Object Detection in Deep Neural Networks Differs from Humans in the Periphery

Anne Harrington[1,2] **Vasha DuTell**[1,2] **Mark Hamilton**[1] **Ayush Tewari**[1]
Simon Stent[3] **William T. Freeman**[1] **Ruth Rosenholtz**[1,2]
[1]MIT CSAIL [2]MIT Brain and Cognitive Sciences [3] Toyota Research Institute
{annekh,vasha}@mit.edu

Abstract

To understand how strategies used by object detection models compare to those in human vision, we simulate peripheral vision in object detection models at the input stage. We collect human data on object change detection in the periphery and compare it to detection models with a simulated periphery. We find that unlike humans, models are highly sensitive to the texture-like transformation in peripheral vision. Not only do models under-perform compared to humans, they do not follow the same clutter effects as humans even when fixing the model task to closely mimic the human one. Training on peripheral input boosts performance on the change detection task, but appears to aid object localization in the periphery much more than object identification. This suggests that human-like performance is not attributable to input data alone, and to fully address the differences we see in human and model detection, farther downstream changes may be necessary. In the future, improving alignment between object detection models and human representations could help us build models with more human-explainable detection strategies.

1 Introduction

As object detection deep neural networks (DNNs) become increasingly accurate, attributing performance to input data and aspects of downstream processing is important for refining and selecting models. Currently, attribution in object detection is better understood in humans where texture-like representations in the periphery are thought to both limit and enable the human ability to locate and identify objects in a scene [Whitney and Levi, 2011, Rosenholtz, 2016].

Leveraging our understanding of human visual processing, we devise a peripheral vision object detection task for both humans and models. We ask humans to detect objects in their periphery on COCO images and then stimulate peripheral vision in object detection models at the input stage to compare model detection strategies to humans. We simulate peripheral vision with an image transformation, allowing us to understand how input data effects the similarity between human and model performance. Our analyses are flexible to any detection DNN that takes images as an input, allowing us to evaluate a variety of architectures from CNNs to transformers.

Our results reveal key differences in how sensitive models and humans are to a peripheral vision transformation and uncover that DNN performance is not responsive to the same clutter effects that humans are. Even though many detection DNNs have better baseline performance on large objects than small, we do not observe the same pattern when simulating peripheral vision suggesting that there is a deeper, downstream mis-alignment between human and DNN behavior that cannot be attributed to input data alone. These patterns hold for a DNN trained with a simulated periphery, with trained models showing improved object localization in the periphery, but interestingly not identification. Overall, by quantifying DNN and human alignment on a peripheral vision task, we are able to better understand strategies used in machine object detection.

ATTRIB: Workshop on Attributing Model Behavior at Scale at NeurIPS 2023.

2 Background

Peripheral vision describes the process in which human vision represents the world with decreasing fidelity at greater eccentricities, i.e. farther from the point of fixation. Over 99% of the human visual field is represented by peripheral vision. While it is thought to be a mechanism for dealing with capacity limits from the size of the optic nerve and visual cortex, peripheral vision has also been shown to serve as a critical determinant of human performance for a wide range of visual tasks [Whitney and Levi, 2011, Rosenholtz, 2016]. Peripheral vision has been successfully modeled as a loss of information in representation space [Rosenholtz et al., 2012b, Freeman and Simoncelli, 2011], where models like TTM [Rosenholtz et al., 2012b] perform a texture-processing-like computation of local summary statistics within pooling regions that grow with eccentricity and tile the visual field (see Fig. 6 for an example). This model has been tested to well predict human performance on an extensive number of behavioral tasks, including peripheral object recognition, visual search, and a variety of scene perception tasks [Ehinger and Rosenholtz, 2016].

Properties of peripheral vision have been used to explain representational and behavioral qualities of models. Notably, prior work has studied the peripheral vision effect of crowding. Crowding is the degradation of peripheral performance in the face of clutter, often demonstrated as the inability to identify a target object when flanked (surrounded) by other objects. Crowding in humans is complex, however, depending not merely on the presence or spacing of nearby objects, but on the features and complexity of local image regions [Vater et al., 2022]. Standard object recognition CNNs have been shown to have very different crowding effects compared humans [Lonnqvist et al., 2020], and often perform worse at recognition under crowded conditions than human-inspired CNNs [Volokitin et al., 2017]. In addition, robustness to adversarial noise has been linked the texture-like representations of peripheral vision [Harrington and Deza, 2022], and robustness to occlusion and other biases were seen in a scene classification model trained on peripheral vision-like images [Deza and Konkle, 2020]. All of these studies, however, are limited to object recognition models. Peripheral vision is critical for task where visual context matters such as detection. In order to explain model behavior in the context of peripheral vision, it is critical to use richer datasets and tasks like detection.

3 Object Change Detection Task

3.1 Human Experiment

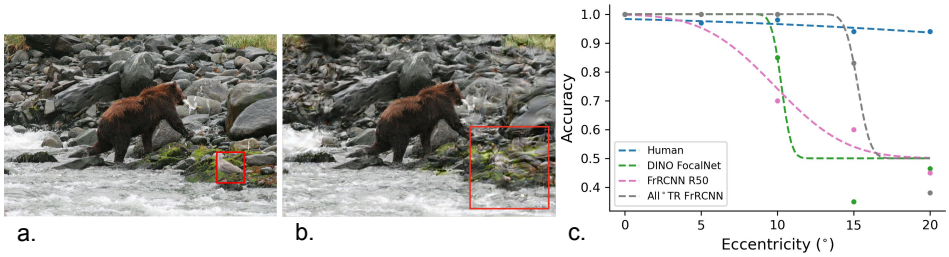

Figure 1: **Example Easy Object Detection. (a)** Original image with target object bounding box, and **(b)** TTM transform for 15° (240 pixels) with extended bounding box (used to perform machine object detection task). **(c)** Human accuracy (blue), compared with accuracy for a pre-trained (pink) and trained on TTM (gray) Faster RCNN R50 model, and a DINO FocalNet model (green). Psychometric curves are fit with an inverse cumulative normal distribution.

