Emergent Properties of Foveated Perceptual Systems

Arturo Deza[1,2]
Center for Brains, Minds and Machines
Massachusetts Institute of Technology[1]
deza@mit.edu

Talia Konkle[2]
Department of Psychology
Harvard University[2]
talia_konkle@harvard.edu

Abstract

We introduce foveated perceptual systems, inspired by human biological systems, and examine the impact that this foveation stage has on the nature and robustness of subsequently learned visual representation. Specifically, these *two-stage* perceptual systems first foveate an image, inducing a texture-like encoding of peripheral information, which is then inputted to a convolutional neural network (CNN) and trained to perform scene categorization. We find that: 1– Systems trained on foveated inputs (Foveation-Nets) have similar generalization as networks trained on matched-resource networks without foveated input (Standard-Nets), yet show greater cross-generalization. 2– Foveation-Nets show higher robustness than Standard-Nets to scotoma (fovea removed) occlusions, driven by the first foveation stage. 3– Subsequent representations learned in the CNN of Foveation-Nets weigh center information more strongly than Standard-Nets. 4– Foveation-Nets show less sensitivity to low-spatial frequency information than Standard-Nets. Furthermore, when we added biological and artificial augmentation mechanisms to each system through simulated eye-movements or random cropping and mirroring respectively, we found that these effects were amplified. Taken together, we find evidence that foveated perceptual systems learn a visual representation that is distinct from non-foveated perceptual systems, with implications in generalization, robustness, and perceptual sensitivity. These results provide computational support for the idea that the foveated nature of the human visual system might confer a functional advantage for scene representation.

1 Introduction

In the human visual system, incoming light is sampled with different resolution across the retinal area of which only 0.01% is occupied by the fovea (our center of gaze) despite consuming 10% of the resources in primary visual cortex [2]. *Foveation* – which is defined as this spatially varying visual sensing, arises from a combination of factors such as photoreceptor density variation in the retina [32], cortical magnification [13] and retinal ganglion cell convergence [26]. One account for this foveated array is related purely to sensory efficiency (biophysical constraints) [18, 9], 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 movable high-acuity fovea, rather than a non-movable uniform resolution retina when given a limited number of photoreceptors as suggested in Akbas & Eckstein [1].

However, it is also possible that foveation plays a functional role at the *representational level*, which can confers perceptual advantages [12]. For example, R.T. Pramod & Arun [30] found that simply blurring the image in the periphery gave a slight increase in object recognition by reducing false positives. Similarly, Cheung et al. [5] found that a recurrent neural network performing a visual search task in a cluttered environment naturally developed a spatially-varying receptive field array by

Preprint. Under review.

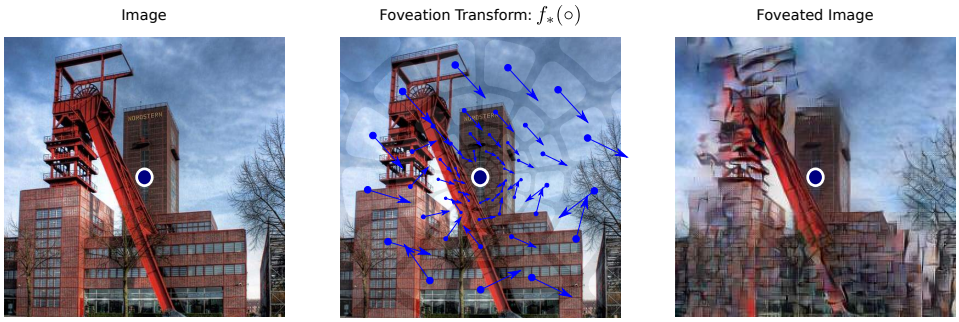

Figure 1: A cartoon illustrating how a foveated image is rendered producing a human visual *metamer* via the foveated feed-forward style transfer model of Deza et al. [8]. Here, each receptive field is locally perturbed with noise in its latent space in the direction of their equivalent texture representation (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.

re-distributing its sensors to the central area to improve its visual acuity (*functionality*) while also minimizing number of eye-movements (*efficiency*).

Expanding on the functional view, Rosenholtz [27] has argued that peripheral vision is not just an impoverished (blurred) version of foveal vision, but instead carries out a different kind of processing analogous to texture representation that is useful for global scene recognition [3, 28, 10, 29]. One approach for exploring foveal-vs-peripheral distinctions was taken by by Wu et al. [33], who directly introduced a bifurcating foveal-peripheral pathway in a neural network. This network showed boosted object detection performance – an interesting result despite not having texture priors induced in the periphery. In this paper, we take a different approach, directly examining the impact of these texture-like computations in the periphery, and whether they convey a functional perceptual advantage [27, 3, 28, 10, 29, 35, 22]. Specifically, we will introduce *foveated perceptual systems*: these are two-stage systems that have a foveation stage followed by a deep convolutional neural network. Further, we will compare these models' perceptual biases to their non-foveated counterpart through a set of experiments: scene categorization accuracy, robustness to occlusion, spatial frequency sensitivity and image-region sensitivity.

Notice that our approach differs from previous work in a critical way. That is, we will mimic foveation over images using a transform that simulates *visual crowding* [20, 24] in the periphery as shown in Figure 1 [10, 8, 31] rather than gaussian blurring [30] or compression [16]. This has the effect of capturing some of the image statistics that are known to be preserved in human peripheral vision that resemble texture computation [22] as previously argued in Rosenholtz [27]. Investigating whether these types of texture-like representations are *functionally* useful in the human visual periphery is still an open research question [7, 24, 31, 3, 35] – and shedding light on this question via computational means is the main goal of our paper.

2 Foveated Perceptual Systems

Here we define perceptual systems as *two-stage* with a spatial transform stage (stage 1: $f(\circ)$), that is relayed to a deep convolutional neural network (stage 2: $g(\circ)$). Note that the first transform stage is a *fixed* operation over the input image, while the second stage has *learnable* parameters. More formally we will define the general perceptual system $S(\circ)$, with input retinal image I as:

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

2.1 Stage 1: Foveation Transform

To model the computations of a foveated early visual system, we employed the recently proposed Metamer model of Deza et al. [8] (henceforth *Foveation Transform*). This model is inspired by the metamer synthesis model of Freeman & Simoncelli [10], where image metamers are constructed by locally matching texture statistics [25] in the visual periphery. Analogously, the Deza et al. [8] Foveation Transform uses a foveated feed-forward style transfer [15] network to latently perturb the image in the direction of its locally matched texture statistic – see Figure 1 for a visualization. When

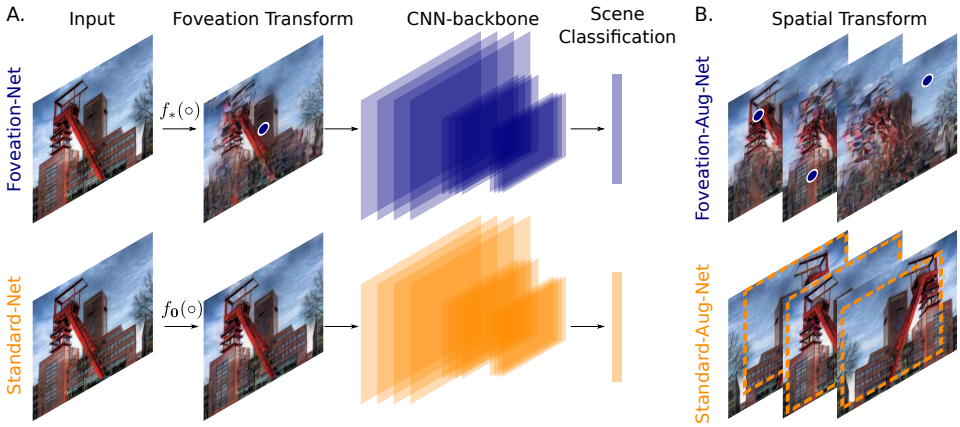

Figure 2: A. Foveation-Net (top row) and Standard-Net (bottom row), where each system receives an image as an input, applies a foveation transform, which is then relayed to a CNN architecture for scene classification. B. Foveation-Aug-Net enables eye-movements, while Standard-Aug-Net performs random cropping, resizing and horizontal mirroring.

calibrating the distortions correctly, the resulting procedure leads to a perturbed image that a human observer is unable to discriminate from a reference image.

