Illusions as features: the generative side of recognition

Tahereh Toosi[1]*, **Kenneth D. Miller**[1]
[1]Center for Theoretical Neuroscience
Zuckerman Mind Brain Behavior Institute
Columbia University

Abstract

Visual illusions have long been considered perceptual mistakes, highlighting a perceived gap between biological and artificial vision. Here, we challenge this view by revealing that robust deep neural networks (DNNs) trained for object recognition implicitly contain a generative model capable of representing illusory contours and shapes. This finding suggests that illusions are not errors, but emergent properties of efficient visual processing. We uncover a mathematical correspondence between optimization for robust pattern recognition and optimization for pattern generation. This insight provides a potential explanation for how visual systems, primarily tuned for pattern recognition, can flexibly generate internal representations, including illusory percepts. Using a robust object recognition model (ResNet50) trained on ImageNet, we demonstrate that the propagated errors during inference approximate the gradient of log conditional probability $p(x|y)$, directly linking recognition error to learned priors. By repurposing the computational graph conventionally used for learning, we query this implicit generative model without additional optimization. When presented with classical illusion stimuli, our model generates representations that mirror perceptual experiences in biological vision. For a Kanizsa square input, edge-like patterns emerge in the perceived 'white square' area. With Rubin's vase, the network produces face-like or vase-like patterns depending on its training (VGGFace vs. ImageNet). These induced activities in early layers capture experimental findings of illusory contours and shapes in early visual areas across species. Our work reconciles the views of the visual cortex as both a pattern recognition and a generative model in a unified framework and clarifies the theoretical basis supporting the effectiveness of robust classifiers in producing stimuli that are perceptually-aligned, a method extensively employed in neuroscience. By demonstrating that robust pattern recognition networks inherently embody generative capabilities, we provide insights into how the brain might integrate prior knowledge with sensory input. This suggests that visual illusions, far from being mistakes, are indicators of the visual system's ability to generate and manipulate internal representations—a feature crucial for efficient visual processing in complex, ambiguous environments.

1 Introduction

Illusions are widely believed to be perceptual mistakes, and illusions strike a remarkable gap between humans and AI. Intriguingly, neurons actually elicited activity in response to induced activity very reliably in superficial cortical layers of early visual areas, between species, pointing to a shared

*Correspondence to: tahereh.toosi@columbia.edu

Workshop on Scientific Methods for Understanding Deep Learning, NeurIPS 2024.

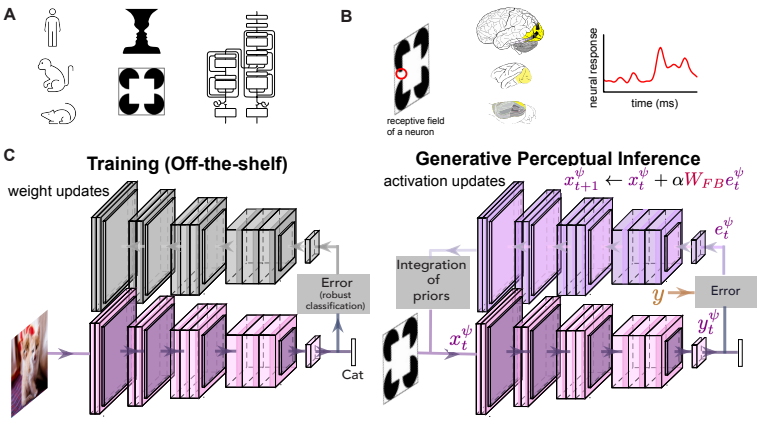

Figure 1: **A** Illusions highlight a striking gap between biological and artificial intelligence [1, 15] **B** Research in animals and humans suggests that activity in early visual areas represents the perceptual state when viewing a visual illusion, and these activities build up over time [9, 13] **C** Training network: Parameters of the feedforward DNN are tuned for robust object classification, e.g. 1000-way classification on Imagenet. Feedback network shown in gray is the error propagation network used in BackPropagation (BP). **C** Neural activation updates according to GPI in the same DNN. Activations x^{ψ}_t in the network are iteratively updated by the integrating the adjusted propagated error e^{ψ}_t through the feedback W_{FB}. Error can be computed as mean square difference to the pure sensory activation, or a target value in last layer both denoted by y. For simplicity, the normalizations for adjustment of the values are not included. α is the learning rate for activation updates

preserved neural mechanism between different brains at different scales and levels of intelligence (Figure 1).

