 41, 17, 6]. Such algorithms can be tricked into mislabeling or missing an object by placing it in an unfamiliar context (Fig. 1).

Here systematically and quantitatively investigated the mechanisms by which contextual information is integrated into visual recognition. We focus on three fundamental aspects of context: [A] the interaction between object size and the amount of contextual information; [B] the geometry, resolution, and content of contextual information; [C] the temporal dynamics of contextual modulation and the interaction between bottom-up and recurrent computations during contextual modulation. By systematically measuring the effect of context in 10 human psychophysics experiments (Fig. 2, Fig. S9, S10, S11), we gain a quantitative understanding of where, when, and how context modulates recognition. Moreover, the human data provides a quantitative benchmark and constrain to test (but not train) computational models.

Inspired by the neuroscience of human vision, we propose Context-aware Two-stream Attention network

(CATNet). The proposed model makes inferences about the target object by guiding attention towards regions with informative contextual cues and object parts via dynamic integration of foveal (object) and peripheral (context) vision, and automatically learning contextual reasoning strategies. We test CATNet and state-of-the-art in-context object recognition models on the same exact psychphysics tasks *without re-training the models for each experiment*. CATNet surpasses other computational models in these experiments and shares remarkable similarity with human recognition abilities.

2. Related Works

2.1. Role of Context in Human Visual Recognition

Many behavioral studies [4, 20] have focused on comparing congruent versus incongruent context conditions: objects appearing in a familiar background can be detected more accurately and faster than objects in an unusual scene (Fig. 1). Several qualitative demonstrations showed that context can help visual processing [2, 7, 25, 1], during recognition tasks [2], detection tasks [7, 25], working memory [18, 1], and visual search [23]. Here we systematically tested the three fundamental properties of context to quantitatively model where, when and how contextual information modulates recognition.

2.2. Role of Context in Computer Vision

Contextual reasoning about objects and relations is critical to machine vision. Some studies show deep nets for object recognition, trained on natural image datasets, *e.g*. ImageNet [28], indeed rely implicitly but strongly on context [19, 8]. These algorithms can fail when objects are placed in an incongruent context ([6, 17, 12]) (Fig. 1).

Many exciting successes of computer vision methods can be partly ascribed to capitalizing on the statistical correlations between contextual information and object labels. Here we briefly and non-exhaustively introduce context-aware computational models in various applications. Qualitative analyses based on the statistical summary of object relationships, have provided an effective source of information for perceptual inference tasks, such as object detection ([46, 34, 24, 47, 32]), scene classification ([21, 48, 52]), semantic segmentation ([52]), and visual question answering ([44]).

Classical approaches, *e.g*. Conditional Random Field (CRF), reason jointly across multiple computer vision tasks in image labeling, scene classification [21, 52, 29, 10], object detection and semantic segmentation [33]. Several graph-based methods incorporating contextual information, combined with neural network architectures, have been successfully applied in object priming [46], place and object recognition [50, 48], object detection [11, 32], and visual question answering [44]. Recent interesting approaches have used deep graph neural networks for contextual inference [26, 13, 15, 5]. These works typically assume that full contextual information is always available. However, in our experiments, we include experimental conditions where partial contextual information is available, such as minimal context, blurred context and only low-level context texture (Figure 2). Breaking away from these previous works where graph optimization is performed globally, our proposed model selects important visual features using an attention mechanism and integrates partial information from both the target object and the context over multiple steps, and, importantly, generalizes to context variations (Section 5). Furthermore, we provide a direct comparison against human benchmark performance.

3. Human psychophysics experiments

We examined three fundamental properties of contextual modulation in visual recognition (Fig. 2) by conducting 10 experiments, schematically illustrated in Fig. 2h, on Amazon Mechanical Turk [49]. We recruited 80 subjects per experiment, yielding a total of $64,000$ trials.

3.1. Experiment setup

The stimuli consisted of 2,259 images spanning 55 object categories from the test set of MSCOCO Dataset [30]. We constrained the size of target objects to four bins (in pixels): Size 1 [16-32], Size 2 [56-72], Size 4 [112-144], and Size 8 [224-288]. Given the stimulus size of 1024×1280 pixels and viewing distance of 0.5 meters, these values correspond to about 1, 2, 4, and 8 degrees of visual angle; but this may vary in MTurk depending on viewing conditions. To avoid any biases and potential memory effects, we took the following precautions: (a) Only one target object was selected per image; (b) Target objects were uniformly distributed over the 4 sizes and 55 categories; (c) Subjects saw at most 2 target objects per category; (d) The trial order was randomized.

