Emergent Properties of Foveated Perceptual Systems

Arturo Deza1,2
Center for Brains, Minds and Machines
Massachusetts Institute of Technology1
deza@mit.edu

Talia Konkle2
Department of Psychology
Harvard University2
talia_konkle@harvard.edu

Abstract

The goal of this work is to characterize the representational impact that foveation operations have for machine vision systems, inspired by the foveated human visual system, which has higher acuity at the center of gaze and texture-like encoding in the periphery. To do so, we introduce models consisting of a first-stage fixed image transform followed by a second-stage learnable convolutional neural network, and we varied the first stage component. The primary model has a foveated-textural input stage, which we compare to a model with foveated-blurred input and a model with spatially-uniform blurred input (both matched for perceptual compression), and a final reference model with minimal input-based compression. We find that: 1) the foveated-texture model shows similar scene classification accuracy as the reference model despite its compressed input, with greater i.i.d. generalization than the other models; 2) the foveated-texture model has greater sensitivity to high-spatial frequency information and greater robustness to occlusion, w.r.t the comparison models; 3) both the foveated systems, show a stronger center image-bias relative to the spatially-uniform systems even with a weight sharing constraint. Critically, these results are preserved over different classical CNN architectures throughout their learning dynamics. Altogether, this suggests that foveation with peripheral texture-based computations yields an efficient, distinct, and robust representational format of scene information, and provides symbiotic computational insight into the representational consequences that texture-based peripheral encoding may have for processing in the human visual system, while also potentially inspiring the next generation of computer vision models via spatially-adaptive computation. Code + Data available here: https://github.com/ArturoDeza/EmergentProperties.

1 Introduction

In the human visual system, incoming light is sampled with different resolution across the retina, a stark contrast to machines that perceive images at uniform resolution. One account for the nature of this foveated (spatially-varying) array in humans is related purely to sensory efficiency (biophysical constraints) [Land & Nilsson 2012; Eckstein 2011], e.g., there is only a finite amount of retinal ganglion cells (RGC) that can relay information from the retina to the Lateral Geniculate Nucleus (LGN) constrained by the thickness of the optic nerve. Thus it is “more efficient” to have a moveable high-acuity fovea, rather than a non-moveable uniform resolution retina when given a limited number of photoreceptors as suggested in [Akbas & Eckstein 2017]. Machines, however do not have such wiring/resource constraints – and with their already proven success in computer vision [LeCun et al. 2015] – this raises the question if a foveated inductive bias is necessary for vision at all.

However, it is also possible that foveation plays a functional role at the representational level, which may confer perceptual advantages – as most computational approaches have mainly focused on saccade planning [Geisler et al. 2006; Mnih et al. 2014; Elsayed et al. 2019; Daucé et al. 2020]. This idea has remained elusive in computer vision, but popular in vision science, and has been explored both psychophysically [Loschky et al. 2019] and computationally [Poggio et al. 2014].
Figure 1: A cartoon illustrating how a biologically-inspired foveated image (texture-based) is rendered resembling a human visual metamer via the foveated feed-forward style transfer model of Deza et al. (2019). Here, each receptive field is locally perturbed with noise in its latent space in the direction of their equivalent texture representation (blue arrows) resulting in visual crowding effects that warp the image locally in the periphery (Balas et al., 2009; Freeman & Simoncelli, 2011; Rosenholtz, 2016). These effects are most noticeable far away from the navy dot which is the simulated center of gaze (foveal region) of an observer under certain viewing conditions.

Cheung et al. (2017); Han et al. (2020). Other works that have suggested representational advantages of foveation include the work of Pramod et al. (2018), where blurring the image in the periphery gave an increase in object recognition performance of computer vision systems by reducing their false positive rate. In Wu et al. (2018)'s GistNet, directly introducing a dual-stream foveal-peripheral pathway in a neural network boosted object detection performance via scene gist and contextual cueing. Relatedly, the most well known example of work that has directly shown the advantage of peripheral vision for scene processing in humans is Wang & Cottrell (2017)'s dual stream CNN that modelled the results of Larson & Loschky (2009) with a log-polar transform and adaptive Gaussian blurring (RGC-convergence). Taken together, these studies present support for the idea that foveation has useful representational consequences for perceptual systems. Further, these computational examples have symbiotic implications for understanding biological vision, indicating what the functional advantages of foveation in humans may be, via functional advantages in machine vision systems.

Importantly, none of these studies introduce the notion of texture representation in the periphery – a key property of peripheral computation as posed in Rosenholtz (2016). What functional consequences does this well-known texture-based coding in the visual periphery have, if any, on the nature of later stage visual representation? Here we directly examine this question. Specifically, we introduce perceptual systems: as two-stage models that have an image transform stage followed by a deep convolutional neural network. The primary model class of interest possesses a first stage image transform that mimics texture-based foveation via visual crowding (Levi, 2011; Pelli, 2008; Doerig et al., 2019a) in the periphery as shown in Figure 1 (Deza et al., 2019), rather than Gaussian blurring (Wang & Cottrell, 2017; Pramod et al., 2018; Malkin et al., 2020) or compression (Patney et al., 2016; Kaplanyan et al., 2019). These rendered images capture image statistics akin to those preserved in human peripheral vision, and resembling texture computation at the stage of area V2, as argued in Freeman & Simoncelli (2011); Rosenholtz (2016); Wallis et al. (2019).

Our strategy is thus to compare in terms of generalization, robustness and bias these foveation-texture models to three other kinds of models. The first comparison model class – foveation-blur models – uses the same spatially-varying foveation operations but uses blur rather than texture based input. The second class – uniform-blur models – uses a blur operation uniformly over the input, with the level of blur set to match the perceptual compression rates of the foveation-texture nets. Finally, the last comparison model class is the reference, which has minimal distortion, and serves as a perceptual upper bound from which to assess the impact of these different first-stage transforms.

Note that our approach is different from the one taken by Wang & Cottrell (2017), who have built foveated models that fit results to human behavioural data like those of Larson & Loschky (2009). Rather, our goal is to explore the emergent properties in CNNs with texture-based foveation on scene representation compared to their controls agnostic to any behavioural data or expected outcome. Naturally, the results of our experimental paradigm is symbiotic as it can shed light into both the importance of texture-based peripheral computation in humans, and could also suggest a new inductive bias for advanced machine perception in scenes.
2 Perceptual Systems

We define perceptual systems as two-stage models with an image transform \( f(\circ) : \mathbb{R}^D \rightarrow \mathbb{R}^D \), that is relayed to a deep convolutional neural network \( g(\circ) : \mathbb{R}^D \rightarrow \mathbb{R}^d \). Note that the first transform stage is a fixed operation over the input image, while the second stage has learnable parameters. In general, the perceptual system \( S(\circ) \), with retinal image input \( I : \mathbb{R}^D \) is defined as:

\[
S(I) = g(f(I))
\]

Such two-stage models have been growing in popularity, and the reasons these models are designed to not be fully end-to-end differentiable is mainly to force one type of computation into the first-stage of a system such that the second-stage \( g(\circ) \) must figure out how to capitalize on such forced transformation and thus assess its \( f(\circ) \) representational consequences (See Figure 2). For example, Parthasarathy & Simoncelli (2020) successfully imposed V1-like computation in stage 1 to explore the learned role of texture representation in later stages with a self-supervised objective, and Dapello et al. (2020) found that fixing V1-like computation also at stage 1 aided adversarial robustness. At a higher level, our objective is similar where we would like to force a texture-based peripheral coding mechanism (loosely inspired by V2; Ziemba et al., 2016) at the first stage to check if the perceptual system (now foveated) will learn to pick-up on this newly made representation through \( g(\circ) \) and make ‘good’ use of it potentially shedding light on the functionality hypothesis for machines and humans.

