

# Putting visual object recognition in context

Mengmi Zhang<sup>1,2</sup>, Claire Tseng<sup>3</sup>, and Gabriel Kreiman<sup>1,2</sup>

{mengmi.zhang@childrens, ctseng@college, gabriel.kreiman@tch}.harvard.edu

<sup>1</sup>Children’s Hospital, Harvard Medical School

<sup>2</sup>Center for Brains, Minds and Machines

<sup>3</sup>Harvard College, Harvard University

## 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 model the role of contextual information in visual recognition, we systematically investigated ten critical properties of where, when, and how context modulates recognition, including the amount of context, context and object resolution, geometrical structure of context, context congruence, and temporal dynamics of contextual modulation. The tasks involved recognizing a target object surrounded with context in a natural image. As an essential benchmark, we conducted a series of psychophysics experiments where we altered one aspect of context at a time, and quantified recognition accuracy. 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 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. All source code and data are publicly available<sup>1</sup>.

## 1. Introduction

The tiny object on the table is probably a spoon, not an elephant. Objects do not appear in isolation. Instead, objects 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

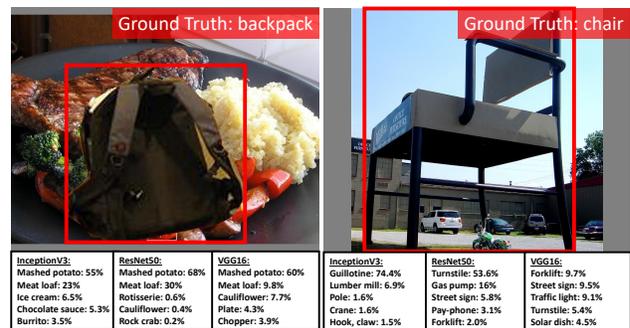


Figure 1. **Misclassification of objects in unfamiliar contexts.** State-of-the-art deep visual recognition networks, such as InceptionV3 [43], ResNet50 [54], and VGG16 [41], make mistakes when the context is incongruent. The top-5 labels and confidence levels by each model are shown at the bottom.

locations. The success in object recognition and detection 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, 42, 18, 6]. Such algorithms can be tricked into mislabeling or missing an object by placing it in an unfamiliar context (Fig. 1).

Here we 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. S1, S7, S9 and S11), we gain a quantitative understanding of where, when, and how context modulates recognition. Moreover, the human data provides a quantitative benchmark to test (but not train) computational models.

<sup>1</sup><https://github.com/kreimanlab/Put-In-Context>

Inspired by the neuroscience of human vision, we propose the Context-aware Two-stream Attention network (CATNet). This 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. The model automatically learns contextual reasoning strategies. In real world applications, models are required to extrapolate to a wide range of different contexts, just like humans. Therefore, we test CATNet and state-of-the-art in-context object recognition models on the same psychophysics tasks *without re-training the models for each experiment*. CATNet surpasses other computational models in these experiments and approximates human recognition abilities despite the enormous amount of extrapolation required.

## 2. Related Works

### 2.1. Role of Context in Human Visual Recognition

Many behavioral studies [4, 21] 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, 26, 1], during recognition tasks [2, 14], detection tasks [7, 26], working memory [19, 1], and visual search [24]. 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. Deep nets for object recognition, trained on natural image datasets, *e.g.* ImageNet [29], rely implicitly but strongly on context [20, 8]. Indeed, these algorithms often fail when objects are placed in an incongruent context ([6, 18, 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 ([47, 35, 25, 48, 33]), scene classification ([22, 49, 53]), semantic segmentation ([53]), and visual question answering ([45]).

Classical approaches, *e.g.* Conditional Random Field (CRF), reason jointly across multiple computer vision tasks in image labeling, scene classification [22, 53, 30, 10],

object detection, and semantic segmentation [34]. Several graph-based methods incorporating contextual information, combined with neural network architectures, have been successfully applied in object priming [47], place and object recognition [51, 49], object detection [11, 33], and visual question answering [45]. Recent interesting approaches have used deep graph neural networks for contextual inference [27, 13, 16, 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 (Fig. 2). Breaking away from these previous works where graph optimization is performed globally, the model proposed here selects relevant visual features using an attention mechanism, and integrates partial information from both the target object and the context over multiple steps. Importantly, the model generalizes to context variations (Sec. 5). Furthermore, we provide a direct comparison against human benchmark performance.

