EMERGENT PROPERTIES OF FOVEATED PERCEPTUAL SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce foveated perceptual systems – a hybrid architecture inspired by human vision, to explore the role of a *texture-based* foveation stage on the nature and robustness of subsequently learned visual representation *in machines*. Specifically, these two-stage perceptual systems first foveate an image, inducing a texture-like encoding of peripheral information – mimicking the effects of *visual crowding* – which is then relayed through a convolutional neural network (CNN) and trained to perform scene categorization. We find that these foveated perceptual systems learn a visual representation that is *distinct* from their non-foveated counterpart through experiments that probe: 1) i.i.d and o.o.d generalization; 2) robustness to occlusion; 3) a center image bias; and 4) high spatial frequency sensitivity. In addition, when using a perceptually driven rate-distortion optimization procedure to compute matched resource systems, we find similarities in pattern of results (with weaker magnitudes) between non-foveated systems and lower-level foveated systems that only implement retinal + V1-like computation. However, when comparing across all matched-resource systems (visual crowding, adaptive gaussian blur and no foveation), several perceptual benefits surge *exclusively* when the crowding operator is *on*. Altogether, these results not only suggest foveation – via visual crowding – as a new inductive bias for machine perception, but also potentially suggest *new symbiotic* computational support for the idea that the foveated nature of the human visual system, mainly peripheral computation that is driven by texture-based encoding linked to area V2, confers a functional advantage for scene perception.

1 INTRODUCTION

In the human visual system, incoming light is sampled with different resolution across the retinal area, 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 ganglion cells that can relay information from the retina to the LGN constrained by the flexibility and 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 even necessary for vision at all.

However, it is also possible that foveation plays a functional role at the *representational level*, which can confer perceptual advantages as has been explored *in humans*. 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; Cheung et al., 2017; Han et al., 2020). Classic symbiotic examples of showing advantages of foveation in humans via showing them in machines 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, and Wu et al. (2018)'s GistNet that directly introduced a dual-stream foveal-peripheral pathway in a neural network to boost object detection performance via scene gist and contextual cueing. Complimentary, 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

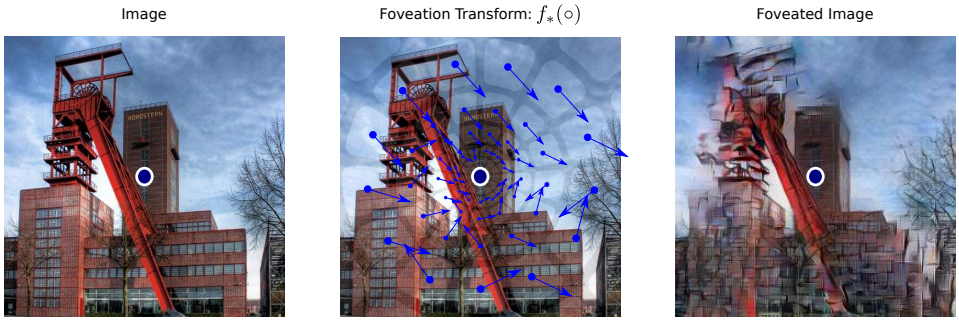

Figure 1: A cartoon illustrating how a foveated image is rendered producing a human visual *metamer* via the foveated feed-forward style transfer model of Deza et al. (2019). Here, each receptive field is locally perturbed with noise in its latent space in the direction of their equivalent texture representation (blue arrows) resulting in *visual crowding* effects in the periphery. These effects are most noticeable far away from the navy dot which is the simulated center of gaze (foveal region) of an observer.

that modelled the known results of Larson & Loschky (2009) with a log-polar transform and adaptive gaussian blurring (foveation stage V1). Indeed, the previous works present support for the *functional* hypothesis of a foveated visual system despite *not introducing texture priors* in the periphery – a key property of peripheral computation as posed in Rosenholtz (2016). Testing whether this texture-based coding of the visual periphery is *functional* in any perceptual system is still an open question.

To address the previous question, we take a higher-level and hybrid representational approach, by directly examining the impact of these *texture-like* computations in the periphery (foveation stage V2), and whether they convey a *functional* perceptual advantage in both humans and machines. To accomplish this, we will introduce *foveated perceptual systems*: these are two-stage systems that have a texture-based foveation stage followed by a deep convolutional neural network. In particular, we will mimic foveation over images using a transform that simulates *visual crowding* (Levi, 2011; Pelli, 2008; Doerig et al., 2019b;a) in the periphery as shown in Figure 1 (Deza et al., 2019) rather than gaussian blurring (Pramod et al., 2018; Wang & Cottrell, 2017) or compression (Patney et al., 2016; Kaplanyan et al., 2019). These rendered images are a type of *visual metamers* that capture image statistics known to be preserved in human peripheral vision that resemble texture computation at the stage of V2 as argued in Freeman & Simoncelli (2011); Rosenholtz (2016); Wallis et al. (2019).