3.2. Detailed description of each experiment

Experiment A: Context quantity.
We investigated the interaction between the object size and the amount of context in two experiments.
Exp A1, Object size. We conjectured that the impact of contextual information would depend on the target object size. We considered 4 object sizes as above. For each size, we introduced either minimal context (tightest rectangular bounding box enclosing the object, Fig. 2b) or full context (the entire image, Fig. 2a).
Exp A2, Amount of context. For each object size, we systematically titrated the amount of contextual information (Fig. 2c). The context-object ratio (CO) is the total image

Figure 2. **Fundamental properties of context and task schematic**. Example image with full context (**a**) and image modifications used in experiments (more examples in Fig. S8). The target location (red box) is always the same across conditions. The correct answer ("mouse") is not shown in the actual experiment). (h) Subjects were presented with a fixation cross (500 ms), followed by a bounding box indicating the target object location (1000 ms). In most experiments (except for Exp C1-3), the image was shown for $T = 200$ ms. After image offset, subjects typed one word to identify the target object.

area *excluding* the target object divided by the object size. We included CO=0 (no pixels surrounding the object), 2, 4, 8, 16, and 128. Some combinations of large sizes and large CO values may not be possible.

Experiment B: Context content.

We studied how context resolution, geometry, and congruency modulated recognition in 5 experiments. Unless stated otherwise, we focused on sizes 1, 2 and 4, minimal and full context.

Exp B1, Blurred context. Human vision shows strong eccentricity dependence (high resolution in the fovea and progressively lower resolution toward the periphery). To quantify the impact of context resolution on recognition, only the context was blurred (Fig. 2d) using a zero-mean Gaussian with standard deviation $\sigma = 2, 4, 8, 16, 32$ pixels (image size = 1024×1280 pixels).

Exp B2, Blurred object. To compare the effect of blurring the context versus the target object, we applied the same Gaussian blurring only to the object itself.

Exp B3, Texture only. We constructed textures constrained by the image statistics [35], and pasted the intact object on them (Fig. 2e). The textures preserve low-level features, but distort high-level features and semantic information.

Exp B4, Jigsaw context. To investigate the impact of the geometrical properties of context, we divided the image into 2×2, 4×4, and 8×8 "jigsaw" pieces (Fig. 2f). The piece containing the target object remained in the same position as in the original image, and the other pieces were randomly scrambled. We discarded cases when the object occupied more than one piece. For size 8, it was not possible to have the 8x8 condition.

Exp B5, (In)congruent context. To examine the importance of context consistency in recognition, we pasted objects in different backgrounds by considering congruent object-context pairs (object and context belong to the same class label), and incongruent object-context pairs (context taken from a different image class label) (Fig. 2g).

Experiment C: Dynamics of contextual modulation.

We investigated the temporal dynamics of contextual effects in 3 experiments.

Exp C1, Exposure time. In experiments **A** and **B**, the image duration T was 200 ms (Fig. 2h). Here we systematically varied T to be 50, 100, or 200 ms (Fig. S9).

Exp C2, Backward masking. Backward masking is a technique commonly used in neuroscience to interrupt visual processing [43]. The mask shown after stimulus offset is purported to block top-down and recurrent computations. We used Portilla masks [35] as in **Exp B3** (Fig. S10). The stimulus exposure times followed those in **Exp C1**.

Exp C3, Asynchronous context presentation. In all experiments above, object and context information were presented synchronously. During natural vision, subjects move their eyes from location P1 to location P2. The information gathered while fixating at P1 acts as a prior temporal context of fixation at P2. To investigate the effect of such prior temporal context in recognition, while conceptually simplifying the problem, we split the image into context-only and object-only parts. First, the context-only part was presented for a duration of 25, 50, 100, or 200 ms. Next, the context was removed, and the object-only part was presented for a duration of 50, 100, or 200 ms (Fig. S11). The synchronous conditions were also included for comparison purposes.

3.3. Performance evaluation and statistics

Most recognition experiments enforced N-way categorization (e.g., [43]). Here we introduced a more unbiased and natural probing mechanism whereby there were no constraints on what words subjects could use to describe the target object (Fig. 2h). To evaluate human performance, we *separately* collected a distribution of ground truth answers for each target object (Mturk subjects *not* participating in the main experiments). Though computational models were evaluated using N-way categorization, we still find it instructive to plot computational results alongside human behavior for comparison purposes. Moreover, relative changes and

