# FINDING BIOLOGICAL PLAUSIBILITY FOR ADVERSARIALLY ROBUST FEATURES VIA METAMERIC TASKS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Recent work suggests that feature constraints in the training datasets of deep neural networks (DNNs) drive robustness to adversarial noise (Ilyas et al., 2019). The representations learned by such adversarially robust networks have also been shown to be more human perceptually-aligned than non-robust networks via image manipulations (Santurkar et al., 2019; Engstrom et al., 2019). Despite appearing closer to human visual perception, it is unclear if the constraints in robust DNN representations match biological constraints found in human vision. Human vision seems to rely on texture-based/summary statistic representations in the periphery, which have been shown to explain phenomena such as crowding (Balas et al., 2009) and performance on visual search tasks (Rosenholtz et al., 2012). To understand how adversarially robust optimizations/representations compare to human vision, we performed a psychophysics experiment using a metamer task similar to Freeman & Simoncelli (2011); Wallis et al. (2019); Deza et al. (2019b) where we evaluated how well human observers could distinguish between images synthesized to match adversarially robust representations compared to non-robust representations and a texture synthesis model of peripheral vision (Texforms (Long et al., 2018)). We found that the discriminability of robust representation and texture model images decreased to near chance performance as stimuli were presented farther in the periphery. Moreover, performance on robust and texture-model images showed similar trends within participants, while performance on non-robust representations changed minimally across the visual field. These results together suggest that (1) adversarially robust representations capture peripheral computation better than non-robust representations and (2) robust representations capture peripheral computation similar to current state-of-the-art texture peripheral vision models. More broadly, our findings support the idea that localized texture summary statistic representations may drive human invariance to adversarial perturbations and that the incorporation of such representations in DNNs could give rise to useful properties like adversarial robustness.

## 1 INTRODUCTION

Texture-based summary statistic models of the human periphery have been shown to explain key phenomena such as crowding (Balas et al., 2009; Freeman & Simoncelli, 2011) and performance on visual search tasks (Rosenholtz et al., 2012) when used to synthesize feature-matching images. These analysis-by-synthesis models have also been used to explain mid-level visual computation (*e.g.* V2) via perceptual discrimination tasks on images for humans and primates (Freeman & Simoncelli, 2011; Ziemba et al., 2016; Long et al., 2018).

However, while summary statistic models can succeed at explaining peripheral computation in humans, they fail to explain foveal computation and core object recognition that involve other representational strategies (Logothetis et al., 1995; Riesenhuber & Poggio, 1999; DiCarlo & Cox, 2007; Hinton, 2021). Modelling foveal vision with deep learning indeed has been the focus of nearly all object recognition systems in computer vision (as machines do not have a periphery) (LeCun et al., 2015; Schmidhuber, 2015) – yet despite their overarching success in a plethora of tasks, they are vulnerable to adversarial perturbations. This phenomena indicates: 1) a critical failure of current artificial systems (Goodfellow et al., 2015; Szegedy et al., 2014); and 2) a perceptual mis-alignment of such systems with humans (Golan et al., 2020; Feather et al., 2019; Firestone, 2020; Geirhos

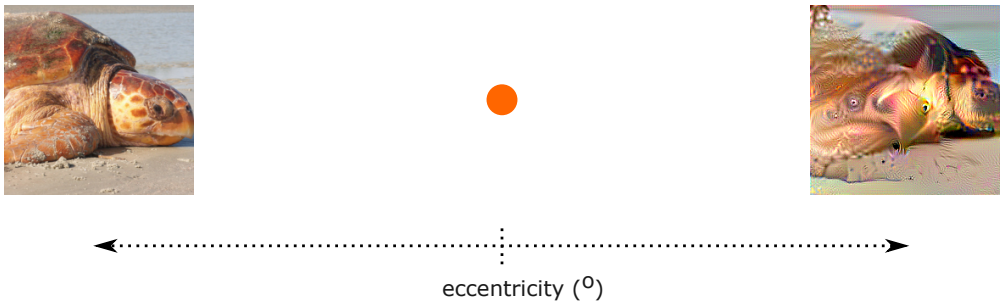


Figure 1: A sample un-perturbed (left) and synthesized adversarially robust (right) image are shown peripherally. When a human observer fixates at the orange dot (center), both images – now placed away from the fovea – are perceptually indistinguishable to each other (i.e. *metameric*). In this paper we investigate if there is a relationship between peripheral representations in humans and learned representations of adversarially trained networks in machines in an analysis-by-synthesis approach. We psychophysically test this phenomena over a variety of images synthesized from an adversarially trained network, a non-adversarially trained network, and a model of peripheral computation as we manipulate retinal eccentricity over 12 humans subjects.

et al., 2021; Funke et al., 2021) – with some exceptions (Elsayed et al., 2018). Indeed, there are many strategies to alleviate these sensitivities to perturbations, such as data-augmentation (Rebuffi et al., 2021; Gowal et al., 2021), biologically-plausible inductive biases (Dapello et al., 2020; Reddy et al., 2020; Jonnalagadda et al., 2021), and adversarial training (Tsipras et al., 2019; Madry et al., 2017). This last strategy in particular (adversarial training) is popular, but has been criticized as being non-biologically plausible – despite yielding some perceptually aligned images when inverting their representations (Engstrom et al., 2019; Santurkar et al., 2019).

Motivated by prior work on summary statistic models of peripheral computation and their potential resemblance to inverted representations of adversarially trained networks, we wondered if these two apparently disconnected phenomena from different fields share any similarities (See Figure 1). Could it be that adversarially trained networks are robust because they encode object representations similar to human peripheral computation? We know machines do not have peripheral computation (Azulay & Weiss, 2019; Deza & Konkle, 2020; Alsallakh et al., 2021), yet are susceptible to a type of adversarial attacks that humans are not. We hypothesize that object representation arising in human peripheral computation holds a critical role for high level robust vision in perceptual systems, but testing this has not been done.