Thus, our strategy in this paper will be to compare these hybrid models' perceptual biases to their non-foveated counterpart through a set of experiments: generalization, robustness to occlusion, image-region bias and spatial frequency sensitivity. A difference from Wang & Cottrell (2017) is that our goal is *not* to implement foveation at the stage of V1 (adaptive gaussian blurring) to fit known results to data (Larson & Loschky, 2009);but rather to *naively* explore the emergent properties of foveation at the stage of V2 *(texture-based)* – an insightful experiment that has never been done before that can potentially yield null results. The hope is that if we find any benefits or relevant differences between these systems, then this could shed light into both the importance of texture-based peripheral computation in humans, and also suggest a new inductive bias for advanced machine perception.

2 Foveated Perceptual Systems

We define perceptual systems as *two-stage* with a foveation transform (stage 1, $f(\circ) : \mathbb{R}^D \rightarrow \mathbb{R}^D$), that is relayed to a deep convolutional neural network (stage 2, $g(\circ) : \mathbb{R}^D \rightarrow \mathbb{R}^d$). 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:

$$S(I) = g(f(I)) \tag{1}$$

Such two stage models have been growing in popularity, and the reasons these models (including our own) 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 if $f(\circ)$ has a representational goal[1]. For

[1]This is a counter-intuitive approach from computer vision systems that are usually end-to-end differentiable.

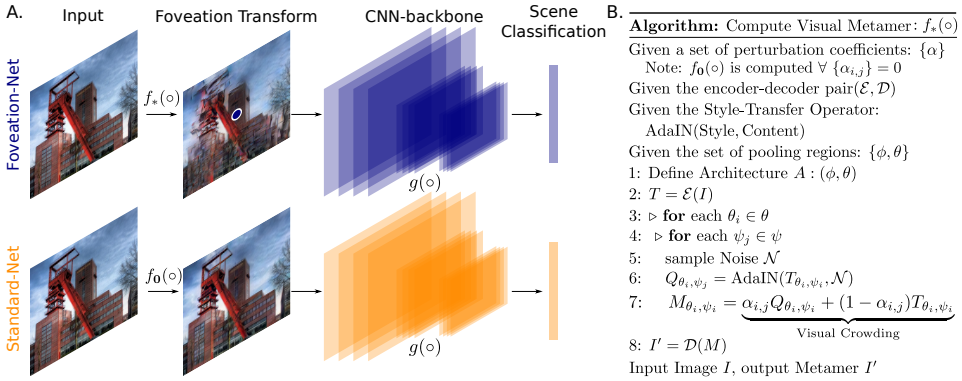

Figure 2: A. Two key Perceptual Systems S: Foveation-Net (top row) and Standard-Net (bottom row), where each system receives an image as an input, applies a foveation transform $(f(\circ))$, which is then relayed to a CNN architecture $(g(\circ))$ for scene classification. Foveation-Net uses a visual crowding model with texture computation (shown on right), while Standard-Nets provides a baseline for a computational upper-bound for all perceptual systems. B. The algorithm of how a foveated image is generated that enables visual crowding (mainly steps 5-7)

.

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 at the first stage to check if the perceptual system through $g(\circ)$ will learn to pick-up on this newly made representation and make 'good' use of it potentially shedding light on the *functionality* hypothesis for machines and humans. This leads us to further motivate our choice for a single-stream foveated-peripheral pathway (vs dual-stream Wang & Cottrell (2017); Wu et al. (2018)), as would like to focus on how $g(\circ)$ will differ purely based on the *learning dynamics* driven by SGD rather than architectural constraints (Deza et al., 2020) for the newly transformed input space (foveated : f_* vs non-foveated : $f_{\mathbf{0}}$) to perform a 20-way scene categorization task.

2.1 Stage 1: Foveation Transform (f)

To model the computations of a foveated visual system, we employed the recently proposed Metamer model of Deza et al. (2019) (henceforth *Foveation Transform*). This model is inspired by the metamer synthesis model of Freeman & Simoncelli (2011), where image metamers are constructed by locally matching texture statistics (Portilla & Simoncelli, 2000; Balas et al., 2009) in the visual periphery. Analogously, the Deza et al. (2019) Foveation Transform uses a foveated feed-forward style transfer (Huang & Belongie, 2017) network to latently perturb the image in the direction of its locally matched texture (see Figure 1). When calibrating the distortions correctly, the resulting procedure leads to a visual metamer (for a human – now forced onto the machine), which is a carefully perturbed image perceptually indistinguishable from its reference image (Feather et al., 2019).

