

# Learning Scene Gist to Improve Object Recognition in Convolutional Neural Networks

A dissertation presented

by

Kevin Wu

to

The School of Engineering and Applied Sciences

in partial fulfillment of the requirements

for the degree of

Masters in Engineering

in the subject of

Computational Science and Engineering

Harvard University

Cambridge, Massachusetts

May 2018

© 2018 -Kevin Wu  
All rights reserved.

## Learning Scene Gist to Improve Object Recognition in Convolutional Neural Networks

### ABSTRACT

Advancements in convolutional neural networks (CNNs) have made significant strides toward achieving high performance levels on object recognition tasks. However, in the real world, objects are almost always presented within scenes with other people and objects. While some approaches utilize information from the entire scene to propose regions of interest, the task of interpreting a particular region or object is still performed independently of other objects and features in the image. Here we demonstrate that a scene’s ‘gist’ can significantly contribute to how well humans can recognize objects. These findings are consistent with the notion that humans foveate on an object and incorporate information from the periphery to aid in recognition. We use a biologically inspired two-part convolutional neural network that models the fovea and periphery to provide a proof-of-principle demonstration that computational object recognition can significantly benefit from the gist of the scene as contextual information. Our model yields accuracy improvements of up to 50% in certain object categories when incorporating contextual gist, while only increasing the original model size by 5%. This proposed model mirrors our intuition about how the human visual system recognizes objects, suggesting specific biologically plausible constraints to improve machine vision and building initial steps towards the challenge of scene understanding.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Background</b>	<b>4</b>
2.1	Types of Contextual Information . . . . .	5
2.1.1	Semantic Context . . . . .	6
2.1.2	Spatial Context . . . . .	6
2.1.3	Scale Context . . . . .	7
2.2	Foveal and Peripheral Vision . . . . .	7
2.2.1	Peripheral Vision . . . . .	8
2.2.2	The Gist of a Scene . . . . .	9
2.3	Role of Context in CNNs . . . . .	9
<b>3</b>	<b>Methods</b>	<b>11</b>
3.1	Dataset . . . . .	11
3.2	Behavioral Experiments . . . . .	12
3.3	Computational Model of Scene Gist . . . . .	14
3.3.1	Fovea sub-network for object recognition . . . . .	14
3.3.2	Periphery sub-network for contextual modulation . . . . .	16
3.3.3	Training . . . . .	17
3.3.4	Evaluation . . . . .	17
<b>4</b>	<b>Results</b>	<b>18</b>
4.1	Human Object Recognition Improves With Context . . . . .	18
4.1.1	Minimal Context Condition . . . . .	18
4.1.2	Full Context Condition . . . . .	19
4.1.3	Contribution of Context Increases For Small Objects . . . . .	20

4.2	GistNet Captures Gist-Like Context . . . . .	22
4.2.1	Understanding When To Use Context . . . . .	23
4.2.2	Gradient-Based Interpretation of Gist . . . . .	24
4.2.3	Robustness to Blurring . . . . .	26
4.2.4	Learning Representations of Semantic Context . . . . .	27
<b>5</b>	<b>Conclusion</b>	<b>30</b>
5.0.1	Limitations . . . . .	31
5.0.2	Application to state-of-the-art models . . . . .	32

THIS THESIS IS DEDICATED TO MY FAMILY, WHO HAS SUPPORTED ME  
ENDLESSLY DURING MY TIME AT HARVARD.

## Acknowledgments

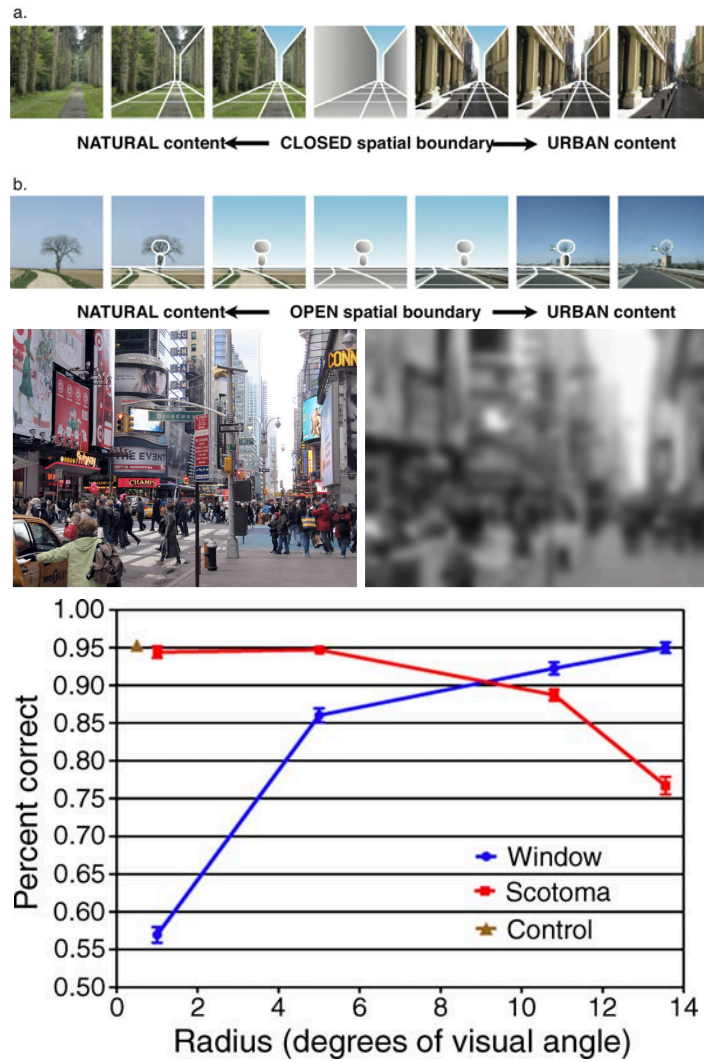
I want to thank Professor Gabriel Kreiman for his steady support of my passion for research in computer vision and neuroscience. I also want to express my gratitude for Professor David Cox on his advice and support for my research and for providing invaluable compute resources for my brother Eric and I. Finally, I thank Pavlos Protopapas and IACS for having me study and research at Harvard in the first place. I really appreciate all the encouragement and guidance you have all provided.

# 1

## Introduction

OBSERVERS CAN RAPIDLY EXTRACT GLOBAL INFORMATION FROM A SCENE, referred to as the image gist [18]. In a few hundred milliseconds, observers reliably ascertain summary scene information, even if specific objects are not recognizable [25]. A prominent feature of the primate visual system is eccentricity-dependent sampling, with a high-resolution foveal region and a lower resolution periphery. The periphery has a decaying density of cells as function of distance from the fovea, and allows for faster approximate perception. With others [3], [31], and [12], we conjecture that low-resolution peripheral information provides an initial approximation of the scene gist. A scene's gist includes properties (e.g., naturalness, openness, roughness, etc.) that represent the dominant spatial structure of a scene [17]. Second-order statistics can be used to compute global features from the image and to classify the scene according to these dimensions, without needing information about specific objects [27].





**Figure 1.1:** Top: A schematic of various spatial properties of a scene from [17]. Middle: Even with significant blurring, observers can ascertain general features of the scene. Bottom: [12] provides evidence that the periphery is more useful than the fovea in determining the gist of a scene.