In order to compare human and model detection strategies, we create an object change detection task that both can perform. We first collected human psychophysics data on an object detection task. We choose a detection rather than a recognition task because humans can guess object identity quite well based on context alone, i.e. even when the object itself is occluded [Wijntjes and Rosenholtz, 2018]. In our detection task, we present two images on every trial, identical except for the presence or absence of a particular object, and ask a human subject to judge which of the two images contained a target object. For the object present images, we choose 26 images from the COCO validation set that have one instance of an object. For the absent image, we remove that object via in-painting (see Appendix Sec. 8.2.1). We selected images with a variety of small-medium objects in different scenes.

In each trial, 10 eye-tracked subjects fixated at a specified location either $5°, 10°, 15°$, or $20°$ away from the target object, and viewed an object present and absent image in random order. Subjects were asked to report which image contained the specified object in a two-interval-forced-choice paradigm (2IFC), viewing 10 present/absent image pairs at each eccentricity (see Appendix Sec. 8.2.2 and 8.2.3 for more details).

We find overall that human object detection performance always degrades progressively with increasing eccentricity (Figure 1, blue line). See Appendix Sec. 8.2.4 for per-image human accuracy. Detection ability is consistently strong at $5°$. However, for some images observers reach near chance performance at $20°$ eccentricity, whereas a few image pairs have objects that are easily detected at all eccentricities. Often, high color contrast between the object and its background and a lack of clutter from other nearby objects made target objects more easily detected in the periphery, leading to better performance, which is consistent with the crowding literature.

4 Machine Experiment

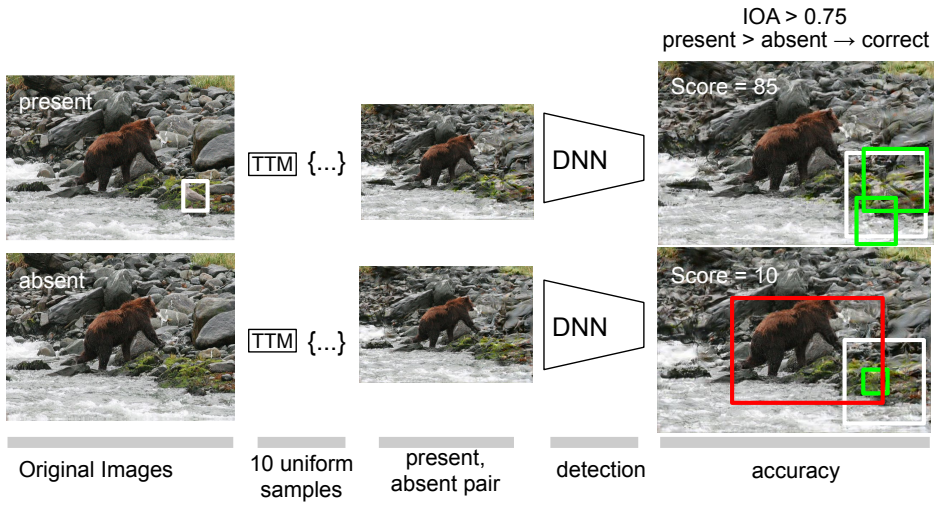

Figure 2: **Workflow for Machine Psychophysics Experiment.** To simulate trials for detection DNNs, we generate 10 TTM transform images for each present/absent image in the experiment. We run inference on the transforms and sum the scores of all box predictions that have an intersection over area (IOA) greater than 0.75%. If the summed score is greater on the present transform, the DNN is recorded as correct for that pairing. White boxes indicate the ground truth and padded ground truth. Green boxes show predictions that meet the area condition, red do not.

To compare human and DNN performance, we have DNN object detection models perform the same two-alternative/interval forced choice task given to human subjects. We start by stimulating peripheral vision in a variety of pre-trained DNNs. We use the Texture Tiling Model [Rosenholtz et al., 2012b], one of the most well tested models of peripheral vision that well predicts human behavior on a variety of tasks (See Fig. 6). We use TTM to transform the object present and absent images used the experiment like human peripheral vision. Because TTM is a stochastic model that is under-constrained compared to image pixel values, we can create 10 different samples of each experiment image to simulate different viewing or experimental trials. By creating 10 samples, this gives us 100 unique image pairings to simulate trials.