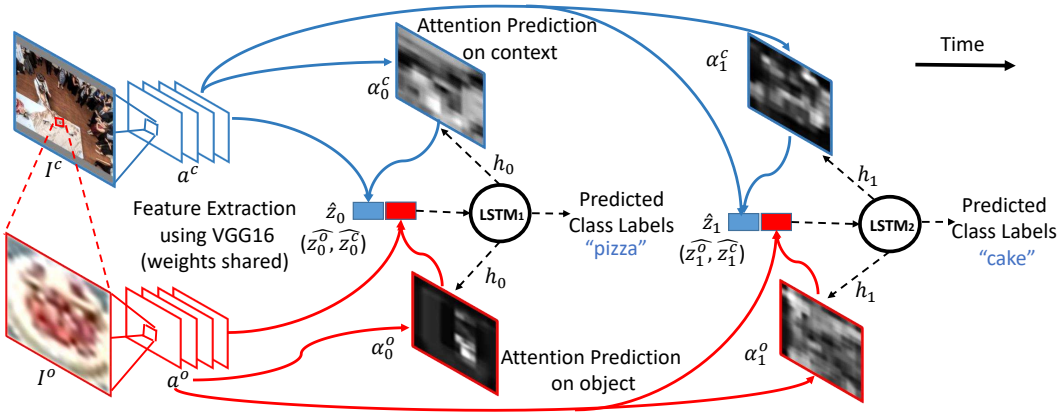

Figure 3. **Architecture overview of Context-aware Two-stream Attention network (CATNet)**. The diagram depicts the iterative modular steps carried out by CATNet over multiple time steps in the context-aware object recognition task. CATNet consists of 3 main modules: feature extraction, attention, and recurrent memory. These three modular steps repeat until a pre-specified number of time steps T_m. For illustrative purposes, only the first and second time steps in a trial are shown here (Section 4 for definition of variables and Fig. S6 and S7 for implementation details of the attention and LSTM modules. CATNet is *only* trained using full context natural images and then it is tested in different conditions specified by each experiment (Section 3.1).

trends in humans can be directly compared to computational results. For human-model, within-human and within-model comparisons, we used the Wilcoxon ranksum test [22], and one-way or two-way ANOVA tests [27] (Supp. Material).

4. Context-aware Two-stream Attention Net

We propose a Context-aware Two-stream Attention network (CATNet), extending previous work on image captioning [51]. CATNet is presented with the stimulus, a natural image where the target object is indicated by a white bounding box. Inspired by the eccentricity dependence of human vision, CATNet has one stream that processes only the target object (I^o, minimal context, Fig. 2b but without the gray background) and a second stream that processes the contextual information in the periphery (I^c, full context, Fig. 2a). The two streams are processed through weight-sharing convolutional neural networks in parallel. I^o is enlarged to be the same size as I^c, such that each convolutional kernel sees I^o at finer-grain details.

CATNet explicitly integrates the fovea and periphery via concatenation and makes a first attempt to predict a class label y_0 out of a pre-defined set of $C = 55$ object classes. Since horizontal and top-down connections pervasive throughout brain cortices presumed to be important for recognition [43], we add a recurrent LSTM module in CATNet to iteratively reason about context. The LSTM module constantly modulates its internal representation of the scene via attention and outputs predicted class labels over multiple time steps t where $t \in \{1, ...T_m\}$. These attention-modulated features maps of I^c and I^o are functions of t. For simplicity in naming conventions, we use superscript to denote c or o in all variables to distinguish visual processes on I^c or I^o respectively and use subscript t to denote time-dependent variables.

4.1. Convolutional Feature Extraction

CATNet takes I^c and I^o as inputs and uses a feed-forward convolutional neural network to extract feature maps a^c and a^o, respectively. We use the VGG16 network [40], pre-trained on ImageNet [14] and fine-tune it at the training stage. To focus on specific parts of the image and select features at those locations, we preserve the spatial organization of features; thus, CATNet uses the output feature maps at the last convolution layer of VGG16. The parameters of both feed-forward feature extractor networks on I^c and I^o are shared. Since I^o is the enlarged version of the target object region in I^c, this results in higher acuity and enhances sensitivity to details of the target object. We describe a_c next but the same ideas apply to a_o.

A feature vector $\mathbf{a_i^c}$ of dimension D represents the part of the image I^c at location i, where $i = 1, .., L$ and $L = W \times H$, and W and H are the width and height, respectively, of the feature map:

$$a^c = \{\mathbf{a_1^c}, ..., \mathbf{a_L^c}\}, \quad \mathbf{a_i^c} \in \mathbb{R}^D \tag{1}$$

4.2. Attentional Modulation

We use a "soft-attention" mechanism as introduced by [3] to compute "the context gist" $\mathbf{\hat{z}_t^c}$ on I_c and "the object gist" $\mathbf{\hat{z}_t^o}$ on I_o (Fig. S6). There are two attention maps on I^c and I^o respectively where each stream has identical architectures but different weight parameters. We describe the context stream of attention but the same principles apply to the object attention map. For each location i in a^c, the attention mechanism generates a positive scalar α_{ti}^c, representing the relative importance of the feature vector $\mathbf{a_{ti}^c}$ in capturing the context gist. α_{ti}^c depends on the feature

vectors $\mathbf{a_i^c}$, combined with the hidden state at the previous step $\mathbf{h_{t-1}}$ of a recurrent network described below:

$$e_{ti}^c = A_h^c \mathbf{h_{t-1}} + A_a^c \mathbf{a_i^c}, \quad \alpha_{ti}^c = \frac{\exp(e_{ti}^c)}{\sum_{i=1}^{L} \exp(e_{ti}^c)} \tag{2}$$

where $A_h^c \in \mathbb{R}^{1\times n}$ and $A_a^c \in \mathbb{R}^{1\times D}$ are weight matrices initialized randomly and learnt during training. Because not all attended regions might be useful for context reasoning, the soft attention module also predicts a gating vector β_t^c from the previous hidden state h_{t-1}, such that β_t^c determines how much the current observation contributes to the context vector at each location: $\beta_t^c = \sigma(W_\beta^c \mathbf{h_{t-1}})$, where $W_\beta^c \in \mathbb{R}^{L\times n}$ is a weight matrix and each element β_{ti}^c in β_t^c is a gating scalar at location i. As also noted by [51], β_t^c helps put more emphasis on the salient objects in the images. Once the attention map α_t^c and the gating scale β_t^c are computed, the model applies the "soft-attention" mechanism to compute $\widehat{\mathbf{z}}_\mathbf{t}^\mathbf{c}$ by summing over all the L regions in the image:

$$\widehat{\mathbf{z}}_\mathbf{t}^\mathbf{c} = \sum_{i=1}^{L} \beta_{ti}^c \alpha_{ti}^c \mathbf{a_i^c} \tag{3}$$

