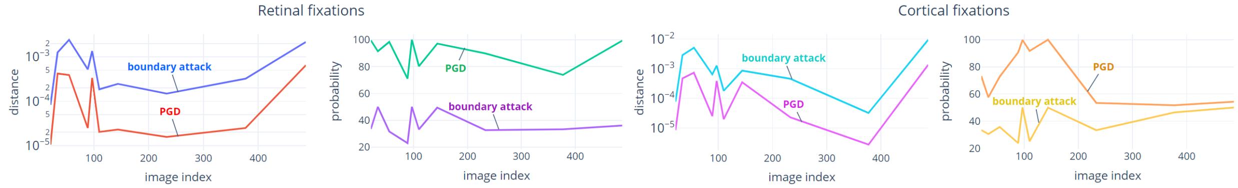
We thank the reviewers for their insightful and positive feedback. We are very encouraged they enjoyed (R2, R4) the idea to take inspiration from the human visual system to increase robustness of DNNs and found it to be unique (R3) and worthy of investigating more. The reviewers appreciated the substantial amount of control experiments and following best practices better than many other papers (R4), including showing scaling between datasets of varying complexities (R1). We are pleased R2 found the ablation studies informative. We appreciate the constructive comments and questions from the reviewers, some on the organization of the text between the main part and the extensive (R4) Supplementary Material, and some suggesting further experimental tests to strengthen the results. We especially would like to thank R4 for suggesting several very relavant papers to cite. As we were constrained by limited time and computational resources, as well as the space to answer, we focus below on the most crucial additional experiments and questions, but will incorporate all feedback in the final version.

(R2, R4) Run PGD with many more steps and smaller step size to truly evaluate robustness. The reviewers raised concerns that our main tests of robustness were against the PGD attack with 5 or 20 steps with a large step size of $\epsilon/3$ (we would like to point out that we have verified robustness to smaller step size of $\epsilon/12$ in the Supp. Mat.). Both reviewers suggested to run attacks with smaller step size and follow (Brendel et al, NeurIPS 2019), where attacks were performed up to a 1000 steps. We have followed this suggestion and performed PGD attacks on the ImageNet10 dataset with step size of $\epsilon/20$ (as suggested by R4) and over a range of steps up to 1000, as shown in the Table below. The drop in robustness is minimal with the additional steps. The trend of robustness seen with the small number of steps was already strongly representative of the general trend, and we believe these additional results help drive the message home that both retinal and cortical sampling are biologically-motivated methods that significantly improve robustness at small perturbations without noticeable computational overhead.

Table 1: PGD with more steps on ImageNet10. Standard: standard ResNet; Coarse: coarse fixations on standard ResNet; Retinal: Retinal fixations model; Cortical: cortical fixations model (Due to time constraints, we were unable to finish running the 1000 step attack at $\epsilon = 0.001$, but up to 500 steps the accuracy stayed at 82%, with matched performance for bigger ϵ).

	STEPS = 20, STEP SIZE = $\epsilon/3$				STEPS = 300, STEP SIZE = $\epsilon/20$				STEPS = 1000, STEP SIZE = $\epsilon/20$			
ϵ	STANDARD	COARSE	RETINAL	CORTICAL	STANDARD	COARSE	RETINAL	CORTICAL	STANDARD	COARSE	RETINAL	CORTICAL
0.02	0%	0%	0.4%	1%	0%	0%	0.2%	0.8%	0%	0%	0.2%	0.8%
0.01	1.6%	1.2%	7.6%	11.8%	1.4%	0.8%	7%	9.4%	1.4%	0.8%	6.8%	9.4%
0.005	20%	16.4%	43.4%	42.8%	19.4%	15.8%	43.4%	42.2%	19.2%	15.4%	43.4%	41.8%
0.001	78.4%	80.8%	86.4%	82%	78.4%	80.8%	86.4%	82%	78.4%	80.8%	86.4%	x%

(R4) Run boundary attack on a larger subset of images and with many more steps. The reviewer suggested that evaluating the boundary attack on a single image for 1000 iterations is insufficient to claim avoiding gradient obfuscation. While we would claim that it served it's purpose together with many other tests reported in the main text and shown in Supp Mat., we have nonetheless ran the attack on more images for up to 10000 steps and compared it to PGD on the same subset. Since these attacks were extremely computationally expensive, we were only able to run them on 10 images for the 10k steps. So far, the trend agrees with the observation made in the paper (see Figure below). In all cases, the size of the perturbation was smaller with PGD, while the probabilities assigned to the mispredicted class for examples generated with 20 steps PGD were much greater than for 10k steps of boundary attack.



(R3) Marginal improvement. The reviewer critiques that our proposed mechanisms only improve the robustness marginally as compared to standard training. While this is true for large perturbations (which both R3 and we note suggests additional mechanisms), the effect at small perturbations is strong: at $\epsilon = 0.005$ the adversarial performance more than doubles over the baselines both in CIFAR10 and ImageNet10.

(R1) Comparison to adversarial training. While the improvement from both mechanisms is lower than that of adversarial training, at $\epsilon = 0.005$ on CIFAR10 they reach approximately the halfway point between standard models and those adversarially trained, as seen in Figure 3 in the paper. However, unlike adversarial training, which is very computationally expensive, our methods pose negligible computational overhead and actually lead to improvement of standard performance on some of the datasets with the retinal fixations – this should be compared to adversarial training, which provably leads to significantly worse standard performance. Additionally, we have shown in the Supp. Mat. that our methods can scale to larger datasets (see the ImageNet100 experiments), while so far adversarial training has been unsuccesful in being scaled up to larger number of classes (hence the suggestion of R2 to try adversarial training on ImageNet is beyond the scope of this work).