2.1 Stage 1: Image Transform

To model the computations of a texture-based foveated visual system, we employed the model of Deza et al. (2019) (henceforth Foveated-Texture Transform). This model is inspired by the metamer synthesis model of Freeman & Simoncelli (2011), where new images are rendered to have locally matching texture statistics (Portilla & Simoncelli, 2000; Balas et al., 2009) in greater size pooling regions of the visual periphery with structural constraints. Analogously, the Deza et al. (2019) Foveation Transform uses a foveated feed-forward style transfer network to latently perturb the image in the direction of its locally matched texture (see Figure 1). Altogether, \( f : \mathbb{R}^D \rightarrow \mathbb{R}^D \) is a convolutional auto-encoder that is non-foveated when the latent space is unperturbed: \( f_0(I) = D(\mathcal{E}(I)) \), but foveated \( \circ \Sigma \) when the latent space is perturbed via localized style transfer: \( f_\Sigma(I) = D(\mathcal{E}_\Sigma(\mathcal{I})) \), for a given encoder-decoder \( (\mathcal{E}, D) \) pair.

Note that with proper calibration, the resulting distorted image can be a visual metamer (for a human), which is a carefully perturbed image perceptually indistinguishable from its reference image (Freeman & Simoncelli, 2011; Rosenholtz et al., 2012; Feather et al., 2019; Vacher et al., 2020). However, importantly in the present work, we exaggerated the strength of these texture-driven distortions
Figure 3: A. Two perceptually matched-resource controls to Foveation-Texture are introduced. Middle-Right, orchid: uniform blurring emulating a matched-resource non-foveated visual system (Uniform-Blur); Far-Right, seagreen: adaptive gaussian blurring (Foveation-Blur) emulating a matched resource blur-based foveated system. B. A Rate-Distortion Optimization procedure is summarized where we find the hyper-parameters of the new matched-resource image transforms \( \{f_0(\cdot), \hat{f}_\ast(\cdot)\} \) to Foveation-Texture via expected SSIM matching over the validation set.

(beyond the metameric boundary), as our aim here is to understand the implications of this kind of texturized peripheral input on later stage representations (e.g. following a similar approach as Dapello et al. (2020)). By having an extreme manipulation, we reasoned this would accentuate the consequences of these distortions, making them more detectable in our subsequent experiments.

### 2.2 Stage 2: Convolutional Neural Network backbone

The transformed images (stage 1) are passed into a standard convolutional neural network architecture. Here we tested two different base architectures: AlexNet (Krizhevsky et al., 2012), and ResNet18 (He et al., 2016). The goal of running these experiments on two different hierarchically local architectures is to let us examine the consequences across all image transforms (with our main focus towards texture-based foveation) that are robust to these different network architectures. Further, this CNN backbone \( g : \mathbb{R}^D \rightarrow \mathbb{R}^d \) should not be viewed in the traditional way of an end-to-end input/output system where the input is the retinal image \( I \), and the output is a one-hot vector encoding a \( d \)-class-label in \( \mathbb{R}^d \). Rather, the CNN \( g \) acts as a loose proxy of higher stages of visual processing (as it receives input from \( f \)), analogous to the 2-stage model of Lindsey et al. (2019).

### 2.3 Critical Manipulations: Foveated vs Non-Foveated Perceptual Systems

Now, we can define the first two of the four perceptual systems that will perform 20-way scene categorization: Foveation-Texture, receives an image input, applies the foveation-texture transform \( f_\ast(\cdot) \), and relays it through the CNN \( g(\cdot) \). Similarly, Reference performs a non-foveated transform \( f_0(\cdot) \), where images are sent through the same convolutional auto-encoder \( D(E(I)) \) of \( f_\ast(\cdot) \), but with the parameter that determines the degree of texture style transfer set to 0 – producing an upper-bounded, compressed and non-foveated reference image – then relayed through the CNN \( g(\cdot) \). Both of these systems are depicted in Figure 2 (A). As the foveation-texture model has less information from the input, relative to the reference networks, we next designed two further comparison models which have a comparable amount of information after the input stage, but with different amounts of blurring in the stage 1 operations. To create matched-resources systems, our broad approach was to use a Rate-Distortion (RD) optimization procedure (Ballé et al., 2016) to match information between the stage 1 operations, given the SSIM (Wang et al., 2004) image quality assessment (IQA) metric.

Specifically, to create matched-resource Uniform-Blur, we identified the standard deviation of the Gaussian blurring kernel (the ‘distortion’ \( \mathcal{D} \)), such that we could render a perceptually resource-matched Gaussian blurred image – w.r.t Reference – that matches the perceptual transmission ‘rate’ \( \mathcal{R} \) of Foveation-Texture via the SSIM perceptual metric (Wang et al., 2004). This procedure yields a model class with uniform blur across the image, but with matched stage 1 information content as the
Foveation-Texture. And, to create matched-resource Foveation-Blur, we carried out this same RD optimization pipeline per eccentricity ring (assuming homogeneity across pooling regions at the same eccentricity), thus finding a set of blurring coefficients that vary as a function of eccentricity. This procedures yielded a different matched-resource model class, this time with spatially-varying blur. Figure 3(B) summarizes our solution to this problem. Details of the RD Optimization are presented in Appendix A.

Ultimately, it is important to note that the selection of the perceptual metric (SSIM in our case), plays a role in this optimization procedure, and sets the context in which we can call a network ‘resource-matched’. We selected SSIM given its monotonic relationship to distortions to human perceptual judgements, symmetric upper-bounded nature, sensitivity to contrast, local structure and spatial frequency, and popularity in the Image Quality Assessment (IQA) community. However to anticipate any possible discrepancy in the interpretability of our future results, we additionally computed the Mean Square Error (MSE), MS-SSIM, and 11 other IQA metrics as recently explored in Ding et al. [2020] to compare all other image transforms to the Reference on the testing set. Our logic is the following: if the MSE is greater(+) for Foveation-Texture compared to Foveation-Blur and Uniform-Blur, then the current distortion levels place Foveation-Texture at a resource ‘disadvantage’ relative to the other transforms, and any interesting results would not only hold but also be strengthened. This same logic applies to the other IQA metrics contingent on their direction of greater distortion. Indeed, these patterns of results were evident across IQA metrics – except those tolerant to texture such as DISTS [Ding et al., 2020] – as shown in Table 1 and Appendix C.

| Image Transforms | SSIM (Matched) | MS-SSIM (+) | MSE (↑) | Mutual Information (↑) | NLDI (↑) | DISTS (↑) |
|------------------|----------------|------------|--------|------------------------|---------|-----------|
| Reference        | 0.98 ± 0.01    | 0.97 ± 0.02| 0.00   | 7.39 ± 0.52            | 0       | 0         |
| Foveation-Texture| 0.58 ± 0.12    | 0.20 ± 0.03| 0.40 ± 0.42| 1.40 ± 0.42          | 0.75 ± 0.16| 0.20 ± 0.03|
| Uniform-Blur     | 0.57 ± 0.15    | 0.36 ± 0.03| 0.45 ± 0.13| 1.85 ± 0.18          | 0.40 ± 0.09| 0.36 ± 0.03|
| Foveation-Blur   | 0.58 ± 0.15    | 0.36 ± 0.03| 1.84 ± 0.96| 0.45 ± 0.11          | 0.35 ± 0.03| 0         |

Table 1: Comparing Image Transforms wrt Reference. Arrows indicate direction of greater distortion.

3 Experiments

Altogether, the 4 previously introduced perceptual systems help us answer three key questions that we should have in mind throughout the rest of the paper: 1) Foveation-Texture vs Reference will tell us how a texture-based foveation mechanism will compare to its perceptual upper-bound – shedding light into arguments about computational efficiency. 2) Foveation-Texture vs Foveation-Blur will tell us if any potentially interesting pattern of results is due to the type/stage of foveation. This will help us measure the contributions of the adaptive texture coding vs adaptive gaussian blurring; 3) Foveation-Texture vs Uniform-Blur will tell us how do these perceptual systems (one foveated, and the other one not) behave when allocated with a fixed number of perceptual resources under certain assumptions – potentially shedding light on why biological organisms like humans have foveated texture-based computation in the visual field instead of uniform spatial processing like modern machines.