On the other hand, it has long been hypothesized that the integration of perceptual priors with sensory inputs, known as perceptual inference, facilitates the brain's interpretation of ambiguous or complex stimuli by leveraging previously acquired knowledge, stored as internal models, to enhance current sensory processing. However, the neural mechanisms that implement such a generative internal model remain elusive. In contrast, the view of the brain as a pattern recognition system has been successful in predicting neural activity patterns using DNNs, offering a mechanistic mapping to the stage of processing along the ventral pathway [18]. However, these models struggle with degraded stimuli or visual illusions [5, 1]. Thus, although there exists models proposed to explain illusions [12], object recognition abilities [18], or integration of priors, there is no single model that explains all of them in a single framework backed by theory. Here, we show that viewing illusions as out-of-distribution stimuli, and perceptual integration of priors as a variant of solving the inverse problem in the score-based generative model embedded in robust classifier offers a unified image computable model explaining all three hallmarks of sensory processing in a unified framework.

We hypothesize that the network, during pattern recognition training, constructs an implicit generative internal model about the data distribution (e.g., natural image prior when trained on ImageNet). Using this internal model, the network can be queried for priors to aid perception when encountering degraded or unusual stimuli. Previous work empirically showed using the gradients in an adversarially trained DNNs, one can generate images [10]. Here, first we show the theory behind the link between recognition errors (gradients) and the score function (used to accumulate perceptual priors) and then show that using a simple variant of Langevin dynamics, illusory contours and shapes appear in early layers of a robust DNN.

2 Generative Inference in Pattern Recognition Networks

Theory: The Duality of Adversarial Training and Score-Based Generation We establish a fundamental connection between adversarial training of classifiers, as introduced by [10], and generative modeling through the lens of the Hyvärinen score [7]. We consider adversarial training as an effective way to arrive at a robust to input noise solution, as adversarial training optimizes to be

robust to worst-case input noise. We posit that adversarial training implicitly minimizes an analogue of the first term in the Hyvärinen score, effectively encouraging the model to learn a smoother probability distribution over the input space. Previously, [10, 14] empirically showed that a network trained for adversarial robustness exhibits properties often attributed to generative models; here, we provide the theoretical framework offering insights and extensions to the previous empirical results.

The Hyvärinen score [7], also known as the Hyvärinen loss function used in score-based generative modeling, is defined as:

$$H(p) = \int \left(\frac{1}{2} \|\nabla \log p(x)\|^2 + \Delta \log p(x) \right) dx \tag{1}$$

where $p(x)$ is the probability density function, ∇ is the gradient operator, and Δ is the Laplacian operator. Our focus is on the first term, that encourages the score function $\nabla \log p(x)$ to have small magnitude.

Adversarial training [10] aims to solve the following min-max problem:

$$\min_{\theta} \mathbb{E} \left[\max_{\delta \in S} L(\theta, x + \delta, y) \right] \tag{2}$$

where θ are the model parameters, x is the input, y is the true label, δ is the adversarial perturbation, S is the set of allowed perturbations, and L is the loss function.