In summary, $f : \mathbb{R}^D \rightarrow \mathbb{R}^D$ is nothing more than a convolutional auto-encoder that is non-foveated when the latent space is un-perturbed: $f_{\mathbf{0}}(I) = \mathcal{D}(\mathcal{E}(I))$, but foveated $(\circ_\Sigma)$ when the latent space is perturbed via localized style transfer: $f_*(I) = \mathcal{D}(\mathcal{E}_\Sigma(I))$, for a given encoder-decoder $(\mathcal{E}, \mathcal{D})$ pair.

2.2 Stage 2: Convolutional Neural Network backbone (g)

The foveated images (stage 1) are passed into a standard 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 architectures is to let us focus on consequences of foveation that are independent of network architecture choices. 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 proxy of higher stages of visual processing (as it receives input from f), analogous to the 2-stage model of Lindsey et al. (2019).

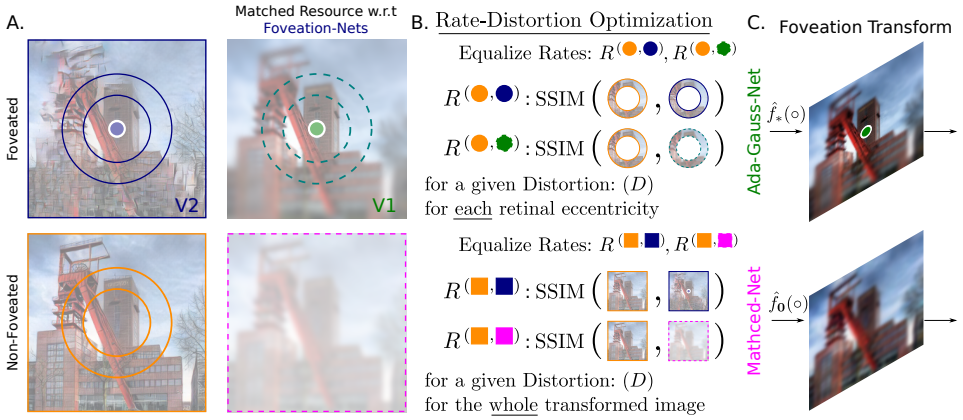

Figure 3: A. Two matched-resource controls to Foveation-Net are introduced. Top, green: adaptive gaussian blurring (Ada-Gauss-Net) emulating V1-like foveation; Bottom, magenta: uniform blurring emulating a matched-resource non-foveated visual system (Matched-Net). B. A Rate-Distortion Optimization procedure is summarized where we compute the matched-resource images that will define the new foveation transforms: $\{\hat{f}_*, \hat{f}_0\}$. C. The foveation transforms of both perceptual systems Ada-Gauss-Net (top) and Matched-Net (bottom) do *not* model visual crowding.

2.3 Critical Manipulations: Foveated vs Non-Foveated Perceptual Systems

Now, we can define two perceptual systems as the focus of our experiments that must perform 20-way scene categorization: Foveation-Net, receives an image input, applies the foveated transform $f_*(\circ)$, and later relays it through the CNN $g(\circ)$; and Standard-Net, that performs a non-foveated transform $f_0(\circ)$ that does not impose any effects of visual crowding – recall that the image is still sent through the convolutional auto-encoder $\mathcal{D}(\mathcal{E}(I))$. Both of these systems are depicted in Figure 2(A.).

Anticipating the potential differences we may find between Foveation-Nets and Standard-Nets, lies the question: Will these differences be driven by *any* type of spatially-varying processing, or the *specific* type of foveation transformation in itself (texture-based coding at stage V2 vs adaptive gaussian blurring at stage V1). Thus, we created a matched-resource V1-foveation transform using a Rate-Distortion (RD) optimization framework inspired from Ballé et al. (2016) by estimating the matched distortion image for a given communication rate between the two types of foveation (V1 vs V2) as shown in Figure 3. Here, R is the 'transmission rate' computed via the expected SSIM (Wang et al., 2004) between the compressed reference image and the image metamers that are used as input for Foveation-Net, and D is the 'distortion strength' which is the std (σ) of the Gaussian used to filter the image via $(\mathcal{G}_\sigma)$ modelling V1-computation per each retinal log-polar eccentricity ring.

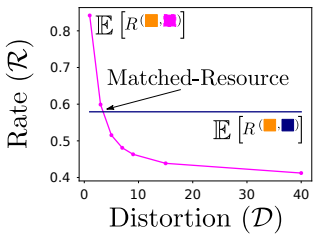

Figure 4: RD-Optimization.