During scene understanding, peripheral information can be used to propose regions of interest for active sampling, and the eyes can then quickly foveate on these regions for high-resolution interpretation. The interplay

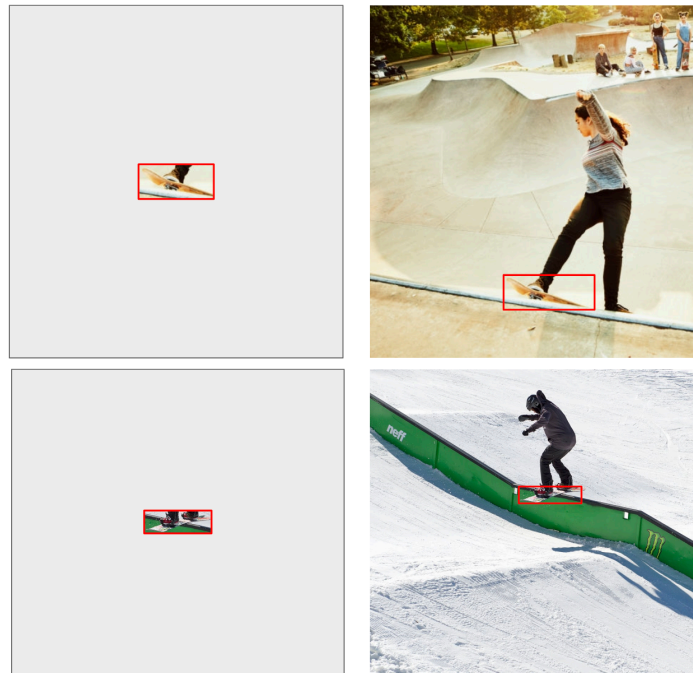
between foveal and peripheral information may enable faster recognition of objects within a scene with a significantly reduced number of cells. Recently, region-base convolutional neural networks (R-CNN) like have been used for object detection and classification within scenes. These architectures [20] mirror elements of active sampling via sequential foveation by creating region proposals on the image, followed by object recognition in each region. Those region proposals cut down on the cost of having to integrate information from the whole scene during each object classification. Yet, these models lack critical components of contextual information provided by interactions between the fovea and the periphery which are characteristic of human vision: (i) a low resolution and rapid peripheral system, (ii) interactions between periphery and foveal information, and (iii) global sharing of information learned across foveations. Using global features from the scene gist may reduce the need for additional region proposals, aiding recognition of all objects within the same scene and enforcing all objects in a scene to be influenced by the same prior during inference. While several previous architectures have demonstrated the usefulness of context in recognition and detection, they have focused on high resolution contextual information, semantic context, and object co-occurrences. The role of low-resolution and global gist-like features in object recognition using deep convolutional neural networks is still poorly understood. In this paper, we focus on building a biologically-inspired system that incorporates global gist into object recognition. We propose a model that incorporates foveal/peripheral interactions, compare the results with behavioral measurements, and provide proof-of-principle evidence that a computational architecture that provides gist level information to the foveal recognition machinery can improve recognition accuracy.

# 2

## Background

TRADITIONAL METHODS OF UNDERSTANDING OBJECT RECOGNITION appeal to object-specific features to ascertain information. For example, an object's color, texture, shape, and intensity can all capture unique elements of an object to a certain extent. However, objects in the real world are rarely presented in isolation. Common sources of variability include occlusion, variable lighting, and differing orientations. Contextual information can play a key role in overcoming these combinatoric complexities.

The role of context for visual recognition is intuitive to most observers. For example, determining the difference between a snowboard and a skateboard can be aided by the presence of either pavement or snow, especially when the object itself is partially occluded or small (Figure 2.1). In addition, objects which lack many discerning features often require context to be identified. However, the contributions of context are hard to pinpoint given the varying definitions of context. In this section, we will cover the different types of contextual information, its role in human and computer vision, and the basis



**Figure 2.1:** A snowboard and skateboard presented with and without context.

of a scene's 'gist'.

## 2.1 Types of Contextual Information

Contexts exist in many forms – object recognition can be aided by small clues such as neighboring objects, as well as full-scale spatial layouts of natural scenes. Biederman et al. [5] characterizes an object's belonging in scenes by its interposition, support, probability, position, and familiar size. Galleguillos et al. [7] crystallizes these context types for visual recognition by three types: *semantic*, *spatial*, and *scale*.

*Semantic* context is 'the likelihood of an object to be found in some scenes but not others'.

*Spatial* context is 'the likelihood of finding an object in some position and not others'.

*Scale* context is 'based on the scales of an object with respect to others'.

### 2.1.1 Semantic Context

Continuous exposure to the real world allow biological visual systems to anticipate and expect order and continuity between objects in scenes. A scene can be generally understood as a meaningful interpretation of an image as a whole, as opposed to a specific object. For example, we may be used to seeing pots and pans in kitchens, while we would not expect to see a motorcycle in a bedroom. Both kitchens and bedrooms serve as latent variables that connect objects which occur together with high frequency. The importance of scene-level contextual information have been shown to be important for human object recognition. Palmer [19] finds that object recognition improves if previously presented with a semantically consistent stimulus, and impaired when presented with an inconsistent stimulus.

We can more formally understand semantic context through Bayes' formula, with an object's scene serving as a prior:

$$P(\text{object}_i|\text{scene}_k) = \frac{P(\text{scene}_k|\text{object}_i)P(\text{object}_i)}{P(\text{scene}_k)}$$

### 2.1.2 Spatial Context

An object's spatial context is understood as its relative position in the scene. For example, we are used to seeing ceiling fans above us, whereas we expect grass to be below us.

In general, object's spatial location can be similarly formulated as a Bayesian prior:

$$P(\text{object}_i|\text{position}_k) = \frac{P(\text{position}_k|\text{object}_i)P(\text{object}_i)}{P(\text{position}_k)}$$

,

where  $\text{position}_k$  refers to a location mapping in 2D.

Calculating the exact location of objects in a scene is difficult, since scenes may be presented at various angles and scales. Nonetheless, previous work has been presented that aim to capture these relationships. Shotton et al.

[22] uses a classifier  $\lambda_i$  which maps the relationships between object class  $c$  and pixel index  $i$ :

$$\lambda_i(c_i, i; \theta_\lambda) = \log \theta_\lambda(c_i, \hat{i})$$

, where  $\hat{i}$  is normalized to account for varying scales of images, and  $\theta_\lambda$  refer to model parameters.

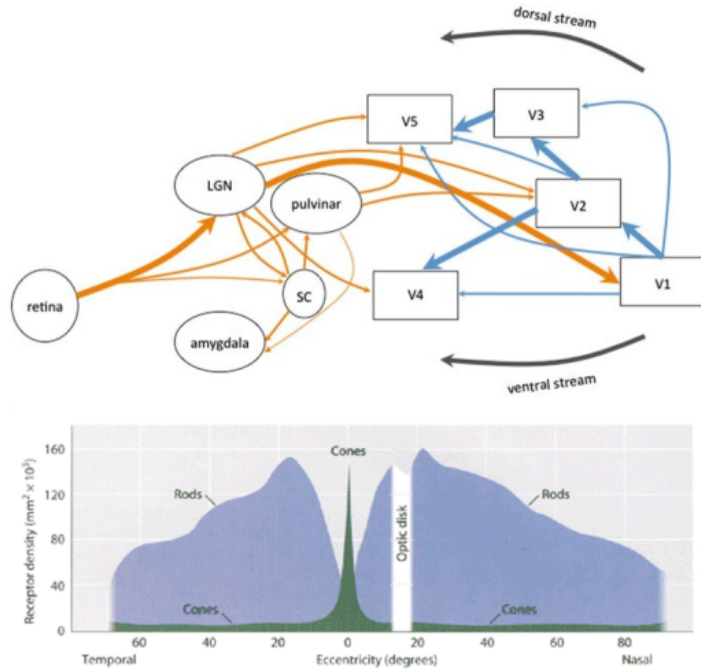
### 2.1.3 Scale Context

In natural scenes, objects adhere to constraints on absolute size and relative scale. The relative proportions of objects within a scene serve to regularize the hypothesis space for each object. For example, when trying to discern objects placed on a dinner plate, we would intuitively expect not to see objects such as chairs, vases, and lamps even though they might share semantic context with plates since they would be too large to fit on a plate.

Previous works which incorporate scale context include Torralba [26], which defines objects in an image as:  $O = \{o, x, \sigma\}$ , where  $o$  is the object class,  $x$  is the spatial location, and  $\sigma$  is the scale relative to the whole image.

## 2.2 Foveal and Peripheral Vision

Object recognition in humans is generally characterized by a stream of connected components in the brain that pass information to each other. First, the retina perceives visual stimuli in four processing stages: photoreception, transmission to bipolar cells, transmission to ganglion cells, and transmission along the optic nerve. Next, the optic nerve serves to compress and carry information to the lateral geniculate nucleus (LGN) in the thalamus. This information is then passed onto the ventral stream, which is understood to be responsible for object identification and recognition.