## 3. Human psychophysics experiments

We examined the three fundamental properties of contextual modulation in recognition (Fig. 2, S1, S7, S9, S11): [A] context amount, [B] context content, [C] context dynamics. We conducted 10 psychophysics experiments, schematically illustrated in Fig. 2h, on Amazon Mechanical Turk [50]. We recruited 80 subjects per experiment, yielding a total of 64,000 trials (Sec. 5).

**Experiment setup:** The stimuli consisted of 2,259 images spanning 55 object categories from the test set of the MSCOCO Dataset [31]. We constrained the size of target objects to four bins : Size 1 [16-32 pixels], Size 2 [56-72], Size 4 [112-144], and Size 8 [224-288]. These bins refer to the number of pixels in the object regardless of their sizes in physical world. Given the image size of  $1024 \times 1280$  pixels, and viewing distance of  $\approx 0.5$  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.

**Performance evaluation and statistics:** Most recognition experiments enforce N-way categorization (*e.g.*, [44]). Here we introduced a more unbiased probing mechanism whereby there were no constraints on the words used to describe the target object (Fig. 2h, Sec. S6.1). To evaluate human performance, we *separately* collected a distribution of ground truth answers for each target object

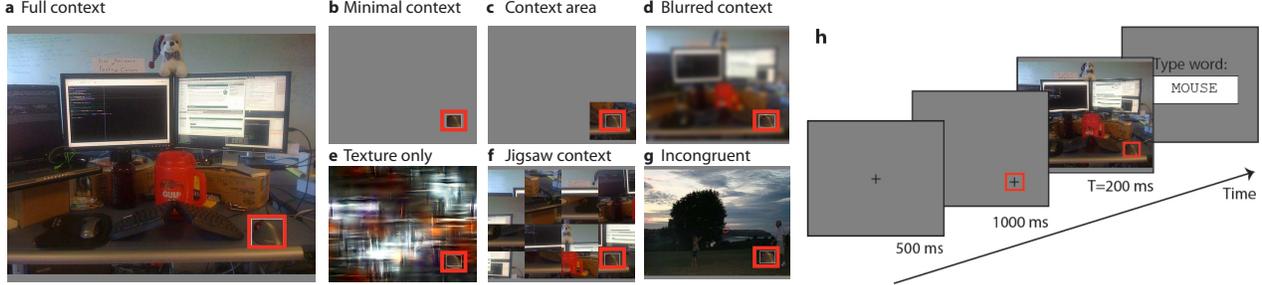


Figure 2. **Fundamental properties of context and task schematic.** Example image with full context (a) and image modifications used in experiments (more examples in Fig. S1). The target location (red box) is always the same across conditions. The correct answer (“mouse”) is not shown in the actual experiment. (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, Fig. S7, S9 and S11), the image was shown for  $T = 200$  ms. After image offset, subjects typed one word to identify the target object.

with infinite viewing time and full context (Mturk subjects *not* participating in the main experiments). Answers in the main experiments were correct if they matched any of the ground truth responses.

Although computational models (Sec. 4) were evaluated using N-way categorization, we find it instructive to plot model results alongside human behavior for comparison purposes. We also show human-model correlations to describe their relative trends across conditions. We use the Wilcoxon ranksum test [23], and one-way/two-way ANOVA tests [28] (Sec. S6.2) for statistical comparisons.

#### 4. Context-aware Two-stream Attention Net

We propose a Context-aware Two-stream Attention network (CATNet), extending work on image captioning [52]. CATNet is presented with 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), and a second stream devoted to contextual information in the periphery ( $I^c$ , full context, Fig. 2a). The two streams are processed in parallel through weight-sharing convolutional neural networks.  $I^o$  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.