For each present/absent pairing, we input a transformed image to the object detection model with low detection threshold (0.01) to get proposed bounding boxes and object scores. We then determine if the proposed box overlaps with a padded ground truth box of the target object; we pad the ground truth box by half the width of a pooling region to account for position uncertainty introduced in human peripheral vision and TTM (see Figure 1 a and b for an example of $15°$ padding). To measure how strongly the DNN predicts there is an object in padded box region, we sum the total scores of all objects (regardless of predicted class) that overlap at least 75 percent (intersection of area with respect to the proposed box). We score the model as correct on a trial if the total object scores for the present image are greater than the absent. We score incorrect if the absent is greater and give a

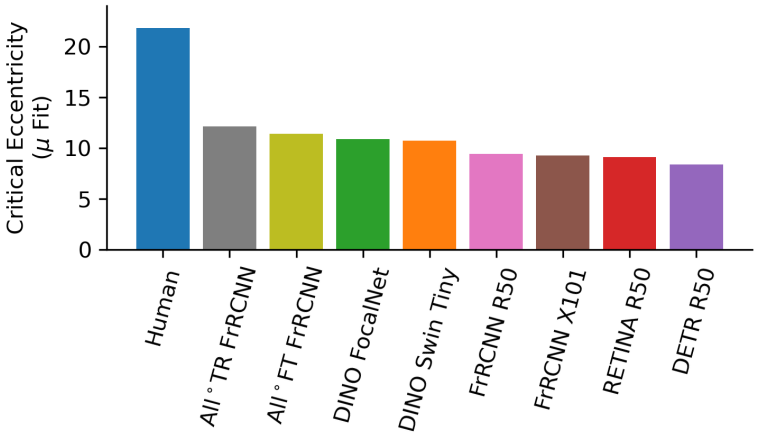

Figure 3: **Detection Critical Eccentricity for Humans and DNNs.** Fitting a psychometric function to the psychophysics performance, we report the average fall-off point of performance in the periphery (μ) over all experiment images.

half score if present and absent are equal. We take the average over all the present-absent pairing accuracies for each eccentricity. See Appendix Sec. 8.3 for pseudo-code and Figure 2 for the general workflow.

To keep the comparison between DNNs and humans fair, we do not enforce that the model must predict the correct object identity when scoring predictions at each trial. Because we use a forced-choice paradigm, humans subjects can give a correct response by simply detecting the presence of any object at the approximate right location, rather the specified one. Although we specify an object class to human subjects, this strategy is likely to happen when peripheral information is poor.

We evaluate a range of object detection DNNs from CNN-based to transformers. These include: DINO FocalNet[Zhang et al., 2022] , DINO Swin Tiny [Zhang et al., 2022], Faster-RCNN-R50 [Ren et al., 2015], Faster-RCNN-X101 [Ren et al., 2015], RetinaNet-R50 [Lin et al., 2017], Detr-R50 [Carion et al., 2020].

5 Humans Out-Perform Models at Peripheral Object Detection

Like the human observers, DNNs' response accuracies are highly image-dependent, with some pairs resulting in poor performance for all models. While human performance falls gradually for most images, DNN object detectors can often retain good accuracy for the 5° eccentricity TTM transforms, but many show sharp falloffs in accuracy to chance performance soon after (See Figure 1 for a representative example).

To quantitatively compare performance, we fit both human and DNN performance data across eccentricity to a psychometric function for each image. We use a reverse cumulative Gaussian distribution which determines the critical (75% correct, halfway between perfect and chance performance) threshold eccentricity by the mean of the distribution (μ), and the performance falloff rate by (σ).

For all images tested, humans outperform all object detection models, with critical thresholds more than 5° greater than detection models (Figure 3). We find generally weak correlations between between DNN and human performance for critical eccentricity (μ) (see Appendix Fig. 12). Among the pre-trained models, DINO detectors have the closest critical eccentricity to humans and have the strongest correlation.

5.1 Training on Peripheral Vision Images

To reduce the gap between human and DNN performance and understand the impact of training on periphery vision inputs, we fine-tune and train a ResNet-50 backbone Faster-RCNN-FPN detection model on images transformed by a variant of TTM – which we call uniform TTM (see Fig. 6 and Appendix Sec. 8.1). Uniform TTM removes the need to select a fixation point in a image and instead models a single distance in the periphery everywhere in an image making it feasible to

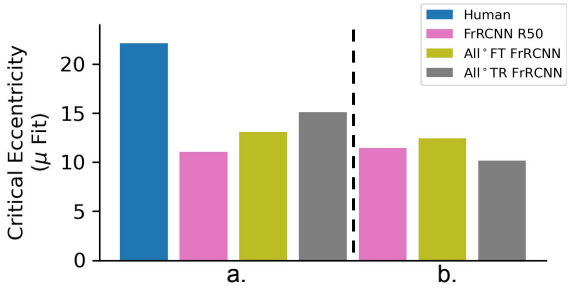

Figure 4: **Detection vs Recognition.** **(a)** Fine-tuning and training on uniform TTM increases baseline Faster-RCNN-R50 performance at the object detection psychophysics task. **(b)** Machine psychophysics performance on recognition. We pass the classification head of Faster-RCNN models the groundtruth bounding box and score performance based on recognition in that region.

render the COCO dataset like peripheral vision. For fine-tuning (plotted as All° FT RCNN) we start from detectron2's [Wu et al., 2019] pre-trained model and use a 3x iteration scheme with a lowered learning rate. When training from scratch (plotted as All° FT RCNN), we use the default 3x training configuration in detectron2.

We find that training a model from scratch using uniform TTM plus original COCO images ($0°$) produces the best performing model in the psychophysics evaluation (see Fig. 3) and has the highest average precision on COCO transformed by uniform TTM (see Appendix Fig. 1). The model trained with uniform TTM has a critical eccentricity of nearly $5°$ greater than the pre-trained baseline (Figure 3) The fine-tuned model, however, under-performs the trained model which we suspect is because of the lowered learning rate during training and a decrease in baseline AP (See Appendix Table 1).

