Neural Networks for Efficient Bayesian Decoding of Natural Images from Retinal Neurons

Nikhil Parthasarathy[*]
Stanford University
nikparth@gmail.com

Eleanor Batty[*]
Columbia University
erb2180@columbia.edu

William Falcon
Columbia University
waf2107@columbia.edu

Thomas Rutten
Columbia University
tkr2112@columbia.edu

Mohit Rajpal
Columbia University
mr3522@columbia.edu

E.J. Chichilnisky[†]
Stanford University
ej@stanford.edu

Liam Paninski[†]
Columbia University
liam@stat.columbia.edu

Abstract

Decoding sensory stimuli from neural signals can be used to reveal how we sense our physical environment, and is valuable for the design of brain-machine interfaces. However, existing linear techniques for neural decoding may not fully reveal or exploit the fidelity of the neural signal. Here we develop a new approximate Bayesian method for decoding natural images from the spiking activity of populations of retinal ganglion cells (RGCs). We sidestep known computational challenges with Bayesian inference by exploiting artificial neural networks developed for computer vision, enabling fast nonlinear decoding that incorporates natural scene statistics implicitly. We use a decoder architecture that first linearly reconstructs an image from RGC spikes, then applies a convolutional autoencoder to enhance the image. The resulting decoder, trained on natural images and simulated neural responses, significantly outperforms linear decoding, as well as simple point-wise nonlinear decoding. Additionally, the decoder trained on natural images performs nearly as accurately on a subset of natural stimuli (faces) as a decoder trained specifically for the subset, a feature not observed with a linear decoder. These results provide a tool for the assessment and optimization of retinal prosthesis technologies, and reveal that the neural output of the retina may provide a more accurate representation of the visual scene than previously appreciated.

1 Introduction

Neural coding in sensory systems is often studied by developing and testing encoding models that capture how sensory inputs are represented in neural signals. For example, models of retinal function are designed to capture how retinal ganglion cells (RGCs) respond to diverse patterns of visual stimulation. An alternative approach – decoding visual stimuli from RGC responses – provides a complementary method to assess the information contained in RGC spikes about the visual world [30, 37]. Understanding decoding can also be useful for the design of retinal prostheses, by providing a measure of the visual restoration that is possible with a prosthesis [25].