**Figure 2.2:** Top: A modular visualization of the human visual system from [29]. Bottom: The distribution of rods and cones across degrees of eccentricity from [2].

### 2.2.1 Peripheral Vision

Not all input is received equally by the retina. The fovea contains the highest possible visual acuity, containing a higher density of cones than anywhere else in the retina. The retinal ganglion receptor cells have the smallest receptive fields. This allows observers to process highly detailed objects.

At the same time, the density of retinal ganglion cells decay linearly away from the center of the retina. The region outside the fovea is referred to as the periphery. The formulation of the density as a function of eccentricity is found in [30]:

$$d_{gf}(r, k) = d_{gf}(0) \times \left[ a_k \left( 1 + \frac{r}{r_{2,k}} \right)^{-2} + (1 - a_k) \exp \left( - \frac{r}{r_{e,k}} \right) \right]$$

, where  $d_{gf}(0)$  is the density at  $r = 0$ ,  $r_{e,k}$  is scale factor of the exponential,

$a_k$  is the weighting of the first term, and the meridian is indicated by  $k$ .

This progressive decay allows human observers to sample from the entire scene without having to process input at the same resolution at each spatial location.

### 2.2.2 The Gist of a Scene

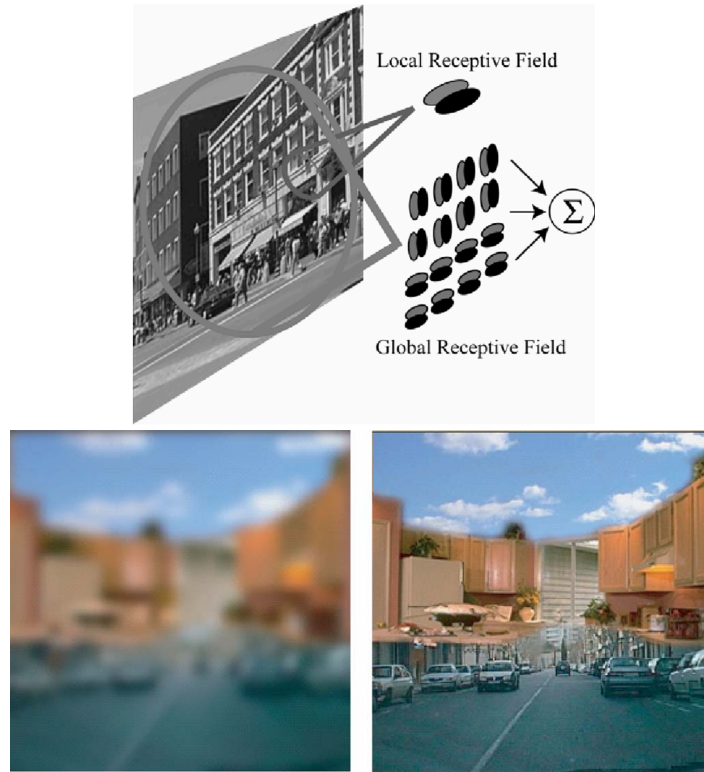
Observers can rapidly extract global information from a scene, referred to as the image gist [16]. In just a glance, the visual system can form representations with enough information to recognize the scene as well as a few objects. It has been shown that humans can recognize the gist of a scene with over 80% accuracy in as little as 36 milliseconds [1]. This representation can involve all levels of processing from low level colors and spatial frequencies to higher level properties such as spatial relationships and objects.

Scene gist can be understood as both *conceptual* and *perceptual*. The verbal description of an image, such as a semantic description of the scene (eg. 'the floor of a rainforest') is conceptual. Meanwhile, perceptual image properties such as spatial frequency, color, and texture make up the structure of a scene.

## 2.3 Role of Context in CNNs

Several neural network architectures incorporating contextual information have been previously proposed. Statistical correlations between low-level features of context and objects have been used for context-based object priming [26]. Additionally, global contextual features can act as priors for place and object recognition [28]. Face detection has been shown to benefit from a separate network that detects cooccurring bodies [32]. Context can also be incorporated by concatenating predictions made using larger bounding boxes around the same object [6] or through a a Recurrent Neural Network (RNN) that moves laterally across an image and updates information at each step [4]. Integrating information at two different resolutions can improve action recognition [10]. A two-part convolutional model that concatenates an RCNN with a contextual network can improve object detection [8]. How-





**Figure 2.3:** Top: Global features of a scene are extracted from the entire image, while local features can be extracted from specific regions. Bottom: Despite being given only global features, observers can confidently describe the scene of an image. This effect speaks to the strong use of priors humans use through the gist of a scene [18].

ever, so far these models have focused on integrating contextual information using a fixed set of convolutional filters and pooling operations for all types of contextual information. In our computational model, we focus on how changes in how information is processed by CNNs can lead to improvements in object recognition.

# 3

## Methods

THE ROLE OF A SCENE’S GIST ON OBJECT RECOGNITION is measured both psychophysics experiments and in computational models. First, starting with human observers, we conduct a set of experiments that isolate the effect of context and restrict the amount of time observers are allowed to view the image. We then observe the relative effects of adding context and use them as a baseline for benchmarking our computational models. For our computational model, we use a two-stream convolutional neural network, where one stream is fed just the object and the other stream is fed just the context.

### 3.1 Dataset

We analyzed images from the MS-COCO dataset, which has been widely used for object-in-scene benchmarks [13]. This data has a higher diversity of categories within each image and instances per category when compared

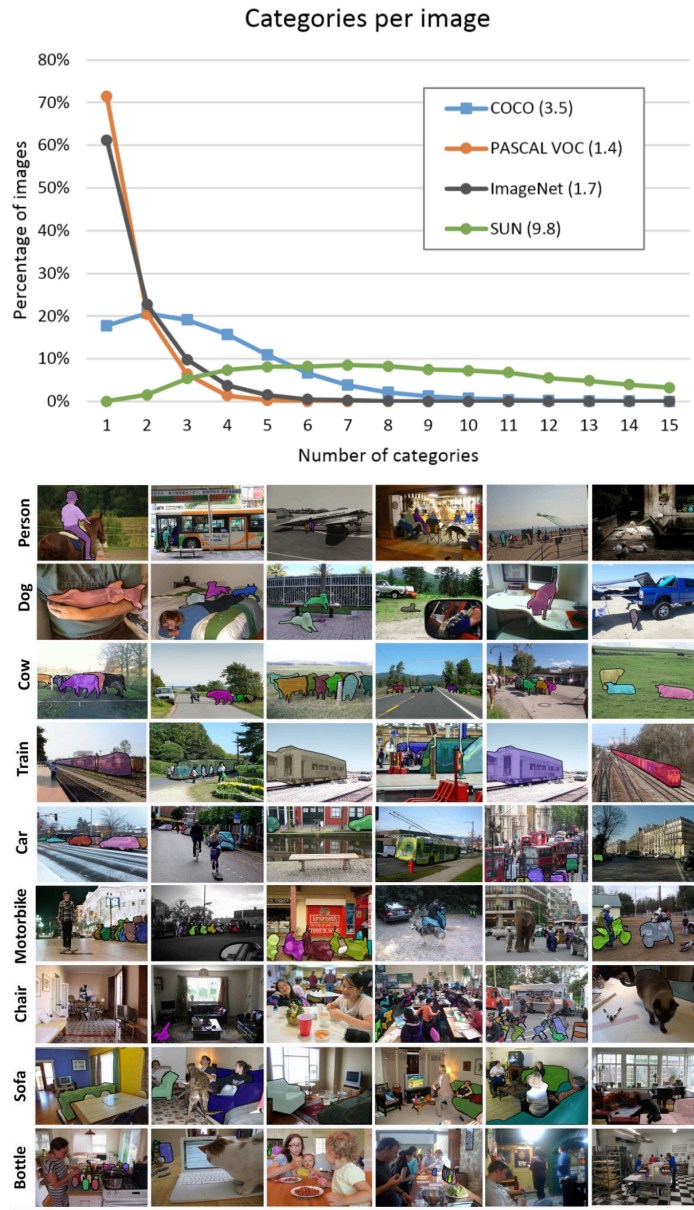
to other object detection and recognition datasets like ImageNet, PASAL VOC, making it desirable to study the effects of context.