Horizontal and top-down connections are presumed to be important for recognition [44]. We added a recurrent LSTM module in CATNet to iteratively reason about context. The LSTM module changes its internal representation of the scene via attention, and predicts class labels over multiple time steps  $t$  where  $t \in \{1, \dots, T_m\}$ . We use superscripts  $c$  or  $o$  to distinguish processes on  $I^c$  or  $I^o$  and subscript  $t$  for 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 [41], pre-trained on ImageNet [15] 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  $i = 1, \dots, L$  and  $L = W \times H$ , and  $W$  and  $H$  are the width and height, respectively, of the feature map:

$$a^c = \{\mathbf{a}_1^c, \dots, \mathbf{a}_L^c\}, \quad \mathbf{a}_i^c \in \mathbb{R}^D \quad (1)$$

##### 4.2. Attentional Modulation

We use a “soft-attention” mechanism [3] to compute “the context gist”  $\hat{\mathbf{z}}_t^c$  on  $I_c$ , and “the object gist”  $\hat{\mathbf{z}}_t^o$  on  $I_o$  (Fig. S13). 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  $\alpha_{ti}^c$ , representing the relative importance of the feature vector  $\mathbf{a}_i^c$  in capturing the context gist.  $\alpha_{ti}^c$  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:

$$e_{ti}^c = A_h^c \mathbf{h}_{t-1} + A_a^c \mathbf{a}_i^c, \quad \alpha_{ti}^c = \frac{\exp(e_{ti}^c)}{\sum_{i=1}^L \exp(e_{ti}^c)} \quad (2)$$

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  $\beta_t^c$  from the previous hidden state  $h_{t-1}$ , such that  $\beta_t^c$  determines how much the current observation contributes

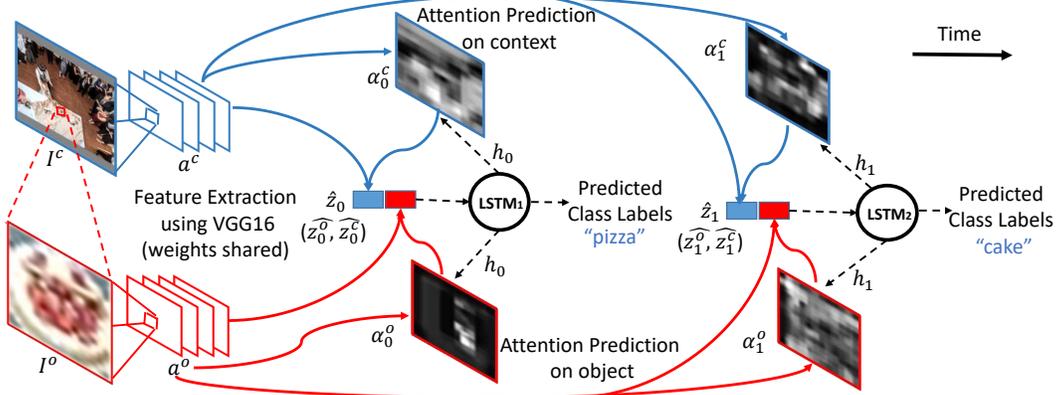


Figure 3. **Architecture overview of Context-aware Two-stream Attention network (CATNet)**. The diagram depicts the iterative modular steps carried out by CATNet over multiple time steps in the context-aware object recognition task. CATNet consists of 3 main modules: feature extraction, attention, and recurrent memory. These three modular steps repeat for a pre-specified number of time steps  $T_m$ . For illustrative purposes, only the first and second time steps are shown here (see Sec. 4 for definition of variables and Fig. S13-S14 for implementation 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 (Sec. 3 and 5).

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  $\beta_{ti}^c$  in  $\beta_t^c$  is a gating scalar at location  $i$ .  $\beta_t^c$  helps put more emphasis on the salient objects in the images [52]. Once the attention map  $\alpha_t^c$  and the gating scale  $\beta_t^c$  are computed, the model applies the “soft-attention” mechanism to compute  $\hat{\mathbf{z}}_t^c$  by summing over all the  $L$  regions in the image:

$$\hat{\mathbf{z}}_t^c = \sum_{i=1}^L \beta_{ti}^c \alpha_{ti}^c \mathbf{a}_i^c \quad (3)$$

We define  $\hat{\mathbf{z}}_t = (\hat{\mathbf{z}}_t^c, \hat{\mathbf{z}}_t^o)$  as concatenation of  $\hat{\mathbf{z}}_t^c$  and  $\hat{\mathbf{z}}_t^o$ , which is used as input to an LSTM module. The attention module is smooth and differentiable, and CATNet learns all the weights end-to-end via back-propagation.