[*],[†]Equal contributions

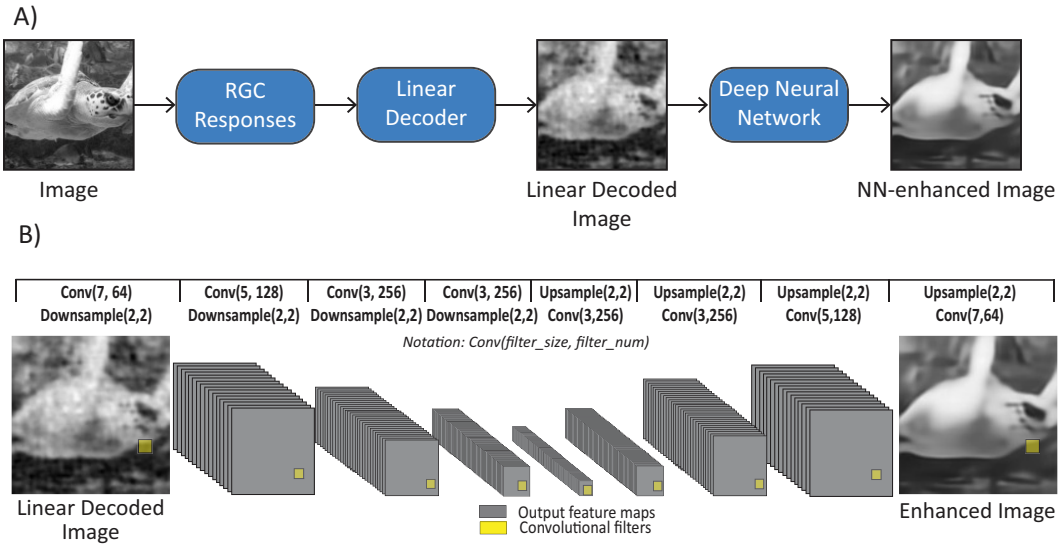

Figure 1: Outline of approach. A) The original image is fed through the simulated neural encoding models to produce RGC responses on which we fit a linear decoder. A deep neural network is then used to further enhance the image. B) We use a convolutional autoencoder with a 4 layer encoder and a 4 layer decoder to enhance the linear decoded image

The most common and well-understood decoding approach, linear regression, has been used in various sensory systems [28, 40]. This method was shown to be successful at reconstructing white noise temporal signals from RGC activity [37] and revealed that coarse structure of natural image patches could be recovered from ensemble responses in the early visual system [33]. Other linear methods such as PCA and linear perceptrons have been used to decode low-level features such as color and edge orientation from cortical visual areas [15, 5]. For more complex natural stimuli, computationally expensive approximations to Bayesian inference have been used to construct decoders that incorporate important prior information about signal structure [24, 26, 29]. However, despite decades of effort, deriving an accurate prior on natural images poses both computational and theoretical challenges, as does computing the posterior distribution on images given an observed neural response, limiting the applicability of traditional Bayesian inference.

Here we develop and assess a new method for decoding natural images from the spiking activity of large populations of RGCs, to sidestep some of these difficulties. Our approach exploits inference tools that approximate optimal Bayesian inference, and emerge from the recent literature on deep neural network (DNN) architectures for computer vision tasks such as super-resolution, denoising, and inpainting [18, 39]. We propose a novel staged decoding methodology – linear decoding followed by a (nonlinear) DNN trained specifically to enhance the images output by the linear decoder – and use it to reconstruct natural images from realistic simulated retinal ganglion cell responses. This approach leverages progress in DNN architecture to more fully incorporate natural image priors in the decoder. We show that the approach substantially outperforms linear decoding, and performs well on an important subclass of natural images (faces) without additional training. These findings provide a potential tool to assess the fidelity of retinal prostheses for treating blindness, and provide a substantially higher lower bound on how accurately real visual signals may be represented in the brain.

2 Approach

To decode images from spikes, we use a linear decoder to produce a baseline reconstructed image, then enhance this image using a more complex nonlinear model, namely a static nonlinearity or a DNN (Figure 1). There are a few reasons for this staged approach. First, it allows us to cast the decoding problem as a classic image enhancement problem that can directly utilize the computer vision literature on super-resolution, in-painting, and denoising. This is especially important for the

construction of DNNs, which remain nontrivial to tune for problems in non-standard domains (e.g., image reconstruction from neural spikes). Second, by solving the problem partially with a simple linear model, we greatly reduce the space of transformations that a neural network needs to learn, constraining the problem significantly.

In order to leverage image enhancement tools from deep learning, we need large training data sets. We use an encoder-decoder approach: first, develop a realistic encoding model that can simulate neural responses to arbitrary input images, constrained by real data. We build this encoder to predict the average outputs of many RGCs, but this approach could also be applied to encoders fit on a cell-by-cell basis [4]. Once this encoder is in hand, we train arbitrarily complex decoders by sampling many natural scenes, passing them through the encoder model, and training the decoder so that the output of the full encoder-decoder pipeline matches the observed image as accurately as possible.

2.1 Encoder model: simulation of retinal ganglion cell responses

For our encoding model, we create a static simulation of the four most numerous retinal ganglion cell types (ON and OFF parasol cells and ON and OFF midget cells) based on experimental data. We fit linear-nonlinear-Poisson models to RGC responses to natural scene movies, recorded in an isolated macaque retina preparation [8, 11, 13]. These fits produce imperfect but reasonable predictions of RGC responses (Figure 2 A). We averaged the parameters (spatial filter, temporal filter, and sigmoid parameters) of these fits across neurons, to create a single model for each of four cell types. To deal with static images, we then reduced these models to static models, consisting of one spatial filter followed by a nonlinearity and Poisson spike generation. The outputs of the static model are approximately equal to summing the spikes produced by the full model over the image frames of a pulse movie: gray frames followed by one image displayed for multiple frames. Spatial filters and the nonlinearity of the final encoding model are shown in Figure 2 B and C.

