We thank the reviewers for their constructive comments. We are pleased that even the most critical reviewers found the claims to be of "substantial significance" *if* they are validated with key controls. *We have completed all of those controls*; while some were in the original submission but not sufficiently highlighted, others are new analyses following reviewer suggestions. In all cases, the controls support our claims and we would happily include these improvements in the final manuscript. We thank the reviewers for their willingness to reconsider their reviewer score in this light.

VOneBlock describes V1 responses better than CNNs. We agree with R1 and R3 that *if* the VOneBlock is not better than CNNs at describing V1 responses, it would undermine our claims. In lines 155-158 we write "the VOneBlock outperformed all tested ImageNet-trained CNNs in explaining responses in the V1 dataset used", and provided the value of explained variance (0.387 ± 0.007). Here, we update Fig.1, showing VOneResNet50 surpasses all standard ImageNet-trained CNNs. We are confident in this result since we searched over all layers in a large pool of CNNs and did not find any coming close to the V1 explained variance of the VOneBlock. Further, **our results are consistent with Cadena et al 2019**: our GFB has parameters constrained by empirical data resulting in a better model of V1; when we use the parameters of the GFB in Cadena et al, we obtain a much lower explained variance (0.296 ± 0.005, marked on x-axis of Fig.R1).

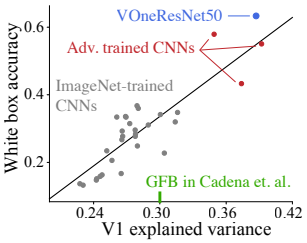

Fig R1: Accuracy in weak white box attack vs V1 explained variance; VOneBlock outperforms standard CNNs

Stochasticity during the attack is not the main source of adversarial robustness. R1 and R4 are concerned that stochasticity makes VOneNets artificially appear more robust. However, the majority of the VOneNets' robustness does not originate from stochasticity during the attacks, but rather from training downstream layers with V1-like features and neural stochasticity. *When we turn off stochasticity completely during attack/inference, VOneResNet50 retains* 70% *of the adversarial robustness gains of the model with stochasticity on, and significantly outperforms ResNet50* (Fig.4). Meanwhile, when we add stochasticity to ResNet50, using an affine transformation to scale the activations to the same magnitude as in the VOneBlock (lines 148-151), we find substantially smaller improvements in robustness (Fig.3B).

Further adversarial attack optimization does not overturn results. We follow R1's suggestions to further verify our attacks, addressing R4's concerns of gradient quality as well. Increasing the attack step size (α) and the number of gradient samples (k) only marginally increases the attack effectiveness. A grid search over α and k (Fig.R.2; PGD with $\|\delta\|_\infty = 1/255$) reveals that at the most effective step size ($\alpha = \epsilon/8$), increasing gradient samples beyond $k = 16$ no longer improves the attack. At $k = 128$, VOneResnet50 accuracy is only reduced from 29.1% to 26.0%, remaining a large margin above ResNet50 at 0.8%. We do not find using 5-10 random restarts improves beyond the most effective k and α settings. Thus, these controls do not qualitatively change our results. Finally, because our model has a stochastic boundary, we focus on PGD and avoid boundary targeted attacks, as the most effect should come from staying as far from the boundary as possible.

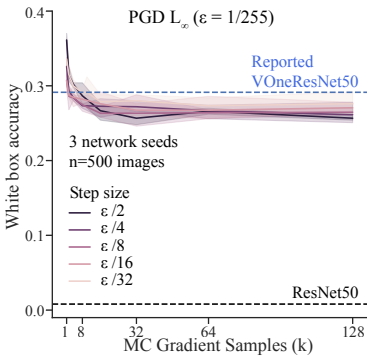

Fig R2: Attack optimization only marginally decreases VOneResNet50 accuracy

Common corruptions provide additional insights. R1 and R3 question the inclusion of common corruptions. While common corruption scores generally do not correlate with V1 similarity, we include them because robustness to imperceptible adversarial attacks should not come at the cost of performance on other perturbations that humans easily deal with, as is the case for adversarially trained models which underperform on this benchmark relative to the baseline and VOneNets. Further, common corruptions provide insights into the role of different components in the VOneBlock in dealing with specific stationary perturbation statistics. **Common corruptions and white box attacks are properly separated**: results are shown separately in Tables 1, 2, B.2, C.4, and C.5; white box only in Fig.1, 2, and 4; and scores are only combined in Fig.3.

Mechanisms contributing to robustness. We agree with R1 and R3 that the mechanisms behind our results warrant further exploration, but believe our study provides a novel and important characterization of how low level visual processing relates to robust object recognition behavior. With ablations, we characterize how different components of the VOneBlock impact performance on a variety of perturbation statistics. Lines 212-218 discuss the effect of high SF Gabors on various corruptions as well as complex cells on adversarial attacks, and Fig.4 shows that the improved adversarial robustness largely stems from training downstream layers with a stochastic V1 front-end.

Other questions from R1 and R3. $\pm$ refers to SD. As stated in the appendix, we will release code and model weights. To the best of our knowledge, the correlation between adversarial robustness and V1 similarity has not been shown before. The CNN layer that best predicts V1 responses is usually not the first (except for the VOneNets). We used parameters from the foveal region of V1 (eccentricity < 5deg). Our model has a field-of-view of 8deg (resolution = 224px/8deg; eccentricity < 4deg). The VOneBlock and all tested CNNs explained less than 40% of V1 response variance which is in line with the literature and is a possible leeway to make progress on robustness. One of the comparison models (ANT3$\times$3+SIN) is trained with input noise and does not show improvements in adversarial robustness.