### 4.3. Recurrent Connections using LSTM

A long short-term memory (LSTM) network predicts the class label  $y_t$  based on the previous hidden state  $\mathbf{h}_{t-1}$  and the gist vector  $\hat{\mathbf{z}}_t$  for  $I^o$  and  $I^c$  [55] (Fig. S14). 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 (Section S3). To compare CATNet and human performance for different exposure time  $T$  (Exp. C), we set one LSTM time step to be 25 ms, and considered the CATNet predicted labels at the corresponding number of time steps  $T_m = T/25$ .

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  $\mathbf{h}_t$ :

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

where  $L_h \in \mathbb{R}^{C \times n}$  is a matrix of learnt parameters initialized randomly. We discuss alternative convolutional LSTM connections in Section S3.

### 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  $x$ :  $LOSS = \sum_{t=1}^{T_m} (-\log(P(y_t|x)))$ . Predicting labels at every time step allows us to assess the effect of image exposure time in **Exp C** (Fig. 2h, S7, S9, S11 and Sec. 5.3). Besides, using ground truth labels at every time step empirically helped CATNet converge faster during training.

We used all MSCOCO training set images for training and validation. On every image, each object was selected as the target, always shown in full context. Only at the testing stage, we varied the context based on different experimental conditions. Importantly, *none of the human behavioral experiments were used to train the model*. Both  $I^c$  and  $I^o$  were  $400 \times 400$  pixels. We set the total number of time steps  $T_m = 8$  for training CATNet. Further implementation details are provided in Sections S3 and S4.

### 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 [37, 38] in object detection. These models were adapted to the context-aware object recognition task (Sec. S5).

To study the role of attention, the two-stream architecture, and recurrent connections, we introduced ablated versions of CATNet (Sec. S5). Starting from original VGG16 object recognition network [41] pre-trained on ImageNet [15], 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.

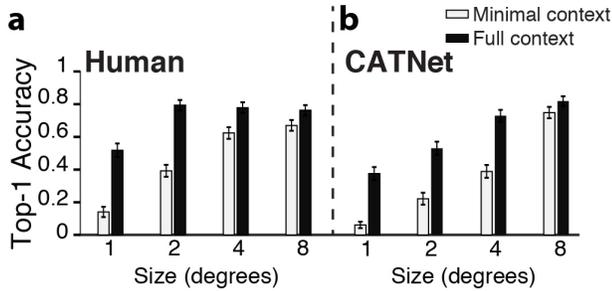


Figure 4. **Context improves recognition, particularly for small objects (Exp A1).** Top-1 accuracy increases with object size (Fig. 2a-b). Contextual information facilitates recognition, particularly for small target objects, for humans (a) and CATNet (b). Here and in subsequent figures, error bars denote SEM. In b, chance level is 1/55 (see text for details).

## 5. Results

### 5.1. Exp A: Amount of context

#### 5.1.1 Object size matters (Exp A1)

We conjectured that the impact of contextual information would depend on the target object size. We considered 4 object sizes (Sec. 3). For each size, we introduced either minimal context (rectangular bounding box enclosing the object, Fig. 2b), or full context (entire image, Fig. 2a).

For the minimal context condition (Fig. 2b), human performance improved monotonically as a function of object size from  $0.140 \pm 0.031$  to  $0.670 \pm 0.035$  (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. The enhanced performance due to contextual modulation strongly depended 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 performance when the target objects are small and hard to recognize.

#### 5.1.2 The amount of context matters (Exp A2)

For each object size, we systematically titrated the amount of contextual information (Fig. 2c). The context-object ratio (CO) is the total image 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 object sizes and large CO values were not possible.

We quantified how the amount of contextual information impacts recognition by titrating the context object ratio (CO) from 0 to 128 (Exp A2, Fig. S2). The amount of context was important both for humans (one-way ANOVA:

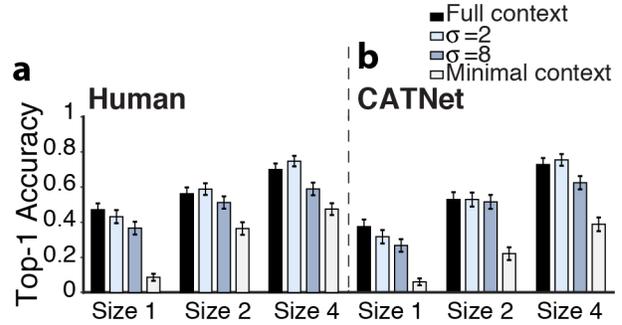


Figure 5. **Contextual facilitation persists even after small amounts of blurring (Exp B1).** A large amount of context blurring (Fig. 2d) is required to disrupt contextual facilitation for humans (a) and CATNet (b). Only  $\sigma = 2$  and  $\sigma = 8$  are shown here (see Fig. S3 for intermediate  $\sigma$  values).