To better understand the impact training on uniform TTM has on the psychophysics performance, we additionally evaluate object recognition in the machine psychophysics (Figure 4). We give the classification head of Faster-RCNN-R50 models the padded ground truth bounding box of the target object. We then score models based on which image, present or absent, has the highest classification probability for the target object. Unlike the detection version of the task, we find that training from scratch performs worse than baseline. This could indicate the trained model improved more at localizing objects rather identifying them in the periphery.

5.2 Effects of Object Size and Clutter

Since both human and DNN performance strongly varied by image, we looked for image properties that might predict performance, and asked if these had similar effects for humans and computer vision DNNs. Examining critical eccentricity as a function of object size, humans have a higher critical eccentricity for larger objects; that is, human performance increases with progressively larger target objects (Figure 5 a). Surprisingly, this relationship does not appear to hold for any object detection model, even the ones trained on uniform TTM.

Human object detection performance in the periphery is known to be strongly mediated by the amount of clutter. One measure of clutter is the number of objects near the target. To test if this holds true for detection models in our experiment, we used the number of ground truth COCO annotations in the image as a proxy for clutter (note clutter can be present in specific sub-regions of an image, and that COCO annotations do not label all objects in many scenes). As expected, human performance decreases as images become more cluttered (Fig 5 b). Performance in object detection models does not show a strong relationship with clutter. This is true even for models trained on uniform TTM, which should reflect the degrading effect of clutter on the peripheral representation, according to TTM.

6 Discussion

To compare detection strategies between humans and DNNs under peripheral vision conditions, we simulate peripheral vision and create a psychophysics experiment. Our results expose a gap in performance between humans and computer vision DNNs in the periphery. When we restrict DNNs with human peripheral vision, detection performance is brittle, degrading sharply while human

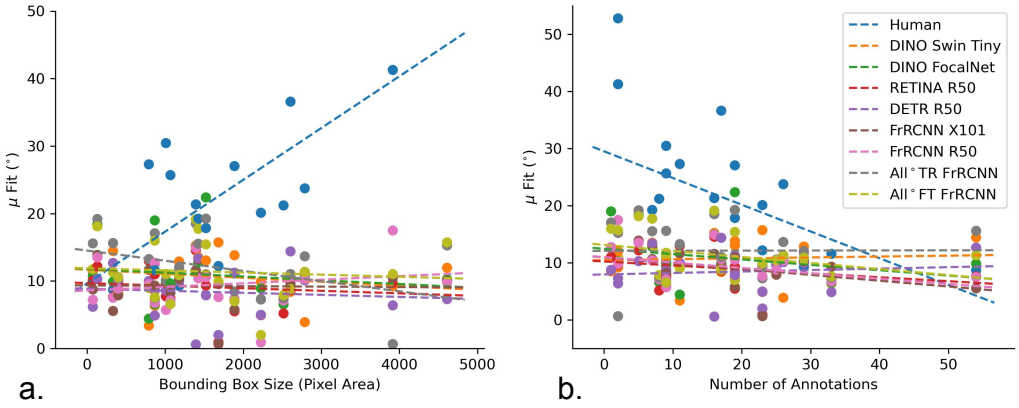

Figure 5: **Object Size and Clutter. a:** Object size predicts critical eccentricity for human observers (blue), but remains low for large objects in all object detection models tested. **b:** In humans (blue), critical eccentricity is highest for images with few objects, with performance decreasing as images become more crowded with objects. This relationship is not observed for tested object detection models, where critical eccentricities remain low.

performance falls off smoothly. Although we do not test human performance on TTM transform images, previous work demonstrated that TTM is a close match to human peripheral vision [Ehinger and Rosenholtz, 2016], so differences in performance cannot be explained by stimuli alone.

In our experiments, we see that DNNs that are more accurate at baseline tend to have a closer detection fall-off to humans in the periphery. This suggests that working to improve models at baseline is not at odds with the goal of training detection models that behave more like humans. However, we also see that even better baseline models like DINO do not display human like performance patterns under clutter. Our result suggests that to get more human explainable mistake patterns in detection models, model representations might need to be further constrained downstream during training.

In our experiment training on images transformed to simulate peripheral vision, we see that training improves DNN performance on the object change detection task. However, the boost in performance we see appears to primarily be attributable to improved object localization, rather than class identification (Fig. 4). Favoring localization over identification aligns with a major goal of peripheral vision – guiding fixation. Training with transformed inputs alone seems to help align models to human object localization patterns, but more work is needed to fully understand this relationship.

In the future, it will be interesting to test more biologically inspired detection models to understand the role those mechanisms play in driving how humans handle clutter and effectively detect objects in the periphery. While we do not train state-of-the-art models on TTM directly, future work in that direction could help further understand how detection models perform under restricted viewing conditions. Overall, by comparing DNNs and humans on a peripheral vision task, our work brings us closer to understanding how input constraints, architecture, and baseline accuracy influence strategies used in object detection.

7 Acknowledgements

This work was funded by the Toyota Research Institute, CSAIL MEnTorEd Opportunities in Research (METEOR) Fellowship, US National Science Foundation under grant number 1955219, and National Science Foundation Grant BCS-1826757 to PI Rosenholtz. The authors acknowledge the MIT SuperCloud Reuther et al. [2018] and Lincoln Laboratory Supercomputing Center for providing HPC resources that have contributed to the research results reported within this paper.