We then tiled the image space (128 x 128 pixels) with these simulated neurons. For each cell type, we fit a 2D Gaussian to the spatial filter of that cell type and then chose receptive field centers with a width equal to 2 times the standard deviation of the Gaussian fit rounded up to the nearest integer. The centers are shifted on alternate rows to form a lattice (Figure 2 D). The resulting response of each neuron to an example image is displayed in Figure 2 E as a function of its location on the image. The entire simulation consisted of 5398 RGCs.

2.2 Model architecture

Our decoding model starts with a classic linear regression decoder (LD) to generate linearly decoded images I^{LD} [37]. The LD learns a reconstruction mapping $\hat{\theta}$ between neural responses X and stimuli images I^{ST} by modeling each pixel as a weighted sum of the neural responses: $\hat{\theta} = (X^T X)^{-1} X^T I^{ST}$. X is augmented with a bias term in the first column. The model inputs are m images, p pixels and n neurons such that: $I^{ST} \in R^{m \times p}, X \in R^{m \times (n+1)}, \hat{\theta} \in R^{(n+1) \times p}$. To decode the set of neural responses X we compute the dot product between $\hat{\theta}$ and X: $I^{LD} = X\hat{\theta}$.

The next step of our decoding pipeline enhances I^{LD} through the use of a deep convolutional auto-encoder (CAE). Our model consists of a 4-layer encoder and a 4-layer decoder. This model architecture was inspired by similar models used in image denoising [12] and inpainting [35, 22]. In the encoder network E, each layer applies a convolution and downsampling operating to the output tensor of the previous layer. The output of the encoder is a tensor of activation maps representing a low-dimensional embedding of I^{LD}. The decoder network D inverts the encoding process by applying a sequence of upsampling and convolutional layers to the output tensor of the previous layer. This model outputs the reconstructed image I^{CAE}.

We optimize the CAE end-to-end through backpropagation by minimizing the pixelwise MSE between the output image of the CAE: $I^{CAE} = D(E(I^{LD}))$ and the original stimuli image I^{ST}.

The filter sizes, number of layers, and number of filters were all tuned through an exhaustive grid-search. Specific architecture details are provided in Figure 1.

