# Perceptual Regularization from Human Visual System Models Improves Adversarial Robustness

*Anonymous Submission*

**Keywords:** adversarial robustness, deep learning, human visual system, perceptual regularisation

**Abstract.** Deep learning models often fail under adversarial perturbations that are imperceptible to humans [1]. This vulnerability highlights a gap between machine vision and human perception. We propose a neuroscience-inspired method that leverages a computational model of the Lateral Geniculate Nucleus (LGN) to guide neural networks toward perceptually aligned gradient maps. Our approach, called *LGN-Aware Regularisation (LGN-AR)*, integrates two key components: (i) an LGN-based gradient regularizer that emphasises structurally relevant image features, and (ii) a noise stability term that encourages robustness to natural variations.

We evaluate LGN-AR on CIFAR-10, CIFAR-100, and MNIST under adversarial settings, including FGSM, PGD [2], and AutoAttack (AA). On **CIFAR-10** with a ResNet-18 backbone, the model achieves **79%** accuracy under FGSM ($\epsilon = 1/255$) and **35%** under PGD-10 ($\epsilon = 8/255$). On **CIFAR-100**, it sustains **49%** under FGSM and **17%** under PGD. On **MNIST**, the model maintains **over 97%** accuracy against PGD perturbations. Under **AutoAttack on CIFAR-10**, LGN-AR achieves **57%** accuracy for $\ell_2$-bounded perturbations ($\epsilon = 0.5$) and **23%** under $\ell_\infty$-bounded perturbations ($\epsilon = 8/255$). These results confirm that LGN-AR improves robustness across datasets and attack types while remaining computationally efficient. Gradient analyses further show improved perceptual alignment, with lower LPIPS [3] and better NIQE scores compared to standard models.

To assess generalisation, we conducted preliminary evaluations on ImageNet using both ResNet-18 and ViT architectures. With LGN-AR, ResNet-18 sustained **48%** top-1 accuracy under $\ell_\infty$ PGD-10 ($\epsilon = 4/255$), compared to **43%** for the baseline. On the ViT model, LGN-AR improved robustness under AutoAttack by **2–3%** across both $\ell_2$ and $\ell_\infty$ settings. These early results suggest that LGN-AR extends effectively beyond small-scale datasets, offering consistent robustness improvements across architectures.

Our findings highlight that perceptual priors from human vision [4] can act as effective regularizers for robustness. LGN-AR offers a lightweight alternative to adversarial training, scales across datasets and architectures, and can be combined with existing defences for further gains. *Future work will explore integrating LGN-AR with adversarial training to further enhance robustness on large-scale benchmarks.*

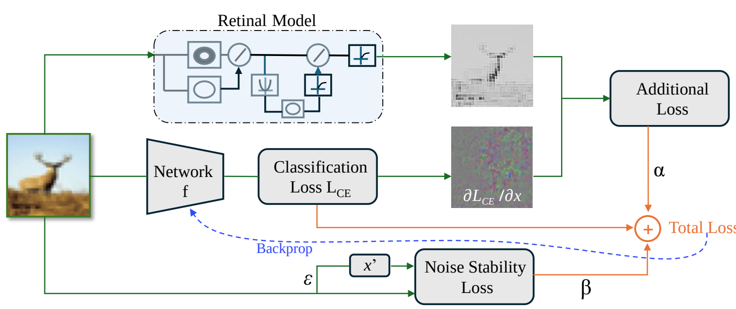


Figure 1: Overview of the LGN-AR framework combining cross-entropy loss, LGN-aligned gradient regularizer, and a noise stability term.

## References

1. Goodfellow, I., Shlens, J., & Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. In ICLR.
2. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards Deep Learning Models Resistant to Adversarial Attacks. In ICLR.
3. Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In CVPR.
4. Fan, J. E., Hernandez, J., Hsieh, P. J., Miller, K. D., & Simoncelli, E. P. (2021). Perceptual Straightening of Natural Videos. *Nature Neuroscience*.