The 2014 training dataset (83K images) was used for training and the 2014 validation dataset (41K images) for testing. The images contain objects that span 80 categories including bicycle, car, dog, clock, etc. (Figure 3.1)

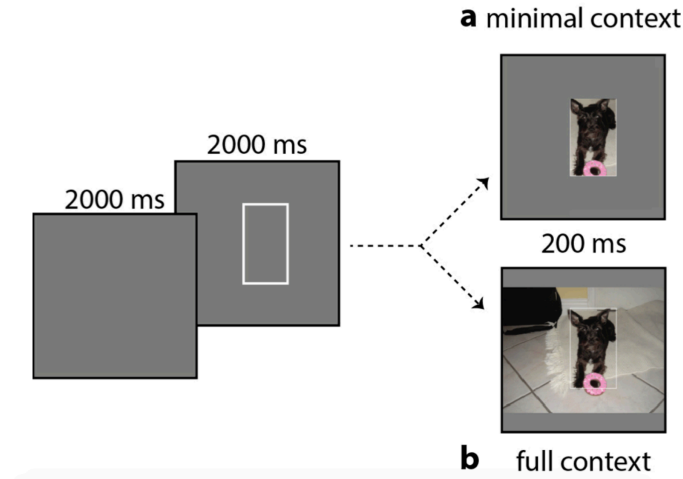
### 3.2 Behavioral Experiments

We designed an psychophysics experiment to isolate the effect of context. The behavioral experiments were performed on Amazon Mechanical Turk, an online task platform. A total of 1,000 objects were randomly selected from the MS-COCO dataset, with one object selected from each image. Each object was shown to observers under two possible conditions: (i) minimal context (object with minimal bounding box and (ii) full context (entire scene. (Figure 3.2) All the images were shown in color. On average, the minimal context images were 154x151 pixels in size whereas the full context images were 468x585 pixels in size.

In each trial, subjects were shown a blank gray screen (2 seconds), followed by a white bounding box indicating the location of the upcoming object of interest (2 seconds), and then the image with the white bounding box for 200 ms (Figure 3.2). In order to minimize eye movements and thus fix the observer’s fovea to the object of interest, we limited the exposure time to 200 ms. Each image, in each of the two possible context conditions, was labeled by three subjects, and each subject was asked to provide up to three labels for the object inside the box. Each response was manually checked and compared to the corresponding ground truth label. If two of the three subjects’ labels matched with the ground truth label for a given object, the image was scored as correctly recognized. To prevent variable viewing conditions across subjects, we set a fixed frame height of 650 pixels and restricted frame and image resizing due to differences in browser window size or screen resolution. Each sequence was saved as a GIF file and rendered by the subject’s browser after full loading.



**Figure 3.1:** Top: A distribution of the number of categories per image across several datasets. Bottom: Sample images and their segmentations from MS-COCO dataset. [13]



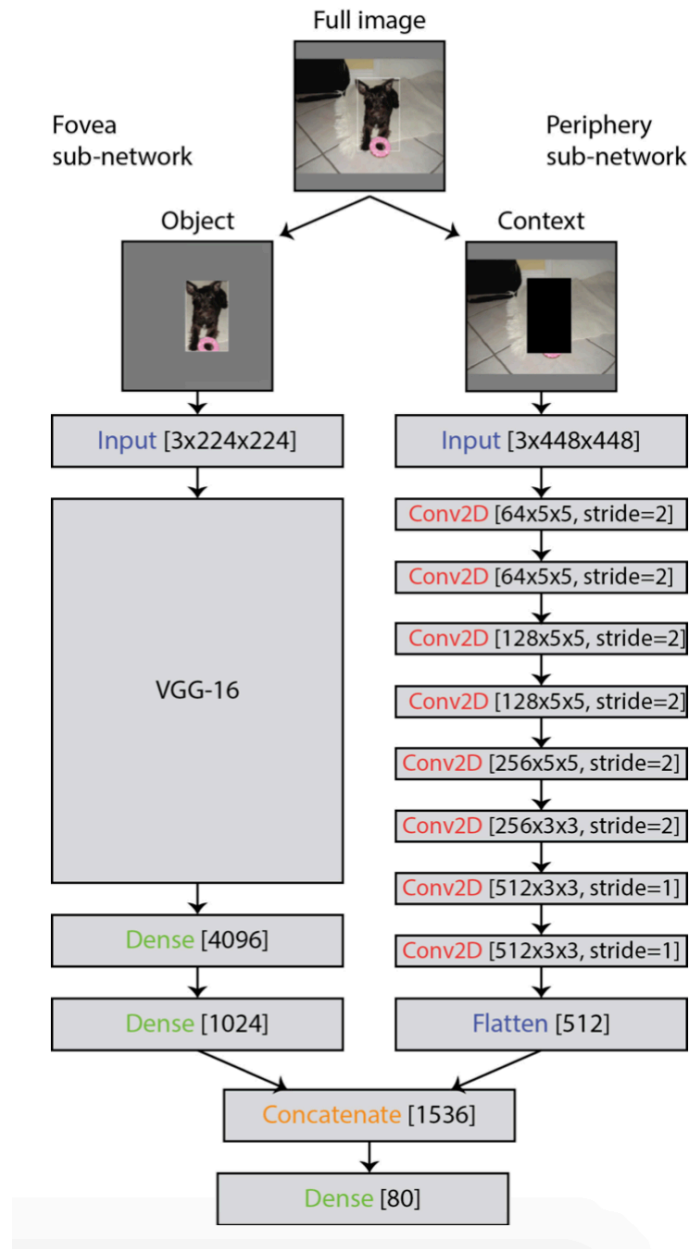
**Figure 3.2:** A schematic of the psychophysics experiment. Each observer is given an object either with or without context for 200 ms after being given a fixation cue beforehand. Minimal context is shown on top, while full context is shown on the bottom.

### 3.3 Computational Model of Scene Gist

The architecture consists of two sub-networks, a foveal network and a peripheral network (Figure 3.3).

#### 3.3.1 Fovea sub-network for object recognition

We used a modified version of the VGG-16 architecture as our baseline model for object recognition. This architecture has been shown to perform well on datasets like ImageNet [23]. We follow the convolutional layers of VGG-16 with two fully connected layers of length 4096 and 1024. While the original implementation of VGG-16 had two fully connected layers of length 4096 followed by one layer of 1000 for classification, we reduce the size of one of the length 4096 layers to 1024 save computational costs. We finally finish with a final classification layer of length 80. We constrained the input size to be  $3 \times 224 \times 224$ , which is the default input dimension for VGG-16.



**Figure 3.3:** GistNet architecture.

### 3.3.2 Periphery sub-network for contextual modulation

#### Designing a computational model of the periphery

Computational models of gist need to be able to capture useful information about a scene at a relatively low computational cost. To encourage the model to pick up gist-like features, we look to the properties of human peripheral vision that are distinct from the fovea – namely, a smaller number of units with larger receptive fields. These properties result in lower visual acuity and reduced sensitivity to detail. We capture similar effects of larger receptive fields in convolutional neural networks with larger kernel sizes, where values in the activation space are derived from a larger field of information from the feature space. We omitted max-pooling operations found in VGG-16 to better preserve spatial information in a scene – by extracting information from regions of highest activity, max-pooling operations reduce information about spatial structure [21]. Instead, we used larger stride sizes in earlier layers to reduce the dimensionality of context and preserve spatial information [9], [14].