We define $\widehat{\mathbf{z}}_\mathbf{t} = (\widehat{\mathbf{z}}_\mathbf{t}^\mathbf{c}, \widehat{\mathbf{z}}_\mathbf{t}^\mathbf{o})$ as concatenation of $\widehat{\mathbf{z}}_\mathbf{t}^\mathbf{c}$ and $\widehat{\mathbf{z}}_\mathbf{t}^\mathbf{o}$, which is used as input to the LSTM module described next. The attention module is smooth and differentiable, and CATNet can learn all the weight matrices in an end-to-end fashion via back-propagation.

4.3. Recurrent Connections using LSTM

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 I^o and I^c. Our implementation of LSTM closely follows [54] (Fig. S7). 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.

To compare CATNet and human performance when exposure time T changes (**Exp. C**), we set one time step in the LSTM to correspond to 25 ms and considered the predicted class labels of CATNet at the corresponding number of time steps $T_m = T/25$ as the answers.

To predict the class label y_t for the target object, the LSTM computes a classification vector where each entry denotes a class probability given the hidden state h_t:

$$y_t = \arg\max_c p(y_c), \quad p(y_c) \propto L_h \mathbf{h_t} \tag{4}$$

where $L_h \in \mathbb{R}^{C\times n}$ is a matrix of learnt parameters initialized randomly.

4.4. Training and Implementation Details

We trained CATNet end-to-end by minimizing the cross entropy loss between the predicted label y_t at each time step t and the ground truth label x:

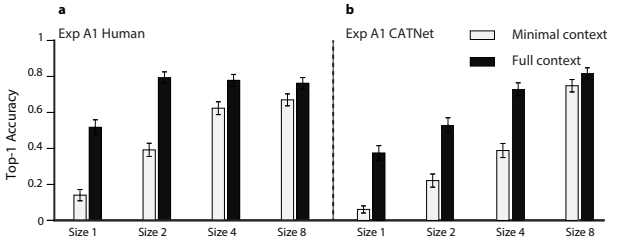

Figure 4. **Contextual modulation is stronger for smaller target objects (Exp A1)**. Top-1 accuracy increases with object sizes (Fig. 2a-b) and contextual information increases accuracy, particularly for small target objects, for humans and CATNet.

$$LOSS = \sum_{t=1}^{T_m} (-\log(P(y_t|x))) \tag{5}$$

We used all images from the MSCOCO training set for training and validating all models. On every training image, each object can be selected as the target object and they are always in shown in full context. Only at the testing stage, we vary the context based on different conditions in each experiment as described in Section 3.1. Importantly, none of the human behavioral experiments are used to train the model. The input image size (both I^c and I^o) was 400×400 pixels. We set the total number of time steps $T_m = 8$ for training CATNet. Further implementation details are provided in the Supp. Material.

Data and code availability: All source code, and the data from the psychophysics experiments will be released publicly upon publication.

4.5. Competitive baselines and ablated models

We compared the results of CATNet against several competitive baselines, such as DeepLab-CRF [9] in semantic segmentation and YOLO3 [36, 37] in object detection. These models were adapted to the context-aware object recognition task (Supp. Material).

To study the role of attention, the two-stream architecture, and recurrent connections, we also introduced a series of ablated versions of CATNet. Starting from original VGG16 object recognition network [40] pre-trained on ImageNet [14] (VGG16 on cropped objects), we added in one component at a time and evaluated their incremental performance change. These models include VGG16 + binary mask, two-stream VGG16, VGG16 + attention, and VGG16 + attention + LSTM.

5. Results

5.1. Object and context size matter

For the minimal context condition (Fig. 2b), human performance improved monotonically as a function of

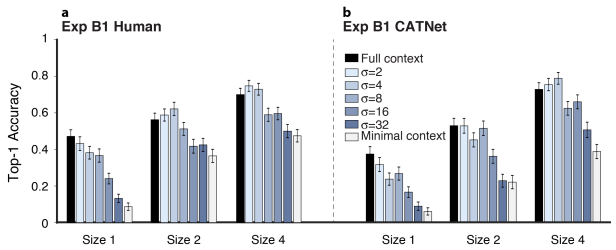

Figure 5. **Contextual facilitation persists even after small amounts of blurring (Exp B1)**. A large amount of context blurring (Fig. 2d) is required to disrupt the recognition enhancement for humans and CATNet.