2.2 Stage 2: Convolutional Neural Network backbone

The foveated images (stage 1) are passed into a standard neural network architecture. Here we tested two different base architectures: AlexNet [17], and ResNet18 [14]. 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 should not be viewed in the traditional way of an end-to-end input/output system where the input is the retinal image, and the output is a class-label. Rather, the CNN acts as a proxy of higher stages of visual processing.

Now, we can define two critical perceptual systems as the focus of our experiments: Foveation-Net and Standard-Net in addition to their augmented variants: Foveation-Aug-Net and Standard-Aug-Net which are fully described in the rest of this subsection and also in Figure 2.

Foveation-Net: We adjusted the parameters of the foveation 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) $s = 0.4$. Thus, when foveation is *on* at the previous scaling factor, we will abbreviate f as f_*.

Standard-Net: We use the same foveation transform at the foveation stage for Standard-Net but set the scaling factor set to $s = 0$. In this way, any potential effects of the compression/expansion operations of the foveation stage in the image are matched between the Foveation-Nets and Standard-Nets. Thus, the only difference after stage 1 is whether the image statistics were texturized in increasingly large pooling windows (Foveation-Nets), or not (Standard-Nets). Extending our notation, no foveation in stage one will be abbreviated as $f_{\mathbf{0}}$.

Foveation-Aug-Net: Eye-movements can be seen as a biologically motivated type of data-augmentation strategy. To test this, we created an enhanced version of Foveation-Net where the deep neural network at stage 2 receives a foveated image of a variable fixation point from one of 9 points from a fixed 3×3 grid.

Standard-Aug-Net: In analogy to the previous augmentation condition, Standard-Aug-Net enhances our Standard-Net model with artificial data-augmentation regimes such as as random cropping ($0.7 - 1.0$ of area), resizing, and a 0.5 chance of horizontally flipping the image at training.

3 Methods

Task: All models were trained on the task of 20-way scene categorization. The scene categories were selected from the Places2 dataset [34], with 4500 images per category in the training set, 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.

Training: Convolutional neural networks of the stage 2 of each perceptual system were trained which resulted in a total of 40 networks: 10 Foveation-Nets, 10 Standard-Nets, 10 Foveation-Aug-Nets, 10 Standard-Aug-Nets. All systems were paired such that their stage 2 architectures started with the same random weight initialization prior to training. These networks were trained via backpropagation with an SGD optimizer (learning rates: $\eta = 0.001$ AlexNet, $\eta = 0.0005$ ResNet18), batch size: 128, and for 270 (AlexNet) and 90 (ResNet18) epochs. All results shown in the main body of the paper are reported for AlexNet as the second stage CNN backbone, and the full set of analogous results on ResNet18 are included in the Supplementary Material.

Testing: Foveation-Nets and Foveation-Aug-Nets were only evaluated on center fixation images at inference time. Similarly, Standard-Nets and Standard-Aug-Nets were tested on non-foveated images (the output of the foveation transform with scaling factor set to zero); and no data-augmentation was performed at inference for any of the networks.

All networks performed mean and standard contrast normalization both at training and testing with values set to mean: $(0.485, 0.456, 0.406)$, std: $(0.229, 0.224, 0.225)$.

4 Experiments

4.1 Generalization and Cross-Generalization

We first examined scene categorization accuracy 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. On the other hand, this same type of image encoding could confer a functional advantage, in which case Foveation-Nets would do better. What we actually found is that Foveation-Nets and Standard-Nets had the exact same scene classification performance: Average accuracies were $65.48 \pm 3.54\%$ for Foveation-Nets and $65.22 \pm 3.32\%$ for Standard-Nets, with no statistically significant difference ($t(9) = -0.309$, $p = 0.763$). Similar performance was also found for our augmented variants: Foveation-Aug-Nets: $65.95 \pm 2.31\%$, Standard-Aug-Nets $66.17 \pm 4.30\%$, n.s. ($t(9) = 0.210, p = 0.838$). Thus, despite the lossier input through the foveation stage, it did not have a major impact on the categorization capacity of the networks.

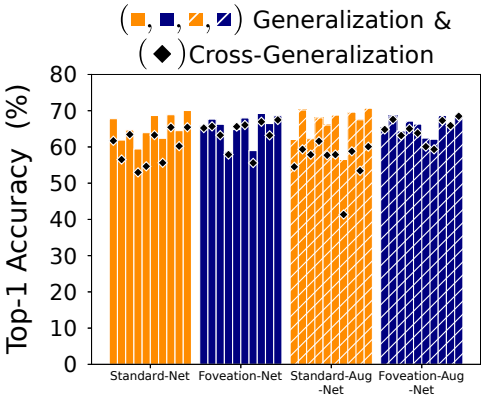

Figure 3: Scene Categorization Accuracy.

However, we found that these networks are not learning the same visual representations over scenes. When we inputted foveated images to Standard-Nets, and vice versa we found asymmetry in cross generalization. That is, Foveation-Nets can classify full-resolution scenes with only a slight cost ($\Delta = 1.78 \pm 1.26\%$), while Standard-Nets have more difficulty classifying foveated scenes ($\Delta = 5.30 \pm 2.03\%$). This second order difference ($\Delta_2 = 5.30\% - 1.78\% = 3.52\%$) was *amplified* when considering our augmented variants: Foveation-Aug-Nets ($\Delta = 1.40 \pm 0.95\%$) and Standard-Aug-Nets ($\Delta = 9.92 \pm 3.14\%$), with $p < 0.001$ for all Δ's, and $\Delta_2^{\text{Aug}} = 9.92\% - 1.40\% = 8.52\%$. Indeed, Figure 3 shows that Foveation-Aug-Nets (64.55 ± 2.90; navy) have less of a decrease in performance than Standard-Aug-Nets ($56.21 \pm 5.31\%$; gold) and that this difference is statistically significant $t(9) = -5.71, p < 0.001$. These results indicate that forcing texture representation in the periphery learns relevant structure as it generalizes to full-resolution image statistics, while the full-resolution trained networks have representations that do not show this same degree of out-of-distribution (o.o.d.) robustness.

