From Images to Perception: Emergence of Perceptual Properties by Reconstructing Images

Pablo Hernández-Cámara[a*], **Jesus Malo**[a], **Valero Laparra**[a]
[a] Image Processing Lab, Universidad de Valencia, Paterna, Spain
[*] Corresponding author: pablo.hernandez-camara@uv.es

Abstract

A number of scientists suggested that human visual perception may emerge from image statistics, shaping efficient neural representations in early vision. In this work, a bio-inspired architecture that can accommodate several known facts in the retina-V1 cortex, the PerceptNet, has been end-to-end optimized for different tasks related to image reconstruction: autoencoding, denoising, deblurring, and sparsity regularization. Our results show that the encoder stage (V1-like layer) consistently exhibits the highest correlation with human perceptual judgments on image distortion despite not using perceptual information in the initialization or training. This alignment exhibits an optimum for moderate noise, blur and sparsity. These findings suggest that the visual system may be tuned to remove those particular levels of distortion with that level of sparsity and that biologically inspired models can learn perceptual metrics without human supervision.

Keywords: Perceptual Representation; Visual Perception; Bio-Inspired Models; Self-Supervised Learning; Autoencoder;

Introduction

On the one hand, following the classical Efficient Coding Hypothesis (H. B. Barlow et al., 1961; H. Barlow, 2001), many researchers have shown that certain behaviors of biological vision systems can be derived from the statistical regularities of natural images. Examples include: (1) color coding based on one achromatic and two chromatic broad-band spectral sensitivities (Buchsbaum & Gottschalk, 1983) and their associated nonlinearities (von der Twer & MacLeod, 2001; Laparra et al., 2012); (2) the emergence of achromatic and spatio-chromatic frequency analyzers (Olshausen & Field, 1996), including their bandwidth (Atick, Li, & Redlich, 1992), adaptation (Gutmann et al., 2014), and nonlinear responses (Schwartz & Simoncelli, 2001). On the other hand, autoencoders also capture the statistics of the images they are trained to reconstruct. Recent works have shown that denoising and deblurring autoencoders can reproduce the human contrast sensitivity function (Li et al., 2022) and exhibit humanlike color illusions (Gomez-Villa et al., 2020). Similarly, compression autoencoders develop non-Euclidean metrics aligned with human judgments of image distortion (Hepburn et al., 2022), while networks trained on low- and mid-level vision tasks also induce perceptually aligned distortion metrics (Kumar et al., 2022; Hernández-Cámara et al., 2025).

These findings raise the question: Can biologically inspired architectures of the early visual system, such as PerceptNet (Hepburn et al., 2020), learn perceptual distances without explicit perceptual supervision? In this work, we train PerceptNet on tasks including image reconstruction, denoising, deblurring, and sparsity regularization. We then analyze how these objectives influence the human alignment with perceptual judgments. Our results show that the strongest correlation with human evaluations arises at the V1 stage of PerceptNet. Notably, this alignment displays optimal values for moderate noise, blur, and sparsity levels.

By demonstrating that autoencoders can learn human-like perceptual properties, our study offers insights into both computational and neurobiological mechanisms of vision. Furthermore, it suggests that bio-inspired architectures may enable perceptual metrics that generalize across tasks without requiring perceptual human-labeled data.

Methods

We base our approach on PerceptNet, a biologically inspired model designed to mimic the visual pathway up to the primary visual cortex (Hepburn et al., 2020). To train the model in a self-supervised fashion, we implement an autoencoder architecture by using PerceptNet as an encoder and appending a PerceptNet inverse version as the decoder. This inverse model mirrors the original PerceptNet but replaces pooling operations with upsampling and divisive normalization with inverse divisive normalization, which performs multiplicative scaling instead of division. Both the encoder and decoder parameters are learned jointly during training.

We use approximately 200,000 natural images sampled from the ImageNet dataset to train the models. Each model is trained until convergence, with hyper-parameters adjusted for each objective to optimize performance. We train the model with four different objectives:

- Image reconstruction: Model trained to minimize the mean squared error (MSE) between input and reconstructed images.
- Denoising: Model trained to reconstruct the clean version of Gaussian noise-corrupted images, parameterized by the noise standard deviation (σ).
- Deblurring: Model trained to reconstruct the clean version of blurred images (by convolving them with a Gaussian kernel), parameterized by the standard deviation of the convolution kernel (σ).
- Sparsity: Model trained to reconstruct images while encouraging sparse representations by adding an L_1 penalty on the mean absolute value of the encoder activations. The sparsity is parameterized by scaling the L1 penalty with a hyperparameter (λ).