Thus, the challenge we now face is how to compare an adversarially trained neural network model to current models of peripheral/mid-level visual processing – and ultimately to human observers as the objective ground-truth. However, determining such perceptual parameterizations is computationally intractable. Inspired by recent works that test have tested summary statistic models via metameric discrimination tasks (Deza et al., 2019b; Wallis et al., 2016; 2017; 2019), we can evaluate how well the adversarially robust CNN model approximates the types of computations present in human peripheral vision with a set of rigorous psychophysical experiments wrt synthesized stimuli.

Our solution consists of performing a set of experiments where we will evaluate the rates of human perceptual discriminability as a function of retinal eccentricity across the synthesized stimuli from an adversarially trained network vs synthesized stimuli from models of mid-level/peripheral computation. If the decay rates at which the perceptual discriminability across different stimuli are similar, then this would suggest that the transformations learned in an adversarially trained network are related to the transformations done by models of peripheral computation – and thus, to the human visual system. It is worth noting that although adversarially robust representations have been shown to be more human-perceptually aligned (Ilyas et al., 2019; Engstrom et al., 2019; Santurkar et al., 2019), they still look quite different when placed in the foveal region from the original reference image (Feather et al., 2021). However, our eccentricity-varying psychophysical experiments are motivated by empirical work that suggests that the human visual periphery represents input in a texture-like scrambled way, that can appear quite different than how information is processed in the fovea (Nandy & Tjan, 2012; Rosenholtz, 2016; Stewart et al., 2020; Herrera-Esposito et al., 2021).

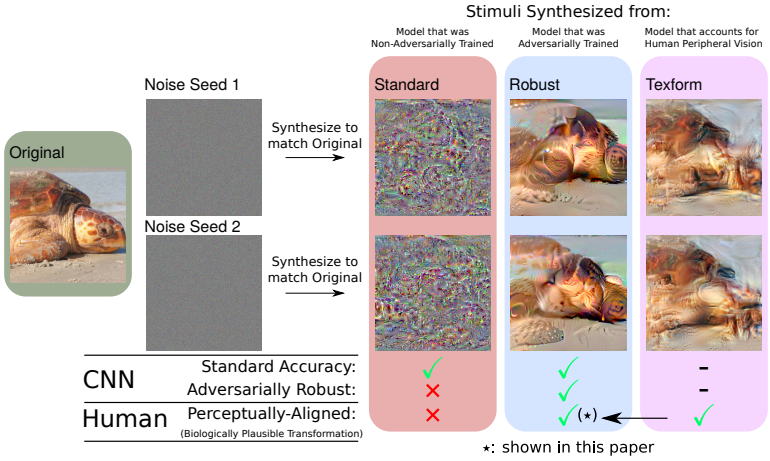


Figure 2: A sub-collection of different synthesized stimuli used in our experiments that shows the differences across (columns) and within (rows) perceptual models. The original stimuli is shown on the left, while two parallel Noise Seeds, give rise to different samples for the Standard, Robust and Texform stimuli. Critically, an adversarially trained network – which was used to synthesize the Robust stimuli (Engstrom et al., 2019) – has implicitly learned to encode a structural prior with localized texture-like distortions similar to the physiologically motivated Texforms that account for several phenomena of *human peripheral computation* (Freeman & Simoncelli, 2011; Rosenholtz et al., 2012; Long et al., 2018). However, Standard stimuli, which are images synthesized from a network with Regular (Non-Adversarial) training have no resemblance to the original sample. In this paper we evaluate how similar these models are, via their derived stimuli, with a set of controlled human psychophysics experiments where we vary the retinal eccentricity of the stimuli.

## 2 Synthesizing Stimuli as a Window to Model Representation

Suppose we have the functions $g_{\text{Adv}}(\circ)$ and $g_{\text{Standard}}(\circ)$ that represent the adversarially trained and standard (non-adversarially) trained neural networks; how can we compare them to human peripheral computation if the function $g_{\text{Human}}(\circ)$ is computationally intractable?

One solution is to take an analysis-by-synthesis approach and to synthesize a collection of stimuli that match the feature response of the model we'd like to analyze – this is also known as feature inversion (Mahendran & Vedaldi, 2015; Feather et al., 2019). If the inverted features (stimuli) of two models are perceptually similar, then it is likely that the learned representations are also aligned. For example, if we'd like to know what is the stimuli $x'$ that produces the same response to the stimuli $x$ for a network $g'(\circ)$, we can perform the following minimization:

$$x' = \arg\min_{x_0} [||g'(x) - g'(x_0)||_2] \tag{1}$$

In doing so, we find $x'$ which should be different from $x$ for a non-trivial solution. This is known as a metameric constraint for the stimuli pair $\{x, x_0\}$ wrt to the model $g'(\circ) : g'(x) = g'(x')$ s.t. $x \neq x'$ for a starting pre-image $x_0$ that is usually white noise in the iterative minimization of Eq.1. Indeed, for the adversarially trained network of Ilyas et al. (2019); Engstrom et al. (2019); Santurkar et al. (2019), we can synthesize robust stimuli wrt to the original image $x$ via:

$$\tilde{x} = \arg\min_{x_0} [||g_{\text{Adv}}(x) - g_{\text{Adv}}(x_0)||_2] \tag{2}$$

which implies – if the minimization goes to zero – that:

$$||g_{\text{Adv}}(x) - g_{\text{Adv}}(\tilde{x})||_2 = 0 \tag{3}$$

Recalling the goal of this paper, we'd like to investigate if the following statement is true: *"a transformation resembling peripheral computation in the human visual system can closely be approximated by an adversarially trained network"*, which is formally translated as: $g_{\text{Adv}} \sim g^{r_*}_{\text{Human}}$ for some retinal eccentricity $(r_*)$, then from Eq. 3 we can also derive:

$$||g^{r_*}_{\text{Human}}(x) - g^{r_*}_{\text{Human}}(\tilde{x})||_2 = 0 \tag{4}$$

However, $g_{\text{Human}}(\circ)$ is computationally intractable, so how can we compute Eq.4? A first step is to perform a psychophysical experiment such that we find a retinal eccentricity $r_*$ at which human observers can not distinguish between the original and synthesized stimuli – thus behaviourally proving that the condition above holds, without the need to directly compute $g_{\text{Human}}$.