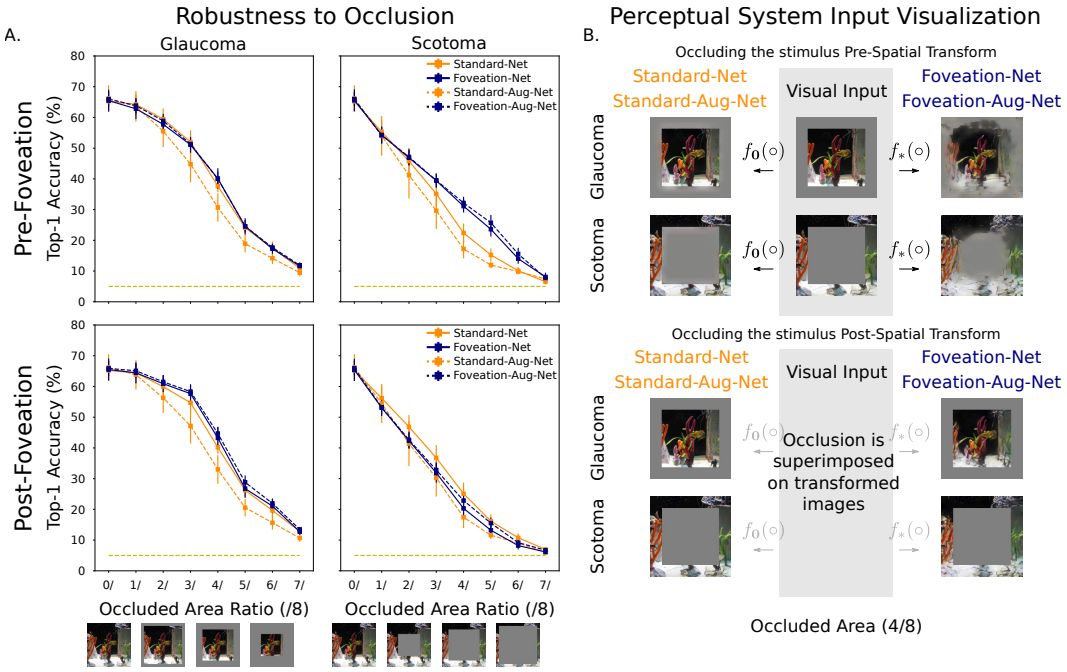

Figure 4: A. The Robustness curves for image occlusion. Top: Here we notice that Foveation-Nets are more robust to scotomal occlusion than Standard-Nets visualized by the navy curve's shift – an effect that is amplified through augmentation. Bottom: Interestingly when areas are equalized for both perceptual systems, this effect is flipped. B. A visualization of how the visual input is transformed for each perceptual system via the Foveation transform. Here we see how this potentially induces occlusion tolerance for the foveated systems by increasing visual scene area due to *visual crowding*.

4.2 Robustness to Occlusion

We next examined how the foveated and standard models 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 increasing size. For the peripheral-occlusion condition, we superimposed a gray box over the image, 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 4 (top) shows a summary of these results. Overall Foveation-Nets showed more robustness to central occlusion but not peripheral occlusion compared to its Standard-Net counterpart (central occlusion: difference in area under the curve (Δ-AUC): 3.135 ± 2.693 (mean $\pm$ std.), $t(9) = -3.493, p < 0.01$; peripheral occlusion: difference in area under the curve (Δ-AUC): -0.164 ± 2.538, $t(9) = 0.194$, n.s.). Upon augmentation, the difference in robustness to central occlusion is amplified for the Foveation-Aug-Net vs the Standard-Aug-Net Δ-AUC: 3.383 ± 3.329, $t(9) = -3.049, p < 0.05$. Additionally, for peripheral occlusion, Foveation-Aug-Nets show higher robustness compared to Standard-Aug-Net Δ-AUC: 6.257 ± 3.855,$t(9) = -4.869, p < 0.001$ – an effect mainly motivated by a drop in the robustness of the the latter. Interestingly, Standard-Aug-Nets are also less robust to peripheral occlusion than Standard-Nets Δ-AUC: -3.130 ± 3.376,$t(9) = 2.782, p < 0.05$.

To further explore these results, we visualized what the image inputs look like after the foveation stage, for both the Human-Aug-Nets and Machine-Aug-Nets. Figure 4 (right) reveals critical information: The effect of the stage 1 operation for the foveated systems *reduces* the occluded area in the periphery – a stark contrast in comparison to Standard-Nets that only received a compressed version of the occluded image.

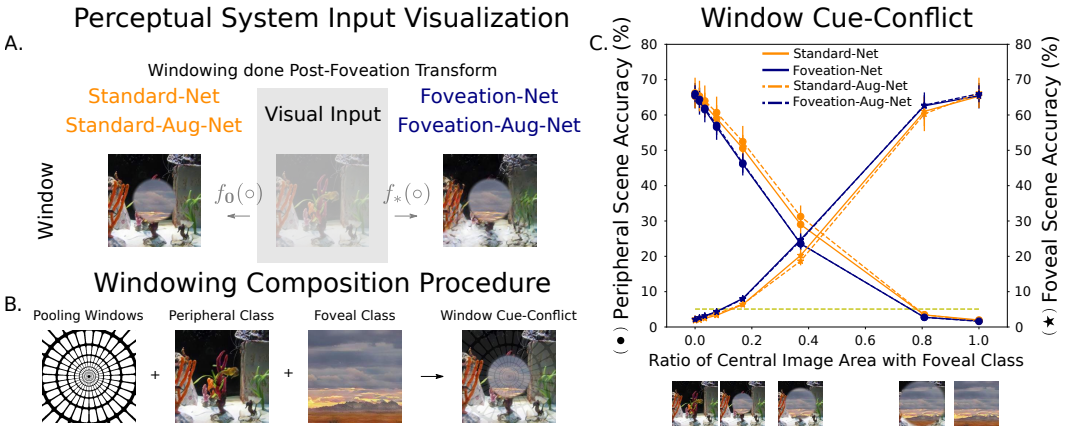

Figure 5: A. A visualization of the stimuli type that is shown to each perceptual system consists of a smoothed composition of a foveal image (*e.g.* badlands) mixed with a peripheral image (*e.g.* aquarium). B. The composition procedure relies on using the same log-polar pooling windows over which the foveated images were rendered. C. Window Cue-Conflict experiment results: Foveated Perceptual Systems – with and without augmentation – show stronger biases to classify scenes by the information shown in the foveal region.

What factor is accounting for the higher robustness of the foveated systems? Is the benefit driven by the learned texturized representations in the subsequent deep convolutional neural network, or is it driven by greater computable visual area, that has been filled in from the foveation stage? To examine this question, we occluded either the center or periphery of the images, but *after* the stage 1 computation (see Figure 4 (bottom)). In this way, the area of occlusion was matched for the second stage classification task. The results reveals that, for central field occlusion, Standard-Nets now perform better than Foveation-Nets (Δ-AUC: -2.777 ± 2.717,$p < 0.05$). This means that the stage one operation was the main force underlying robustness to occlusion. Interestingly, there is no difference in robustness and overall accuracy when considering the augmented systems (Δ-AUC: 1.424 ± 4.123, n.s.).