References

N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

A. Deza and T. Konkle. Emergent properties of foveated perceptual systems. *arXiv preprint arXiv:2006.07991*, 2020.

K. A. Ehinger and R. Rosenholtz. A general account of peripheral encoding also predicts scene perception performance. *Journal of Vision*, 16(2):13–13, 2016.

J. Freeman and E. P. Simoncelli. Metamers of the ventral stream. *Nature neuroscience*, 14(9): 1195–1201, 2011.

A. Harrington and A. Deza. Finding biological plausibility for adversarially robust features via metameric tasks. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=yeP_zx9vqNm`.

T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

B. Lonnqvist, A. D. Clarke, and R. Chakravarthi. Crowding in humans is unlike that in convolutional neural networks. *Neural Networks*, 126:262–274, 2020.

S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

A. Reuther, J. Kepner, C. Byun, S. Samsi, W. Arcand, D. Bestor, B. Bergeron, V. Gadepally, M. Houle, M. Hubbell, M. Jones, A. Klein, L. Milechin, J. Mullen, A. Prout, A. Rosa, C. Yee, and P. Michaleas. Interactive supercomputing on 40,000 cores for machine learning and data analysis. In *2018 IEEE High Performance extreme Computing Conference (HPEC)*, pages 1–6. IEEE, 2018.

R. Rosenholtz. Capabilities and limitations of peripheral vision. *Annual review of vision science*, 2: 437–457, 2016.

R. Rosenholtz, J. Huang, and K. A. Ehinger. Rethinking the role of top-down attention in vision: Effects attributable to a lossy representation in peripheral vision. *Frontiers in psychology*, 3:13, 2012a.

R. Rosenholtz, J. Huang, A. Raj, B. J. Balas, and L. Ilie. A summary statistic representation in peripheral vision explains visual search. *Journal of vision*, 12(4):14–14, 2012b.

R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022.

C. Vater, B. Wolfe, and R. Rosenholtz. Peripheral vision in real-world tasks: A systematic review. *Psychonomic bulletin & review*, 29(5):1531–1557, 2022.

A. Volokitin, G. Roig, and T. A. Poggio. Do deep neural networks suffer from crowding? *Advances in neural information processing systems*, 30, 2017.

D. Whitney and D. M. Levi. Visual crowding: A fundamental limit on conscious perception and object recognition. *Trends in cognitive sciences*, 15(4):160–168, 2011.

M. W. Wijntjes and R. Rosenholtz. Context mitigates crowding: Peripheral object recognition in real-world images. *Cognition*, 180:158–164, 2018.

Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection, 2022.

8 Appendix

8.1 Texture Tiling Model

Figure 6: **The Texture Tiling Model**. **(a)** Original image. **(b)** TTM transformed image assuming the green dot is the fixation point. **(c)** TTM modified to uniformly render an image at one distance in the periphery. 15° is shown here.

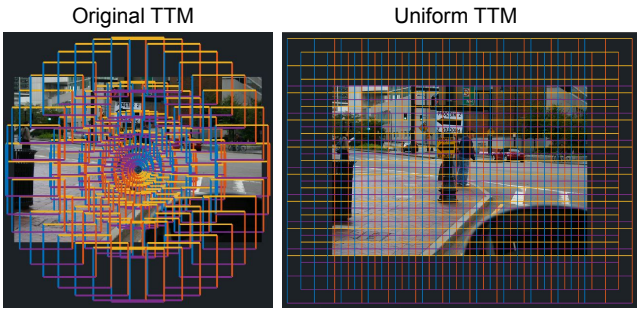

Figure 7: **Pooling Regions in Original and Uniform Texture Tiling Models (TTM).** Original TTM [Rosenholtz et al., 2012a] is foveated, so its pooling regions are small around the fixation point and grow farther from the fixation. We adapt TTM to use a fixed pooling region size everywhere in the image (Uniform TTM). The size is determined by the distance in the periphery being modeled.

In the original TTM model, the pooling region size is determined by a pooling rate, r, and a distance from the fovea, d. For the uniform version, we fix d for a certain eccentricity rather than varying it like the original model. We set the overlap between pooling regions to be 60%, and we arrange the uniform pooling regions in a rhombic lattice to make it as close as possible to original TTM. We use the same synthesis procedure as original TTM (matching statistics for each pooling region iteratively from noise). The uniform TTM transforms take between $2-3$ hours to synthesize on 1 CPU core (compared to the original TTM transforms, which take 6 hours on 1 core). Closer eccentricities like 5° take longer to run than large ones because the pooling region size is small. For training, we create uniform TTM transforms for 5,10,15, or 20°. For all TTM transforms, we assume that there are 16 pixels per degree, which is standard for original TTM.

When changing to uniform pooling, we also change the ordering of pooling region optimization from foveated TTM. Foveated TTM alternates spiraling from fovea to edge of periphery and back. This caused artifacts in uniformly-tiled TTM. Therefore, we opted for a randomly ordered optimization of the pooling regions, eliminating the optimization artifacts.