More generally, we'd like to compare the *psychometric functions* between stimuli generated from a standard trained network (standard stimuli), an adversarially trained network (robust stimuli), and a model that captures peripheral and mid-level visual computation (texform stimuli (Freeman & Simoncelli, 2011; Long et al., 2018)). Then we will assess how the psychometric functions vary as a function of retinal eccentricity. If there is significant overlap between psychometric functions between one model wrt the model of peripheral computation; then this would suggest that the transformations developed by such model are similar to those of human peripheral computation. We predict that this will be the case for the adversarially trained network $(g_{\text{Adv}}(\circ))$. Formally, for any model $g$, and its synthesized stimuli $x_g$ – as shown in Figure 2, we will define the psychometric function $\delta_{\text{Human}}$, which depends on the eccentricity $r$ as:

$$\delta_{\text{Human}}(g; r) = ||g^r_{\text{Human}}(x) - g^r_{\text{Human}}(x_g)||_2 \tag{5}$$

where we hope to find:

$$\delta_{\text{Human}}(g_{\text{Adv}}; r) = \delta_{\text{Human}}(g_{\text{Texform}}; r); \forall r. \tag{6}$$

### 2.1 Standard and Robust Model Stimuli

To evaluate robust vs non-robust feature representations, we used the ResNet-50 models of Santurkar et al. (2019); Ilyas et al. (2019); Engstrom et al. (2019). We used their models so that our results could be interpreted in the context of their findings that features may drive robustness. Both models were trained on a subset of ImageNet (Russakovsky et al., 2015), termed Restricted ImageNet (Table 1). The benefit of Restricted ImageNet, stated by Ilyas et al.; Engstrom et al., is models can achieve better standard accuracy than on all of ImageNet. One drawback is that it is imbalanced across classes. Although the class imbalance was not problematic for comparing the adversarially robust model to standard-trained one, we did ensure that there was a nearly equal number of images per class when selecting images for our stimulus set to avoid class effects in our experiment (i.e. people are better at discriminating dog examples than fishes independent of the model training).

Using their readily available models, we synthesized robust and standard model stimuli using an image inversion procedure (Mahendran & Vedaldi, 2015; Gatys et al., 2015; Santurkar et al., 2019; Engstrom et al., 2019; Ilyas et al., 2019). We used gradient descent to minimize the difference between the representation of the second-to-last network layer of a target image and an initial noise seed as shown in Figure 9. Target images were randomly chosen from the test set of Restricted ImageNet. We chose $100$ target images for each of the $9$ classes and synthesized a robust and standard stimulus for 2 different noise seeds. $5$ target images were later removed as they were gray-scale and could not also be rendered as Texforms with the same procedure as the majority. All stimuli were synthesized at a size of $256 \times 256$ pixels, this was equivalent to $6.67 \times 6.67$ degrees of visual angle (d.v.a.) when performing the psychophysical experiments (See A.6 for calculation).

### 2.2 Texform Stimuli

Texforms (Long et al., 2018) are object-equivalent rendered stimuli from the Freeman & Simoncelli (2011); Rosenholtz et al. (2012) models that break the metameric constraint to test for mid-level visual representations in Humans. These stimuli – initially inspired by the experiments of Balas et al. (2009) – preserve the coarse global structure of the image and its localized texture statistics (Portilla & Simoncelli, 2000). Critically, we use the texform stimuli – *voiding the metameric constraint* – as a perceptual control for the robust stimuli, as the texforms incarnate a sub-class of biologically-plausible distortions that loosely resemble the mechanisms of human peripheral processing.

As the texform model has 2 main parameters which are the scaling factor $s$ and the simulated point of fixation $z$, we must perform a perceptual optimization procedure to find the set of texforms $\hat{x}$ that match the robust stimuli $\tilde{x}$ as close as possible (w.r.t to the original image) *before* testing their discriminability to human observers as a function of eccentricity. To do this, we used the accelerated texform implementation of Deza et al. (2019a) and generated 45 texforms with the *same* collection of initial noise seeds as the robust stimuli to be used as perceptual controls. Similar to Deza & Konkle

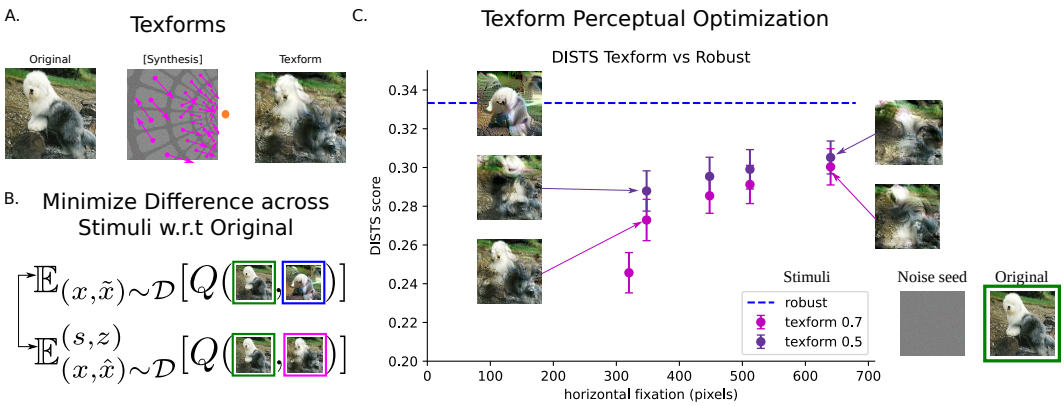


Figure 3: (A.) A cartoon depicting the texform generating process where log-polar receptive fields are used as areas over which localized texture synthesis is performed – imitating the type of texture-based computation found in the human periphery and area V2. (B.) The perceptual optimization framework where the goal is to find the set of texform parameters $(s_*, z_*)$ over which the loss is minimized to match the levels of distortions of the robust stimuli *before* performing human psychophysics. (C.) The texform perceptual optimization pipeline results show the DISTS scores (Ding et al., 2020) of texforms synthesized across different scaling factors and fixations points compared to adversarially robust stimuli synthesized from the same noise seed across 45 images (5 per RestrictedImageNet class selected randomly). Error bars indicate two standard errors from the mean.