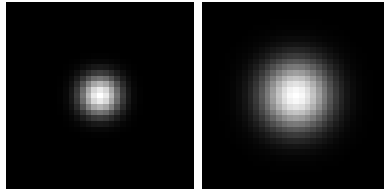
#### Receptive field size as a function of kernel size

To better understand the effect of kernel sizes on larger receptive field sizes, we calculated the *path distribution* of CNNs with different kernel sizes. We define the *path distribution* to be the number of paths that each pixel in the input image has to reach a neuron in the final layer of the CNN. We postulate that the more paths that an input pixel has to reach the final neuron, the more ability it has to influence its value.

Figure 3.4 shows an example of how filter size increases the receptive field size of neurons in a CNN.

#### Implementation

The periphery network uses an 8-layer fully-convolutional neural network structure. The context input is 3x448x448 in size and contains the entire scene, minus the object (the minimal bounding box is replaced with zero



**Figure 3.4:** A 2D plot of the distribution of paths to a single neuron over 8 convolutional layers. Here, we use the number of paths as a proxy of understanding the receptive field of neurons. Left: The path distribution when using a 3x3 filter size. Right: The path distribution when using a 5x5 filter size.

values). The first 5 layers have a 5x5 kernel size, followed by 3 layers with 3x3 kernel size, all with ReLU activation. The first 6 layers of the network have a stride of 2, while the last 2 layers have a stride of 1. The flattened layer is concatenated with the penultimate layer in the foveal network. Finally, this concatenated layer is followed by a dense layer of length 80 for classification. In total, this gist model adds 5.7M parameters to the baseline VGG-16 model, which is less than 5% of the total number of parameters in our baseline model (121M).

### 3.3.3 Training

Both models were trained over 1M single-batch iterations using the Adam optimizer [11] with a starting learning rate of  $10^{-6}$ . The fovea subnetwork was given pre-trained weights from ImageNet, while the periphery subnetwork was not since the architecture deviates from VGG-16.

### 3.3.4 Evaluation

Models trained on the MS-COCO dataset have typically been benchmarked on metrics like mean average precision (mAP) to measure both detection and recognition. However, here we evaluated recognition alone given that a region of interest has already been determined. As such, we focused on category-wise prediction accuracy as our primary metric for our computational models. We provide top-1, top-3, and top-5 accuracy to be consistent with measures of recognition rates on datasets like ImageNet and CIFAR-10.



# 4

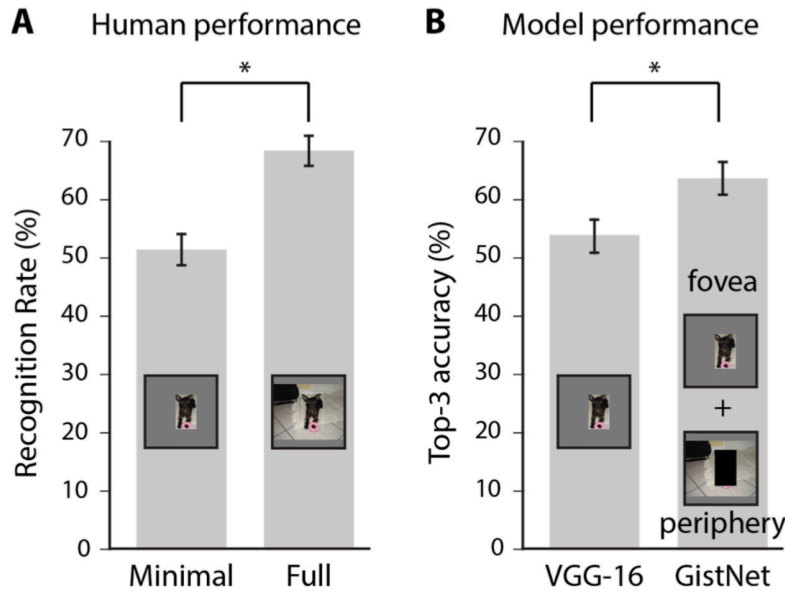
## Results

WE PRESENT THREE MAIN RESULTS: First, we present the observed positive effect of context on human recognition. Second, we show that our computational model improves with context, but with a relatively lower margin. Finally, we display how the features learned by the peripheral sub-network exhibit traits consistent with scene gist processed by the periphery in the human visual system.

### 4.1 Human Object Recognition Improves With Context

#### 4.1.1 Minimal Context Condition

Humans showed 51.4% performance (95% confidence interval of [48.13, 54.6]%) in the minimal context condition (Figure 4.1). Whereas the computational models were forced to classify objects into 80 possible categories, in the behavioral experiments subjects were free to use any word to de-






**Figure 4.1:** (A) Object recognition performance in the behavioral experiment based on  $n=1000$  images for minimal context (left) or full context (right). Error bars denote 95% confidence intervals. (B) Top-3 accuracy for the same set of 1000 images for the VGG-16 (left) and GistNet model with full context

scribe the images. Hence, there is no clear definition of chance levels for the psychophysics experiments. However these results show that subjects performed reasonably well in this task given the constraints and provide a baseline to evaluate recognition performance under limited exposure of small unsegmented objects with minimal context.

#### 4.1.2 Full Context Condition

When subjects were presented with full context images, their performance increased to 68.5% (95% confidence interval of [65.0, 71.8]%, Figure 4.1). Here, we determine a correct prediction to mean two out of three subjects agree on the same correct ground truth. The recognition rates for one out of three subjects correctly labeling an object without and with context are 68.5% and 88.1%, respectively.

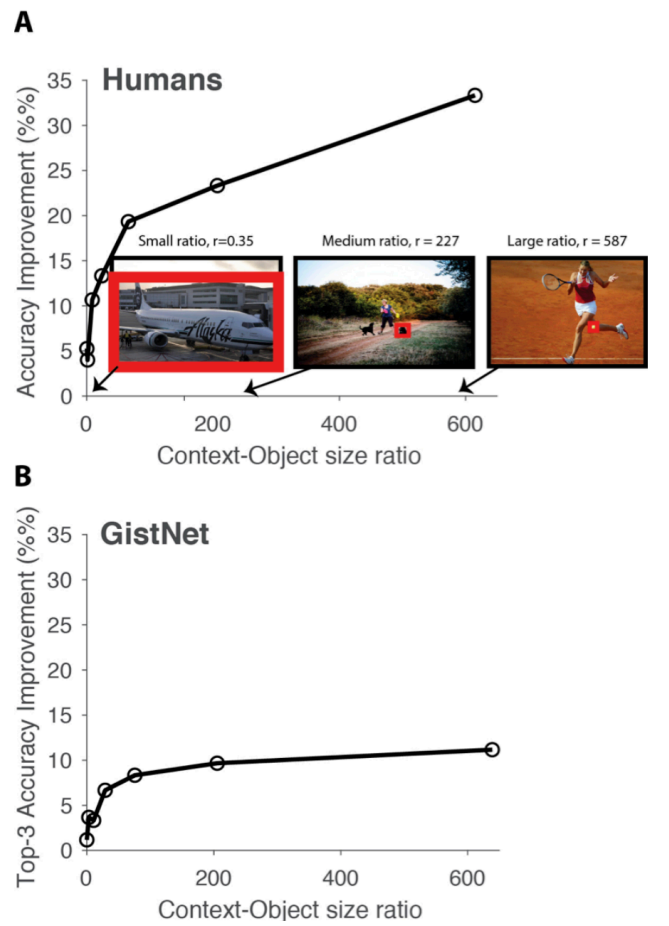
	Humans		Models		Ground Truth
	minimal context	full context	VGG-16	GistNet	
<b>A</b>  image_id: 413616	house frisbee disk	hat frisbee disk	pizza	frisbee	frisbee
<b>B</b>  image_id: 65969	boat water bottle canoe	vase vase+flower vase	boat	suitcase	vase
<b>C</b>  image_id: 544299	appliance machine sofa	toaster toaster toaster	clock	toaster	toaster

**Figure 4.2:** Three example images showing the minimal context bounding box (red), the descriptions from 3 subjects with minimal or full context, the VGG-16 and Gist-Net labels and the ground truth labels. (A) Context helps the model, humans got it right without context; (B) Context helped humans, models got it wrong. (C) Context helps both humans and models.