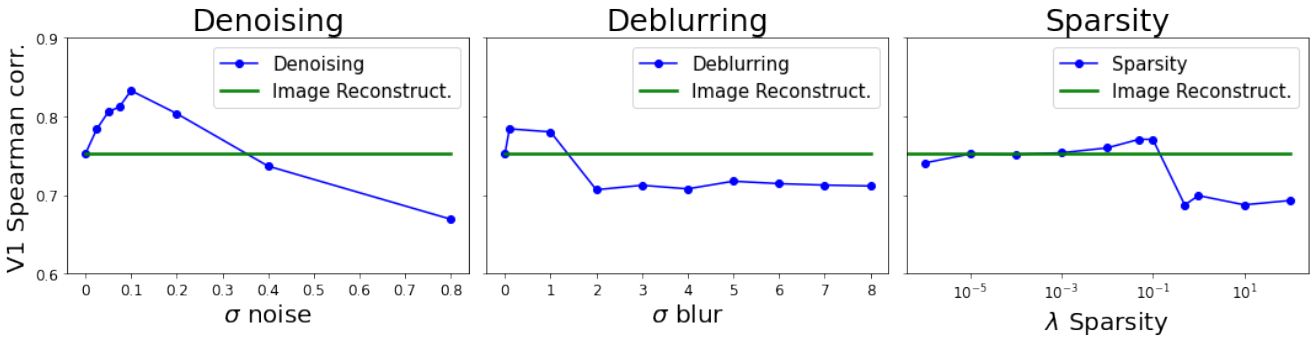

Figure 1: TID2013 Spearman correlation at the end of the encoder (V1-like layer) as a function of the different training parameters for denoising (left), deblurring (center) and sparsity (right).

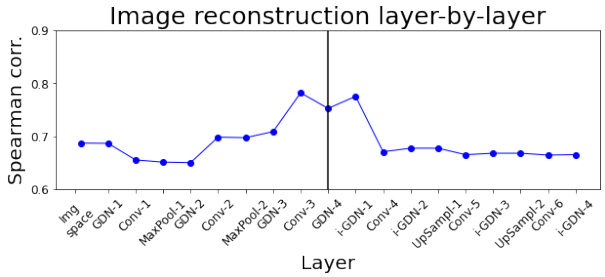

Figure 2: TID2013 Spearman correlation layer-by-layer when training the model to reconstruct natural images.

To evaluate perceptual alignment, we compute the correlation between differences in activations at each model layer and human Mean Opinion Score (MOS) from TID2013, a standard image quality assessment (IQA) database (Ponomarenko et al., 2015). This allows us to analyze how the different objectives affect the emergence of human-aligned perceptual representations at the different layers.

Results

We first examine how the correlation with human MOS varies across the model's layers under the simpler image reconstruction objective. Figure 2 reveals a clear peak in correlation in the final layers of the encoder, corresponding to the V1-like stage of PerceptNet.

We then analyze the effect of different training objectives by focusing on the encoder output. Figure 1 left shows how the correlation increases with noise level, reaching a maximum at approximately $\sigma = 0.1$, after which it declines. This suggests that moderate noise levels improve perceptual alignment, but excessive noise reduces it.

Figure 1 center shows that a small blur with $\sigma \leq 1$ produces an increase in correlation, but a stronger blur reduces it. This suggests that the model benefits from learning to reverse slight degradations, but heavy blur impairs its ability to align with human perception.

Finally, figure 1 right shows that although sparsity has less effect than the previous goals, a moderate sparsity improves correlation, but higher levels ($\lambda > 0.1$) reduce it. This indicates a trade-off where sparsity enhances representations up to a point but reduces performance if over-enforced.

These findings suggest that perception emerges from efficient coding strategies, where the brain balances information preservation with noise suppression. The non-monotonic effects indicate that perception is optimized through an intermediate level of regularization rather than extreme constraints.

Conclusions

Our study demonstrates that biologically inspired models can develop perceptual representations aligned with human vision through self-supervised learning alone, without perceptual supervision. Specifically, we show that PerceptNet, when trained as an autoencoder with appropriate regularization, exhibits emergent perceptual properties that strongly correlate with human judgments.