$F(7, 5097) = 31, p < 10^{-15}$ ), and for 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 the human psychophysics measurements. These experiments demonstrate that the context *quantity* can strongly enhance recognition; which we refer to as *context modulation* for short throughout the rest of the text.

### 5.2. Exp B: Context Content

We studied how context resolution, geometry, and congruency modulate recognition in 5 experiments, focusing on object sizes 1/2/4, with minimal/full context.

#### 5.2.1 Blurred context is sufficient (Exp B1)

Due to the strong eccentricity dependence of human vision, the periphery has less resolution than the fovea (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 quantify the impact of context resolution, we blurred the context (Fig. 2d) using a zero-mean Gaussian with standard deviation  $\sigma = 2, 4, 8, 16, 32$  pixels (image size =  $1024 \times 1280$  pixels) (Exp B1, Fig. 2d). Each subject saw all blurred conditions, with different images.

Accuracy dropped from levels indistinguishable from the full resolution condition when  $\sigma \leq 8$  pixels to the minimal context condition levels when  $\sigma = 32$  pixels (Fig. 5, one-way ANOVA:  $F(4, 2933) = 28, p < 10^{-15}$ , Fig. S3). 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. The effects of blurring were also captured by CATNet, where contextual modulation disappeared only when using large  $\sigma$  values (one-way ANOVA:  $F(4, 2354) = 2, p < 0.05$ ). Similar to Exp A1-A2, humans outperformed CATNet on small objects.

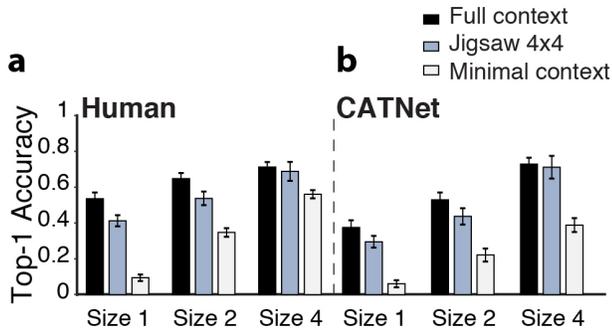


Figure 6. **Large geometrical context re-arrangements disrupt contextual facilitation (Exp B4).** Scrambling context pieces (Fig. 2f) reduced facilitation only when many small context pieces were changed, both for humans (a) and CATNet (b). Only the 4x4 condition is shown here (see Fig. S6 for other conditions).

We compared the effect of blurring the context versus the target object by applying the same Gaussian blurring (Exp. B1) only to the object itself (Exp. B2, Fig. S4). Although the number of pixels affected by blurring the target object was much smaller than blurring the context (for a fixed  $\sigma$ ), modifying the object led to larger accuracy drops, for object sizes 2 and 4, both for humans and CATNet.

### 5.2.2 Contextual effects rely on spatial configuration

The relative position of objects and features in an image also affects recognition; *e.g.*, the sky is often at the top. To evaluate the impact of contextual configuration, we randomly scrambled the images into  $2 \times 2$ ,  $4 \times 4$ , or  $8 \times 8$  "jigsaw" pieces while the piece with the target object remained in the same position (Exp B4, Fig. 2f). We did not consider cases when the object occupied more than one piece. Both humans and CATNet relied on the spatial configuration of context over all object sizes (Fig. 6, humans: one-way ANOVA:  $F(3, 2182) = 58, p < 10^{-15}$ ; CATNet: one-way ANOVA:  $F(3, 1787) = 29, p < 10^{-15}$ , Fig. S6). The inconsistent spatial configuration of contextual information in the  $4 \times 4$  and  $8 \times 8$  configurations led to a reduction in accuracy. Interestingly, the  $2 \times 2$  configuration was not different from the unscrambled full context condition, probably because each large piece already contained sufficient contextual information, and context reasoning decreases with distance to the target [56].