The PGD adversarial training involves K iterative steps:

$$x_{t+1} = \Pi_S(x_t + \alpha \nabla_x L(\theta, x_t, y)) \tag{3}$$

where Π_S is projection onto the allowed perturbation set S, and α is the step size. After K steps, the model parameters are updated to minimize the loss at the adversarial point:

$$\min_{\theta} L(\theta, x_K, y) \tag{4}$$

Using Taylor expansion for each PGD step:

$$L(\theta, x_t + \alpha \nabla_x L(\theta, x_t, y), y) = L(\theta, x_t, y) + \alpha \|\nabla_x L(\theta, x_t, y)\|^2 + O(\alpha^2) \tag{5}$$

The cumulative effect over K steps, ignoring higher-order terms, is:

$$L(\theta, x_K, y) \approx L(\theta, x_0, y) + \alpha \sum_{t=0}^{K-1} \|\nabla_x L(\theta, x_t, y)\|^2 \tag{6}$$

where x_0 is the original input. Therefore, adversarial training implicitly minimizes:

$$\min_{\theta} \left(L(\theta, x_0, y) + \alpha \sum_{t=0}^{K-1} \|\nabla_x L(\theta, x_t, y)\|^2 \right) \tag{7}$$

For a well-trained classifier, the negative log-likelihood of the true class should approximate the loss:

$$L(\theta, x, y) \approx -\log p_\theta(y|x) \tag{8}$$

By Bayes' rule:

$$\log p_\theta(y|x) = \log p_\theta(x|y) + \log p(y) - \log p(x) \tag{9}$$

Therefore:

$$L(\theta, x, y) \approx -\log p_\theta(x|y) - \log p(y) + \log p(x) \tag{10}$$

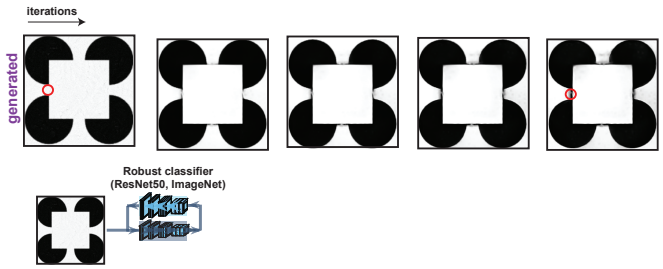

Figure 2: Generated inferred activity by GPI in response to Kanizsa square illusion in ResNet50, adversarially trained on 1000-way classification on ImageNet [10]. The generated activity mirrors the finding on animal studies in (Figure 1)

Taking the gradient with respect to x:

$$\nabla_x L(\theta, x, y) \approx -\nabla_x \log p_\theta(x|y) + \nabla_x \log p(x) \tag{11}$$

Since $\log p(y)$ is independent of x, its gradient is zero. For a robust classifier trained on natural images, we expect $\nabla_x \log p(x)$ to be small in regions of high data density. Therefore:

$$\nabla_x L(\theta, x, y) \approx -\nabla_x \log p_\theta(x|y) \tag{12}$$

Substituting this back into the adversarial training objective:

$$\min_\theta \left(L(\theta, x_0, y) + \alpha \sum_{t=0}^{K-1} \|\nabla_x \log p_\theta(x_t|y)\|^2 \right) \tag{13}$$

The second term now directly parallels the first term of the Hyvärinen score:

$$\frac{1}{2}\|\nabla_x \log p(x)\|^2 \tag{14}$$

This reveals that adversarial training implicitly learns the score function of the conditional data distribution $p(x|y)$. When averaged over classes, this approximates learning the score of the full data distribution:

$$\mathbb{E}_y[\|\nabla_x \log p(x|y)\|^2] \approx \|\nabla_x \log p(x)\|^2 \tag{15}$$

This shows that adversarial training implicitly minimizes an analogue of the first term in the Hyvärinen score. The second term that involves the divergence of the model's core function is computationally intensive and has been eliminated or approximated in the practical settings for high-dimensional data [17, 16]. The formal analogy we established to generative models explains perceptually aligned gradients in robust neural networks [14, 8, 3] and the subsequent widespread use of these networks in neuroscience to generate stimuli for human and animal experiments [6, 4, 2].