Perceptual Systems trained on these images (which define a new function $\hat{f}_*$), provide a matched-resource control to Foveation-Net that we dub Ada-Gauss-Net. This same optimization procedure can be extended to the entire image to also develop a matched-resource control for a non-foveated visual system (Matched-Net, $\hat{f}_0$: Figure 4), as one could argue that Standard-Nets will be at an advantage to Foveation-Nets since their foveal regions are identical and the retinal ganglion cells have not been properly re-distributed (Cheung et al., 2017) & Figure5(A.).

Altogether these 4 perceptual systems help us answer 2 key questions: 1) Foveation-Net (V2) vs Ada-Gauss-Net (V1) 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 crowding mechanism and texture computation vs adaptive gaussian blurring; 2) Foveation-Net (V2) vs Matched-Net will tell us how do these perceptual systems (one foveated, and the other one not) behave when allocated with a fixed number of resources. Naturally this line of thought can also aid us in comparing Ada-Gauss-Net vs Matched-Net; Ultimately, Standard-Net provides us a computational upper-bound and a reference for all model systems as it is not resource-matched and non-foveated. These 2 questions have the general purpose of asking if visual crowding serve a *representational goal*? – an open question in vision.

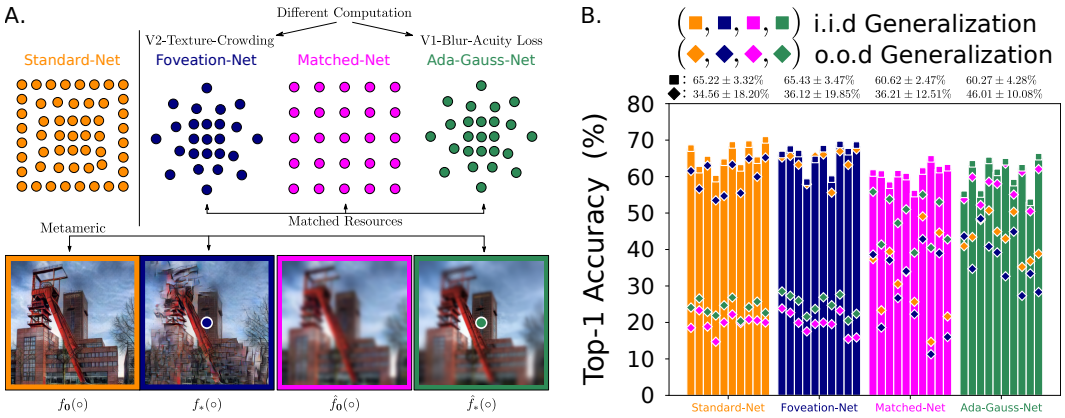

Figure 5: A. A legend illustrating the 4 key perceptual systems in all our experiments with their differences and similarities: Standard-Net, Foveation-Net, Matched-Net and Ada-Gauss-Net. B. Scene Categorization Accuracy: Foveation-Net has the highest i.i.d generalization while Ada-Gauss-Net has the greatest o.o.d generalization – both perceptual systems are *foveated* (spatially-varying).

3 EXPERIMENTS

In the following section we will discuss results found with regards to i.i.d. & o.o.d. generalization, robustness to occlusion, central image bias and spatial frequency sensitivity. All results shown in this section use AlexNet as $g(\circ)$ for the second stage of each perceptual system. *Furthermore we show the same pattern of results for ResNet18 as $g(\circ)$ independent of SGD-induced learning dynamics*. These extended results and all statistical testing can be accessed in the Supplementary Material.

3.1 GENERALIZATION

We first examined the generazation capacity for our 4 network types. Given that Foveation-Nets receive distorted inputs in the periphery one could expect them to do worse compared to Standard-Nets (albeit Standard-Net possessing more perceptual resources) and worse or on par with Matched-Nets that is also non-foveated like Standard-Nets but with perceptually matched resources. On the other hand, it is possible that the texture-based image encoding could confer a functional advantage, in which case Foveation-Nets would do better. If this is indeed the case, we would also see an advantage of Foveation-Nets over Ada-Gauss-Net; but if their performance is the same, we can conclude that *any* type of spatially-varying processing operator can enhance scene categorization performance. Naturally, a somewhat contrived but interesting experiment that we will also conduct is the out-of-distribution (o.o.d) case in which each perceptual system is shown stimuli type that they never encountered during training – in particular stimuli that is the input to the *other* perceptual systems.