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

### 5.2.3 Low-level contextual properties do not lead to facilitation (Exp B3)

Given that the moderately blurred context still facilitated recognition (Fig. 5), we asked whether low-level texture features could also enhance performance. We constructed textures constrained by the image statistics [36], and pasted

the intact object on them in their original locations (Exp B3, Fig. 2e). The textures preserve low-level features, but distort high-level features and semantic information.

Low-level texture features did not facilitate object recognition for either humans or CATNet (Fig. S5). 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.

### 5.2.4 Incongruent context impairs recognition

Given that low-level textures did not help (and could even hurt recognition), and inspired by Fig. 1 and related experiments, we next studied recognition when objects were removed from their original images and placed in the same location but in different images with either a congruent context (object and context belong to the same class label) or incongruent context (context taken from a different image class label, Fig. 2g).

Congruent context enhanced recognition for small object sizes compared to the minimal context condition both for humans and CATNet (Fig. 7). Although congruent context typically share similar correlations between objects and scene properties, pasting the object in a congruent context led to weaker enhancement. This lower contextual facilitation may be due to the erroneous relative size between objects, unnatural boundaries created by pasting, or contextual cues specific to each image. CATNet was relatively oblivious to these effects and performance in the congruent condition was closer to that in the original full context condition.

In stark contrast, incongruent context 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,  $p < 0.001$ ).

## 5.3. Exp C: Dynamics of Contextual Modulation

The previous sections characterized *spatial* aspects of contextual modulation. The temporal dynamics of recognition places strong constraints to interpret the flow of bottom-up and top-down visual processes [46, 44, 39]. Next, we conducted 3 experiments to investigate the dynamics of contextual effects on recognition.

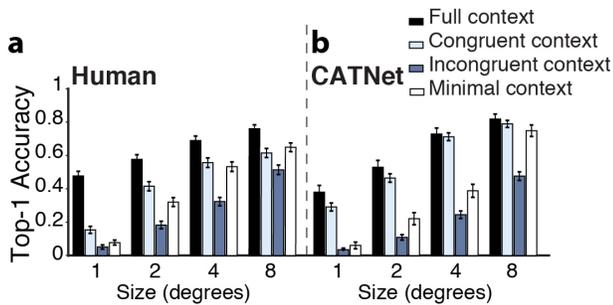


Figure 7. **Incongruent context impairs recognition.** Pasting the target objects in different but congruent contexts facilitates recognition. Pasting the target objects in incongruent contexts (Fig. 2g) impairs recognition, both for humans and CATNet.

### 5.3.1 Contextual modulation is fast (Exp C1)

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. S7, Exp. C1). Human performance was largely unaffected by the image duration (Fig. S8). 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 recognition for CATNet.

### 5.3.2 Backward masking weakens contextual modulation (Exp C2)

The rapid computations in Exp C1 are thought to involve largely bottom-up processing [40, 17]. Despite the short exposure, additional computations could take place after stimulus offset. The next experiment sought to interrupt those computations using backward masking (Exp C2, Fig. S9). Backward masking is commonly used in neuroscience to interrupt visual processing [44]. The mask shown after stimulus offset is purported to block top-down and recurrent computations. We used Portilla masks [36] as in Exp B3 (Fig. S9) and stimulus exposure times followed Exp C1.

Backward masking did not change accuracy in the minimal context condition (Fig. S10). The recognition enhancement in the full context condition was impaired when the mask was introduced after 50-100 ms exposure, but not with 200 ms, consistent with previous studies [44]. In sum, contextual modulation is fast and is likely to involve recurrent computations that can be interrupted by masking.

### 5.3.3 Brief exposure to context is sufficient for facilitation (Exp C3)

In all the experiments above, object and context information were presented synchronously. During natural vision, subjects move their eyes from a given location P1 to another 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  $T_1 = 25, 50, 100,$  or  $200$  ms. Next, the context was removed, and the object-only part was presented for a duration  $T_2 = 50, 100,$  or  $200$  ms (Exp. C3, Fig. S11). The corresponding synchronous conditions were also included for comparison purposes.