(2020) we minimize the perceptual dissimilarity $\mathcal{Z}$ to find $(s_*, z_*)$ over this subset of images that we will later use in the human psychophysics ($\sim$ 900 texforms):

$$(s_*, z_*) = \underset{(s,z)}{\arg\min}\, \mathcal{Z} = ||\mathbb{E}_{(x,\tilde{x})\sim\mathcal{D}}[Q(x,\tilde{x})] - \mathbb{E}^{(s,z)}_{(x,\hat{x})\sim\mathcal{D}}[Q(x,\hat{x})]||_2 \tag{7}$$

for an image quality assessment (IQA) function $Q(\circ,\circ)$. We selected DISTS in our perceptual optimization setup given that it is the IQA metric that is most tolerant to texture-based transformations (Ding et al., 2020; 2021). A cartoon illustrating the texform rendering procedure, the perceptual optimization framework and the respective results can be seen in Figure 3. In our final experiments (See Next Section) we used texforms rendered with a simulated scale of 0.5 and horizontal simulated point of fixation placed at 640 pixels. Critically, this value is *immutable* and texforms (like robust stimuli) will not vary as a function of eccentricity to provide a fair discriminability control in the human psychophysics. For a further discussion on texforms and their biological plausibility and/or synthesis procedure, please see Supplement A.2.

## 3 Human Psychophysics: Discriminating between stimuli as a function of retinal eccentricity

We designed two human psychophysical experiments: the first was a an oddity task similar to Wallis et al. (2016), and the second was a matching, two-alternative forced choice task (2AFC). Two different tasks were used to evaluate how subjects viewed synthesized images both only in the periphery (oddity) and those they saw in the fovea (matching 2AFC). The oddity task consisted of finding the oddball stimuli out of a series of 3 stimuli shown peripherally one after the other (100ms) masked by empty intervals (500ms) while holding center fixation. Chance for the oddity task was 1 out of 3 $(33.3\%)$. The matching 2AFC task consisted of viewing a stimulus in the fovea (100ms) and then matching it to two candidate templates in the visual periphery (100 ms) while holding fixation. A 1000 ms mask was used in this experiment and chance was 50%.

For both experiments, we also had interleaved trials where observers had to engage in a Original stimuli vs Synthesized stimuli task, or a Synthesized stimuli vs Synthesized stimuli discrimination task (two stimulus pairs synthesized from *different* noise seeds to match model representations). The goal of these experimental variations (called *'stimulus roving'*) was two-fold: 1) to add difficulty to

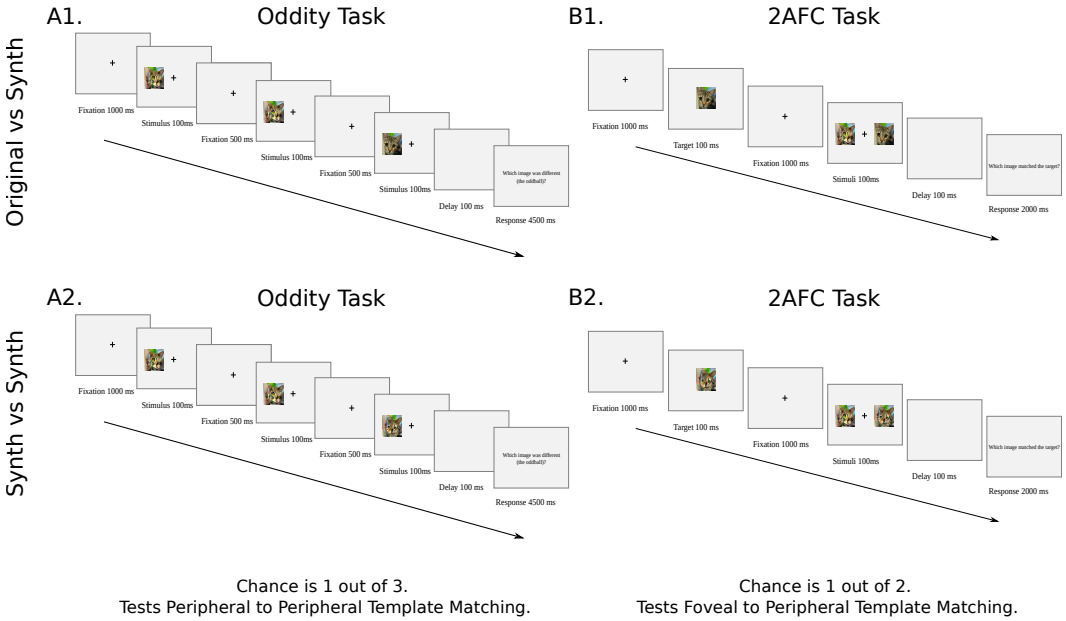


Figure 4: A schematic of the two human psychophysics experiments conducted in our paper. The first (A1.,A2.) illustrates an Oddity task where observers must determine the 'oddball' stimuli without moving their eyes for very brief presentation times (100 ms) that are masked which do not allow for eye-movements or feedback processing. The second experiment (B1.,B2.) shows the 2 Alternative Forced Choice (AFC) Matching Tasks where observers must match the foveal template to 2 potential candidates on the left or right of the image. All trials are done while observers are instructed to remain fixating at the center of the image. These two experiments test different mechanisms of visual processing. In particular the Oddity task examines discriminability of purely peripheral-to-peripheral representations, while the 2AFC task evaluates discriminability of foveal-to-peripheral representation. Differences across rows indicate the type of interleaved trials shown to the observers: (1) Original vs Synthesized, and (2) Synthesized vs Synthesized. Critically these are image perceptual discrimination tasks *not* image categorization tasks.

the tasks thus reducing the likelihood of ceiling effects; 2) to gather two psychometric functions per family of stimuli, which portrays a better description of each stimuli's evoked perceptual signatures.