Dataset: All previously introduced models were trained to perform 20-way scene categorization. Scene categories were selected from the Places2 dataset [Zhou et al., 2017], and were re-partitioned into a new 4500 images per category for training, 250 per category for validation, and 250 per category for testing. The categories included were: aquarium, badlands, bedroom, bridge, campus, corridor, forest path, highway, hospital, industrial area, japanese garden, kitchen, mansion, mountain, ocean, office, restaurant, skyscraper, train interior, waterfall. Samples of these scenes coupled with their image transforms can be seen in Figure 4.

Networks: Training: Convolutional neural networks of the stage 2 of each perceptual system were trained which resulted in 40 image-transform based networks per architecture (AlexNet/ResNet18):
Figure 5: Scene Categorization Accuracy of AlexNet and ResNet18 as $g(\circ)$. We observe the following: Foveation-Texture has greater i.i.d. generalization than other matched-resource systems across both network architectures; Uniform-Blur’s o.o.d. generalization interacts with the architecture (performing worse for ResNet18, but highest for AlexNet); Foveation-Blur maintains high o.o.d. generalization independent of network architecture. Confusion Matrices can be seen in Appendix H.

10 Foveation-Texture, 10 Reference, 10 Uniform-Blur, 10 Foveation-Blur; totalling 80 trained networks to compute relevant error bars shown in all figures (standard deviations, not standard errors) and to reduce effects of randomness driven by the particular network initialization. All systems were paired such that their stage 2 architectures $g(\circ)$ started with the same random weight initialization prior to training. Testing: The networks of each perceptual system were tested on the same type of image distribution they were trained on. Learning Dynamics: Available in Appendix G.

3.1 Texture-based foveation provides greater i.i.d. generalization than Blur-based foveation

How well does the foveation-texture stage classify scene images (i.i.d. generalization) compared to the other matched-resource models that use blurring and the reference? The results can be seen in Figure 5. Each bars’ height reflects overall accuracy for each of the 10 neural network backbone runs ($g(\circ)$) per system, with a square marker at the top indicating the i.i.d. accuracy. We found that Foveation-Texture had similar i.i.d. performance to the Reference – which is the undistorted perceptual upper bound, and greater performance than both Uniform-Blur and Foveation-Blur. Thus the compression induced by foveated-texture generally maintains scene category information.

We next performed a contrived experiment where we tested how well each perceptual system could classify the stage 1 outputs of the other models. For example, we showed a set of foveated blurred images to a network trained on foveated texture images. This experiment is in essence a test of out-of-distribution (o.o.d.) generalization. The results of these tests are also shown in Figure 5. For each model, the classification accuracy for the inputs from the other stage 1 images is indicated by the height of the different colored diamonds, where the color corresponds to the stage 1 operation.

This experiment yielded a rather complex set of patterns, that even differed depending on the architecture (AlexNet vs ResNet18 as $g(\circ)$). Generally, the Foveation-Texture model had a similar profile of generalization as the Reference model. However, the networks trained with different types of blur (Uniform-Blur & Foveated-Blur) in some cases showed very high o.o.d. generalization – though once again this is contingent on $g(\circ)$.

Unraveling the underlying causes to understand this last set of results sets the stage for our experiments in the rest of this section. So far it seems like Foveation-Texture has learned to properly capitalize the texture information in the periphery and still out-perform all other matched-resource systems even if heavily penalized under several IQA metrics (Table 1) – highlighting the critical differences in texture vs blur for scene processing. As for the interaction of Uniform-Blur with $g(\circ)$, is is likely that the residual connections are counter-productive to o.o.d. generalization (or it has overfit). Interestingly, humans have a combination of texture and adaptive-gaussian based peripheral computation (Ehinger & Rosenholtz 2016). so future work should look into the effects of continual learning, joint-training or a combined image transform (Texture + Blur) to merge gains of both i.i.d and o.o.d generalization.
3.2 Texture-based foveated systems preserve greater high-spatial frequency sensitivity

We next examined whether the learned feature representations of these models are more reliant on low or high pass spatial frequency information. To do so, we filtered the testing image set at multiple levels to create both high pass and low pass frequency stimuli and assessed scene-classification performance over these images for all models, as shown in Figure 6. Low pass frequency stimuli were rendered by convolving a Gaussian filter of standard deviation $\sigma = [0, 1, 3, 5, 7, 10, 15, 40]$ pixels on the foveation transform $(f_0, \hat{f}_0, f_*, \hat{f}_*)$ outputs. Similarly, the high pass stimuli was computed by subtracting the reference image from its low pass filtered version with $\sigma = [\infty, 3, 1, 0.7, 0.55, 0.45, 0.4]$ pixels and adding a residual. These are the same values used in the experiments of Geirhos et al. (2019).

We found that Foveation-Texture and Reference trained networks were more sensitive to High Pass Frequency information, while Foveation-Blur and Uniform-Blur were selective to Low Pass Frequency stimuli. Although one may naively assume that this is an expected result – as both Foveation-Blur and Uniform-Blur networks are exposed to a blurring procedure – it is important to note that: 1) the foveal resolution has been preserved between Foveation-Texture and Foveation-Blur (See Fig. 4), thus high spatial frequency sensitivity could have still predominated in Foveation-Blur but it did not (though see Fig. 6 A2/B2 where these high pass Gabors are still learned, implying that higher layers in $g(\cdot)$ overshadow their computation); and 2) Foveation-Texture could have also learned to develop low spatial frequency sensitivity given the crowding/texture-like peripheral distortion, but this was not the case (likely due to the weight sharing constraint embedded in the CNN architecture Elsayed et al., 2020). Finally, the robustness to low-pass filtering of Foveation-Blur suggests that foveation via adaptive gaussian blurring may implicitly contribute to scale-invariance as also shown in Poggio et al. (2014); Cheung et al. (2017); Han et al. (2020).

3.3 Texture-based foveation develops greater robustness to occlusion

We next examined how all perceptual systems could classify scene information under conditions of visual field loss, either from left to right (left2right), top to bottom (top2bottom), center part of
Overall we found that, across all types of occlusion the Foveation-Texture modules have greater robustness to occlusion than both the Foveation-Blur and Uniform-Blur models. Further, the Foveation-Texture models have nearly equivalent performance to the Reference. In contrast, both models with blurring, whether uniformly or in a spatially-varying way, were far worse at classifying scenes under conditions of visual field loss. These results highlight that the texture-based information content captured by the foveation-texture nets preserves scene category content in dramatically different way than simple lower-resolution sampling – perhaps using the texture-bias (Geirhos et al., 2019) in their favor; as humans too use texture as their classification strategy for scenes (Renninger & Malik, 2004).

In addition, the Foveation-Texture model is not overfitting. As recent work has suggested an Accuracy vs Robustness trade-off where networks trained to outperform under the i.i.d. generalization condition will do worse under other perceptual tasks – mainly adversarial (Zhang et al., 2019) – we did not observe such trade-off and a greater accuracy did not imply lower robustness to occlusion.

3.4 Foveated systems learn a stronger center image bias than non-foveated systems

It is possible that foveated systems weight visual information strongly in the foveal region than the peripheral region as hinted by our occlusion results (the different rate of decay for the accuracy curves in the Scotoma and Glaucoma conditions). To resolve this question, we conducted an experiment where we created a windowed cue-conflict stimuli where we re-rendered our set of testing images with one image category in the fovea, and another one in the periphery (all aligned with a different class systematically; ex: aquarium with badlands). We also had an additional condition where the conflicting cue was now square-like and uniformly and randomly paired with a conflicting scene class and more finely sampled. We then systematically varied the fovea-periphery visual area ratios & re-examined classification accuracy for both the foveal and peripheral scenes (Figure 8).

We found that the Foveation-Texture and Foveation-Blur transform imposed the networks $g(\phi)$ to learn to weigh information in the center of the image stronger than Reference & Uniform-Blur for scene categorization. A qualitative way of seeing this foveal-bias is by checking the foveal/peripheral
ratio where these two accuracy lines cross. The more leftward the cross-over point (⊗), the higher the foveal bias (highlighted through the vertical bars). This result was unexpected as we initially predicted that \( g(\cdot) \) would weigh the peripheral information stronger as it has been implicitly regularized through a distortion. However this was not the case and our findings are similar to Wang & Cottrell (2017) who showed this foveal bias on a foveated system with adaptive blur with a dual-stream neural network. Thus, these results indicate that the spatially varying computation from center to periphery is mainly responsible for the development of a center image bias even with a weight sharing constraint. Furthermore, it is possible that one of the functions of any spatially-varying coding mechanisms in the visual field is to enforce the perceptual system to attend on the foveal region – avoiding the shortcut of learning to attend the entire visual field if unnecessary (Geirhos et al., 2020).