8.2 Human Psychophysics Experiment

8.2.1 Present / Absent Experiment Image Pairs

To create pairs of images where a given object was both present and absent, we used images from the MS-COCO validation set [Lin et al., 2014] (such that they would be novel to both humans and to trained networks). We found images in landscape orientation where an object from a COCO object category appeared and was labeled only once in the image, and the object was detected with at least 50% confidence in the original image with the detectron2 [Wu et al., 2019] object detection model (faster_rcnn_r50_fpn). From this set, we hand-selected 26 images that spanned a range of conditions that would affect the difficulty of the peripheral detection task (object identity and size, variation in luminance and color contrast from background, amount of crowding around object, etc). We then

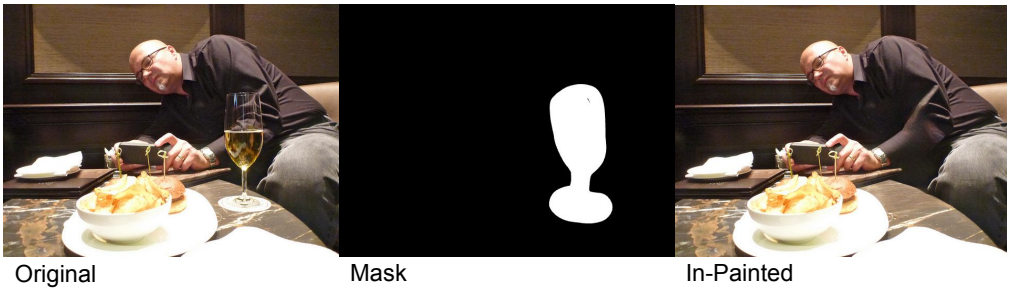

Figure 8: **Example of LaMa [Suvorov et al., 2022] inpainting on COCO validation set image**. We use inpainting to create the object absent version of each image in the psychophysics experiment. Here, the wine-glass is removed with few artifacts.

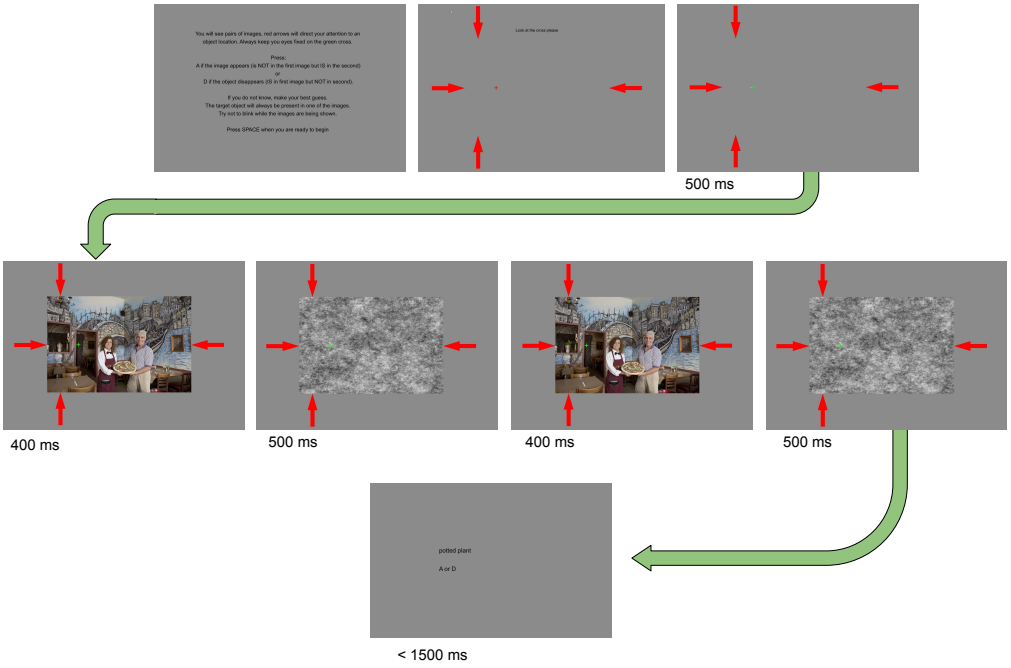

Figure 9: Human psychophysics experiment trial. Subjects complete a 2IFC (2 interval forced choice) task where they determine if a target object appears or disappears in a sequence. Red arrows indicate the location of the target object. Subjects fixate at a cross that is placed at $5, 10, 15$ or 20° away from the object.

used the LaMa image in-painting model to inpaint the chosen object [Suvorov et al., 2022], with a hand-drawn in-painting mask rather than the entire bounding box, as to avoid in-painting nearby objects in crowded scenes. In addition to the in-painted images, for each COCO image we also created a size matched $1/f$ pink noise mask to eliminate any motion transients and after-image effects during the experiment. These 26 image pairs were used for our object detection experiments. Note that figures reflect 24 images, as 2 images were removed from analysis because of poor psychometric curves fits (see Figures **??** and **??** to view the final image set).

8.2.2 Experimental Setup

All participants provided informed consent prior to participation, in compliance with the Common Rule (45 CFR 46), and this study was assessed as exempt from review by MIT's Institutional Review Board, pursuant to 45 CFR 46.101(b)(2). Participants took approximately 2 hours to complete the study and were paid a $\$40$ Amazon gift card for their participation.