We had 12 participants complete both the oddity and matching 2AFC experiments. The oddity task was always performed first so that subjects would never have foveated on the images before seeing them in the periphery. We had two stimulus conditions (1) robust & standard model images and (2) texforms. Condition 1 consisted of the inverted representations of the adversarially robust and standard-trained models. The two model representations were randomly interleaved since they were synthesized with the same procedure. Condition 2 consisted of texforms synthesized with a fixed and perceptually optimized fixation and scaling factor which yielded images closest in structure to the robust representations at foveal viewing (robust features have no known fixation and scaling – which is why partly we evaluate multiple points in the periphery. Recall Figure 3). We randomly assigned the order in which participants saw the different stimuli. More details found in A.6.

The main results of our 2 experiments can be found in Figure 5, where we show how well Humans can discriminate per type of stimuli class and task. Mainly, human observers achieve near perfect discrimination rates for the Standard stimuli wrt to their original references, but near chance levels when discriminating to another synthesized sample. This occurs for both experimental paradigms (Oddity + 2AFC), suggesting that the network responsible for encoding standard stimuli is a poor model of human peripheral vision given no interaction with retinal eccentricity.

However, we observe that Humans show similar perceptual discriminability rates for Robust and Texform stimuli – and that these vary in a similar way as a function of retinal eccentricity. Indeed, for both of these stimuli their perceptual discrimination rates follow a sigmoidal decay-like curve

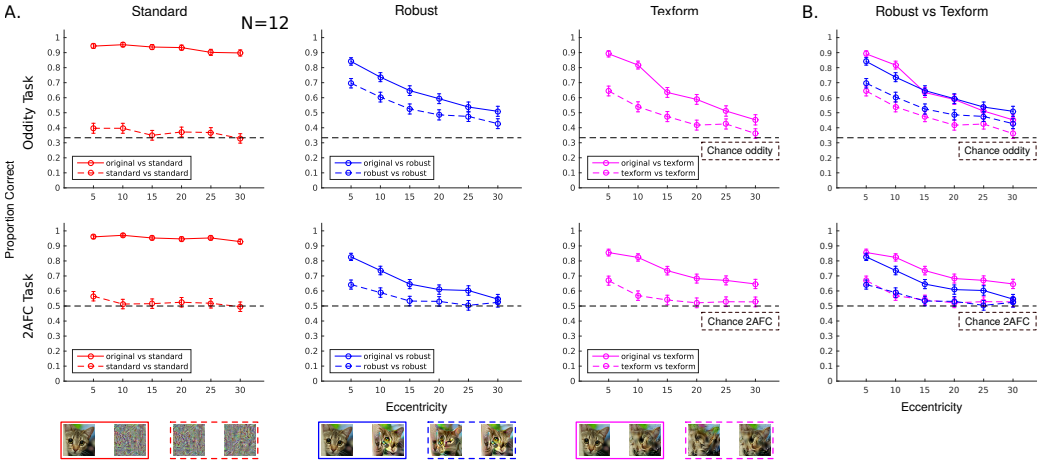


Figure 5: Pooled observer results of both psychophysical experiments are shown (top and bottom row). (A.) Left: we see that observers perfectly discriminate the original image wrt the standard stimuli, in addition to chance performance when comparing against synthesized stimuli. Critically there is no interaction of the standard stimuli with retinal eccentricity which suggests that the model used to synthesize such stimuli is a poor model of peripheral computation. Middle: Human observers do worse at discriminating the robust stimuli wrt the original as a function of eccentricity and also between synthesized robust samples. Given this decay in perceptual discriminability, it would suggest that the adversarially trained model used to synthesize robust stimuli does capture aspects of peripheral computation. This effect can also be seen on the texforms (Right) – which have been extensively used as stimuli from derived models that capture peripheral and V2-like computation. (B.) Superimposed human performance for Robust and Texform stimuli. Errorbars are computed via bootstrapping and represent the 95% confidence interval.

when comparing the stimuli to the original, and also between synthesized samples. The similarity between the blue and magenta curves from Figure 5 suggests that if the texform stimuli do capture some aspect of peripheral computation, then – by transitivity – so do the adversarial stimuli which were rendered from an adversarially trained network. These results empirically verify our initial hypothesis that we set out to test in this paper. A superposition of these results in reference to the Robust stimuli for a better interpretation can also be seen in Figure 5 (B.).

### 3.1 Simulated Fovea/Periphery Image Quality Assessment (IQA) across stimuli

Some distortions are more perceptually noticeable than others for humans and neural networks (Berardino et al., 2017; Martinez-Garcia et al., 2019) – so how do we assess which model better accounts for peripheral computation, if there are many distortions (derived from the synthesized model stimuli) that can potentially yield the same perceptual sensitivity in a discrimination task?

Our approach consists of computing two IQA metrics (DISTS & MSE) over the entire psychophysical testing set over 2 opposite levels of a Gaussian Pyramid decomposition (Burt & Adelson, 1987). This procedure checks which stimuli presents the greatest distortion (MSE), and yet yields greater perceptual invariance (DISTS). A Gaussian Pyramid decomposition was selected as it stimulates the frequencies preserved given changes in human contrast sensitivity and cortical magnification factor from fovea to periphery (Anstis, 1974; Geisler & Perry, 1998). These two metrics were one that is texture-tolerant and perceptually aligned (DISTS), and another that is a non-perceptually aligned metric: Mean Square Error (MSE). Both IQA metrics were computed in pixel space for both the Original vs Synthesized and Synthesized vs Synthesized conditions.

Results are explained in Figure 6, where Standard Stimuli yields low perceptual invariance to the original image at both levels of the Gaussian Pyramid, but robust and texform stimuli have a similar degree of perceptual invariance. Critically, robust stimuli are slightly more distorted via MSE than texform stimuli suggesting that the adversarially trained model has learned to represent peripheral computation better than the texform model by maximizing the perceptual null space and throwing away more useless low-level image features (hence achieving greater Mean Square Error).