For the peripheral field loss, the same pattern of results is found for both the pre-foveation and post-foveation stage with no difference between Standard-Nets and Foveation-Nets (Δ-AUC:1.173 ± 1.80,$t(9) = -1.922$, n.s.), and a maintained difference for the Augmented variants (Δ-AUC: 5.257 ± 3.526,$t(9) = -4.473, p < 0.01$). This suggests that foveation-stage is generally aiding visual performance when the center is removed as the foveation stage fills in the image from the periphery to the fovea, but not the other way around – potentially due to bigger pooling regions in the periphery where image is already removed *a priori* in the glaucoma condition.

4.3 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. 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 and examined classification accuracy for both the foveal and peripheral scene (Figure 5).

We found that Foveation-Nets maintained higher accuracy for the foveal scene class than do Standard-Nets, as the amount of competing peripheral information increased. Similarly, Standard-Nets maintained higher accuracy for the peripheral scene class than Foveation-Nets, as the amount of competing foveal information increases. 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.

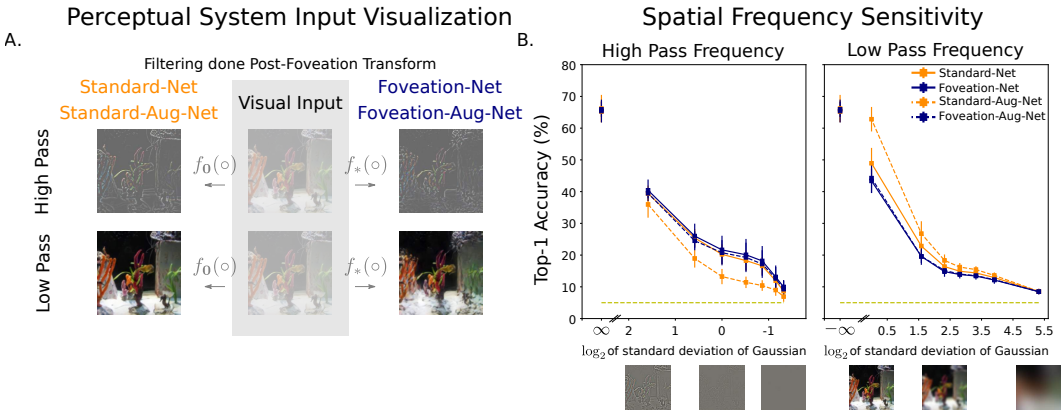

Figure 6: A. Sample images from the foveated and non-foveated images and how they change as a function of spatial frequency filtering at the post-foveation stage. B. Foveation-Nets have greater sensitivity to high spatial frequency filtered stimuli than Standard-Nets. Conversely, they show less tolerance for low pass frequency stimuli. This trend is amplified for their augmented variants.

To quantify these differences, we computed a paired t-test over the difference in AUC considering Standard-Nets vs Foveation-Nets. Standard-Nets shows more robust peripheral scene classification accuracy ($\bullet$): Δ-AUC: 4.550 ± 3.164, $(t(9) = 4.314, p < 0.01)$, which was amplified for the augmented variants Δ-AUC: 6.422 ± 3.541, $(t(9) = 5.441, p < 0.001)$. And, Foveation-Nets show more robust foveal scene classification accuracy ($\star$) than Standard-Nets:Δ-AUC:3.188 ± 1.178, $t(9) = 8.119, p < 0.0001$, which is amplified for the augmented variants Δ-AUC: $3.851 \pm 1.151, t(9) = 10.034, p <$ 1e-5.

Interestingly, there is no variation between any system and its augmented counterpart *i.e.* Foveation-Nets vs Foveation-Aug-Nets: Δ-AUC:$-0.185 \pm 2.344, t(9) = -0.237$, n.s.), and Standard-Nets vs Standard-Aug-Nets: Δ-AUC:$-2.058 \pm 5.261, (t(9) = -1.173$, n.s.). These additional results suggest that the foveation mechanism itself is mainly responsible for the development of a center image bias, rather than the data-augmentation scheme such as eye-movements or blur through the resizing.

4.4 Spatial Frequency Sensitivity

We next examined whether Foveation-Nets and Standard-Nets learn feature representations that are more reliant on low or high spatial frequency information, at the second stage of visual processing (post-foveation). 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 or f_*) outputs. Similarly, the high pass stimuli was computed by subtracting the reference image from its low pass filtered version with $\sigma = [\infty, 3, 1.5, 1, 0.7, 0.55, 0.45, 0.4]$ pixels. These are the same values used in the experiments of Geirhos et al. [11].

The results indicate that Foveation-Nets show nominally more sensitivity to the information in high-pass frequency stimuli vs Standard-Net, but this effect was not statistically significant (Δ-AUC: 6.229 ± 19.285, $t(8) = -0.969$, n.s., (Figure 6). Interestingly, the complementary effect is clear in the low-pass frequency: Standard-Nets show relatively more robustness recognizing scene categories in blurry images, Δ-AUC: $-10.952 \pm 6.912, t(8) = 4.753, p = 0.001$. When comparing the augmented versions of each perceptual system this tendency was greatly amplified (high-pass Δ-AUC: 26.253 ± 17.713,$t(8) = -4.447, p < 0.01$; low-pass Δ-AUC: -33.290 ± 3.468,$t(8) = 28.800, p <$ 1e-9).

Given that Foveation-Nets have less sensitivity to low-spatial frequency content, as well as a central visual bias, this raises questions about how these results are related. Future experiments can address whether this bias is present uniformly across the image, or whether it is amplified in the center of the visual field.

5 Discussion

Here we designed a two-stage foveated system; examined the impact on the learned representation for classification, generalization, and robustness to occlusion; and probed the sensitivity to different image content in the central vs periphery and high vs low spatial frequencies. We find that the Foveation-Nets had similar generalization as networks trained on matched-resource Standard-Nets, but showed greater cross-generalization to full spatial resolution images. The Foveation-Nets show higher robustness than scotomal occlusions, largely driven by the foveation stage itself, which effectively is filling-in scene-category relevant information into the lesioned visual field. Interestingly, we found that the foveation stage induced the encoder networks to rely *relatively more* on central information than Standard-Nets. Relatedly, the Foveation-Nets showed representations that were less reliant on low-spatial frequency information than the Standard-nets, with an equal or slightly stronger reliance on high-spatial frequency image content. Thus, the foveation stage clearly induced differences in the subsequently learned representation, which were less about texture in the periphery than we anticipated at the outset.

There are three results to highlight about this work. The first relates to role of central vs peripheral information for scene classification. Interestingly, our occlusion experiments arrive at similar conclusions with some behavioural research. For example, Larson & Loschky [19] also experimented with glaucoma and scotoma-like stimuli probing humans with their ability to perform scene recognition. They found that humans can indeed perform scene recognition in the absence of foveal information, however when equalizing occluded areas (foveal and peripheral), they find that the foveal area is *more* critical for scene gist than the periphery – a result consistent with our post-transform occlusion and cue-conflict experiments.

The second observation is that there is little to no variation between Foveation-Nets and Foveation-Aug-Nets across all experiments. This pattern of data is compatible with human psychophysics showing that scene recognition can happen in a glance, without requiring multiple fixations (scene gist) [23], and that scenes can also be recognized independently of point of fixation or retinal eccentricity [4].