4 Discussion

The present work was designed to probe the impact of foveated texture-based input representations in machine vision systems. To do this we specifically compared the learned perceptual signatures in the second-stage of visual processing across a set of networks trained on other image transforms. We found that when comparing Foveation-Texture to their matched-resource models that differed in computation: Foveation-Blur (foveated w/ adaptive gaussian blur) and Uniform-Blur (non-foveated w/ uniform blur) – that peripheral texture encoding did lead to specific representational signatures, particularly greater i.i.d generalization, preservation of high-spatial frequency sensitivity, and robustness to occlusion – even as high as its perceptual upper bound (Reference). We also found that foveation (in general) seems to induce a focusing mechanism, servicing the foveal/central region – whereas neither a perceptually upper-bounded system (Reference) or a non-foveated compressed system (Uniform-Blur) did not develop as strongly.

Figure 8: Foveated Perceptual Systems – independent of the computation type (Foveation-Texture, Foveation-Blur) – show stronger biases to classify hybrid scenes with the foveal region; a result also observed in humans (Larson & Loschky, 2009).

The particular consequences of our foveation stage raises interesting future directions about what computational advantages could arise when trained on object categorization (Pramod et al., 2018) coupled with eye-movements (Akbas & Eckstein, 2017; Deza et al., 2017), as objects are typically centered in view and have different hierarchical/compositional priors than scenes (Zhou et al., 2014; Deza et al., 2020), in addition to different processing mechanisms (Renninger & Malik, 2004; Ehinger & Rosenholz, 2016). We are currently exploring the impact of these foveated texture-based representational signatures on shape vs texture bias for object recognition similar to Geirhos et al. (2019) and Hermann et al. (2020), and assessing their interaction with scene representation.

Further, a future direction is investigating the effects of texture-based foveation to adversarial robustness. Motivated by the recent work of Dapello et al. (2020) which has shown promise of adversarial robustness via enforcing stochasticity and V1-like computation by obeying the Nyquist sampling frequency of these filters w.r.t the image (Serre et al., 2007) in addition to a natural gamut of orientations and frequencies as studied in De Valois et al. (1982), it raises the question of how much we can further push for robustness in hybrid perceptual systems like these, drawing on even more biological mechanisms. Works such as Luo et al. (2015) and recently Reddy et al. (2020), Kiritani & Ono (2020) have already taken steps in this direction by coupling fixations with a spatially-varying retina. However, the representational impact of texture-based foveation on adversarial robustness, and its symbiotic implication for human vision still remains an open question.
References

Akbas, E. and Eckstein, M. P. Object detection through search with a foveated visual system. *PLoS computational biology*, 13(10):e1005743, 2017.

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

Ballé, J., Laparra, V., and Simoncelli, E. P. End-to-end optimized image compression. *arXiv preprint arXiv:1611.01704*, 2016.

Cheung, B., Weiss, E., and Olshausen, B. Emergence of foveal image sampling from learning to attend in visual scenes. *International Conference on Learning Representations (ICLR)*, 2017.

Dapello, J., Marques, T., Schrimpf, M., Geiger, F., Cox, D. D., and DiCarlo, J. J. Simulating a primary visual cortex at the front of cnns improves robustness to image perturbations. *BioRxiv*, 2020.

Daucé, E., Albiges, P., and Perrinet, L. U. A dual foveal-peripheral visual processing model implements efficient saccade selection. *Journal of Vision*, 20(8):22–22, 2020.

De Valois, R. L., Yund, E. W., and Hepler, N. The orientation and direction selectivity of cells in macaque visual cortex. *Vision research*, 22(5):531–544, 1982.

Deza, A. and Eckstein, M. Can peripheral representations improve clutter metrics on complex scenes? In *Advances in Neural Information Processing Systems*, pp. 2847–2855, 2016.

Deza, A., Peters, J. R., Taylor, G. S., Surana, A., and Eckstein, M. P. Attention allocation aid for visual search. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 220–231, 2017.

Deza, A., Jonnalagadda, A., and Eckstein, M. P. Towards metamerism via foveated style transfer. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BJzbG20cFQ

Deza, A., Liao, Q., Banburski, A., and Poggio, T. Hierarchically local tasks and deep convolutional networks. *CBMM Memo*, 2020.

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

Ding, K., Ma, K., Wang, S., and Simoncelli, E. P. Comparison of Image Quality Models for Optimization of Image Processing Systems. *arXiv e-prints*, art. arXiv:2005.01338, May 2020.

Doerig, A., Bornet, A., Choung, O. H., and Herzog, M. H. Crowding reveals fundamental differences in local vs. global processing in humans and machines. *bioRxiv*, 2019a. doi: 10.1101/744268. URL https://www.biorxiv.org/content/early/2019/08/23/744268

Doerig, A., Bornet, A., Rosenholtz, R., Francis, G., Clarke, A. M., and Herzog, M. H. Beyond bouma’s window: How to explain global aspects of crowding? *PLoS computational biology*, 15(5):e1006580, 2019b.

Eckstein, M. P. Visual search: A retrospective. *Journal of vision*, 11(5):14–14, 2011.

Eckstein, M. P., Koehler, K., Welbourne, L. E., and Akbas, E. Humans, but not deep neural networks, often miss giant targets in scenes. *Current Biology*, 27(18):2827–2832, 2017.

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

Elsayed, G., Kornblith, S., and Le, Q. V. Saccader: Improving accuracy of hard attention models for vision. In *Advances in Neural Information Processing Systems*, pp. 700–712, 2019.

Elsayed, G., Ramachandran, P., Shlens, J., and Kornblith, S. Revisiting spatial invariance with low-rank local connectivity. In *International Conference on Machine Learning*, pp. 2868–2879. PMLR, 2020.
Feather, J., Durango, A., Gonzalez, R., and McDermott, J. Metamers of neural networks reveal divergence from human perceptual systems. *Advances in Neural Information Processing Systems*, 32:10078–10089, 2019.

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

Fridman, L., Jenik, B., Keshvari, S., Reimer, B., Zetzsche, C., and Rosenholtz, R. Sideeye: A generative neural network based simulator of human peripheral vision. *arXiv preprint arXiv:1706.04568*, 2017.

Gatys, L. A., Ecker, A. S., and Bethge, M. Texture synthesis using convolutional neural networks. *arXiv preprint arXiv:1505.07376*, 2015.

Gatys, L. A., Ecker, A. S., and Bethge, M. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2414–2423, 2016.

Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., and Wichmann, F. A. Generalisation in humans and deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 7538–7550, 2018.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bygh9j09KX

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *arXiv preprint arXiv:2004.07780*, 2020.

Geisler, W. S. and Perry, J. S. 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.

Geisler, W. S., Perry, J. S., and Najemnik, J. Visual search: The role of peripheral information measured using gaze-contingent displays. *Journal of Vision*, 6(9):1–1, 2006.

Han, Y., Roig, G., Geiger, G., and Poggio, T. Scale and translation-invariance for novel objects in human vision. *Scientific Reports*, 10(1):1–13, 2020.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hermann, K. L., Chen, T., and Kornblith, S. The origins and prevalence of texture bias in convolutional neural networks. *Neural Information Processing Systems*, 2020.

Huang, X. and Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1501–1510, 2017.

Kaplanyan, A. S., Sochenov, A., Leimkühler, T., Okunev, M., Goodall, T., and Rufo, G. Deepfovea: neural reconstruction for foveated rendering and video compression using learned statistics of natural videos. *ACM Transactions on Graphics (TOG)*, 38(6):1–13, 2019.

Kiritani, T. and Ono, K. Recurrent attention model with log-polar mapping is robust against adversarial attacks. *arXiv preprint arXiv:2002.05388*, 2020.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Land, M. F. and Nilsson, D.-E. *Animal eyes*. Oxford University Press, 2012.