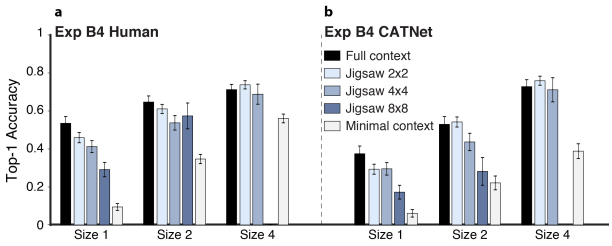

Figure 6. **Large geometrical context re-arrangements disrupts contextual enhancement (Exp B4)**. Scrambling context pieces (Fig. 2f) reduces the contextual enhancement only when many small context pieces are changed, both for humans and CATNet.

object size from 0.14 ± 0.031 to 0.67 ± 0.035 (**Exp A1**, Fig. 4, one-way ANOVA: $F(3, 5097) = 215$, $p < 10^{-15}$). This effect was readily captured by the CATNet model (one-way ANOVA: $F(3, 4368) = 304$, $p < 10^{-15}$). Adding full contextual information (Fig. 2a) led to a large improvement in performance both for humans and CATNet. Contextual modulation strongly depends on object size: the performance ratio between the full context and minimal context conditions was 4.7 and 2.5 (humans and CATNet, respectively) for object size 1, whereas the ratio was 1.1 and 1.05 (humans and CATNet, respectively) for object size 8. Contextual information greatly facilitates recognition when the target objects are small and hard to recognize.

We further quantified how the amount of contextual information impacts recognition by titrating the context object ratio (CO) from 0 to 128 (**Exp A2**, Fig. S1). The amount of context is important both for humans (one-way ANOVA: $F(7, 5097) = 31$, $p < 10^{-15}$) and CATNet (one-way ANOVA: $F(7, 4368) = 23$, $p < 10^{-15}$).

Across all the CO ratios, humans outperformed CATNet for small object sizes and CATNet outperformed humans for the largest object size. Of note, CATNet was *never* trained or fine-tuned with any human data. These experiments demonstrate that the context *quantity* has a strong impact on recognition.

5.2. Blurred context is sufficient for recognition

Due to strong eccentricity dependence of human vision, peripheral information has less resolution than the fovea. In fact, the resolution drops so sharply that humans are legally blind in the far periphery. We conjectured that low resolution context could be sufficient to facilitate recognition. To test this conjecture, we applied different amounts of blurring in the context (**Exp B1**, Fig. 2d).

Human recognition accuracy dropped with the amount of blurring from levels indistinguishable from the full resolution condition when $\sigma \leq 8$ pixels all the way to levels indistinguishable from the minimal context condition when $\sigma = 32$ pixels (Fig. 5, one-way ANOVA: $F(4, 2933) = 28$, $p < 10^{-15}$). Interestingly, there was a wide range of blurring that led to robust context modulation, consistent with the notion that humans do not require full resolution context for recognition. The effects of blurring were also captured by CATNet, where contextual enhancement disappeared only when using large σ values (one-way ANOVA: $F(4, 2354) = 2$, $p < 0.05$). Similar with the results in **exp A1** and **exp A2**, humans outperformed CATNet on small objects.

We also compared the effects of blurring the object itself without blurring the context (**Exp. B2**, Fig. S2). Although the total number of pixels affected by blurring the target object is much smaller than blurring the context (for a fixed σ), modifying the object led to larger accuracy drops, for object sizes 2 and 4 both for humans and CATNet.

5.3. Contextual effects rely on spatial configuration

Another important aspect of context is the relative position of objects and features in the image; *e.g.*, the sky is often at the top under natural viewing conditions. To evaluate how the spatial configuration of context impacts recognition accuracy, we scrambled the images into various numbers of jigsaw pieces while the piece containing the target object remained in the same position as in the original image (**Exp B4**, Fig. 2f). Both humans and CATNet relied on the spatial configuration of context over all object sizes (humans: one-way ANOVA: $F(3, 2182) = 58$, $p < 10^{-15}$; CATNet: one-way ANOVA: $F(3, 1787) = 29$, $p < 10^{-15}$). The inconsistent spatial configuration of contextual information in the 4×4 and 8×8 configurations led to a reduction in recognition accuracy. Interestingly, accuracy in the 2×2 configuration was not significantly different from the unscrambled full context condition, probably because each large piece itself already contains sufficient contextual information or the effect of context reasoning decreases with increasing distance to the target object [55].

CATNet was more robust to the distorted spatial configurations: recognition accuracy differed from the full-context condition only for the 8×8 configuration (for 2×2 and 4×4, two-tailed ranksum test, $p \geq 0.12$).

5.4. Bad context is worse than no context

Given that the moderately blurred context still retained its effects on recognition (Fig. 5), we asked whether the contextual effects could still be elicited using low-level texture features from the images. We tested this possibility by pasting objects on Portilla textures constrained by the image statistics (**Exp B3**, Fig. 2e).

Low-level texture features did not facilitate object recognition for either humans or CATNet (Fig.S3). In fact, human performance was actually slightly impaired when objects were embedded within these textures compared to the minimal context condition (two-tailed ranksum test, all object sizes, $p < 0.04$). For CATNet, low-level texture features improved recognition with respect to minimal context only for object size 1, but the effect was much smaller than when using full contextual information.