Surprisingly, even 25 ms exposure to context was sufficient to trigger contextual modulation (Fig. S12). For small objects, contextual facilitation was larger with increased context exposure, reaching the levels of the synchronous condition for 100 ms context exposure. In sum, a previous saccade, which typically last 200 ms, provides enough contextual information that can be held in memory and enhance recognition of a minimal context object. Even shorter exposure to context already enhances recognition.

## 5.4. Comparison with other models

We have focused on the CATNet model introduced in Fig. 3. Several other computational models incorporate some form of contextual information (Sec. 2). We compared CATNet versus two state-of-the-art models incorporating contextual information for semantic segmentation: (1) Deeplab [10]), and (2) object detection (YOLO3, [38]). The average accuracy across all conditions in all the variations of Experiments A and B are shown in Table 1. Details about performance of these models are shown in Fig. S16-S22 (Deeplab), and Fig. S23-S29 (YOLO3).

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 demonstrate the strong contextual facilitation repeatedly observed in every experiment (Fig. 4, 5, 6, 7).

Similarly, even though YOLO3 has a dedicated recognition module after region proposal, it failed to incorporate contextual information when recognizing small objects. We also emphasize again that *all* computational models, including CATNet, performed worse than humans on small objects in every experiment, suggesting that it is necessary to come up with better ways of reasoning about context in computer vision tasks.

At least partly, the baseline models struggle with small objects due to lack of scale tolerance. In addition to absolute accuracy, we also report the correlation between each model and human performance across conditions for each experiment in Table 1. The correlation between the algorithms and humans reflects how each model is affected by different conditions. Not only did the baselines show lower accuracy, but they also showed lower correlation with human performance than CATNet.

Accuracy	A1	A2	B1	B2	B3	B4	B5
Humans	<b>0.58</b>	<b>0.58</b>	<b>0.47</b>	<b>0.49</b>	<b>0.39</b>	<b>0.48</b>	0.43
CATNet	0.48	0.48	0.42	0.39	0.34	0.41	<b>0.44</b>
DeepLab [10]	0.52	0.45	0.37	0.42	0.31	0.38	0.39
YOLO3 [38]	0.26	0.25	0.13	0.14	0.13	0.13	0.19
Correlation	A1	A2	B1	B2	B3	B4	B5
CATNet	0.89	<b>0.89</b>	<b>0.95</b>	0.87	0.89	<b>0.92</b>	<b>0.93</b>
DeepLab [10]	<b>0.90</b>	0.83	0.86	<b>0.88</b>	<b>0.90</b>	0.81	0.91
YOLO3 [38]	0.75	0.78	0.74	0.78	0.75	0.66	0.87

Table 1. **Performance and correlations between humans and models for Experiments A and B.** See Sec. 3 for definitions of evaluation metrics. Best is in bold. See Table S1 for additional comparisons.

## 5.5. Ablation reveals critical model components

To distill how different components of CATNet contribute to integrating contextual information, we considered modified versions with ablated components (Tab. S1). We first evaluated pre-trained VGG16 [41]. The accuracy of VGG16 was essentially at chance, particularly for small objects (Fig. S30-S36), 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 the binary mask increased performance with respect to VGG16, accuracy was still well below CATNet (Figs. S37-S43), suggesting that the attentional mechanism to weigh the different features plays an important role. We therefore implemented an attention module (Sec. 4, VGG16+attention), which led to a large performance boost (Figs. S44-S50), consistent with previous work showing the efficiency of attention in vision tasks [32]. In Fig. S15, 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 [32], 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. Additional improvement in performance was achieved by incorporating a two-stream module (Figs. S51-S57), and an LSTM module (Figs. S58-S64).

## 6. Discussion

We quantitatively studied the role of context in visual recognition in human observers and computational models in tasks requiring identification of target objects in natural settings. 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. The notion of full context used here and in most computer vision databases is arbitrarily defined by the person taking the picture (as opposed to a true full image which could be defined by the entire human visual field). Thus, Experiment

A provides a direct titration of how different amounts of context impact recognition (Fig. 4, Fig. S2).