### 4.1.3 Contribution of Context Increases For Small Objects

The effect of context is not the same for all objects. In our study, we observed that context has a much smaller effect when the object is already provided in high-resolution, whereas small objects benefit much from having context.

The proportion of context in an image correlates with how much context improves object recognition performance. We compute the ratio  $r$  between the number of pixels in the context (pixels in image - pixels in object) and the number of pixels in the object. Object recognition improvement with context increased approximately logarithmically with  $r$ . At a ratio of  $r \sim 200$ , the improvement was as large as 30%.



**Figure 4.3:** Context Helps More With Smaller Objects. We measure the effect of context as a function of object size to be logarithmically increasing. (A) shows accuracy improvements for humans, while (B) shows improvements for the computational model.

**Table 4.1:** Accuracy on GistNet on MS-COCO 2014 validation set.

Average Accuracy on 41K images from 2014-val			
	Top-1 ACC	Top-3 ACC	Top-5 ACC
Object Only	35.3%	55.2%	65.1%
Object + Gist	<b>41.1%</b>	<b>61.7%</b>	<b>71.1%</b>

**Table 4.2:** GistNet vs. Other Models

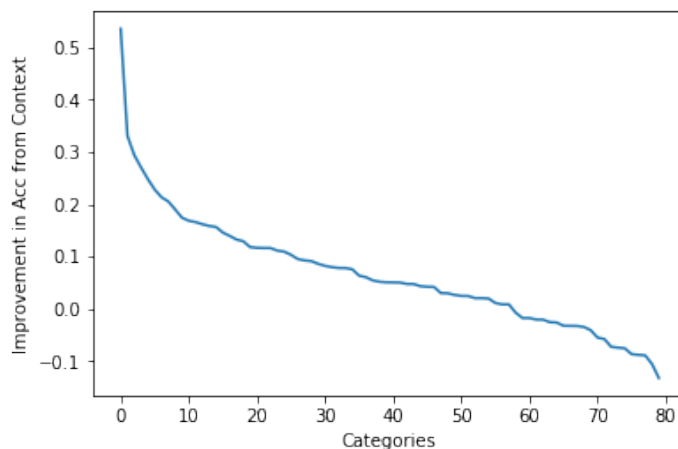
Comparisons With Other Models				
	Top-1 ACC	Top-3 ACC	Top-5 ACC	# Parameters
VGG-16 (Only Object)	35.3%	55.2%	65.1%	<b>121M</b>
VGG-16 (+10% crop)	37.2%	57.2%	65.9%	<b>121M</b>
VGG-16 (+25% crop)	38.7%	58.2%	66.9%	<b>121M</b>
Spotlight Net	39.4%	60.8%	70.2%	244M
GistNet	<b>41.1%</b>	<b>61.7%</b>	<b>71.1%</b>	127M

## 4.2 GistNet Captures Gist-Like Context

We trained both the VGG-16 model and the GistNet model for object classification. In the minimal context condition, the VGG-16 model achieved a top-3 accuracy of 55.2% (95% confidence interval [54.7, 55.7]%, chance = 1.25%, (Figure 4.1). Introducing the full contextual information improved average top-3 category accuracy by 6.5%. Table 4.1 reports top-1, top-3 and top-5 accuracy for the minimal context and full context conditions. Using other architectures instead of VGG-16 with the minimal context condition did not change the conclusion. For example, top-1 performance was 38.8% for ResNet50, improving upon VGG16 at 35.3% but still below GistNet at 41.1%.

As noted above, we cannot quantitatively compare the computational and behavioral results. At a qualitative level, the computational model captures the behavioral improvement in accuracy when contextual information is incorporated. Figure 4.2 shows several examples illustrating a variety of contextual effects where adding gist-like information can help humans, the model, or both. Adding slightly more contextual information to VGG-16 improves its performance but does not reach the performance of GistNet.

Using the Spotlight network model, which also contains a second stream processing whole context [8], yields lower performance than GistNet. Additionally, Spotlight net requires almost twice as many parameters as GistNet. A breakdown by object category shows that gist can improve recognition rates by as much as 30-50% in categories like 'fork' and 'spoon' (Table 4.3). Context does not always help, such as in categories like 'couch' and 'apple'. Top-3 accuracy increases in 58 out of 80 (72.5%), of categories. Figure 4.4 shows the distribution of net change in accuracy across categories.



**Figure 4.4:** The net positive contribution of context on accuracy rate. We observed that 58/80 categories improved accuracy as a result of including contextual gist.

Similar to the behavioral experiments, GistNet also revealed a logarithmic increase in the improvement due to context with an increasing ratio of context to object (Figure 4.3). Performance in the GistNet model improved by as much as 10% in images with a 200:1 context to object size ratio.

#### 4.2.1 Understanding When To Use Context

We are interested in when context is useful. By analyzing the KL-Divergence of the softmax prediction layer from a uniform distribution, we may have a proxy for a network’s certainty. Intuitively, since the softmax layer is

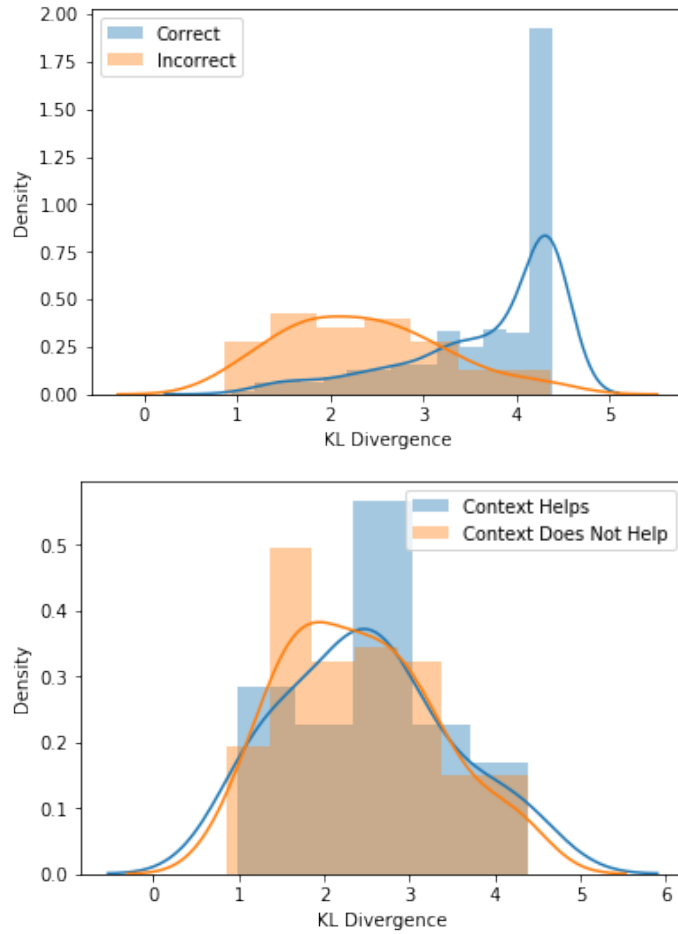
**Table 4.3:** Recognition accuracy for the 4 categories showing the largest positive context effect (top) and the 4 categories showing the largest negative context effect (bottom), based on 41k total MS-COCO images in the 2015 validation set. (95% confidence intervals in parenthesis).

Category-wise Accuracy				
Category	Spoon	Fork	Mouse	Racket
Object Only	21.1% (18.3, 23.4)	7.4% (5.9, 8.9)	65.1% (31.0, 38.2)	55.7% (52.9, 58.5)
Object + Gist	74.7% (72.2, 77.2)	40.5% (37.7, 43.3)	64.0% (60.4, 67.6)	82.7% (80.5, 84.9)
Category	Couch	Apple	Keyboard	Cell Phone
Object Only	53.2% (50.6, 55.8)	63.1% (58.8, 67.4)	65.6% (62.3, 69.0)	40.3% (37.9, 42.6)
Object + Gist	39.9% (37.4, 42.4)	52.4% (48.0, 56.8)	56.6% (53.0, 60.1)	31.5% (29.9, 33.7)

constrained to sum to 1, then lower values would indicate that there is less weight on any particular prediction, while higher values mean that there are particular classes that are given much greater weight. In Figure 4.5, we observe the divergences for cases when the object-only network produces correct and incorrect predictions differs qualitatively. Meanwhile, we there is no clear difference between the divergence for when context helps from when it does not. We postulate that KL-Divergence alone is not a strong enough indicator for knowing when to add context, and this would be an interesting area for further research.