Given that low-level textures did not help (and could even hurt recognition), and inspired by Fig. 1, we next studied recognition when objects were removed from their original images and placed in the same location but in different images: congruent contexts (images with same class labels) or incongruent contexts (images with different class labels, Fig. 2g).

Congruent contexts enhanced recognition for smaller object sizes compared to the minimal context condition both for humans and CATNet (Fig. 7). The facilitation elicited by congruent context was lower than that in the original full context. Although congruent contexts typically share similar correlations between objects and scene properties, pasting the object in a congruent context did not lead to the same enhancement. This may be due to the erroneous relative size between objects, the unnatural boundaries created by pasting, or important contextual cues specific to each image. Interestingly, CATNet was relatively oblivious to these effects and performance in the congruent condition was closer to that in the original full context condition.

In high contrast with these observations, incongruent contexts consistently degraded recognition performance below the minimal context condition. Across all object sizes, subjects showed higher accuracy for objects in congruent versus incongruent contexts (one-way ANOVA: $F(1, 2530) = 92$, $p < 10^{-15}$). Accuracy was lower for incongruent context than minimal context (two-tailed ranksum test, $p = 0.0005$). Similarly, CATNet recognition accuracy also positively correlated with congruent context (one-way ANOVA: $F(1, 2977) = 515$, $p < 10^{-15}$) and was degraded by incongruent context (for all object sizes, two-tailed ranksum test, $p < 0.001$).

5.5. Temporal dynamics of contextual modulation

The dynamics of recognition places strong constraints to interpret the flow of bottom-up and top-down visual processes [45, 43, 38]. We conducted 3 experiments to examine the dynamics of contextual effects on recognition.

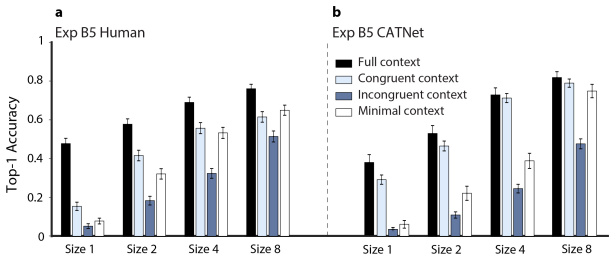

Figure 7. **Incongruent context impairs recognition.** Pasting the target objects in different but congruent contexts facilitates recognition. Pasting the target objects in incongruent contexts (Fig. 2g) impairs recognition, both for humans and CATNet.

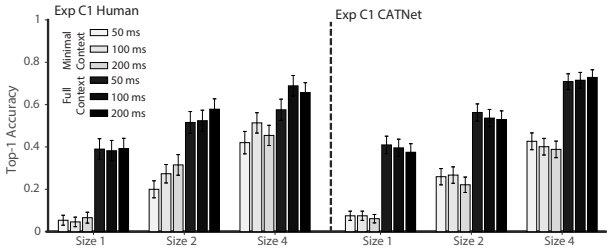

Figure 8. **Stimulus exposure time has little effect in recognition (Exp C1).** Exposure time was varied from 50 to 200 ms. Exposure time of 50 ms is sufficient to get the "gist" of context.

First, we varied the exposure time T (Fig. 2h) from 50 to 200 ms (**Exp. C1**). Interestingly, human performance was largely unaffected by the image duration (Fig. 8). To assess the role of exposure time in CATNet, each computational time step was mapped to 25 ms (Sec 4.3). Consistent with human behavior results, exposure time had no effect on object recognition for CATNet.

Exp C1 shows that context modulation occurs within a short stimulus presentation duration. Such rapid computations are typically thought of as involving largely bottom-up processing [39, 16]. Despite the short exposure, there could be additional computations that take place after stimulus offset. The next experiment sought to interrupt those computations using backward masking, where presentation of the stimulus is rapidly followed by Portilla mask [35] (**Exp C2**, Fig. S10).

Accuracy in the minimal context condition was not changed by backward masking (Fig. S4). The recognition enhancement in the full context condition was impaired when the mask was introduced after 50-100 ms exposure to the image, but not with longer exposures, consistent with previous studies [43]. Overall, these results show that contextual modulation is fast and involves recurrent computations.

In natural vision, subjects interpret a scene by moving their eyes in ballistic saccades; thus, contextual information is often available *before* processing an object. When fixating on a given object, subjects already have prior contextual information from the previous fixations. To

approximate this process and study contextual reasoning with semi-realistic temporal priors, we designed an experiment where the context and target object were shown asynchronously: context was presented for 25, 50, 100 or 200 ms *before* showing the minimal context image (**Exp. C3**, Fig. S11). Surprisingly, even 25 ms exposure to context was sufficient to trigger contextual modulation (Fig. S5). For small objects, contextual facilitation was larger with increased context exposure, reaching the levels of the synchronous condition for 200 ms. In sum, a previous saccade, which typically last 200 ms, provides contextual information that can be held in memory and enhance recognition of a minimal context object, and even shorter exposure to context already enhances recognition.