Result 1: Figure 5(B.) shows that a foveated system with texture-based computation in the periphery (Foveation-Net) *has greater i.i.d generalization* than its matched resource non-foveated system (Matched-Net) and a foveated system with adaptive gaussian blurring (Ada-Gauss-Net); and also performs on par with an un-matched resource non-foveated system (Standard-Net) – an unexpected victory for Foveation-Nets. For an o.o.d generalization task where systems are shown o.o.d images of other, we find that *drops in performance for foveated perceptual systems are lower compared to their non-foveated counter-part* (Foveation-Net vs Standard-Net; Ada-Gauss-Net vs Matched-Net); and that overall Ada-Gauss-Nets *has greater o.o.d generalization compared to all other systems*, while Foveation-Net struggles to cross-generalize to low pass spatially transformed images compared to Ada-Gauss-Net – a result that ties to the experiments of Section 3.4.

3.2 ROBUSTNESS TO OCCLUSION

We next examined how all perceptual systems could classify scene information under conditions of visual field loss, either from the center part of the image (scotoma), or the periphery (glaucoma). This manipulation lets us examine the degree to which the representations are relying on central *vs* peripheral information to classify scene categories. For the foveal-occlusion conditions, we superimposed a central gray square on each image from the testing image set, with 8 levels of

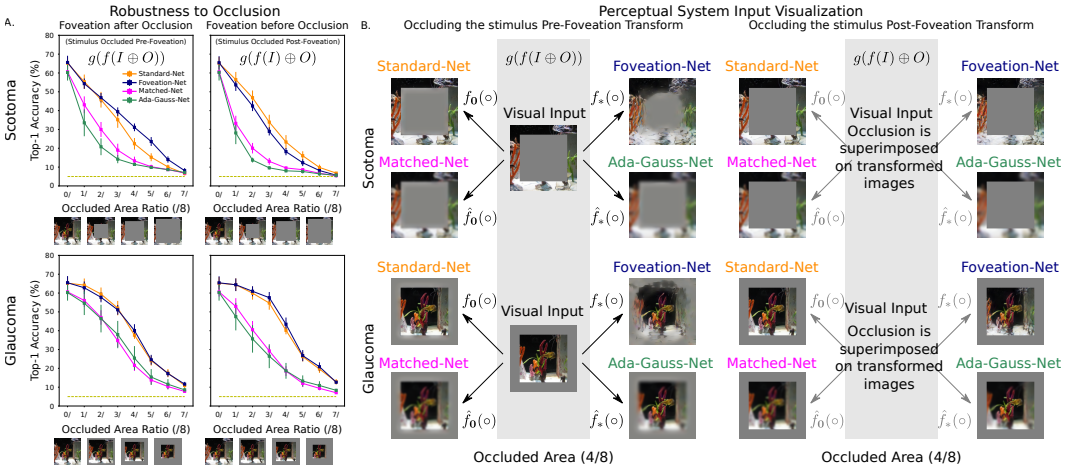

Figure 6: A. The Robustness curves for image occlusion $(\circ \oplus O)$ when the Foveation Transform is On (as the stimuli is shown before the transform), and when the Foveation Transform is Off (stimuli shown post-transform). Foveation-Nets performs has greater robustness to occlusion across all matched-resource systems, and also bumps it's performance due to *visual crowding*. B. A visualization of how the visual input is transformed for each perceptual system given their respective Foveation transform (f). Left: The effects of crowding are noticeable for Foveation-Nets (but not Ada-Gauss-Nets) as it increases visual area. Right: Even without the aid of visual crowding, the learned texture representation yields greater robustness with matched occluded areas for Foveation-Nets.

increasing size. For the peripheral-occlusion condition, we superimposed a gray box over the outter image boundaries, with area-matched levels of occlusion. These images were then passed through all of the trained models to compute scene categorization performance. Accuracy was measured at each level of occlusion and area under the curve was computed as an index of robustness to occlusion.

Figure 6 (top) shows a summary of these results. Overall Foveation-Nets showed more robustness to central/scotoma occlusion to all other systems: un-matched non-foveated (Standard-Nets), matched-resource foveated system with adaptive gaussian blurring (Ada-Gauss-Net), and matched-resource non-foveated systems (Matched-Net). This result is similar to peripheral/glaucoma occlusion with the exception of Foveation-Nets being on par with non-resource matched Standard-Net – which is still quite impressive. To further explore these results, we visualized what the image inputs look like after the foveation stage, across all systems. Figure 6 (B. left) reveals critical information:

Result 2a: The effect of the stage 1 crowding operation $(f_*(\circ))$ of Foveation-Net *exclusively reduces* the occluded area in the periphery – a stark contrast compared to all other systems. By filling in the image with texture statistics due to the foveation stage (f_*) – Foveation-Net (but not Ada-Gauss-Net) are able to boost classification performance by using greater visual information and/or the peripheral texturized representations that Standard-Nets & Matched-Nets do not have access to.