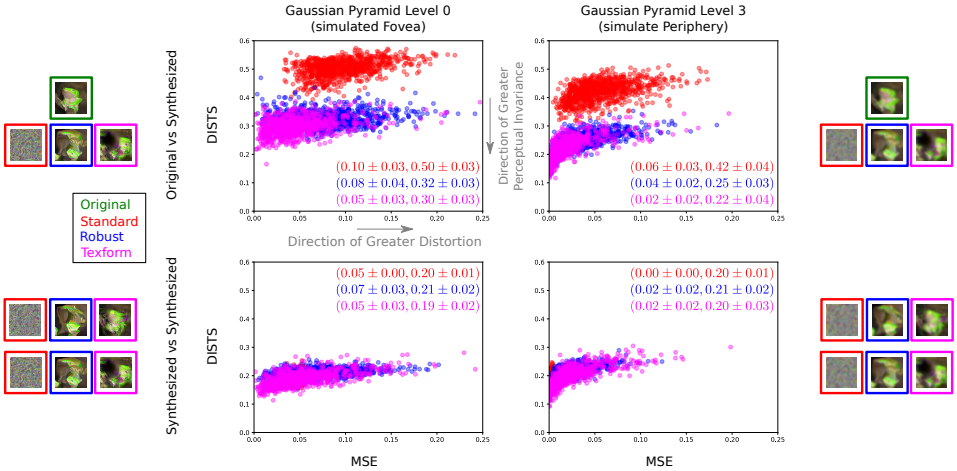


Figure 6: Here we evaluate how the different stimuli differ to each other wrt to the original (top row) or synthesized samples (bottom row) via two IQA metrics: DISTS and MSE. This characterization allows us to compare which model discards more information (MSE) while yielding a greater degree of model based perceptual invariance. We find that Texform and Robust stimuli are similar terms of both IQA scores, suggesting their models compute the same transformations. This is observed at the 0th level (simulated fovea) and 3rd level (simulated periphery) of the Gaussian Pyramid.

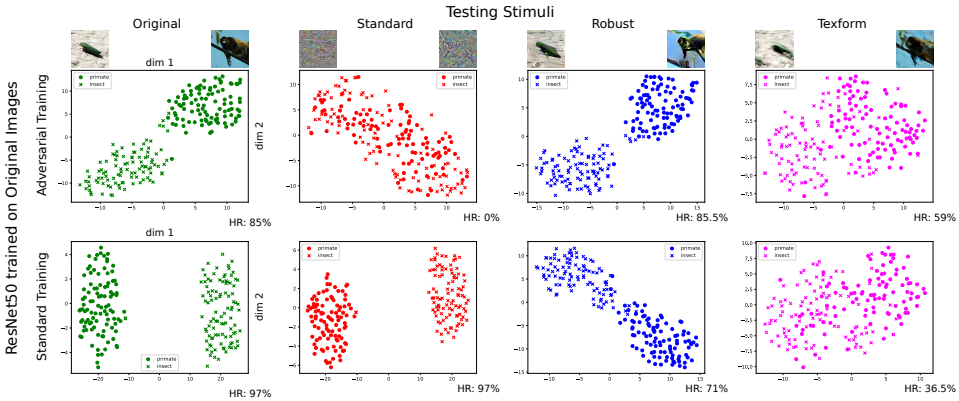


Figure 7: Here we show a 2D projection using t-SNE (Van der Maaten & Hinton, 2008) to visualize the outputs of the last layer of the Adversarially trained network (that was used to synthesize the Robust Stimuli), and the Standard trained network (that was used to synthesize the Standard stimuli), both on a family of different stimuli: Original, Standard, Robust and Texform. The Adversarially trained network – similar to the human – can not distinguish between 2-class Standard Stimuli (unlike the Standard Network that has a near perfect 2-class hit rate). Most importantly, the Adversarially trained network yields a near double hit rate on Texform classification wrt the Standard trained network. This suggests that the Adversarially trained network has a representation that is more perceptually aligned to models of Peripheral Computation than the Standard trained model.

## 4 Discussion

One of the fundamental questions in modern computer vision is understanding what are the principles that give rise to adversarial stimuli – which show a striking perceptual divergence between man and machine (Feather et al., 2019; Golan et al., 2020; Geirhos et al., 2021; Funke et al., 2021). While it may seem theoretically impossible to escape from adversarial stimuli (Gilmer et al., 2018) – perhaps our efforts in the community should focus on understanding the biologically plausible mechanisms *(if any)* of the solutions that grant some level of adversarial robustness aligned with human error. In this paper we focused on potentially linking the representations learned from an adversarially trained network and human peripheral computation via a series of psychophysical experiments through a family of stimuli synthesized from these models.

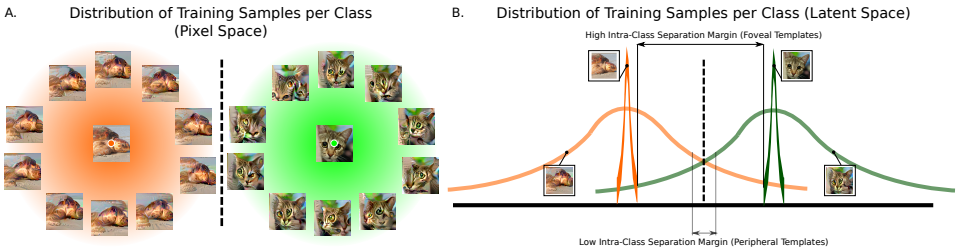


Figure 8: A cartoon depicting a conjecture of how peripheral computation may induce adversarial robustness. In (A.) we see a family of perceptually equidistant peripheral templates from the original foveal template (center dot) constructed with the adversarially robust model used to perform peripheral encoding. In (B.) we observe the same templates projected from a high-dimensional space into a uni-dimensional space. We also see that the greater *covariances* only induced by peripheral templates lead to greater adversarial robustness during learning in a perceptual system – despite having equal Intra-Class *means* (for both foveal or peripheral templates). This suggests that peripheral computation may implicitly act as a *natural visual regularizer* of learned representations.