5.6. Comparison with other models

Thus far, we focused on presenting the results of the CATNet model introduced in Fig. 3. As discussed in Section 2, such as [50, 48], other computational models have been proposed to incorporate some form of contextual information. We compared CATNet versus two state-of-the-art models incorporating contextual information for semantic segmentation (deeplab [10]), and object detection (yolo3 [10]). Details about performance of these models are shown in Fig. S13 and S14. Although deeplab and yolo3 leverage on global context information, CATNet outperformed both models, especially on small objects. For example, deeplab performed almost as well as CATNet on large objects but it failed to recognize small objects and demonstrate the strong contextual facilitation repeatedly observed in every experiment (Fig. 4, 5, 6, 7). These observations also hold true for yolo3. Even though yolo3 has a dedicated object recognition module after region proposal, it failed to take contextual information into account when recognizing small objects. We also note again that all computational models, including CATNet, performed worse than humans on small objects in every experiment, which suggests that it is necessary to come up with more intelligent ways of reasoning about context in existing computer vision tasks.

5.7. Ablation reveals critical model components

We also compared CATNet versus many other baselines, including modified versions of CATNet with ablated components. To gauge performance based on visual features in the whole image without focusing on the target object location, we evaluate pre-trained VGG16 [40] as a lower bound. As expected, the accuracy of VGG16 was essentially at chance, particularly for small objects (Fig. S15), confirming that in-context object recognition is not a trivial visual feature mapping task and requires focusing on the target object location. Next, we concatenated the natural stimulus with a binary mask indicating the target object location (VGG16+binarymask). Although this increased performance, accuracy was still well below CATNet (Fig. S16), suggesting that the attentional mechanism to weigh the different features plays an important role. To evaluate this, we implemented an attention module (Section 4, VGG16+attention). This led to a large performance boost, consistent with previous work showing the efficiency of attention in computer vision tasks [31]. In Fig. S12, we provide visualization examples of predicted attention maps on context and target objects respectively. CATNet learns to focus on informative context regions for recognition. Consistent with previous work [31], attention on target objects is sparse and focuses on object edges or the minimal context regions surrounding the target rather than on visual features on the targets themselves. We make further comparisons with a VGG16 version that includes an LSTM module and also with a two-stream version of VGG16 in Fig. S18 and S19.

6. Discussion

Here we quantitatively studied the role of context in visual object recognition in human observers and computational models in a task that involved recognizing target objects in various contexts. We investigated three critical properties of context: quantity, quality, and dynamics. Contextual facilitatory effects were particularly pronounced for small objects and increased with the amount of peripheral information. Consistent with the eccentricity dependence of human vision, facilitation was not affected by small amounts of blurring, or geometrical rearrangements that left intact information near the target object. Congruent contextual information typically enhanced visual recognition, while incongruent context impairs. Contextual effects could not be accounted for by low-level properties of the image. Interestingly, such contextual modulation happened fast, and could even be elicited in an asynchronous fashion where the context is shown before the target object, but they could be impaired by rapid interruption via backward masking.

To investigate how far we are from human-level in-context object recognition, we evaluated competitive methods in computer vision and introduced a recurrent neural network model (CATNet). CATNet combines a feed-forward visual stream module that extracts image features in a dynamic fashion with an attention module to prioritize different image locations, and integrates information over time, producing a label for the target object. Surprisingly, even though the model lacks the expertise that humans have in interacting with objects in their context, the model adequately demonstrated human-like behavioral characteristics under different context conditions and reaches almost human-level performance in a series of in-context object recognition

tasks. However, there are still significant gaps between models and humans, particularly when recognizing small objects within context and even large objects out of context. These results introduce benchmarks to integrate object recognition and scene understanding, and provide initial steps to understand human visual recognition and improve current intelligent computer vision systems.

References

[1] Elissa Aminoff, Nurit Gronau, and Moshe Bar. The parahippocampal cortex mediates spatial and nonspatial associations. *Cerebral Cortex*, 17(7):1493–1503, 2006. 2

[2] Mark E Auckland, Kyle R Cave, and Nick Donnelly. Nontarget objects can influence perceptual processes during object recognition. *Psychonomic bulletin & review*, 14(2):332–337, 2007. 2

[3] Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu. Multiple object recognition with visual attention. *arXiv preprint arXiv:1412.7755*, 2014. 4

[4] Moshe Bar and Elissa Aminoff. Cortical analysis of visual context. *Neuron*, 38(2):347–358, 2003. 2

[5] Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, et al. Interaction networks for learning about objects, relations and physics. In *Advances in neural information processing systems*, pages 4502–4510, 2016. 2

[6] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 456–473, 2018. 1, 2

[7] Irving Biederman, Robert J Mezzanotte, and Jan C Rabinowitz. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive psychology*, 14(2):143–177, 1982. 2

[8] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *arXiv preprint arXiv:1904.00760*, 2019. 2

[9] 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. 5

[10] 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, 2018. 2, 8

[11] Xinlei Chen, Li-Jia Li, Li Fei-Fei, and Abhinav Gupta. Iterative visual reasoning beyond convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7239–7248, 2018. 2

[12] Myung Jin Choi, Antonio Torralba, and Alan S Willsky. Context models and out-of-context objects. *Pattern Recognition Letters*, 33(7):853–862, 2012. 1, 2