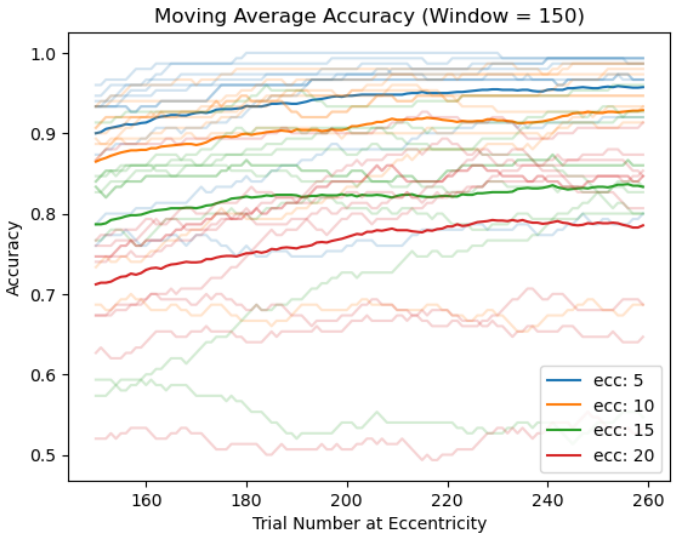

Figure 10: **Learning Over Experiment.** Despite including practice trials before the experiment and excluding correct/incorrect feedback during the experiment, human responses did exhibit a small learning effect over the course of the experiment. Accuracy is plotted against x-axis ordered by the number of times subject had seen any image at a given eccentricity. Bold lines show moving window averaged accuracy over all subjects, and pale lines show individual subject's data.

12 subjects participated in the human psychophysics experiment. We discarded the data from 2 subjects due to a computer malfunction and difficulty eye-tracking with a strong contact lens prescription. The remaining subjects consisted of 4 Male, 5 Female, and 1 Non-binary subjects ranging in age from 19 to 31. All had self-reported normal or corrected to normal visual acuity with contact lenses, with no history of eye surgery. 2 subjects had corrective lenses for myopia with a correction less than -1.25, but did not normally wear glasses or contacts (and did not during the experiment); We included these subjects as the viewing distance was only 82cm.

Subjects were seated and head placed in the chinrest of an EyeLink 1000 in tower-mount configuration. Subjects were 82 cm from a monitor screen, and their left eye position tracked. Nine-point calibration was performed and validated to within 1 degree at each point. The experiment allowed for fixation within 2 degrees, displaying a small dot on the screen for real-time feedback of measured fixation location. Subjects were asked to pause the trial block to re-calibrate if measured fixation did not reflect fixation location, or they had difficulty with the system recognizing their fixation.

8.2.3 Experimental Paradigm

The experiment consisted of a 2IFC (two interval forced choice) task where subjects report which out of two images contains a target object. Each subject saw 26 image pairs 10 times at 4 different fixation locations $(5, 10, 15, 20^\circ)$ away from a target object, where the fixation location was at the vector computed from the target object location towards the center-point of the image. The order of each presentation was randomized across the whole experiment. Each subject saw the present/absent image pair 5 times present-first and 5 times absent-first. No correct/incorrect feedback was given to the subject.

Subjects maintained fixation on a cross presented at either 5, 10, 15, or 20 degrees from the object location, and were eye-tracked to ensure fixation was maintained within 2 degrees. Attention was directed to the object location with latitude/longitude arrows. After presentation, subjects were given the original COCO object category, and prompted to report which image contained the the object by reporting if the object 'appeared' (was in the 2nd image but not the first) or 'disappeared' (was in the 1st image but not the second).

Each trial waited to begin until the subject fixated on a cross before proceeding. If the subject broke fixation anytime an image or mask is shown, the trial was aborted and shuffled to the end of

experiment. Each image was shown for 400ms, and a size-matched pink noise mask was shown after for 500ms to eliminate visible flicker of appearance of disappearance. Subjects were given a response window of 1.5 seconds, and the image pair shuffled to the end of the experiment in a time-out.

The experiment in total was 1040 trials long. Subjects were given a break every 150 trials, re-calibrating after each break and before starting a new block. Before beginning the experiment, subjects completed a practice round consisting of 15 trials of very easy image pairs. 2 subjects needed to do the practice round a second time before they reported being comfortable with the task. The images in the practice round were much larger than those in the actual experiment to make the task easier (10+ degrees). This may have contributed to a learning effect we observed in some subjects where performance improves with the number of trials completed (Figure 10).

8.2.4 Human Psychophysics Results

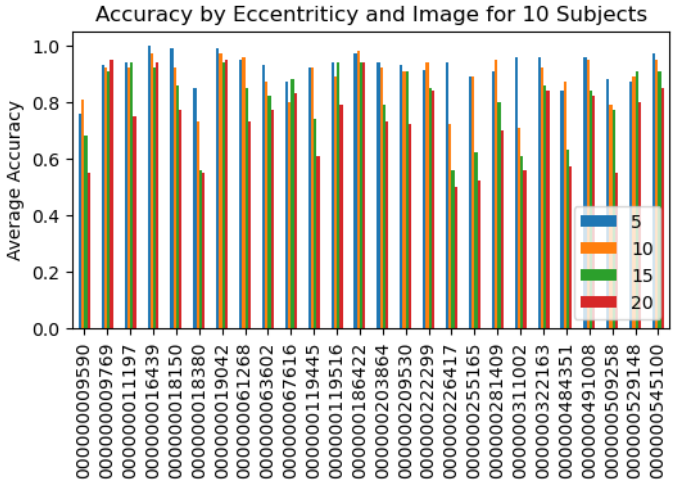

Figure 11: **Per-Image Accuracy over all Subjects.** Human performance was very image-dependent, but decayed as eccentricity increased for all images

Overall, subjects performed well at the task for most images, only reaching chance performance for approximately 30% of the images (Figure 11). Images 000000009769 and 000000067616 were removed from downstream analysis due to poor fitting of psychometric function. The difficulty was extremely image-dependent - subjects reported similar feedback during debriefing that certain images were extremely difficult and that they were guessing (though results show they often performed better than chance, despite this), while other images were extremely easy. The hardest and easiest reported images tended to be those least and most crowded, and those with the most and least background contrast, respectively.