Generative perceptual inference Generative Perceptual Inference (GPI) is an instantiation of Langevian dynamics that leverages the implicit generative model within robust deep neural networks (DNNs) trained for object recognition. GPI operates by iteratively updating activations in the network's early layers, integrating sensory input with learned priors. One variant of GPI goes as follows: The process begins by presenting an input image x to the network, along with a noisy version x'. The difference between their latent representations, $(r' - r)$, is then propagated through the network's feedback connections to the input level. This propagated error is added back to x', and the process is repeated. Mathematically, this can be expressed as $x_{t+1} = x_t + \alpha W_{FB}(r'_t - r_t)$, where W_{FB} represents the feedback weights and α is a learning rate (Fig. 1). By repurposing the computational graph typically used for backpropagation during training, GPI enables the network

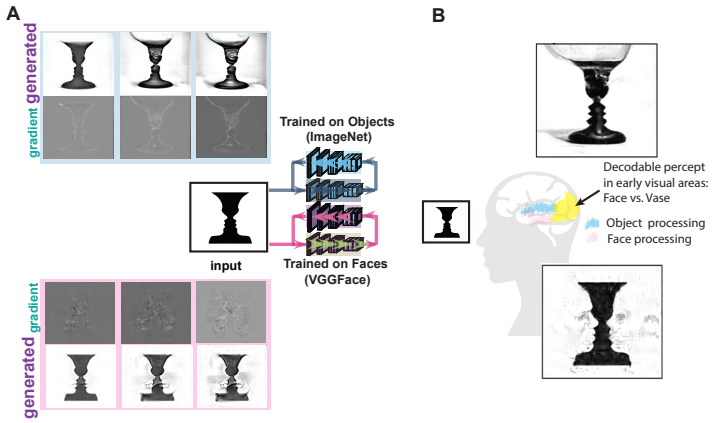

Figure 3: **A** Generated inferred activity by GPI in response to Rubin's vase illusion in two ResNet50, one trained for 1000-way robust classification on ImageNet, the other for 500-way robust face recognition on VGGface2. **B** The distinct generated activities in early visual layers replicate the finding in human brain imaging study [13]. Please note that the same architecture undergoing the same training but without robust training does not show these effects (results not shown here).

to query its learned priors without additional optimization. This mechanism allows robust DNNs to generate representations that mirror perceptual experiences in biological vision, particularly when presented with ambiguous or illusory stimuli such as the Kanizsa square or Rubin's vase. Induced activity in the early layers captured experimental findings on visual illusion representation in early visual areas [9, 11]. GPI confirms the critical role of feedback in integrating priors and inducing illusions.

3 Results

We tested our model on classic visual illusions, such as Kanizsa's square and Rubin's face-vase illusion. For the Kanizsa square, our model generated inferred activity that closely resembled the illusory contours observed in biological visual systems (Figure 2). The model's behavior mirrored the laminar-specific induced patterns observed in mice and monkeys, with distinct activation patterns in different cortical layers (L2/3, L4). In the case of the Rubin's face-vase illusion, we observed that the model's perception could be biased towards either faces or vases, depending on its training (Figure 3). When trained on a dataset of objects (e.g., ImageNet), the model tended to perceive the vase. Conversely, when trained on faces (e.g., VGGFace), it was more likely to perceive faces. This demonstrates how the model's learned priors influence its interpretation of ambiguous stimuli, similar to what is observed in human perception. The ability of robust neural networks in fast recognition was previously known activation by illusory contours and shapes in cortical lamina across species. This aligns with observations in biological visual systems, where prior knowledge and expectations help to maintain stable perception in the face of noisy or ambiguous sensory input.