[13] Wongun Choi and Silvio Savarese. A unified framework for multi-target tracking and collective activity recognition. In *European Conference on Computer Vision*, pages 215–230. Springer, 2012. 2

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4, 5

[15] Zhiwei Deng, Arash Vahdat, Hexiang Hu, and Greg Mori. Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4772–4781, 2016. 2

[16] James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012. 7

[17] Nikita Dvornik, Julien Mairal, and Cordelia Schmid. Modeling visual context is key to augmenting object detection datasets. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 364–380, 2018. 1, 2

[18] Alinda Friedman. Framing pictures: The role of knowledge in automatized encoding and memory for gist. *Journal of experimental psychology: General*, 108(3):316, 1979. 2

[19] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 2

[20] Joshua OS Goh, Soon Chun Siong, Denise Park, Angela Gutchess, Andy Hebrank, and Michael WL Chee. Cortical areas involved in object, background, and object-background processing revealed with functional magnetic resonance adaptation. *Journal of Neuroscience*, 24(45):10223–10228, 2004. 2

[21] Josep M Gonfaus, Xavier Boix, Joost Van de Weijer, Andrew D Bagdanov, Joan Serrat, and Jordi Gonzalez. Harmony potentials for joint classification and segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3280–3287. IEEE, 2010. 2

[22] Tammy Harris and James W Hardin. Exact wilcoxon signed-rank and wilcoxon mann–whitney ranksum tests. *The Stata Journal*, 13(2):337–343, 2013. 4

[23] John M Henderson, Phillip A Weeks Jr, and Andrew Hollingworth. The effects of semantic consistency on eye movements during complex scene viewing. *Journal of experimental psychology: Human perception and performance*, 25(1):210, 1999. 2

[24] Derek Hoiem, Alexei A Efros, and Martial Hebert. Geometric context from a single image. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 654–661. IEEE, 2005. 2

[25] Andrew Hollingworth. Does consistent scene context facilitate object perception? *Journal of Experimental Psychology: General*, 127(4):398, 1998. 2

[26] Hexiang Hu, Guang-Tong Zhou, Zhiwei Deng, Zicheng Liao, and Greg Mori. Learning structured inference neural

networks with label relations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2960–2968, 2016. 2

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

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 2

[29] Lubor Ladicky, Chris Russell, Pushmeet Kohli, and Philip HS Torr. Graph cut based inference with co-occurrence statistics. In *European Conference on Computer Vision*, pages 239–253. Springer, 2010. 2

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2

[31] Drew Linsley, Dan Shiebler, Sven Eberhardt, and Thomas Serre. Learning what and where to attend. 2018. 8

[32] Yong Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. Structure inference net: Object detection using scene-level context and instance-level relationships. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6985–6994, 2018. 2

[33] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2014. 2

[34] Dennis Park, Deva Ramanan, and Charless Fowlkes. Multiresolution models for object detection. In *European conference on computer vision*, pages 241–254. Springer, 2010. 2

[35] Javier Portilla and Eero P Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International journal of computer vision*, 40(1):49–70, 2000. 3, 7

[36] 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. 5

[37] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv*, 2018. 5

[38] Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11):1019, 1999. 7

[39] Thomas Serre, Gabriel Kreiman, Minjoon Kouh, Charles Cadieu, Ulf Knoblich, and Tomaso Poggio. A quantitative theory of immediate visual recognition. *Progress in brain research*, 165:33–56, 2007. 7

[40] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1, 4, 5, 8

[41] Jin Sun and David W Jacobs. Seeing what is not there: Learning context to determine where objects are missing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5716–5724, 2017. 1

[42] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 1

[43] 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, 4, 7

[44] Damien Teney, Lingqiao Liu, and Anton van den Hengel. Graph-structured representations for visual question answering. *arXiv preprint*, 2017. 2

[45] Simon Thorpe, Denis Fize, and Catherine Marlot. Speed of processing in the human visual system. *nature*, 381(6582):520, 1996. 7

[46] Antonio Torralba. Contextual priming for object detection. *International journal of computer vision*, 53(2):169–191, 2003. 2

[47] A Torralba, K Murphy, and WT Freeman. Using the forest to see the trees: Ob-ject recognition in contex. *Comm. of the ACM*, 2010. 2

[48] Antonio Torralba, Kevin P Murphy, and William T Freeman. Contextual models for object detection using boosted random fields. In *Advances in neural information processing systems*, pages 1401–1408, 2005. 2, 8

[49] Amazon Mechanical Turk. Amazon mechanical turk. *Retrieved August*, 17:2012, 2012. 2

[50] 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. 2, 8

[51] 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. 4, 5

[52] Jian Yao, Sanja Fidler, and Raquel Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 702–709. IEEE, 2012. 2

[53] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 1

[54] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*, 2014. 5

[55] Mengmi Zhang, Jiashi Feng, Karla Montejo, Joseph Kwon, Joo Hwee Lim, and Gabriel Kreiman. Lift-the-flap: Context reasoning using object-centered graphs. *arXiv preprint arXiv:1902.00163*, 2019. 6