But is this the main factor accounting for the higher robustness of the foveated systems? In other words: is it driven by greater computable visual area that has been filled in from the foveation stage; Or is the benefit driven by the learned texturized representations in the subsequent deep convolutional neural network? To examine this question, we occluded either the center or periphery of the images, but *after* the stage 1 computation (see Figure 6 (A., right)). In this way, the area of occlusion was matched for the second stage classification task.

Result 2b: Across the matched-resource systems and matched-occluded area condition: Foveation-Net has greater robustness to occlusion than Matched-Net and Ada-Gauss-Net. This puzzling result for both scotomal and glaucomal occlusion, not only suggests that the learned representations in Foveation-Net is *different* than Ada-Gauss-Net, but that Foveation-Net has learned a locally distributed representation vs a *globally distributed* representation (Ada-Gauss-Net). In addition as Foveation-Nets and Ada-Gauss-Nets only differ in peripheral computation, this suggests that a texture-based encoding mechanism provides a representational advantage than an adaptive blurred representation when information is unavailable, occluded or removed.

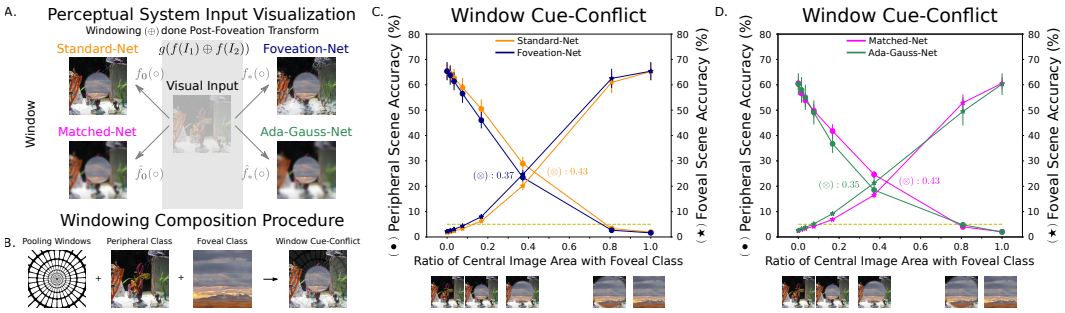

Figure 7: A. A visualization of the stimuli type that is shown to each perceptual system consists of a smoothed composition of a foveal image (*e.g.* badlands) mixed with a peripheral image (*e.g.* aquarium). B. The composition procedure relies on using the same log-polar pooling windows over which the foveated images were rendered. C. Foveated Perceptual Systems – independent of V1/V2-type computation (Ada-Gauss-Net, Foveation-Net) – show stronger biases to classify scenes with the foveal region; a result also observed in humans (Larson & Loschky, 2009).

3.3 Window Cue-Conflict Experiment

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) – and due to the fact that the information in the periphery is texturized/blurred. To assess such difference, 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 cue-conflicting images were paired with a different class (*ex:* aquarium with badlands). We then systematically varied the fovea-periphery visual area ratio & examined classification accuracy for both the foveal and peripheral scenes (Figure 7).

Result 3: We found that Foveation-Nets and Ada-Gauss-Nets maintained higher accuracy for the foveal scene class than do Standard-Nets and Matched-Nets, as the amount of competing peripheral information increased – and vice versa for the peripheral scene class. 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. Thus, Foveation-Nets have learned to weigh information in the center of the image more when categorizing scenes – a similar finding to Wang & Cottrell (2017) indeed as Ada-Gauss-Net vs Matched-Net shows the exact same pattern of results *even under the weight-sharing constraint in the second stage CNN* $(g(\circ))$. In addition, the *foveation mechanism* (Foveation-Nets or Ada-Gauss-Nets) itself is mainly responsible for the development of a center image bias. 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 (Geirhos et al., 2020).

3.4 Spatial Frequency Sensitivity

We next examined differences of learned feature representations that are more reliant on low or high spatial frequency information, at the second stage of visual processing (post-foveation) across all systems. 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 8(A.). Low pass frequency stimuli $(\mathcal{F}_L)$ 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 $(\mathcal{F}_H)$ was computed by subtracting the reference image from its low pass filtered version with $\sigma = [\infty, 3, 1.5, 1, 0.7, 0.55, 0.45, 0.4]$ pixels. These are the same values used in the experiments of Geirhos et al. (2019).