#### 4.2.2 Gradient-Based Interpretation of Gist

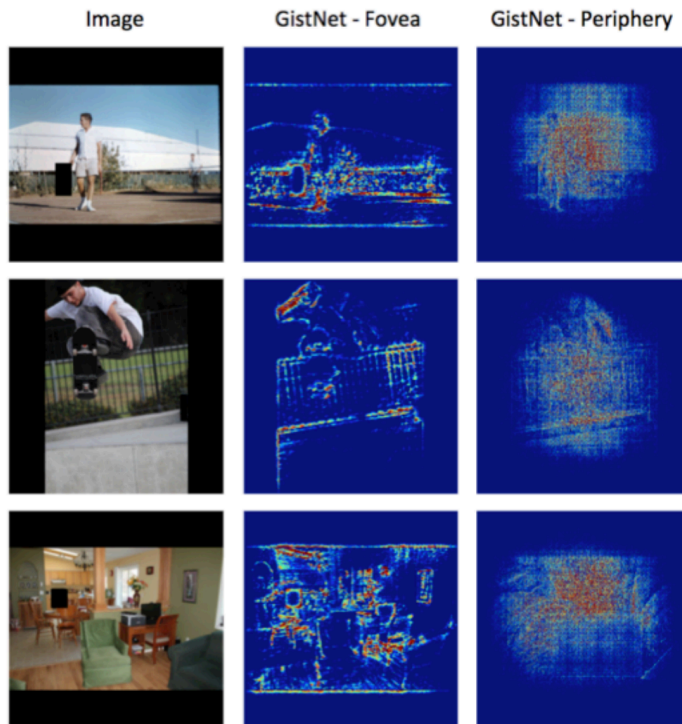
To gain intuition about the image features used by the fovea and periphery sub-networks, we calculated the gradient of the input image with respect to the fully connected layers from the periphery sub-network and the fovea sub-network, called a saliency-map [24]. Figure 4.6 shows example images and the saliency maps for each through the fovea and periphery sub-networks. Qualitatively, while VGG-16 determines detailed lines and edges as important for object recognition, GistNet focuses on broader and more uniform features. As outlined in the introduction, we think of the periphery sub-



**Figure 4.5:** Top: The distributions of KI-Divergences from a Uniform distribution for correct and incorrect predictions for a VGG-16 given only the object. Bottom: The same distributions for GistNet when context helps and when it does not.

network as providing coarse scene gist information using units with coarser receptive fields. This is consistent with our earlier analysis of the larger receptive fields resulting from larger kernel sizes.



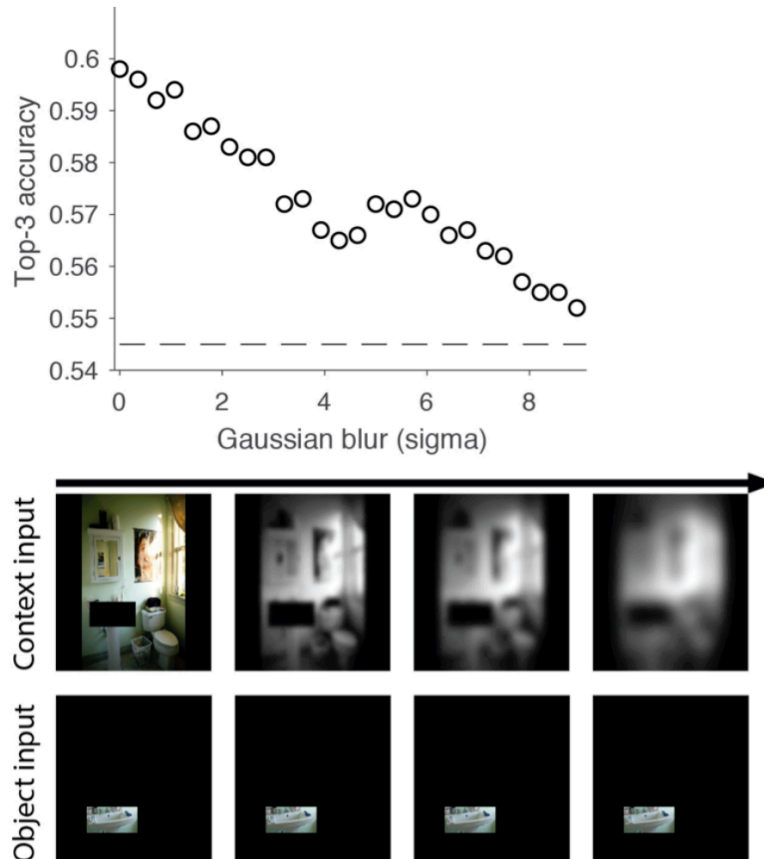


**Figure 4.6:** We calculate the gradient of the input image with respect to the final fully connected layers from both a VGG-16 and GistNet’s periphery sub-network. Both networks are given the whole image (object and context). We can qualitatively show that the periphery sub-network is affected by a much larger coverage on the input image when compared to the VGG-16.

### 4.2.3 Robustness to Blurring

We conjectured that the improvement introduced by GistNet would be robust to significant degrees of scene blurring. This would corroborate our understanding of the human periphery’s larger receptive fields, which captures more global features rather than finer local features. In order to test the extent to which GistNet uses gist-like features to aid in object recognition, we produced predictions at 40 different levels of context degradation using Gaussian blurring. Blurring was applied only to the context and not to the object. Figure 4.7 shows object recognition rates for GistNet vs. baseline accuracy from VGG-16 as a function of the level of Gaussian blurring

introduced. Even when object-distinguishing features are blurred away from the context, GistNet still performs favorably compared to VGG-16.



**Figure 4.7:** Context improves GistNet top-3 accuracy (blue) with respect to VGG-16 (green) even after significant amount of blurring is applied to the context input (x-axis). Bottom: Example image with blurred context and constant object.

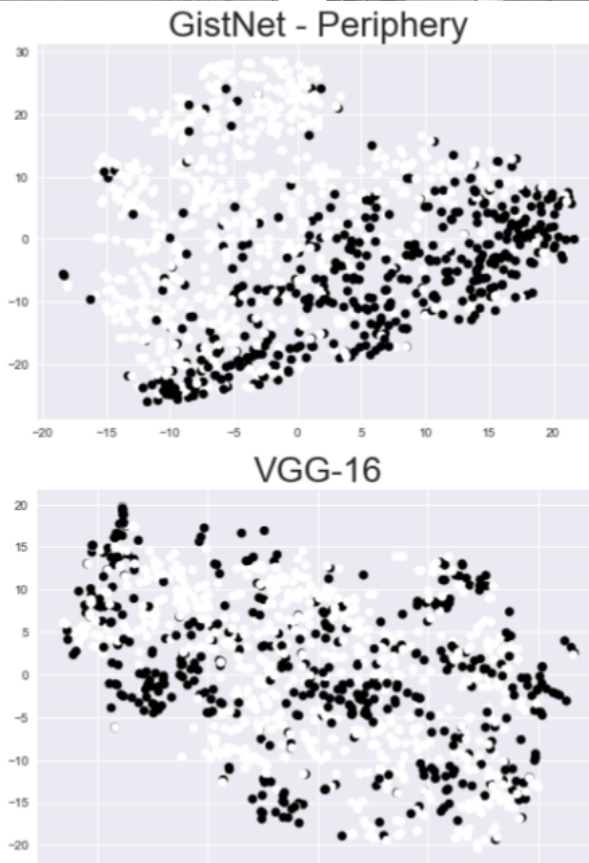
#### 4.2.4 Learning Representations of Semantic Context

We examined whether the peripheral component of GistNet can learn gist-like scene understanding as a byproduct of object recognition. To do so, we first labeled 1,000 images from the MS-COCO dataset as an indoor or

outdoor scene on Amazon Mechanical Turk. Each subject was given a whole image and was given an unlimited amount of time to choose between 'indoor' or 'outdoor' labels on each scene.