A key finding is that the highest alignment with human perception consistently arises at the encoder stage, which corresponds to V1 processing. This suggests that early visual representations in the brain may naturally reflect the statistical properties of the environment when optimized for reconstruction. Interestingly, moderate levels of noise, blur, and sparsity enhance this alignment, while excessive regularization reduces it. These results support that the visual system may be tuned to remove those particular levels of distortion with that level of sparsity and that biologically inspired models can learn perceptual metrics without human supervision.

Our findings complement and extend recent work showing that self-supervised models—such as denoising-deblurring autoencoders that replicate the human contrast sensitivity function (Li et al., 2022) or compression autoencoders that learn human-aligned non-Euclidean metrics (Hepburn et al., 2022)—can capture perceptual properties without explicit supervision. Moreover, our analysis of task-dependent alignment patterns resonates with studies demonstrating that networks trained for low- and mid-level vision tasks can induce humanlike distortion metrics (Hernández-Cámara et al., 2025).

By showing that biologically grounded architectures like PerceptNet can achieve similar alignment, our work provides further insight into the computational principles that may underlie biological vision. It also suggests that bio-inspired perceptual metrics could generalize across tasks and datasets, offering robust, interpretable models of perception without reliance on human annotations.

Acknowledgments

This work was supported in part by MCIN/AEI/FEDER/UE under Grants PID2020-118071GB-I00 and PID2023-152133NB-I00, by Spanish MIU under Grant FPU21/02256 and in part by Generalitat Valenciana under Projects GV/2021/074, CIPROM/2021/056, and by the grant BBVA Foundations of Science program: Maths, Stats, Comp. Sci. and AI (VIS4NN). Some computer resources were provided by Artemisa, funded by the EU ERDF through the Instituto de Física Corpuscular, IFIC (CSIC-UV).

References

Atick, J. J., Li, Z., & Redlich, A. N. (1992). Understanding retinal color coding from first principles. *Neural computation*, *4*(4), 559–572.

Barlow, H. (2001). Redundancy reductionrevisited. *Network: computation in neural systems*, *12*(3), 241.

Barlow, H. B., et al. (1961). Possible principles underlying the transformation of sensory messages. *Sensory communication*, *1*(01), 217–233.

Buchsbaum, G., & Gottschalk, A. (1983). Trichromacy, opponent colours coding and optimum colour information transmission in the retina. *Proceedings of the Royal society of London. Series B. Biological sciences*, *220*(1218), 89–113.

Gomez-Villa, A., et al. (2020). Color illusions also deceive cnns for low-level vision tasks: Analysis and implications. *Vision Research*, *176*, 156–174.

Gutmann, M. U., et al. (2014). Spatio-chromatic adaptation via higher-order canonical correlation analysis of natural images. *PloS one*, *9*(2), e86481.

Hepburn, A., et al. (2020). Perceptnet: A human visual system inspired neural network for estimating perceptual distance. In *2020 ieee international conference on image processing (icip)* (pp. 121–125).

Hepburn, A., et al. (2022). On the relation between statistical learning and perceptual distances. In *International conference on learning representations.* Retrieved from https://openreview.net/forum?id=zXM0b4hi5$_B$

Hernández-Cámara, P., et al. (2025). Dissecting the effectiveness of deep features as metric of perceptual image quality. *Neural Networks*, *185*, 107189.

Kumar, M., et al. (2022). Do better imagenet classifiers assess perceptual similarity better? *arXiv preprint arXiv:2203.04946*.

Laparra, V., et al. (2012). Nonlinearities and adaptation of color vision from sequential principal curves analysis. *Neural Computation*, *24*(10), 2751–2788.

Li, Q., et al. (2022). Contrast sensitivity functions in autoencoders. *Journal of Vision*, *22*(6), 8–8.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*(6583), 607–609.

Ponomarenko, N., et al. (2015). Image database tid2013: Peculiarities, results and perspectives. *Signal processing: Image communication*, *30*, 57–77.

Schwartz, O., & Simoncelli, E. P. (2001). Natural signal statistics and sensory gain control. *Nature neuroscience*, *4*(8), 819–825.

von der Twer, T., & MacLeod, D. I. (2001). Optimal nonlinear codes for theperception of natural colours. *Network: Computation in Neural Systems*, *12*(3), 395.