Consistent with the eccentricity dependence of human vision, facilitation was not affected by small amounts of blurring (Fig. 5, Fig. S3), or geometrical rearrangements that left intact information near the target object (Fig. 6). Congruent contextual information typically enhanced recognition, while incongruent context impaired performance (Fig. 7). Contextual effects could not be accounted for by low-level image properties (Fig. S5). Such contextual modulation happened fast (Fig. S7-S8), and could even be elicited in an asynchronous fashion where the context was shown before the target object (Fig. S11-S12). Contextual modulation was impaired by rapid interruption via backward masking (Fig. S9-S10).

To compare against the benchmark of human-level in-context recognition, we evaluated competitive methods in computer vision, and introduced a recurrent neural network model (CATNet, Fig. 3). CATNet combines a feed-forward visual stream module that dynamically extracts image features with an attention module to prioritize different image locations. CATNet integrates information over time, producing a label for the target object. Surprisingly, even though the model lacks human expertise in interacting with objects in their context, CATNet adequately demonstrated human-like behavioral characteristics, and reached almost human-level performance in a wide range of in-context recognition tasks. However, there are still significant gaps between models and humans, particularly when recognizing small objects within context, and also for large objects out of context.

Even though context is often only *implicitly* incorporated in current algorithms, contextual information is critical to vision applications (*e.g.*, object and action recognition). Dissociating the contributions of object and context helps us better interpret computer vision models. In addition, context can be used to fool current algorithms (*e.g.*, Fig. 1). Thus, the experiments presented here help us understand models' failure cases. Explicitly incorporating contextual cues can further help protect computer vision models against context-based adversarial attacks. These results introduce benchmarks to integrate object recognition and scene understanding, and provide initial steps to understand human visual recognition and improve intelligent computer vision systems.

## Acknowledgements

This work was supported by NIH R01EY026025 and by the Center for Minds, Brains and Machines, funded by NSF STC award CCF-1231216. MZ is supported by postdoctoral fellowship of Agency for Science, Technology and Research. We thank Martin Schrimpf for initial discussions that inspired this work, and Pranav Misra and Kasper Vinken for comments on the paper.

## 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. 3
- [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. 4
- [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, 7, 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] Jodi L Davenport and Mary C Potter. Scene consistency in object and background perception. *Psychological science*, 15(8):559–564, 2004. 2
- [15] 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. 3, 4
- [16] 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
- [17] James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012. 7
- [18] 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
- [19] 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
- [20] 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
- [21] Joshua OS Goh, Soon Chun Siang, Denise Park, Angela Gutches, 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
- [22] 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
- [23] 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
- [24] 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
- [25] 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
- [26] Andrew Hollingworth. Does consistent scene context facilitate object perception? *Journal of Experimental Psychology: General*, 127(4):398, 1998. 2
- [27] 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
- [28] PK Ito. 7 robustness of anova and manova test procedures. *Handbook of statistics*, 1:199–236, 1980. 3
- [29] 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
- [30] 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
- [31] 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
- [32] Drew Linsley, Dan Shiebler, Sven Eberhardt, and Thomas Serre. Learning what and where to attend. 2018. 8
- [33] 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
- [34] 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
- [35] Dennis Park, Deva Ramanan, and Charless Fowlkes. Multiresolution models for object detection. In *European conference on computer vision*, pages 241–254. Springer, 2010. 2
- [36] 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. 6, 7
- [37] 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. 4
- [38] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv*, 2018. 4, 7, 8
- [39] Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11):1019, 1999. 6
- [40] 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
- [41] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1, 3, 4, 8
- [42] 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
- [43] 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
- [44] 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. 2, 3, 6, 7
- [45] Damien Teney, Lingqiao Liu, and Anton van den Hengel. Graph-structured representations for visual question answering. *arXiv preprint*, 2017. 2
- [46] Simon Thorpe, Denis Fize, and Catherine Marlot. Speed of processing in the human visual system. *nature*, 381(6582):520, 1996. 6
- [47] Antonio Torralba. Contextual priming for object detection. *International journal of computer vision*, 53(2):169–191, 2003. 2
- [48] A Torralba, K Murphy, and WT Freeman. Using the forest to see the trees: Object recognition in context. *Comm. of the ACM*, 2010. 2
- [49] 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
- [50] Amazon Mechanical Turk. Amazon mechanical turk. Retrieved August, 17:2012, 2012. 2
- [51] 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
- [52] 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. 3, 4
- [53] 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
- [54] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 1
- [55] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*, 2014. 4
- [56] 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