Result 4: We found that Foveation-Nets and Standard-Nets were more sensitive to High Pass Frequency information, while Ada-Gauss-Nets and Matched-Nets were sensitive to Low Pass Frequency stimuli. This suggests that: 1) foveation via adaptive gaussian blurring (Ada-Gauss-Nets) may implicitly contribute to scale-invariance as also shown in Poggio et al. (2014); Cheung et al. (2017); Han et al. (2020); 2) Foveation-Nets' peripheral texture representation – linked to visual crowding – can potentially enforce a biologically-plausible mechanism of a shape bias as argued in Figure 8.

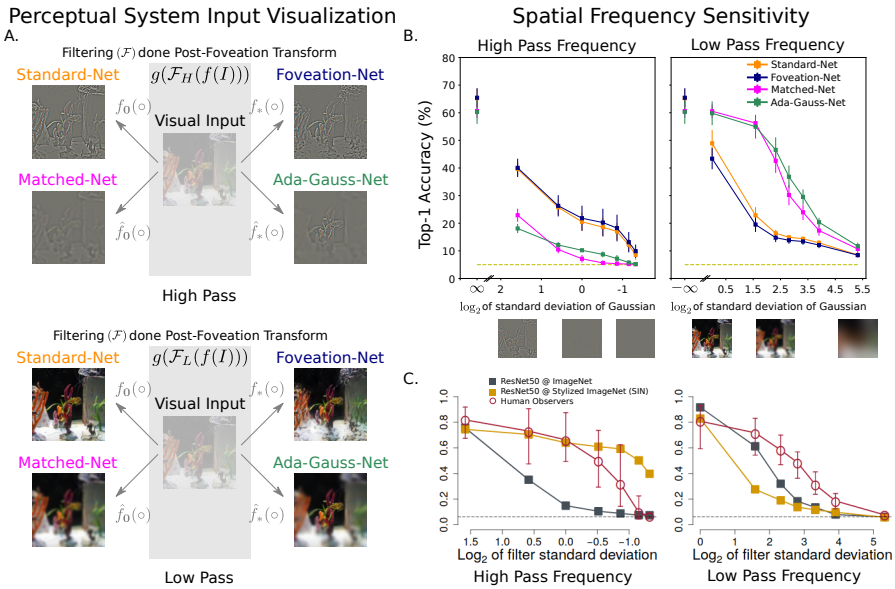

Figure 8: A. Sample images from the foveated and non-foveated images and how they change as a function of spatial frequency filtering at the post-foveation stage. B. Foveation-Nets have nominally greater sensitivity to high spatial frequency filtered stimuli than Standard-Nets, and both of these systems present notably higher sensitivity to high spatial frequencies than Matched-Nets and Ada-Gauss-Nets. Conversely, this pattern is reversed for low pass frequency stimuli. C. This suggests Foveation-Nets's crowding mechanism may naturally enforce a *shape-bias* given it's high pass spatially frequency tolerance as these Spatial Frequency curves show similar trends as Geirhos et al. (2019)'s Stylized ImageNet (SIN) that were trained on objects (adapted and redrawn).

4 Discussion

The present work was motivated by a broader question about the functional role for peripheral texture statistics for scene representation. We wondered if Foveation-Nets showed a type of functional advantage to Standard-Nets in set of perceptual tasks. Surprisingly, this was not the pattern as they (for better) perfomed on par – however, when comparing Foveation-Nets to their matched-resource models that differed in computation: Ada-Gauss-Nets and Matched-Nets – *we do find that peripheral texture encoding does provide a computational advantage* with reference to adaptive gaussian blurring or no spatially-varying computation. We also found that foveation (in general) induces a *focusing mechanism*, servicing the foveal/central region, and that crowding-based computation still preserves *high spatial-frequency selectivity* – this last result likely due to weight-sharing that does not give room to low-spatial frequency tuned filters to naturally emerge as in Wang & Cottrell (2017).

The particular consequences of our foveation stage raises interesting future directions about what will happen when trained on objects, which 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)). Specifically, one intriguing possibility is that our *foveated* representational signatures may induce more shape sensitivity for object recognition (rather than texture sensitivity; Geirhos et al. (2019); Hermann et al. (2020)), and may amplify the perceptual differences we identified across foveated and non-foveated systems.

Finally, a future direction is investigating the effects of crowding 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 matching the Nyquist sampling frequency of these filters (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 can we futher push for robustness in hybrid perceptual systems like these with even *more* biological constraints. Both Luo et al. (2015) and Reddy et al. (2020) have already taken steps in this direction by coupling fixations with a spatially-varying retina, yet investigating the computational role of visual crowding through texture-based computation via it's impact to adversarial robustness still remains an open question.

REFERENCES

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

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.

Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimized image compression. *arXiv preprint arXiv:1611.01704*, 2016.