The final observation is that, in contrast to Foveated-Nets vs Foveation-Aug-Nets, there were substantial consequences of standard augmentation schemes. Specifically, relative to Standard-Nets, the Standard-Aug-Nets show (i) worse cross-generalization, (ii) less robustness to glaucoma and scotoma, (iii) a stronger peripheral bias, and (iv) and greater sensitivity to low-pass frequency. While data-augmentation schemes are used with the hope of improving generalization performance [6, 21, 36], these results suggest that the crop and resize augmentations lead to a generally lower resolution image representation, and potentially less effective representation for out-of-distribution scene recognition, robustness and center image bias.

The present work was motivated by a broader question about the functional role for peripheral texture statistics for scene representation. It could have been the case that Foveation-Nets showed similar or higher scene classification accuracy than Standard-Nets, using a stronger peripheral bias. Surprisingly, this was not the pattern. We instead found that the foveation stage operated more as a *focusing mechanism*, servicing the central visual field, and de-emphasizing low-frequency information. Note that these results do not rule out the possibility of distinct functional consequences for peripheral computations, which continues to be an important question for future investigations. Instead, these particular consequences of our foveation stage raise interesting questions about what will happen when trained on objects, which are typically centered in view. Specifically, one intriguing possibility is that these representational signatures may induce more shape sensitivity (rather than texture sensitivity; [11]), and may amplify the perceptual differences we identified across foveated and non-foveated systems.

Broader Impact: The significance of our results are two-fold for human and machine vision. For human vision, the spatially varying nature of the visual field – that has traditionally been accepted to be a purely metabolic optimization constraint in the machine vision community – is perhaps a necessary computation for high-level robust vision as argued in Rosenholtz [27]. For machine vision, it is likely that the next generation of artificial perceptual systems (including deep neural networks) that already incorporate canonical computations from the human ventral stream such as filtering, half-wave rectification, pooling, and local gain control will benefit from the addition of foveation.

References

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

[2] Azzopardi, P. and Cowey, A. Preferential representation of the fovea in the primary visual cortex. *Nature*, 361(6414):719, 1993.

[3] 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.

[4] Boucart, M., Moroni, C., Thibaut, M., Szaffarczyk, S., and Greene, M. Scene categorization at large visual eccentricities. *Vision Research*, 86:35–42, 2013.

[5] 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.

[6] Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 113–123, 2019.

[7] 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.

[8] 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`.

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

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

[11] 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`.

[12] 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.

[13] Harvey, B. M. and Dumoulin, S. O. The relationship between cortical magnification factor and population receptive field size in human visual cortex: constancies in cortical architecture. *Journal of Neuroscience*, 31(38):13604–13612, 2011.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

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

[21] Li, Y., Hu, G., Wang, Y., Hospedales, T., Robertson, N. M., and Yang, Y. Dada: Differentiable automatic data augmentation. *arXiv preprint arXiv:2003.03780*, 2020.

[22] Long, B., Yu, C.-P., and Konkle, T. Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences*, 115(38): E9015–E9024, 2018.

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

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

[25] 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.

[26] Purves, D., Augustine, G., Fitzpatrick, D., Katz, L., LaMantia, A., McNamara, J., and Williams, S. Functional specialization of the rod and cone systems. *Neuroscience*, 2001.

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

[28] 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.

[29] Rosenholtz, R., Yu, D., and Keshvari, S. Challenges to pooling models of crowding: Implications for visual mechanisms. *Journal of vision*, 19(7):15–15, 2019.

[30] R.T. Pramod, H. K. and Arun, S. Human peripheral blur is optimal for object recognition. *arXiv:1807.08476*, 2018.

[31] 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.

[32] Wässle, H., Grünert, U., Röhrenbeck, J., and Boycott, B. B. Retinal ganglion cell density and cortical magnification factor in the primate. *Vision research*, 30(11):1897–1911, 1990.

[33] 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.

[34] 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.

[35] 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.

[36] Zoph, B., Cubuk, E. D., Ghiasi, G., Lin, T.-Y., Shlens, J., and Le, Q. V. Learning data augmentation strategies for object detection. *arXiv preprint arXiv:1906.11172*, 2019.

6 Supplementary Material

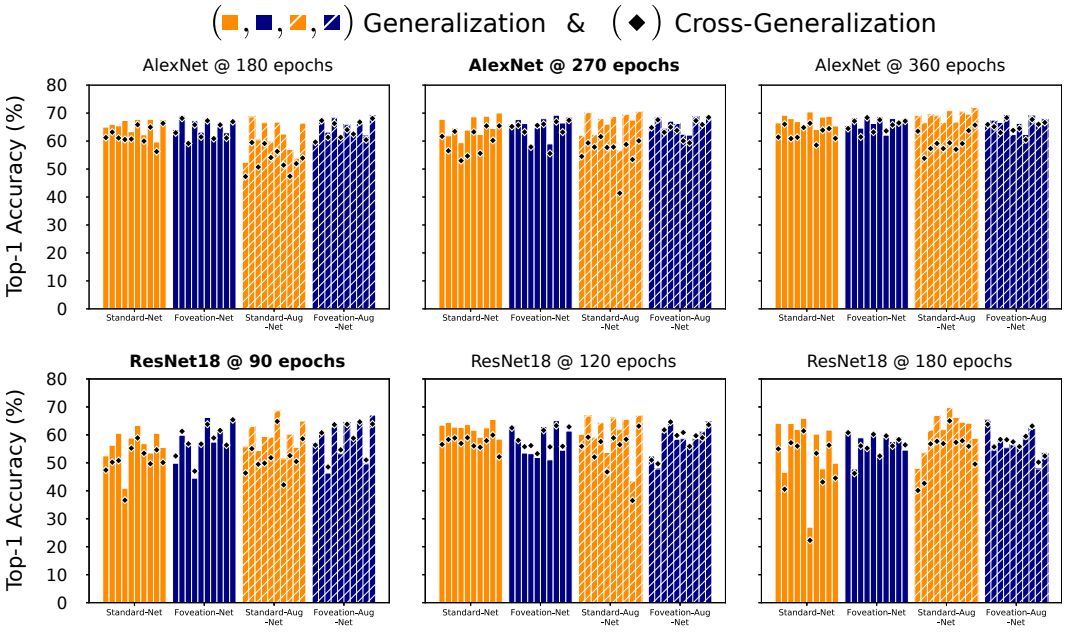

Figure 7: **Generalization:** The full Generalization and Cross-Generalization plots for AlexNet and ResNet18 across multiple epochs of training. We observe that our results do not vary, though selecting the correct epoch for analysis is important before any system begins to overfit (See Figure 8).