To evaluate the type of representations each sub-network creates, we fed each image in its entirety through the foveal and peripheral sub-components and extracted the fully connected layer from each. We used t-distributed stochastic neighbor embedding (t-SNE) [15], a commonly used non-linear dimensionality reduction technique, to visualize the representation layers in two dimensions. After plotting the representations for each, we overlaid the true labels which are unknown to each network at inference.

Qualitatively, the periphery sub-network represents scene-level information much more clearly than VGG-16, as can be appreciated by the better visual separation of the indoor and outdoor labels. Varying the perplexity parameter in tSNE from 5 to 45 in increments of 10 did not produce any appreciable differences in the tSNE visualization in Figure 4.8. A logistic regression classifier trained on the dense layer weights yields an accuracy of 72.2% and 75.1% with VGG-16 and GistNet, respectively, to separate indoor and outdoor images. When training the classifier using the tSNE embedding, the classification accuracy was 61% and 80% with VGG-16 and GistNet, respectively.



**Figure 4.8:** Top: An example of an image labeled outdoor (left) and an image labeled indoor (right). Bottom: a plot of t-SNE embeddings on the fully connected layers for a VGG-16 trained only to recognize objects (top) and the peripheral sub-network of GistNet trained to incorporate context (bottom).

# 5

## Conclusion

OUR RESULTS SUPPORT THE NOTION THAT SCENE GIST CAN IMPROVE OBJECT RECOGNITION IN BOTH HUMANS AND COMPUTERS. We show that a network that is inspired by larger receptive fields and a lower density of receptor cells in the human periphery can learn representations are both useful for object recognition and can exhibit gist-like qualities.

The behavioral experiments show that adding 200 ms of exposure to contextual information can provide a large advantage in object recognition. Qualitatively, this effect is reproduced in the proof-of-principle dual architecture presented, whereby a sub-network processes information within the fovea and a second sub-network provides gist-like features. We posit that the peripheral subnetwork, with larger kernels and wider strides instead of max-pooling, mimics the biological functions of the human peripheral system in creating a gist-like understanding of a scene. This computational model also shows an improvement in object recognition performance when the full context is used, even though it only increases the overall parameter size by

5%. We pursued three approaches towards understanding what aspects of the scene are used by GistNet: (1) We constructed saliency maps to visualize the image areas that were used by each sub-network. While the fovea sub-network finds local edges and lines, GistNet finds more holistic scene information corresponding to gist-like features. This also suggests that the periphery sub-network is not merely learning additional local features to increase the power of the fovea subnetwork. (2) GistNet still outperforms VGG-16 even when provided highly blurred context. This observation provides additional support to the notion that the peripheral component of GistNet learns to extract global features that are preserved through significant blurring. (3) A dimensionality reduction analysis of the penultimate dense layers within VGG-16 and GistNet reveals that the periphery sub-network encodes scene-level information. We introduce the indoor/outdoor label as a perceptual property of the scene. The clear separation of points in our lower dimensional embedding of the representations implies that the periphery subnetwork is able to learn perceptual properties in an unsupervised manner.

### 5.0.1 Limitations

While the computational experiment involved forced-choice 80-way categorization and the human behavioral experiment involved free object naming making direct comparisons difficult to interpret, the contextual effect was larger for humans than the model. One potential source for this difference is that humans could selectively use context during recognition. Intuitively, some objects are easy to recognize and may not require context. Obscured or small objects, on the other hand, may be highly dependent on context for accurate recognition. This observation is supported by the large differences between category performance. Our formulation of GistNet involves a simple concatenation of the peripheral network to an existing object recognition model. Since inference in neural networks is deterministic, the contribution of contextual features will be the same each time. A gating or weighting mechanism in the concatenation layer that determines the "usefulness" of the context before merging can reduce instances where context does not help

or even may hurt recognition. However, measures of network confidence are still underdeveloped as shown in our analysis of KL-Divergence.

### **5.0.2 Application to state-of-the-art models**

We chose to use the VGG-16 model as a baseline and backbone for the fovea sub-network. Yet, the GistNet architecture should be amenable to most existing object recognition models. There exist exceptions for R-CNNs, since feature maps are computed once at a fixed combination of kernel size and strides. Unlike GistNet, R-CNNs only compute features at one type of resolution. Global gist-like features constitute but one aspect of contextual information. Future efforts should also benefit from combining gist with other contextual cues including temporal information, high-level semantic context, and temporal integration via active sampling through multiple saccades.

## Bibliography

- [1] URL <https://www.k-state.edu/psych/vcl/basic-research/scene-gist.html>.
- [2] Peripheral vision. URL <http://www.webexhibits.org/colorart/ag.html>.
- [3] Moshe Bar, Karim S Kassam, Avniel Singh Ghuman, Jasmine Boshyan, Annette M Schmid, Anders M Dale, Matti S Hämäläinen, Ksenija Marinkovic, Daniel L Schacter, Bruce R Rosen, et al. Top-down facilitation of visual recognition. *Proceedings of the National Academy of Sciences of the United States of America*, 103(2):449–454, 2006.
- [4] Sean Bell, C Lawrence Zitnick, Kavita Bala, and Ross Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2874–2883, 2016.
- [5] Irving Biederman. Perceiving real-world scenes. *Science*, 177(4043):77–80, 1972.
- [6] Xinlei Chen and Abhinav Gupta. Spatial memory for context reasoning in object detection.
- [7] Carolina Galleguillos and Serge Belongie. Context based object categorization: A critical survey. *Computer vision and image understanding*, 114(6):712–722, 2010.
- [8] Georgia Gkioxari, Ross Girshick, and Jitendra Malik. Contextual ac-



tion recognition with r\* cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1080–1088, 2015.

- [9] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.
- [10] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.
- [11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [12] Adam M Larson and Lester C Loschky. The contributions of central versus peripheral vision to scene gist recognition. *Journal of Vision*, 9(10):6–6, 2009.
- [13] 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.
- [14] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 4898–4906, 2016.
- [15] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.
- [16] Aude Oliva. Gist of the scene. In *Neurobiology of attention*, pages 251–256. Elsevier, 2005.
- [17] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.

- [18] Aude Oliva and Antonio Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23–36, 2006.
- [19] Stephen E Palmer. The effects of contextual scenes on the identification of objects. *Memory & Cognition*, 3:519–526, 1975.
- [20] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.
- [21] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems*, pages 3859–3869, 2017.
- [22] Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision*, 81(1):2–23, 2009.
- [23] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [24] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.
- [25] Simon Thorpe, Denis Fize, and Catherine Marlot. Speed of processing in the human visual system. *nature*, 381(6582):520, 1996.
- [26] Antonio Torralba. Contextual priming for object detection. *International journal of computer vision*, 53(2):169–191, 2003.
- [27] Antonio Torralba and Aude Oliva. Statistics of natural image categories. *Network: computation in neural systems*, 14(3):391–412, 2003.
- [28] 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.

- [29] Marika Urbanski, Olivier A Coubard, and Clémence Bourlon. Visualizing the blind brain: brain imaging of visual field defects from early recovery to rehabilitation techniques. *Frontiers in integrative neuroscience*, 8:74, 2014.
- [30] Andrew B Watson. A formula for human retinal ganglion cell receptive field density as a function of visual field location. *Journal of Vision*, 14(7):15–15, 2014.
- [31] Christian Wolf and Alexander C Schütz. Trans-saccadic integration of peripheral and foveal feature information is close to optimal. *Journal of Vision*, 15(16):1–1, 2015.
- [32] Chenchen Zhu, Yutong Zheng, Khoa Luu, and Marios Savvides. Cms-rccnn: contextual multi-scale region-based cnn for unconstrained face detection. In *Deep Learning for Biometrics*, pages 57–79. Springer, 2017.