Laparra, V., Ballé, J., Berardino, A., and Simoncelli, E. P. Perceptual image quality assessment using a normalized laplacian pyramid. *Electronic Imaging*, 2016(16):1–6, 2016.
Larson, A. M. and Loschky, L. C. The contributions of central versus peripheral vision to scene gist recognition. *Journal of Vision*, 9(10):6–6, 2009.

Larson, E. C. and Chandler, D. M. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of electronic imaging*, 19(1):011006, 2010.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *nature*, 521(7553):436, 2015.

Levi, D. M. Visual crowding. *Current Biology*, 21(18):R678–R679, 2011.

Lindsey, J., Ocko, S. A., Ganguli, S., and Deny, S. The effects of neural resource constraints on early visual representations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=S1xq3oR5tQ.

Loschky, L. C., Szaffarczyk, S., Beugnet, C., Young, M. E., and Boucart, M. The contributions of central and peripheral vision to scene-gist recognition with a 180 visual field. *Journal of Vision*, 19(5):15–15, 2019.

Luo, Y., Boix, X., Roig, G., Poggio, T., and Zhao, Q. Foveation-based mechanisms alleviate adversarial examples. *arXiv preprint arXiv:1511.06292*, 2015.

Malkin, E., Deza, A., and tomaso a poggio. (CUDA)-optimized real-time rendering of a foveated visual system. In *NeurIPS 2020 Workshop SVRHM*, 2020. URL https://openreview.net/forum?id=ZMsqkUadtZ7.

Mnih, V., Heess, N., Graves, A., et al. Recurrent models of visual attention. In *Advances in neural information processing systems*, pp. 2204–2212, 2014.

Parthasarathy, N. and Simoncelli, E. P. Self-supervised learning of a biologically-inspired visual texture model. *arXiv preprint arXiv:2006.16976*, 2020.

Patney, A., Salvi, M., Kim, J., Kaplanyan, A., Wyman, C., Benty, N., Luebke, D., and Lefohn, A. Towards foveated rendering for gaze-tracked virtual reality. *ACM Transactions on Graphics (TOG)*, 35(6):179, 2016.

Pelli, D. G. Crowding: A cortical constraint on object recognition. *Current opinion in neurobiology*, 18(4):445–451, 2008.

Poggio, T., Mutch, J., and Isik, L. Computational role of eccentricity dependent cortical magnification. *arXiv preprint arXiv:1406.1770*, 2014.

Portilla, J. and Simoncelli, E. P. A parametric texture model based on joint statistics of complex wavelet coefficients. *International journal of computer vision*, 40(1):49–70, 2000.

Pramod, R. T., Katti, H., and Arun, S. P. Human peripheral blur is optimal for object recognition. *arXiv preprint arXiv:1807.08476*, 2018.

Reddy, M. V., Banburski, A., Pant, N., and Poggio, T. Biologically inspired mechanisms for adversarial robustness. *arXiv preprint arXiv:2006.16427*, 2020.

Renninger, L. W. and Malik, J. When is scene identification just texture recognition? *Vision research*, 44(19):2301–2311, 2004.

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

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

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., and Poggio, T. Robust object recognition with cortex-like mechanisms. *IEEE transactions on pattern analysis and machine intelligence*, 29(3):411–426, 2007.
Sheikh, H. R. and Bovik, A. C. Image information and visual quality. *IEEE Transactions on image processing*, 15(2):430–444, 2006.

Shumikhin, M. M. A. *Quantitative measures of crowding susceptibility in peripheral vision for large datasets*. PhD thesis, Massachusetts Institute of Technology, 2020.

Vacher, J., Davila, A., Kohn, A., and Coen-Cagli, R. Texture interpolation for probing visual perception. *Advances in Neural Information Processing Systems*, 33, 2020.

Wallis, T. S., Funke, C. M., Ecker, A. S., Gatys, L. A., Wichmann, F. A., and Bethge, M. Image content is more important than bouma’s law for scene metamers. *eLife*, 8:e42512, 2019.

Wallis, T. S. A., Funke, C. M., Ecker, A. S., Gatys, L. A., Wichmann, F. A., and Bethge, M. A parametric texture model based on deep convolutional features closely matches texture appearance for humans. *Journal of Vision*, 17(12), Oct 2017. doi: 10.1167/17.12.5. URL http://doi.org/10.1167/17.12.5.

Wang, P. and Cottrell, G. W. Central and peripheral vision for scene recognition: A neurocomputational modeling exploration. *Journal of vision*, 17(4):9–9, 2017.

Wang, Z. and Simoncelli, E. P. Translation insensitive image similarity in complex wavelet domain. In *Proceedings.(ICASSP’05)*. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005., volume 2, pp. ii–573. IEEE, 2005.

Wang, Z., Simoncelli, E. P., and Bovik, A. C. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Conference on Signals, Systems & Computers*, 2003, volume 2, pp. 1398–1402. Ieee, 2003.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Wu, K., Wu, E., and Kreiman, G. Learning scene gist with convolutional neural networks to improve object recognition. In *2018 52nd Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–6. IEEE, 2018.

Xue, W., Zhang, L., Mou, X., and Bovik, A. C. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE Transactions on Image Processing*, 23(2):684–695, 2013.

Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., and Jordan, M. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pp. 7472–7482. PMLR, 2019.

Zhang, L., Zhang, L., Mou, X., and Zhang, D. Fsim: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8):2378–2386, 2011.

Zhang, L., Shen, Y., and Li, H. Vsi: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 23(10):4270–4281, 2014.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. Object detectors emerge in deep scene cnns, 2014.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

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

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes] We have focused our experiments on implementing a two-stage model that has a texture-based foveation transform and compared it to a reference model (a perceptual upper bound), and two matched resource systems: one foveated with blur and another one uniformly blurred.
   (b) Did you describe the limitations of your work? [Yes] At the end of each Experiments Sub-Section we provide a mini-discussion of our work and how it fits or does not fit the literature. Mainly we provide limitations in the Discussion at the end (See Section 4).
   (c) Did you discuss any potential negative societal impacts of your work? [No] To our knowledge, there are none.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] We include only one supplementary theoretical result and proof in the Appendix.
   (b) Did you include complete proofs of all theoretical results? [Yes] See above.

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Supplementary Material (that provides access to a URL).
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] These are reported briefly in Section 3 and in more detail throughout the Appendix.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] All experiments were run with paired initial noise seeds to control for matched initial conditions derived from SGD (though the order in which the networks were exposed to images was different). All errorbars report 1 standard deviation, and these can be seen throughout Sections 3.2, 3.3, 3.4.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] These are specified in the Appendix.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We use a re-partition of the Places2 Dataset which is cited.
   (b) Did you mention the license of the assets? [No] Given that to our knowledge the Places2 dataset is widely known and free to use.
   (c) Did you include any new assets either in the supplemental material or as a URL? [No] As everything in the Supplementary Material/URL has been created/derived by us.
   (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [N/A] We did not run any experiments with humans.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] We did not run any experiments with humans, and the scene classes we used were all publicly known and non-offensive places: e.g. ocean.

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] No human subjects were used.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] No human subjects were used.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] No human subjects were used.
Figure 9: A full explanatory diagram of the Rate-Distortion Optimization Procedure inspired from both Ballé et al. (2016) and Deza et al. (2019). The goal is to find the equivalent ‘perceptual transmission rate’ for a given distortion $\sigma$ to find a matched-resource perceptual input for Foveation-Texture that is non-foveated. This optimization pipeline produces Uniform-Blur, a perceptual system that receives as input uniformly blurred images as a way to loosely mimic uniform retinal ganglion cell re-distribution in as if it were to occur in humans. We now have a proper control to evaluate how a foveated texture based model (Foveation-Texture) compares to a non-foveated model (Uniform-Blur) when restricted with the same amount of perceptual resources under the aggregate SSIM matching constraint.

A Description of All Perceptual Systems

Foveation-Texture: We adjusted the parameters of the foveation texture transform to have stronger distortions in the periphery that can consequently amplify the differences between a foveated and non-foveated system. This was done setting the rate of growth of the receptive field size (scaling factor) $s = 0.4$.