8.3 Machine Psychophysics Experiment

We provide pseudo-code for the machine psychophysics procedure. We tried a variety of detection criteria including, enforcing that predictions match the target category, enforcing the size the box predictions to be no more than half or twice the size of the padded ground truth, and taking the average score over all boxes that overlap the padded ground truth. We arrived at the summing approach described in the Algorithm1 because it yielded the highest critical μ scores and showed similar trends in performance to the other approaches.

Algorithm 1 For each object present/absent image in the human experiment, we create 100 pairings of uniform TTM transform images (P and A). We simulate trials by looking at the box predictions ($boxes$) of a detection DNN for each pairing (p, a). We sum the total box scores that overlap with the target object box at least 0.75% IOA (intersection over area). To determine this overlap, we take the target object ground truth box (gt) and pad it with by half a pooling region (pr). If the total score for the present image ($pprob$) is higher than the absent ($aprob$), we record the DNN model as having a correct response. We average over all 100 pairings for final accuracy (acc).

procedure GETMODELACCURACY
  $acc = 0$ ▷ initialize accuracy
  for $p, a \in (P, A)$ **do** ▷ loop through all present/absent pairings for one object
    $pprob$ = GetTargetDetectionScore(p, gt, pr) ▷ get score for target object
    $aprob$ = GetTargetDetectionScore(a, gt, pr)
    if $pprob = aprob$ **then** $acc = acc + 0.5$ ▷ get per trial accuracy
    if $pprob > aprob$ **then** $acc = acc + 1$
  $acc = acc \div trials$ ▷ take average over all trials
function GETTARGETDETECTIONSCORE(im, gt, pr)
  $gtx = gt + 0.5 \times$ size(pr) ▷ expand ground truth bounding box by half pooling region
  $boxes, scores =$ DNN(im) ▷ get box proposals
  $prob = 0$ ▷ initialize total score of overlapping proposals
  for $b, s \in (boxes, scores)$ **do**
    if ioa$(gtx, b) > 0.75$ **then** ▷ check boxes that overlap expanded ground truth
      $prob = prob + s$
  return $prob$ ▷ return sum of overlapping scores

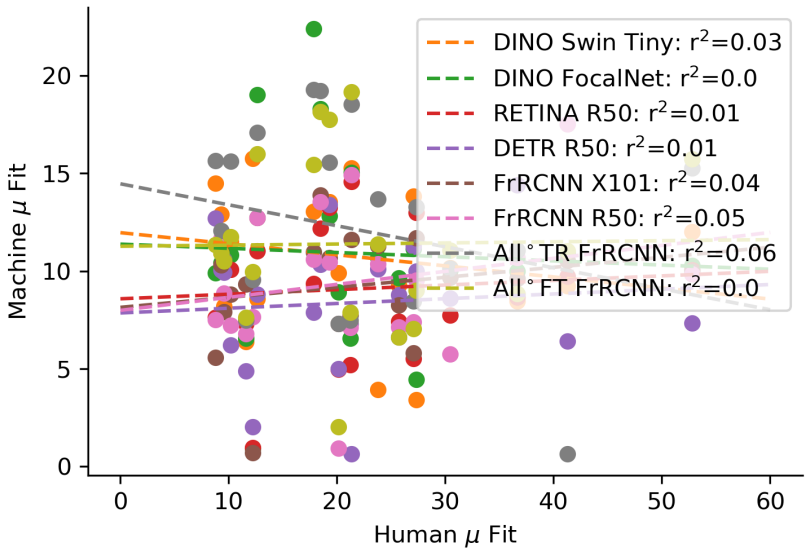

Figure 12: **Correlation between Human and Machine Critical Eccentricity using original TTM in the machine psychophysics.**

8.4 Fine-Tuning Object Detection Models

Model	AP 0°	AP 5°	AP 10°	AP 15°	AP 20°
Faster-RCNN-R50	**36.7**	29.4	19.9	5.9	4.2
All ° FT Faster-RCNN-R50	36.1	**31.8**	27.7	13.9	10.8
All ° Train Faster-RCNN-R50	33.8	30.5	**28.1**	**15.8**	**12.7**

Table 1: **Average Precision (AP) on Uniform TTM-transformed COCO Validation Set**. All models are Faster-RCNN ResNet50 FPN architecture [Ren et al., 2015].

8.5 Training Procedure

We fine-tuned and train from scratch the Faster R-CNN model from the Detectron2 library [Wu et al., 2019] (faster_rcnn_r50_fpn) using a mixture of original training images, and TTM transforms for varying eccentricities. We use 55,000 images from each eccentricity along with original COCO images. Fine-tuning was trained for 180,000 iterations starting from the weights of a pre-trained R-CNN from [Wu et al., 2019]. We set the solver to step at 120,000 and 160,000. We set the base learning rate to 3×10^{-4}. All other training parameters are the same R-CNN training parameters in [Wu et al., 2019] as the baseline model. To train from scratch, we use the same 3x training schedule provided in [Wu et al., 2019] for Faster RCNN R50 FPN models (starting from an ImageNet trained ResNet50 backbone, training for 270,000 iters, 16 images per batch).