We found that stimuli synthesized from an adversarially trained (and thus robust) network are metameric to the original stimuli in the further periphery (slightly above $30\,\mathrm{deg}$) for both Oddity and 2AFC Matching tasks. However, more important than deriving a critical eccentricity for metameric gaurantees across stimuli in Humans – we found a surprisingly similar pattern of results in terms of how perceptual discrimination interacts with retinal eccentricity when comparing the adversarially trained network's robust stimuli with classical models of peripheral computation and V2 encoding (mid-level vision) that were used to render the texform stimuli (Freeman & Simoncelli, 2011; Long et al., 2018; Ziemba et al., 2016; Ziemba & Simoncelli, 2021). Further, this type of eccentricity-driven interaction does not occur for stimuli derived from non-adversarially trained (standard) networks.

More generally, now that we found that adversarially trained networks encode a similar class of transformations that occur in the visual periphery – how do we reconcile with the fact that adversarial training is biologically *implausible* in humans? Recall from the work of Ilyas et al. (2019) that performing *standard training* on robust images yielded similar generalization and adversarial robustness as performing adversarial training on standard images; how does this connect then to human learning if we assume a uniform learning rule in the fovea and the periphery?

We think the answer lies in the fact that as humans learn to perform object recognition, they not only fixate at the target image, but they also look around, and can eventually learn where to make a saccade given candidate object peripheral template – thus learning certain invariances when the object is placed both in the fovea and the periphery (Cox et al., 2005; Williams et al., 2008; Poggio et al., 2014; Han et al., 2020). This is an idea that dates back to Von Helmholtz (1867) as highlighted in Stewart et al. (2020) on the interacting mechanisms of foveal and peripheral vision in humans.

Altogether, this could suggest that spatially-uniform high-resolution processing is redundant and sub-optimal in the *o.o.d.* regime in the way that the visual representation which is computed is independent of point of fixation – as seen classically in adversarially-vulnerable CNNs that are translation invariant and have no foveated/spatially-adaptive computation. Counter-intuitively, the fact that our visual system *is* spatially-adaptive could give rise to a more robust encoding mechanism of the visual stimulus as observers can encode a distribution rather than a point as they move their center of gaze. Naturally, from all the possible types of transformations, the ones that are similar to those shown in this paper – which loosely resemble localized texture-computation – are the ones that potentially lead to a robust hyper-plane during learning for the observer (See Fig. 7 and 8).

Finally, we'd like to add a disclaimer – we use the term *biological plausibility* at a representational level through-out this paper. However, current work is looking into reproducing the experiments carried out in this paper with a physiological component to explore temporal dynamics (MEG) and localization (fMRI) evoked from the stimuli. While it is not obvious if we will find a perceptual signature of the adversarial robust stimuli in humans, we think this novel stimuli and experimental paradigm presents a first step towards the road of linking what is known (and unknown) across texture representation, peripheral computation, and adversarial robustness in humans and machines.

## REFERENCES


Bilal Alsallakh, Narine Kokhlikyan, Vivek Miglani, Jun Yuan, and Orion Reblitz-Richardson. Mind the pad – {cnn}s can develop blind spots. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=m1CD7tPubNy.

Stuart M Anstis. A chart demonstrating variations in acuity with retinal position. *Vision research*, 14(7):589–592, 1974.

Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *Journal of Machine Learning Research*, 20:1–25, 2019.

Benjamin Balas, Lisa Nakano, and Ruth Rosenholtz. A summary-statistic representation in peripheral vision explains visual crowding. *Journal of vision*, 9(12):13–13, 2009.

Alexander Berardino, Valero Laparra, Johannes Ballé, and Eero Simoncelli. Eigen-distortions of hierarchical representations. *Advances in Neural Information Processing Systems*, 30, 2017.

Peter J Burt and Edward H Adelson. The laplacian pyramid as a compact image code. In *Readings in computer vision*, pp. 671–679. Elsevier, 1987.

David D Cox, Philip Meier, Nadja Oertelt, and James J DiCarlo. 'breaking'position-invariant object recognition. *Nature neuroscience*, 8(9):1145–1147, 2005.

Joel Dapello, Tiago Marques, Martin Schrimpf, Franziska Geiger, David D Cox, and James J DiCarlo. Simulating a primary visual cortex at the front of cnns improves robustness to image perturbations. *BioRxiv*, 2020.

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

Arturo Deza, Yi-Chia Chen, Bria Long, and Talia Konkle. Accelerated texforms: Alternative methods for generating unrecognizable object images with preserved mid-level features. In *Conference on Cognitive Computational Neuroscience*, 2019a.

Arturo Deza, Aditya Jonnalagadda, and Miguel P Eckstein. Towards metamerism via foveated style transfer. In *International Conference on Learning Representations*, 2019b.

James J DiCarlo and David D Cox. Untangling invariant object recognition. *Trends in cognitive sciences*, 11(8):333–341, 2007.

K Ding, K Ma, S Wang, and EP Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Comparison of full-reference image quality models for optimization of image processing systems. *International Journal of Computer Vision*, 129(4):1258–1281, 2021.

Gamaleldin F Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alexey Kurakin, Ian J Goodfellow, and Jascha Sohl-Dickstein. Adversarial examples that fool both computer vision and time-limited humans. In *NeurIPS*, 2018.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Brandon Tran, and Aleksander Madry. Adversarial robustness as a prior for learned representations. *arXiv preprint arXiv:1906.00945*, 2019.

Jenelle Feather, Alex Durango, Ray Gonzalez, and Josh McDermott. Metamers of neural networks reveal divergence from human perceptual systems. In *Advances in Neural Information Processing Systems*, pp. 10078–10089, 2019.

Jenelle Feather, Alex Durango, Guillame Leclerc, Aleksander Madry, and Josh McDermott. Adversarial training aligns invariances between artificial neural networks and biological sensory systems. *Cosyne Meeting Abstract*, 2021.

Chaz Firestone. Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences*, 117(43):26562–26571, 2020.