This value ($s = 0.4$) was used instead of $s = 0.5$, given that experiments of Freeman & Simoncelli (2011); Deza et al. (2019) have shown that this scaling factor yields a match with physiology but only when human observers are psychophysically tested between pairs of synthesized/rendered image metamers. Works of Wallis et al. (2017, 2019); Deza et al. (2019); Shumikhin (2020) have suggested that when comparing a non-foveated reference image to its foveated texturized version, the scaling factor is actually much smaller than 0.5 (0.24, or in some cases as small as 0.20; See Table 3). We thus selected a smaller factor of $s = 0.4$ (that is still metameric to a human observer between synthesized pairs), as smaller scaling factors significantly reduced the crowding effects. Ultimately, the selection of this value is not critical in our studies as: 1) we are interested in grossly exaggerating the distortions beyond the human metameric boundary to test if the perceptual system will learn something new or different from the highly manipulated images that use a new family of transformations; 2) we are not making any comparative measurements to human psychophysical experiments where matching such scaling factors would be critical e.g. Deza & Eckstein (2016); Eckstein et al. (2017); Geirhos et al. (2018).

Reference: We use the same image transform at the foveation stage for Reference but set the scaling factor set to $s = 0$. In this way, any potential effects of the compression/expansion operations of the image transform stage in the perceptual system is tightly upper-bounded by Reference over Foveation-Texture. Thus, the only difference after stage 1 is whether the image statistics were texturized in increasingly large pooling windows (Foveation-Texture), or not (Reference) – however note that the texturization procedure comes at a computational cost and modifies the amount of resources allocated in the image.

Indeed, the Reference system does not provide a matched-resource non-foveated control – the Reference model only provides a non-foveated upper bound that removes the effects of crowding that Foveation-Texture does have (See Theorem 1). In fact, the matched-resource control – under certain constraints (See Table 2) – that is also non-foveated is the Uniform-Blur system as described earlier in the paper, and in more detail as follows.
Figure 10: A. The full explanatory diagram of the Rate-Distortion Optimization Procedure adapted for Foveation-Blur. B. The goal is to find the equivalent ‘perceptual transmission rate’ for a given distortion $\sigma$ to find a matched-resource perceptual input for Foveation-Texture that is foveated but with adaptive Gaussian blurring, i.e. we must find the standard deviation of the Gaussian blurring kernel which is computed over a set of eccentricity rings that have been windowed with cosine functions. C. The full Rate-Distortion curves as a function of retinal eccentricity rings.

**Uniform-Blur:** Uniform-Blur provides a non-foveated resource matched control with respect to Foveation-Texture. This perceptual system is essentially computed via finding the optimal standard deviation $\sigma$ of the Gaussian filtering kernel $G_\sigma$ as shown in Figure 4. This distortion image is computed via the convolution $\ast$ of the Gaussian filter $G_\sigma$ with the image $f_0(I)$. Here, Wang et al. (2004)’s SSIM is our candidate perceptual metric as it will take into consideration the luminance, contrast and structural changes locally for the entire image and pool them together for an aggregate perceptual score (and also the rate $R$) that is upper bounded by 1 and correlated with human perceptual judgments. As SSIM operates on the luminance of the image, all validation images over which the RD curve (right) was computed were transformed to grayscale to find the optimal standard deviation ($\sigma = 3.4737$).

It is also worth emphasizing that the previous matching procedure is done over an aggregate family of images in the validation set (hence the use of the expected value $E[\circ]$ in Figure 4). This gives us a single standard deviation that will be used to filter all the images corresponding to the Uniform-Blur transform the same way.

**Foveation-Blur:** Is a foveated perceptual system that receives Rate-Distortion optimized images that have been blurred with different standard deviations of the gaussian kernel $G_\sigma$ as a function of retinal eccentricity. We picked the same eccentricity rings (collection of pooling regions that lie along the same retinal eccentricity) as Foveation-Texture given that we did not want to include a potential effect that is driven by differences in receptive field sizes rather than differences in type of computation. Figure 10 shows the full set of distortion strengths ($\sigma$) of each receptive field ring to match the perceptual transmission rate of the Foveation Texture Transform ($f_\ast(\circ)$).

There are other alternatives to potentially find the set of standard deviation coefficients that are not driven by a rate-distortion optimization procedure. One possibility could have been to find a mapping between pixels and degrees of visual angle as done in Pramod et al. (2018) and derive the coefficients by fitting a contrast sensitivity function given the visual field. While this approach is appealing, the coefficients for object recognition such as in ImageNet Russakovsky et al. (2015) can not be extended to scenes such as Places Zhou et al. (2017). In addition, the coupling of the RD-optimization with SSIM provides a perceptual guarantee to compare Foveation-Blur-Net to either Foveation-Texture or Uniform-Blur.
B Reference as a perceptual Upper Bound

\textbf{Theorem 1.} Reference is a perceptual upper bound, and its generalization performance can be matched, but can not be exceeded (due to possession of maximum image information).

\textit{Proof.} Let \(I' = D(M)\) be the decoded image to be received by the second stage \(g(\phi)\) of any perceptual system, where \(M_{\theta_i,\phi_i} = \alpha_i Q_{\theta_i,\phi_i} + (1 - \alpha_i) T_{\theta_i,\phi_i}\) is the convex combination between structure and texture for the collection of pooling regions \(i\) (Figure 2B.). It can be observed that for Reference the values of \(\alpha\) yield \(\alpha_i = 0, \forall i,\) thus any other system that has at least 1 value of \(\alpha_i \neq 0\) will render a decoded image with a non-zero distortion in pixel space, thus making the resources (amount of information) of Reference greater or equal than any other system with non-zero coefficients (e.g. Foveation-Texture).

\textbf{Remark 1.} An example of a theoretically matched generalization performance system to Reference from another non-zero distortion network is possible if the family of pre-distorted images were based on textures (also see \cite{Gatys2015} Figure 5).

\textbf{Remark 2.} The resulting transformed images from \(f_0(\phi)\) and \(f_\ast(\phi)\) are not diffeomorphic to each other.