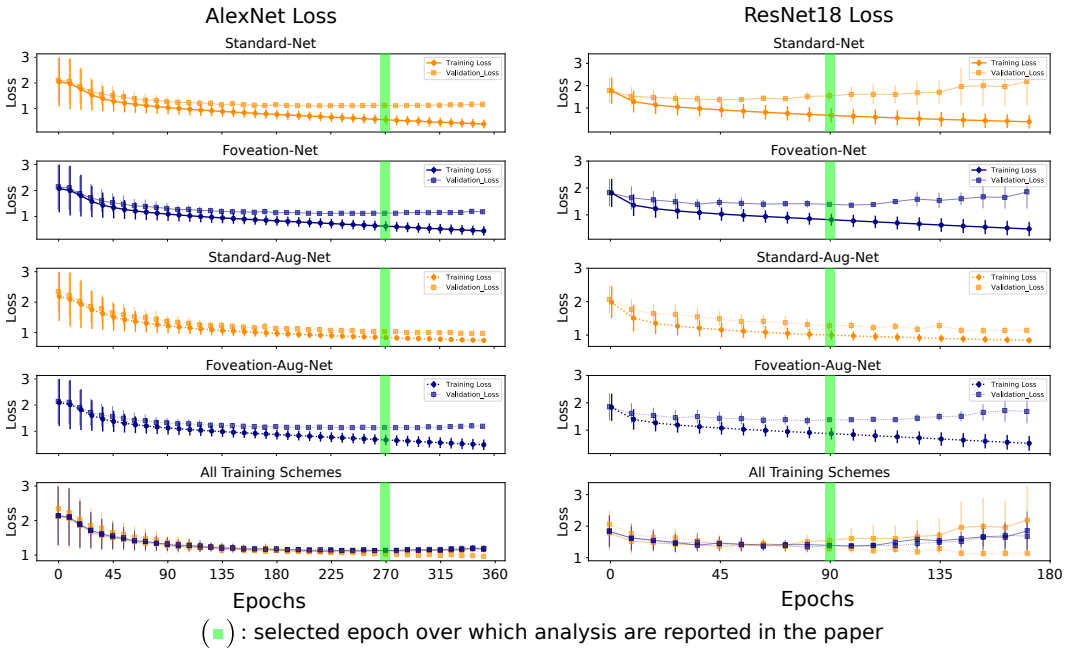

Figure 8: **Learning:** An averaged visualization of loss function convergence as a function of epochs. Each point in the plot is the average across the 10 different network runs from the locally averaged/smooth loss function per each 9 epochs. Notice that the epoch that we pick for each network is an epoch *before* any system begins to overfit. In general, Standard-Aug-Nets seem to be less propense to overfitting – visualized by the longer tail in the validation loss.

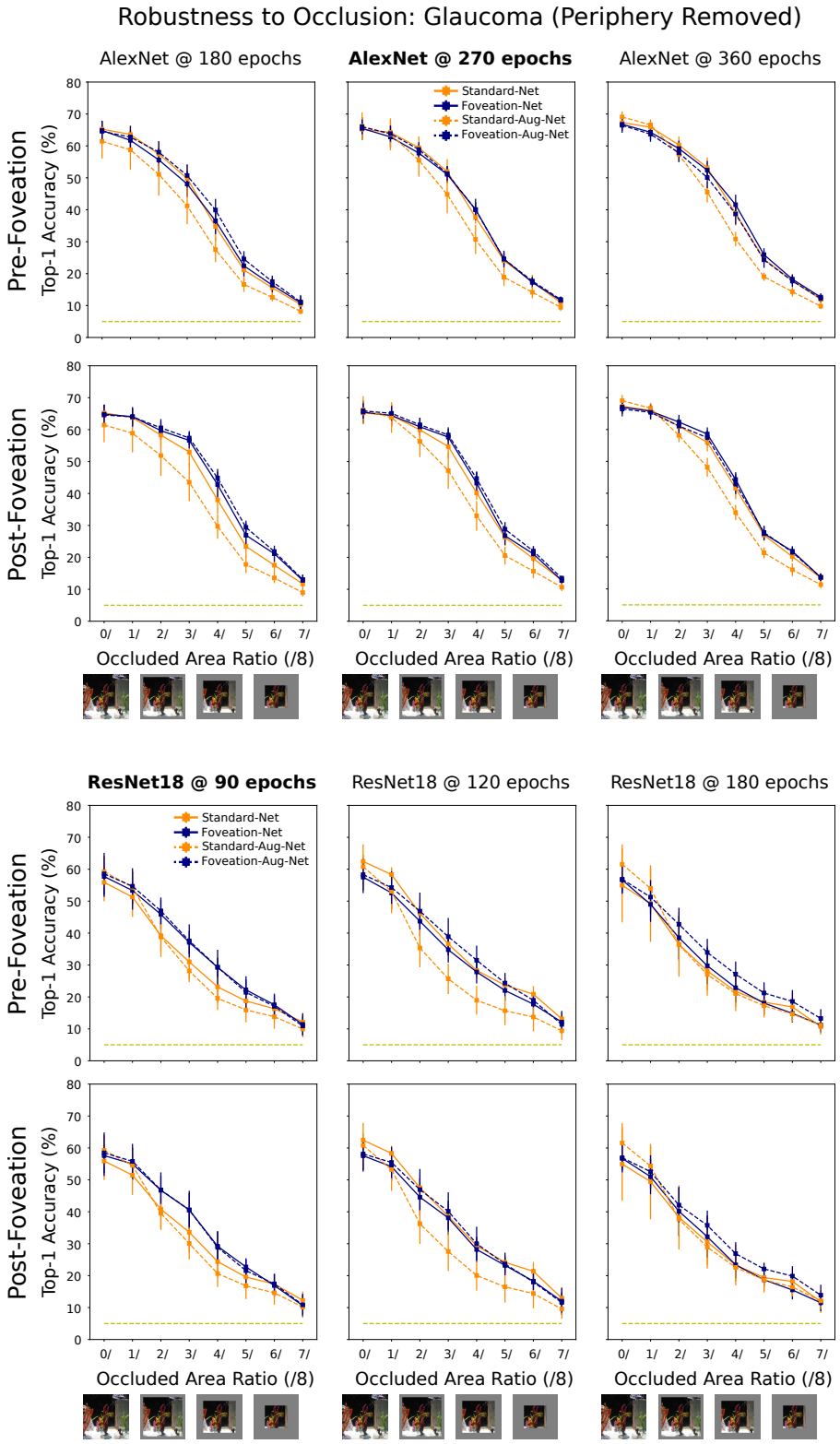

Figure 9: **Robustness to Glaucoma Occlusion (Periphery Removed):** The same pattern of results as reported in the main body of the paper is shown independent of epoch or network architecture for the CNN backbone of each perceptual system. These are the following: no significant differences for Foveation-Net vs Standard-Net at the pre-foveation stage (ResNet18 @ 90 epochs being an exception), and significant differences between Foveation-Aug-Net vs Standard-Aug-Net (pre and post foveation), in addition to Standard-Net vs Standard-Aug-Net.