Muriel Boucart, Christine Moroni, Miguel Thibaut, Sebastien Szaffarczyk, and Michelle Greene. Scene categorization at large visual eccentricities. *Vision Research*, 86:35–42, 2013.

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

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 113–123, 2019.

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.

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

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

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

Arturo Deza, Qianli Liao, Andrzej Banburski, and Tomaso Poggio. Hierarchically local tasks and deep convolutional networks. *CBMM Memo*, 2020.

A. Doerig, A. Bornet, O. H. Choung, and M. H. Herzog. 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`.

Adrien Doerig, Alban Bornet, Ruth Rosenholtz, Gregory Francis, Aaron M Clarke, and Michael H Herzog. Beyond bouma's window: How to explain global aspects of crowding? *PLoS computational biology*, 15(5):e1006580, 2019b.

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

Jenelle Feather, Alex Durango, Ray Gonzalez, and Josh McDermott. Metamers of neural networks reveal divergence from human perceptual systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 32*, pp. 10078–10089. Curran Associates, Inc., 2019. URL `http://papers.nips.cc/paper/9198-metamers-of-neural-networks-reveal-divergence-from-human-perceptual-systems.pdf`.

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

Lex Fridman, Benedikt Jenik, Shaiyan Keshvari, Bryan Reimer, Christoph Zetzsche, and Ruth Rosenholtz. Sideeye: A generative neural network based simulator of human peripheral vision. *arXiv preprint arXiv:1706.04568*, 2017.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bygh9j09KX.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *arXiv preprint arXiv:2004.07780*, 2020.

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.

Ethan Harris, Mahesan Niranjan, and Jonathan Hare. Foveated convolutions: Improving spatial transformer networks by modelling the retina. *1st Workshop of Shared Visual Representations in Human and Machine Intelligence (SVRHM) at NeurIPS*, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Katherine L. Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks. *Neural Information Processing Systems*, 2020.

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

Anton S Kaplanyan, Anton Sochenov, Thomas Leimkühler, Mikhail Okunev, Todd Goodall, and Gizem Rufo. 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.

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

Michael F Land and Dan-Eric Nilsson. *Animal eyes*. Oxford University Press, 2012.

Adam M Larson and Lester C Loschky. The contributions of central versus peripheral vision to scene gist recognition. *Journal of Vision*, 9(10):6–6, 2009.

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

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

Yonggang Li, Guosheng Hu, Yongtao Wang, Timothy Hospedales, Neil M Robertson, and Yongxing Yang. Dada: Differentiable automatic data augmentation. *arXiv preprint arXiv:2003.03780*, 2020.

Jack Lindsey, Samuel A Ocko, Surya Ganguli, and Stephane Deny. A unified theory of early visual representations from retina to cortex through anatomically constrained deep cnns. *arXiv preprint arXiv:1901.00945*, 2019.

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

Yan Luo, Xavier Boix, Gemma Roig, Tomaso Poggio, and Qi Zhao. Foveation-based mechanisms alleviate adversarial examples. *arXiv preprint arXiv:1511.06292*, 2015.

Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.

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

Anjul Patney, Marco Salvi, Joohwan Kim, Anton Kaplanyan, Chris Wyman, Nir Benty, David Luebke, and Aaron Lefohn. Towards foveated rendering for gaze-tracked virtual reality. *ACM Transactions on Graphics (TOG)*, 35(6):179, 2016.

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

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.

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

Manish V Reddy, Andrzej Banburski, Nishka Pant, and Tomaso Poggio. Biologically inspired mechanisms for adversarial robustness. *arXiv preprint arXiv:2006.16427*, 2020.

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

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

Thomas Serre, Lior Wolf, Stanley Bileschi, Maximilian Riesenhuber, and Tomaso Poggio. Robust object recognition with cortex-like mechanisms. *IEEE transactions on pattern analysis and machine intelligence*, 29(3):411–426, 2007.

Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019.

Timothy Tadros, Nicholas C Cullen, Michelle R Greene, and Emily A Cooper. Assessing neural network scene classification from degraded images. *ACM Transactions on Applied Perception (TAP)*, 16(4):1–20, 2019.

T. S. A. Wallis, C. M. Funke, A. S. Ecker, L. A. Gatys, F. A. Wichmann, and M. Bethge. 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`.

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.

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

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

Kevin Wu, Eric Wu, and Gabriel Kreiman. 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.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns, 2014.

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

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.

Barret Zoph, Ekin D Cubuk, Golnaz Ghiasi, Tsung-Yi Lin, Jonathon Shlens, and Quoc V Le. Learning data augmentation strategies for object detection. *arXiv preprint arXiv:1906.11172*, 2019.