C Full set of IQA Metrics

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|}
\hline
(mean±std) & SSIM (Matched) & MSE (↑) & Mutual Information (↓) \\
\hline
Reference & 1.0 & 0.0 & 7.39 ± 0.52 \\
\hline
Foveation-Texture & 0.58 ± 0.11 & 976.75 ± 522.22 & 1.40 ± 0.42 \\
Uniform-Blur & 0.57 ± 0.15 & 458.67 ± 277.13 & 1.86 ± 0.58 \\
Foveation-Blur & 0.58 ± 0.15 & 507.35 ± 302.71 & 1.84 ± 0.56 \\
\hline
(mean±std) & MS-SSIM (Wang et al. 2003) (↓) & CW-SSIM (Wang & Simoncelli 2005) (↓) & FSIM (Zhang et al. 2011) (↓) \\
\hline
Reference & 1.0 & 1.0 & 1.0 \\
\hline
Foveation-Texture & 0.20 ± 0.03 & 0.74 ± 0.05 & 0.76 ± 0.05 \\
Uniform-Blur & 0.36 ± 0.03 & 0.98 ± 0.01 & 0.69 ± 0.09 \\
Foveation-Blur & 0.36 ± 0.03 & 0.98 ± 0.01 & 0.67 ± 0.10 \\
\hline
(mean±std) & VSI (Zhang et al. 2011) (↓) & GMSD (Xue et al. 2013) (↑) & NLPD (Laparra et al. 2016) (↑) \\
\hline
Reference & 1.0 & 0.0 & 0.0 \\
\hline
Foveation-Texture & 0.93 ± 0.02 & 0.19 ± 0.03 & 0.75 ± 0.16 \\
Uniform-Blur & 0.91 ± 0.04 & 0.19 ± 0.03 & 0.40 ± 0.09 \\
Foveation-Blur & 0.91 ± 0.04 & 0.22 ± 0.04 & 0.45 ± 0.11 \\
\hline
(mean±std) & MAD (Larson & Chandler 2010) * (↑) & VIF (Sheikh & Bovik 2006) (↓) & LPIPSvgg (Zhang et al. 2018) * (↑) \\
\hline
Reference & 0.0 & 1.0 & 0.0 \\
\hline
Foveation-Texture & 166.77 ± 19.46 & 0.12 ± 0.03 & 0.35 ± 0.05 \\
Uniform-Blur & 182.19 ± 16.50 & 0.12 ± 0.03 & 0.52 ± 0.07 \\
Foveation-Blur & 185.90 ± 18.60 & 0.16 ± 0.03 & 0.54 ± 0.08 \\
\hline
(mean±std) & DISTS (Ding et al. 2020) * (↑) \\
\hline
Reference & 0.0 & 0.0 & 0.0 \\
Foveation-Texture & 0.20 ± 0.03 \\
Uniform-Blur & 0.36 ± 0.03 \\
Foveation-Blur & 0.35 ± 0.03 \\
\hline
\end{tabular}
\caption{List of Full IQA Metrics from \cite{Ding2020} where we compare Image Transforms \(f(\phi)\) w.r.t. Reference for the \textit{testing} set. Arrows (↑ / ↓) indicate the direction of the \textit{greatest} distortion according to the metric thus values further away from the Reference place a specific transform at a resource disadvantage. We observe matched distortion via virtual ties for SSIM (matched and optimized in the \textit{validation} set), VSI, GMSD FSIM, and VIF; greater distortion (Foveation-Texture at a disadvantage) for MSE, Mutual Information, MS-SSIM, CW-SSIM, NLPD; and lower distortion (Foveation-Texture at an advantage) for MAD, and texture-based tolerance methods such as DISTS and LPIPSvgg – hence implicitly proving that our transform does indeed preserve local texture. Scores were computed over 5000 images. Numbers in bold represent highest/lowest IQA scores; virtual ties were declared if highly overlapping standard deviations are noticeable \textit{e.g.}: FSIM, VIF.}
\end{table}
D Image Transform Samples

Figure 11 is an extension of Figure 4 which shows a collection of randomly sampled images from each one of the 20 scene classes and how they look under each image transform before being fed to each network. Details worth noticing include: 1) Reference images are not full high resolution, and are slightly compressed given the encoder/decoder pipeline of the transform to operate as a tighter upper bound (observable when zooming in); 2) The foveal area is preserved and identical for Reference, Foveation-Texture and Foveation-Blur; 3) The peripheral distortions are more or less apparent contingent on the image structure; 4) All images used in our experiments were rendered at $256 \times 256$ px.
E Differences across other Foveation models

There are currently 4 foveation models that implement texture-like computation in the peripheral field of view as shown in Table 3. We selected the Foveation Texture Transform model of Deza et al. (2019) given that it is computationally tractable to render a foveated image dataset (100'000) at a rate of 1 image/second (rather than hours Freeman & Simoncelli (2011) or minutes Wallis et al. (2017)). We did not use the highly accelerated model of Fridman et al. (2017) (order of miliseconds, that was based on the Texture-Tiling Model of Rosenholtz et al. (2012)) because it was: 1) Not psychophysically tested with human observers thus there is no guarantee of visual metamerism via the choice of texture statistics (although see the recent work of Shumikhin (2020)); 2) But most importantly, it does not provide an upper-bound computational baseline (similar to Reference).

Altogether, we think that re-running our experiments and testing them with all other foveated models such as the before-mentioned is a direction of future work as we would be curious to see the replicability of our pattern of results across other texture-based peripheral models. Naturally, the type of texture-based foveation used will also yield different matched resource systems (Uniform-Blur and Foveation-Blur), as different models rely on texture computation in different ways – and thus will affect the IQA metric scores when performing the perceptual optimization.
### Differences to other Relevant Work

There are several works that have used foveation to show a type of representational advantage over non-foveated systems. Mainly Pramod et al. (2018) with adaptive gaussian blur, and Wu et al. (2018) with scene gist, that have been targeted towards a computational goal in increasing object recognition performance. For scene recognition, only Wang & Cottrell (2017) has successfully modelled known behavioural results of Larson & Loschky (2009) via a dual-stream neural network that uses adaptive gaussian blurring and a log-polar transform. One key difference however is that we are interested in exploring the effects of peripheral texture-base computation that give rise to visual crowding and that is also linked to area V2 in the primate ventral stream – rather than retinal as in Wang & Cottrell (2017) which resembles our control condition: Foveation-Blur.

In general, we are taking a complimentary approach to Wang & Cottrell (2017) & Wu et al. (2018), and a similar one to Pramod et al. (2018) where we a priori do not know of a functional role of texture-based computation or prime ourselves to fit our model to a reference behavioural result. Thus we explore what perceptual properties it may have in comparison to a non-foveated system (Uniform-Blur, Reference) or a foveated system that only implements adaptive gaussian blurring (Foveation-Blur). Table 4 highlights key similarities & differences between these papers and ours.

| Model System Focus | Goal of Foveal-Peripheral Architecture | Spatial Frequency Preference | Weighted Bias Emerges | Robustness to Distortion Type | Critical Radius (Larson & Loschky, 2009) | Out of Distribution Generalization | Accounts for pooling regions | Accounts for visual crowding | Accounts for retinal eccentricity | Accounts for loss of visual acuity | Critical Radius (Larson & Loschky, 2009) | 8 deg | Not Applicable (Objects) | ~ 8.67 deg (Estimated from Fig. 8) |
|--------------------|---------------------------------------|-------------------------------|-----------------------|-------------------------------|----------------------------------------|-----------------------------------|-------------------------------|--------------------------|-----------------------------|-----------------|----------------|-----------------|
| Wang & Cottrell (2017) | Pipeline 1 (Ours)                      | Dual + Gating                 | Coupling the fovea + periphery | log-polar + adaptive gaussian blurring | retinal (Geisler & Perry, 1998)          | -                                | Implicit via adaptive gaussian blurring | -                       | Implicit via cropping | -              | 8 deg | -               | -               |
| Wu et al. (2018)     | Pipeline 2 (Wang & Cottrell, 2017)    | Dual + Concatenation          | Contextual modulation (scene gist) | Region Selection              | retinal (Geisler & Perry, 1998)          | -                                | Implicit via adaptive gaussian blurring | -                       | Implicit via cropping | -              | 8 deg | -               | -               |
| Pramod et al. (2018) | Pipeline 3 (Our)                       | Single                        | Serializing the (single) two-stage model | adaptive gaussian blurring | -                                | -                                | Implicit via adaptive gaussian blurring | -                       | Implicit via cropping | -              | 8 deg | -               | -               |
| (Ours)              |                                       |                               | Visual Metamer w/ texture-distortion | V2 (Freeman & Simoncelli, 2011) | -                                | -                                | Implicit via adaptive gaussian blurring | -                       | Implicit via cropping | -              | 8 deg | -               | -               |

Table 4: A summary set of Foveal-Peripheral CNN model characteristics.
G  Training, Testing and Learning Dynamics

Figure 12: Learning Dynamics visualized via the Validation Accuracy over all epochs for AlexNet and ResNet18 as $g(\circ)$. Left: A1/B1 we see the aggregate Validation Accuracy. Right: A2/B2 the individual Validation Accuracies are shown for each network.
Figure 13: Validation Loss (Cross Entropy) over all epochs for AlexNet and ResNet18 as $g(\phi)$. Left: A1/B1 we see the aggregate Validation Loss. Right: A2/B2 the individual Validation Losses for each network. It is interesting to see that despite re-bound effects in the validation loss, that the validation accuracy continues to increase (See Figure 12).
Figure 14: Training Loss (Cross Entropy) over all epochs for AlexNet and ResNet18 as $g(\circ)$. Left: A1/B1 we see the aggregate training loss. Right: A2/B2 the individual training losses for each network.