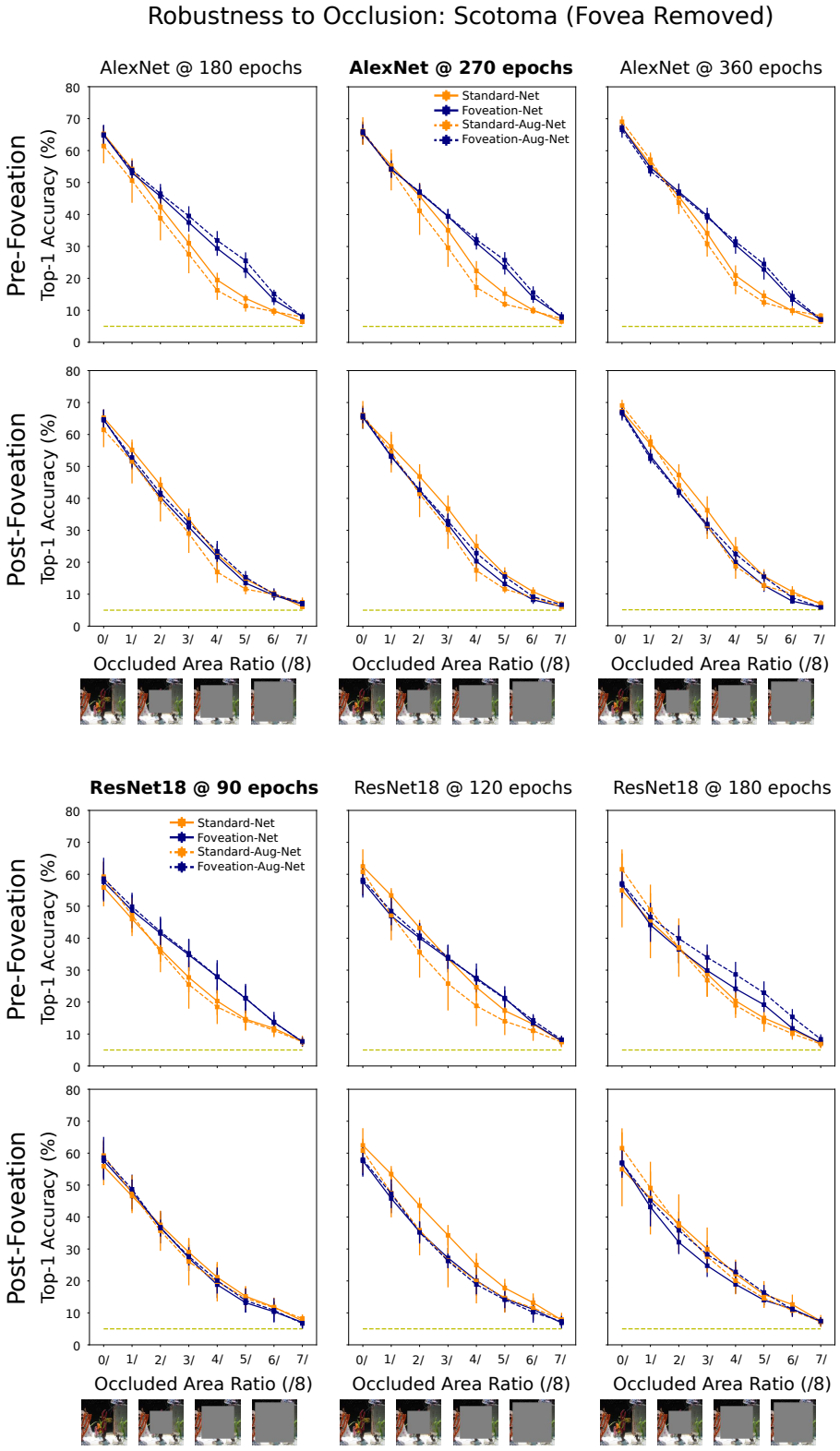

Figure 10: **Robustness to Scotoma Occlusion (Fovea Removed):** Foveation-Nets have greater robustness than Standard-Nets at the pre-foveation stage across networks and epochs (in addition to their augmented variants). This effect goes away at the post-foveation stage. This effect is not obvious for ResNet18 @ 120 epochs and 180 epochs given the difference in performance at testing.

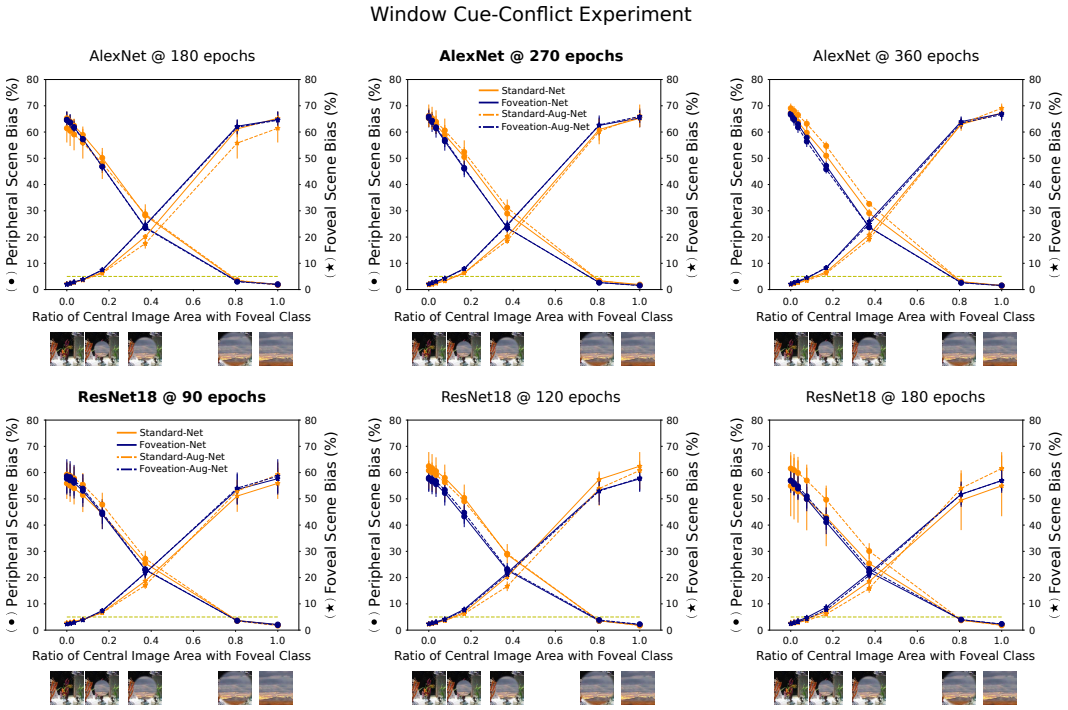

Figure 11: **Window Cue Conflict Experiment:** The pattern or results with regards to a greater central remains independent of the network architecture (stage 2) and the epoch. This can be verified by finding the cross-over points for Foveation-Nets and Foveation-Aug-Nets being placed more leftwards than Standard-Nets and Standard-Aug-Nets. These results are independent of potential perceptual differences at testing time (see AlexNet @ 360 epochs, or ResNet18 @ 120 or 180 epochs), where the cross-over point for Standard-Nets and Standard-Aug-Nets is still shifted more biased towards the right than Foveation-Nets and Foveation-Aug-Nets – implying a greated need for foveal area to arrive to the point of subjective equality (PSE).