4 Conclusion

We demonstrated that a pattern recognition model embodies a generative model that can be queried using GPI. This framework offers insights into how perceptual priors induce illusory contours and shapes in neural networks optimized for object recognition. Our work has several important implications for neuroscience and artificial intelligence. First, it provides a mechanistic explanation for how the brain might implement both fast recognition and flexible cognitive control within the same neural architecture. Second, it offers a new approach to building more brain-like AI systems that can seamlessly integrate bottom-up and top-down information processing. Finally, it suggests new avenues for investigating the neural basis of visual perception, potentially leading to more targeted experiments and interventions in future neuroscientific studies. By bridging the gap between fast pattern recognition and flexible integration of priors, it provides a more complete model of visual processing that aligns closely with observed phenomena in biological systems.

References

[1] N. Baker, G. Erlikhman, P. J. Kellman, and H. Lu. Deep convolutional networks do not perceive illusory contours. *Cognitive Science*, 2018.

[2] J. Feather, G. Leclerc, A. Madry, and J. H. McDermott. Model metamers reveal divergent invariances between biological and artificial neural networks. *Nature Neuroscience*, 26: 2017–2034, 2023. doi: 10.1038/s41593-023-01442-0. URL https://www.nature.com/articles/s41593-023-01442-0.

[3] R. Ganz, B. Kawar, and M. Elad. Do perceptually aligned gradients imply robustness? In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10628–10648. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/ganz23a.html.

[4] G. Gaziv, M. J. Lee, and J. J. DiCarlo. Robustified anns reveal wormholes between human category percepts. *arXiv preprint arXiv:2308.06887*, 2023. URL https://arxiv.org/abs/2308.06887.

[5] R. Geirhos, C. R. M. Temme, J. Rauber, H. H. Schütt, M. Bethge, and F. A. Wichmann. Generalisation in humans and deep neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 7549–7561, Red Hook, NY, USA, 2018. Curran Associates Inc.

[6] C. Guo, M. Lee, G. Leclerc, J. Dapello, Y. Rao, A. Madry, and J. DiCarlo. Adversarially trained neural representations are already as robust as biological neural representations. In *Proceedings of the 39th International Conference on Machine Learning*, pages 8072–8081. PMLR, 2022.

[7] A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005. URL http://jmlr.org/papers/v6/hyvarinen05a.html.

[8] S. Kaur, J. Cohen, and Z. C. Lipton. Are perceptually-aligned gradients a general property of robust classifiers? 2019.

[9] T. S. Lee and M. Nguyen. Dynamics of subjective contour formation in the early visual cortex. *Proceedings of the National Academy of Sciences*, 98(4):1907–1911, 2001. doi: 10.1073/pnas.98.4.1907.

[10] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[11] A. Pak, E. Ryu, C. Li, and A. A. Chubykin. Top-down feedback controls the cortical representation of illusory contours in mouse primary visual cortex. *The Journal of Neuroscience*, 40 (3):648–660, Dec. 2019. ISSN 1529-2401.

[12] Z. Pang, C. B. O'May, B. Choksi, and R. VanRullen. Predictive coding feedback results in perceived illusory contours in a recurrent neural network. *Neural Networks*, 144:164–175, 2021. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2021.08.024.

[13] L. Parkkonen, J. Andersson, M. Hämäläinen, and R. Hari. Early visual brain areas reflect the percept of an ambiguous scene. *Proceedings of the National Academy of Sciences*, 105(51): 20500–20504, 2008. doi: 10.1073/pnas.0810966105.

[14] S. Santurkar, D. Tsipras, B. Tran, A. Ilyas, L. Engstrom, and A. Madry. Image synthesis with a single (robust) classifier. 2019.

[15] H. S. Shahgir, K. S. Sayeed, A. Bhattacharjee, W. U. Ahmad, Y. Dong, and R. Shahriyar. IllusionVQA: A challenging optical illusion dataset for vision language models. 2024.

[16] Y. Song, S. Garg, J. Shi, and S. Ermon. Sliced score matching: A scalable approach to density and score estimation. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, pages 574–584. PMLR, 2019.

[17] P. Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

[18] D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, May 2014. ISSN 1091-6490. doi: 10.1073/pnas.1403112111.