Perceptual Systems were trained with SGD, nestorov momentum, no dampening, weight decay = 0.0005, momentum = 0.9, a batch size of 128. Color Normalization of mean = (0.485, 0.456, 0.406), and std = (0.229, 0.224, 0.225). Systems that used AlexNet as $g(\circ)$ were trained for 120 epochs with a scheduled learning rate, where the initial learning rate of 0.01 was halved after the 30th epoch, and halved again after 60th epoch. Systems that used ResNet18 as $g(\circ)$ were trained for 80 epochs and with an initial learning rate of 0.05, which was multiplied by 0.25 after the first 16 epochs, and then multiplied again by 0.25 after the 32nd epoch. All systems were trained with a cross-entropy loss and received images size of $256 \times 256 \times 3$. No data-augmentation or cropping was used at training or testing.
Figure 15: Generalization Dynamics over a set of multiple epochs for AlexNet and ResNet18 as $g(o)$. 
Figure 16: A sample collection of Confusion Matrices for the first of the 10 randomly initialized networks for each of the 4 perceptual systems with their transforms for both AlexNet and ResNet18 as \( g(\circ) \). We see similar classification patterns between Foveation-Texture and the Reference, and also similar classification strategies between the Foveation-Blur and Uniform-Blur system. The asymmetry in the upper and lower off-diagonal quadrants highlight the differences between Foveation-Texture & Reference vs Foveation-Blur & Uniform-Blur. Each row/column per confusion matrix represents each of the scene classes in alphabetical order. These classes are: aquarium, badlands, bedroom, bridge, campus, corridor, forest path, highway, hospital, industrial area, japanese garden, kitchen, mansion, mountain, ocean, office, restaurant, skyscraper, train interior, waterfall.
Figure 17: Sample image used in a full i.i.d and o.o.d evaluation.
Figure 18: Evolution of AlexNet as $g(\cdot)$ Conv-1 Filters from 1st Random Weight Initialization.
Figure 19: Evolution of ResNet18 as $g(\circ)$ Conv-1 Filters from 1st Random Weight Initialization.
The size of all shown images was $256 \times 256 \times 3$, thus the units of the gaussian filters specified from Section 3.2 are in pixels. For a given Gaussian filtering operation $G_\sigma$ for a given standard deviation $\sigma$, low pass spatial frequency (LF) images were computed via:

$$LF(I^C) = G_\sigma \star I^C$$

for each channel $C$. Similarly, High Pass Spatial Frequency (HF) image stimuli were computed via:

$$HF(I^C) = I^C - G_\sigma \star I^C + \text{mean}^{C}_{\text{val}}$$

where $\text{mean}^{C}_{\text{val}}$ (which we call the residual in the main body of the paper) is the average of image intensity over the held-out validation set for each channel $C$, a small extension from Geirhos et al. (2019) as our image stimuli is in both color and grayscale.
Figure 20: Aggregate Spatial Frequency Sensitivity for AlexNet as $g(\omega)$ after epochs 0, 1, 5, 30, 70, 120.
Figure 21: Individual Spatial Frequency Sensitivity for AlexNet as $g(\omega)$ after epochs 0, 1, 5, 30, 70, 120.
ResNet18 as $g(\sigma)$

Spatial Frequency Sensitivity

Figure 22: Aggregate Spatial Frequency Sensitivity for ResNet18 as $g(\sigma)$ after epochs 0, 1, 4, 20, 60, 80.
Figure 23: Individual Spatial Frequency Sensitivity for ResNet18 as $g(\circ)$ after epochs 0, 1, 4, 20, 60, 80.
High Pass Spatial Frequency (Color)

Figure 24: Sample High Pass Spatial Frequency Color Stimuli.
### High Pass Spatial Frequency (Gray)

| Reference | Foveation-Texture | Uniform-Blur | Foveation-Blur |
|-----------|-------------------|--------------|---------------|

**Figure 25:** Sample High Pass Spatial Frequency Gray Stimuli.
Low Pass Spatial Frequency (Color)

Figure 26: Sample Low Pass Spatial Frequency Color Stimuli.
Low Pass Spatial Frequency (Gray)

Figure 27: Sample Low Pass Spatial Frequency Gray Stimuli.
J Robustness to Occlusion

Figure 28: Left2Right Occlusion Sample Stimuli.
Figure 29: Left2Right Occlusion Sample Stimuli.
Figure 30: Top2Bottom Occlusion Sample Stimuli.
Figure 31: Top2Bottom Occlusion Sample Stimuli.
Figure 32: Scotoma Occlusion Sample Stimuli.
Figure 33: Scotoma Occlusion Sample Stimuli.
Figure 34: Glaucoma Occlusion Sample Stimuli.
Figure 35: Glaucoma Occlusion Sample Stimuli.
Figure 36: Individual Robustness to Occlusion plots for AlexNet as $g(\cdot)$ after epochs 0, 1, 5, 30, 70, 120.
Figure 37: Aggregate Robustness to Occlusion plot for AlexNet as $g(\sigma)$ after epochs 0, 1, 5, 30, 70, 120.
Figure 38: Individual Robustness to Occlusion plots for ResNet18 as $g(x)$ after epochs 0, 1, 4, 20, 60, 80.
Figure 39: Aggregate Robustness to Occlusion plot for ResNet18 as \(g(\circ)\) after epochs 0, 1, 4, 20, 60, 80.
Figure 40: Aggregate and Individual Window Cue-Conflict plots for AlexNet and ResNet18 as \(g(\circ)\) after epochs 0, 1, 5, 30, 70, 120 and 0, 1, 4, 20, 60, 80. respectively
Figure 41: Sample Window Cue Conflict Stimuli.
**L Square Cue-Conflict**

**Square Cue-Conflict**

![Plot Diagram](image)

Figure 42: Aggregate and Individual Square Cue-Conflict plots for AlexNet and ResNet18 as $g(\circ)$ after epochs 0, 1, 5, 30, 70, 120 and 0, 1, 4, 20, 60, 80 respectively
Square Cue Conflict

Figure 43: Sample Square Cue Conflict Stimuli.
Square Cue Conflict

Figure 44: Sample Square Cue Conflict Stimuli.
Differences from Previous Manuscript Versions

[Added; this submission] Improved training and convergence of stage 2 neural networks. AlexNet + ResNet18 now have scheduled learning rates, weight decay and Nesterov momentum when trained with SGD for each image distribution.

[Added; this submission] High Pass and Low Pass Spatial Frequency experiments for grayscale stimuli as suggested in round of review from ICML 2021.

[Added; this submission] Square Uniform cue-conflict experiment to re-verify center image bias as suggested in round of review from ICML 2021.

[Added; this submission] Left2Right & Top2Bottom experiments moved to main body.

[Added; this submission] both Aggregate and Individual plots for each system to qualitatively check for variance in individual network differences.

[Added; this submission] Visualization of filters from the first convolutional layer for each system.

[Added; this submission] Additional use of Mean Square Error, Mutual Information and 10 more IQA metrics from Ding et al. (2020) as supporting Image Quality Assessment metrics to compare to SSIM for Rate-Distortion Optimization as suggested through reviews in ICML 2021.

[Added; for ICML 2021] Sketched proof of Reference being a Perceptual Upper Bound.

[Added; for ICLR 2021] Rate-Distortion Optimization procedure to compute Uniform-Blur and Foveation-Blur.

[Added; for ICLR 2021] Improved written clarity, and re-emphasized focus of paper on Foveation w.r.t Machines (not humans – which caused misinterpretation and rejection from NeurIPS 2020).

[Removed; for ICML 2021] Claims about Foveation-Texture inducing a shape bias (currently parallel work) from Submission to ICLR 2021.

[Removed; for ICLR 2021] Experiments about data-augmentation via eye-movements + classical augmentation schemes such as random cropping + rescaling (parallel work) from Submission to NeurIPS 2020.

[Bug fix; this submission] Even runs were continuations of odd runs in 10 run randomization across networks due to bug w.r.t distributed parallelization, from submission to ICML 2021. Note: General pattern of results did not change, and all curves have been re-plotted.

A recurrent theme in negative reviews has been that the model does not (in its current state) advance the state of the art by beating a baseline. While these hallmarks are pivotal for computer vision, our goal is complimentary, as we would like to model, and understand the representational consequences – beyond accuracy – of spatially-adaptive computation in machines inspired by the foveated visual system of humans.