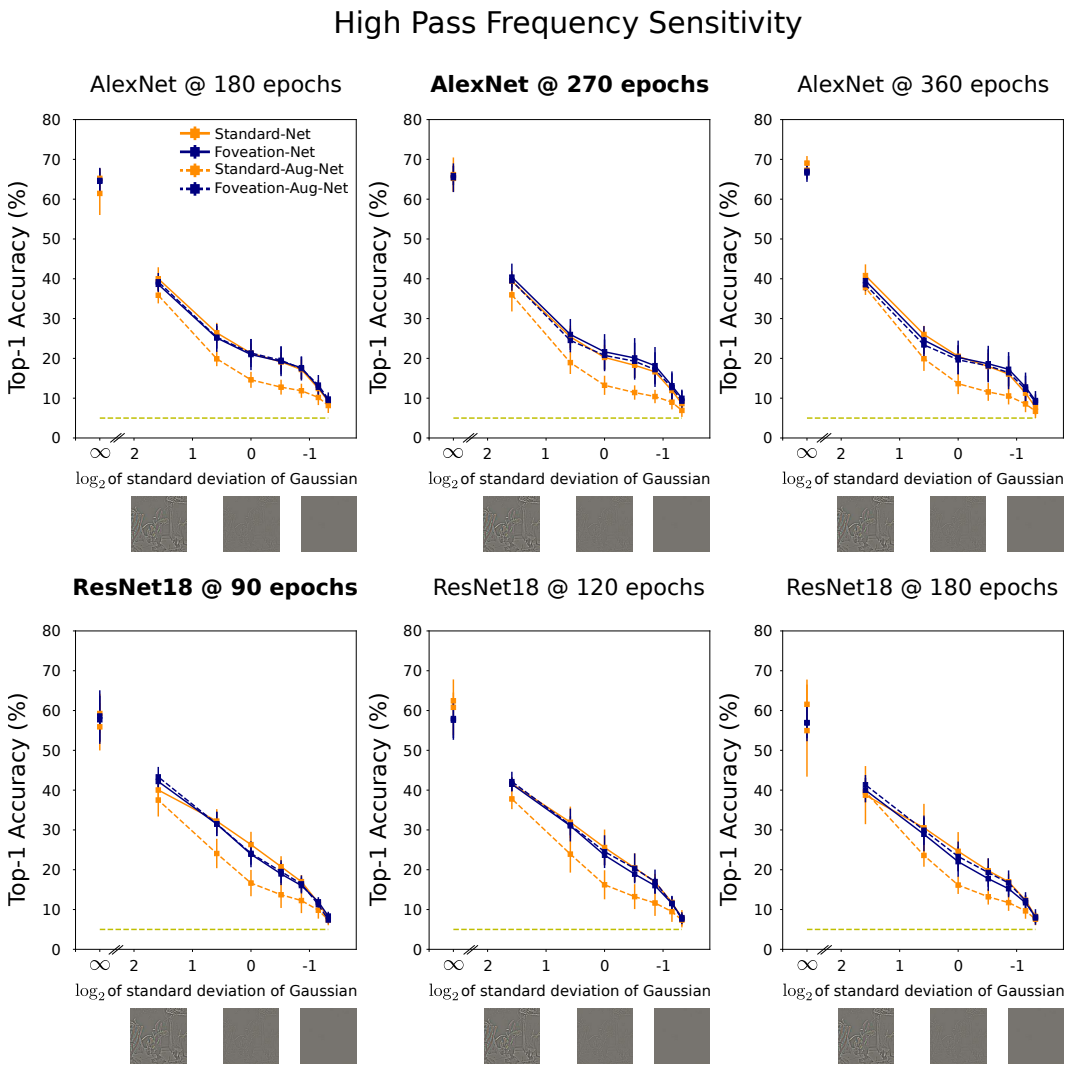

Figure 12: **High Pass Frequency Sensitivity:** There are no notorious differences for high pass frequency sensitivity across network architecture and epochs in comparison to the results reported in the main body of the paper. Specifically, these patterns are: no statistically significant difference between Foveation-Nets and Standard-Nets. No statistically significant difference between Foveation-Nets and Foveation-Aug-Nets. Statistically significant differences between Standard-Nets and Standard-Aug-Nets, and consequently statistically significant differences between Foveation-Aug-Nets and Standard-Aug-Nets.

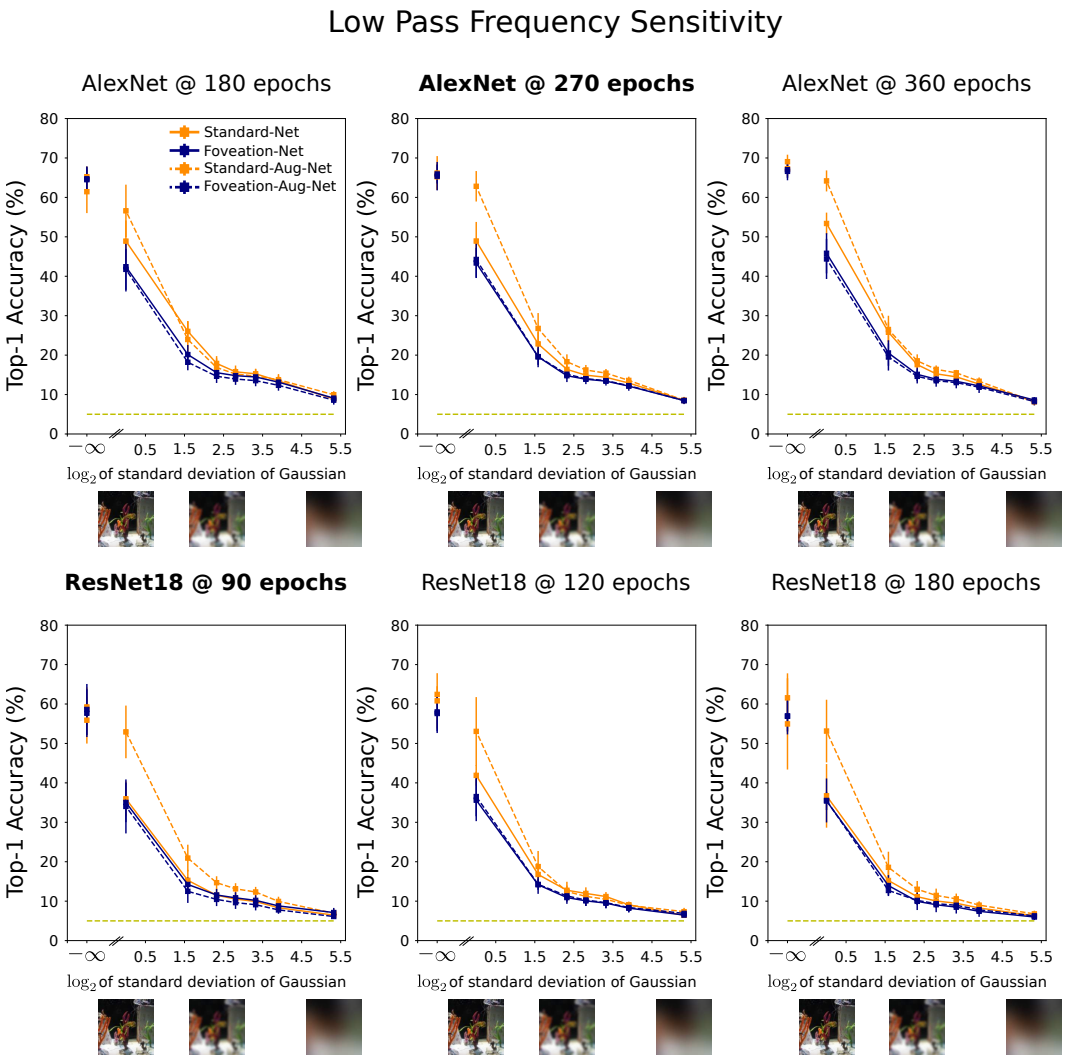

Figure 13: **Low Pass Frequency Sensitivity:** The following family of plots shows non-homogenous differences for Foveation-Nets vs Standard-Nets (significant for AlexNet, varying for ResNet18). In all cases Foveation-Nets vs Foveation-Aug-Nets shows no significant differences. However, there is a great difference (across architectures and epochs) between Standard-Nets vs Standard-Aug-Nets, and consequently Foveation-Aug-Nets vs Standard-Aug-Nets.