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

Christina M Funke, Judy Borowski, Karolina Stosio, Wieland Brendel, Thomas SA Wallis, and Matthias Bethge. Five points to check when comparing visual perception in humans and machines. *Journal of Vision*, 21(3):16–16, 2021.

Leon Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. *Advances in neural information processing systems*, 28:262–270, 2015.

Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. *Neural Information Processing Systems*, 2021.

Wilson S Geisler and Jeffrey S Perry. Real-time foveated multiresolution system for low-bandwidth video communication. In *Human vision and electronic imaging III*, volume 3299, pp. 294–305. International Society for Optics and Photonics, 1998.

Justin Gilmer, Luke Metz, Fartash Faghri, Sam Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow. Adversarial spheres, 2018. URL `https://openreview.net/forum?id=SyUkxxZ0b`.

Tal Golan, Prashant C. Raju, and Nikolaus Kriegeskorte. Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proceedings of the National Academy of Sciences*, 117(47):29330–29337, 2020. ISSN 0027-8424. doi: 10.1073/pnas.1912334117. URL `https://www.pnas.org/content/117/47/29330`.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *International Conference on Learning Representations*, 2015.

Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Andrei Calian, and Timothy A Mann. Improving robustness using generated data. *Advances in Neural Information Processing Systems*, 34, 2021.

Yena Han, Gemma Roig, Gad Geiger, and Tomaso Poggio. Scale and translation-invariance for novel objects in human vision. *Scientific reports*, 10(1):1–13, 2020.

Daniel Herrera-Esposito, Ruben Coen-Cagli, and Leonel Gomez-Sena. Flexible contextual modulation of naturalistic texture perception in peripheral vision. *Journal of vision*, 21(1):1–1, 2021.

Geoffrey Hinton. How to represent part-whole hierarchies in a neural network. *arXiv preprint arXiv:2102.12627*, 2021.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, pp. 125–136, 2019.

Aditya Jonnalagadda, William Wang, and Miguel P Eckstein. Foveater: Foveated transformer for image classification. *arXiv preprint arXiv:2105.14173*, 2021.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Gang Liu, Yann Gousseau, and Gui-Song Xia. Texture synthesis through convolutional neural networks and spectrum constraints. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pp. 3234–3239. IEEE, 2016.

Nikos K Logothetis, Jon Pauls, and Tomaso Poggio. Shape representation in the inferior temporal cortex of monkeys. *Current biology*, 5(5):552–563, 1995.

Bria Long, Chen-Ping Yu, and Talia Konkle. Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences*, 115 (38):E9015–E9024, 2018.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations*, 2017.

Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5188–5196, 2015.

Marina Martinez-Garcia, Marcelo Bertalmío, and Jesús Malo. In praise of artifice reloaded: Caution with natural image databases in modeling vision. *Frontiers in neuroscience*, 13:8, 2019.

Anirvan S Nandy and Bosco S Tjan. Saccade-confounded image statistics explain visual crowding. *Nature neuroscience*, 15(3):463–469, 2012.

Tomaso Poggio, Jim Mutch, and Leyla Isik. Computational role of eccentricity dependent cortical magnification. *arXiv preprint arXiv:1406.1770*, 2014.

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.

Sylvestre-Alvise Rebuffi, Sven Gowal, Dan A Calian, Florian Stimberg, Olivia Wiles, and Timothy Mann. Fixing data augmentation to improve adversarial robustness. *Neural Information Processing Systems*, 2021.

Manish V Reddy, Andrzej Banburski, Nishka Pant, and Tomaso Poggio. Biologically inspired mechanisms for adversarial robustness. *Neural Information Processing Systems*, 2020.

Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11):1019–1025, 1999.

Ruth Rosenholtz. Capabilities and limitations of peripheral vision. *Annual Review of Vision Science*, 2:437–457, 2016.

Ruth Rosenholtz, Jie Huang, Alvin Raj, Benjamin J Balas, and Livia Ilie. A summary statistic representation in peripheral vision explains visual search. *Journal of vision*, 12(4):14–14, 2012.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Shibani Santurkar, Dimitris Tsipras, Brandon Tran, Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Image synthesis with a single (robust) classifier. *Neural Information Processing Systems*, 2019.

Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

Emma EM Stewart, Matteo Valsecchi, and Alexander C Schütz. A review of interactions between peripheral and foveal vision. *Journal of vision*, 20(12):2–2, 2020.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *International Conference on Learning Representations*, 2014.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *International Conference on Learning Representations*, 2019.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Hermann Von Helmholtz. *Handbuch der physiologischen Optik: mit 213 in den Text eingedruckten Holzschnitten und 11 Tafeln*, volume 9. Voss, 1867.

Thomas SA Wallis, Matthias Bethge, and Felix A Wichmann. Testing models of peripheral encoding using metamerism in an oddity paradigm. *Journal of vision*, 16(2):4–4, 2016.

Thomas SA Wallis, Christina M Funke, Alexander S Ecker, Leon A Gatys, Felix A Wichmann, and Matthias Bethge. A parametric texture model based on deep convolutional features closely matches texture appearance for humans. *Journal of vision*, 17(12):5–5, 2017.

Thomas SA Wallis, Christina M Funke, Alexander S Ecker, Leon A Gatys, Felix A Wichmann, and Matthias Bethge. Image content is more important than bouma's law for scene metamers. *ELife*, 8:e42512, 2019.

Mark A Williams, Chris I Baker, Hans P Op De Beeck, Won Mok Shim, Sabin Dang, Christina Triantafyllou, and Nancy Kanwisher. Feedback of visual object information to foveal retinotopic cortex. *Nature neuroscience*, 11(12):1439–1445, 2008.

Corey M Ziemba and Eero P Simoncelli. Opposing effects of selectivity and invariance in peripheral vision. *Nature Communications*, 12(1):1–11, 2021.

Corey M Ziemba, Jeremy Freeman, J Anthony Movshon, and Eero P Simoncelli. Selectivity and tolerance for visual texture in macaque v2. *Proceedings of the National Academy of Sciences*, 113(22):E3140–E3149, 2016.