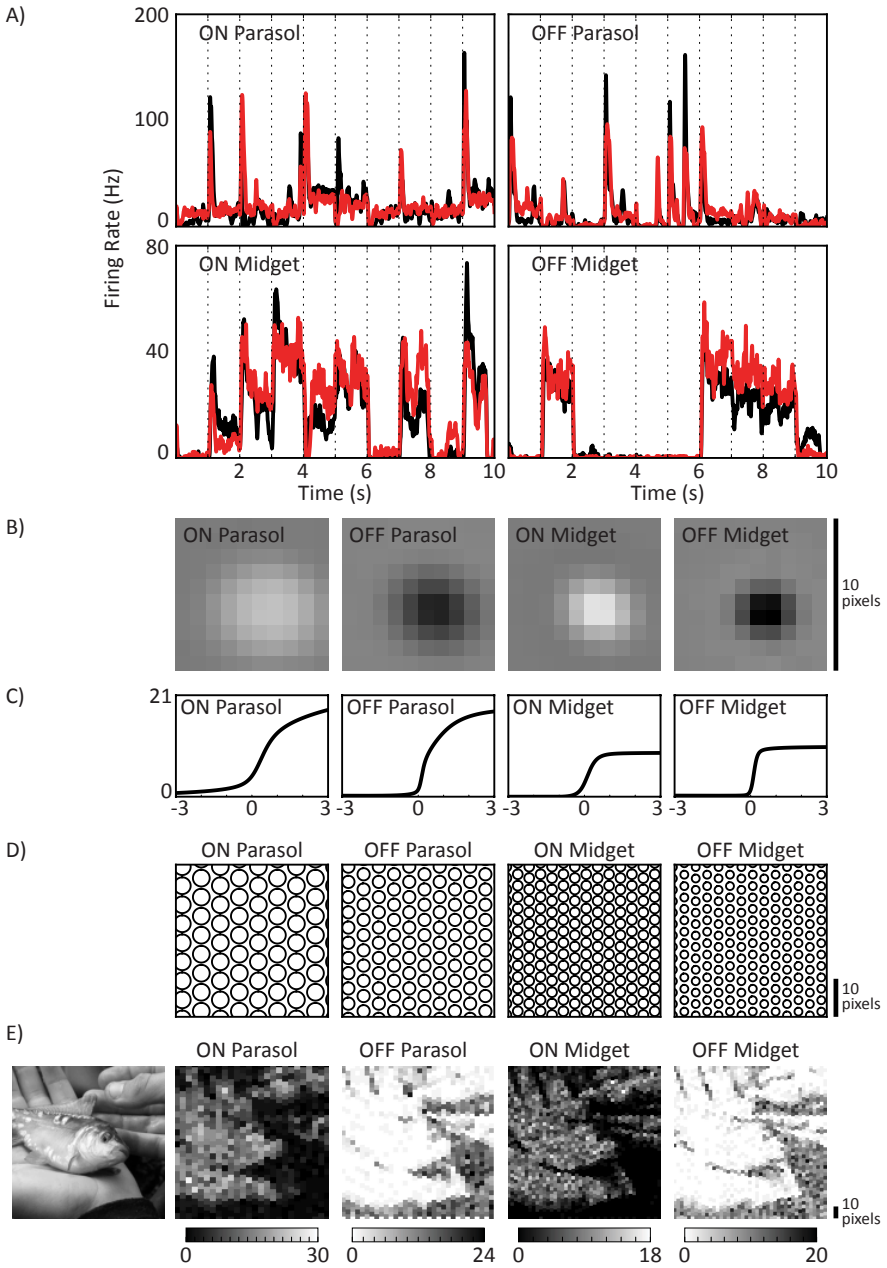

Figure 2: Encoding model. A) Full spatiotemporal encoding model performance on experimental data. Recorded responses (black) vs LNP predictions (red; using the averaged parameters over all cells of each type) for one example cell of each type. The spiking responses to 57 trials of a natural scenes test movie were averaged over trials and then smoothed with a 10 ms SD Gaussian. B) Spatial filters of the simulated neural encoding model are shown for each cell type. C) The nonlinearity following the spatial filter-stimulus multiplication is shown for each cell type. We draw from a Poisson distribution on the output of the nonlinearity to obtain the neural responses. D) Demonstration of the mosaic structure for each cell type on a patch of the image space. The receptive fields of each neuron are represented by the 1 SD contour of the Gaussian fit to the spatial filter of each cell type. E) The response of each cell is plotted in the square around its receptive field center. The visual stimulus is shown on the left. The color maps of ON and OFF cells are reversed to associate high responses with their preferred stimulus polarity.

2.3 Training and Evaluation

To train the linear decoder, we iterate through the training data once to collect the sufficient statistics $X^T X$ and $X^T I^{ST}$. We train the convolutional auto-encoder to minimize the pixelwise MSE P_{MSE} with the Adam optimizer [16] and a learning rate of 0.002. To avoid overfitting, we monitor P_{MSE} changes on a validation set over the course of two epochs. Specifically, we stop training if in epoch $E_i \in i = \{1, 2, ..., n\}$ the validation loss does not improve by more than $0.1\% E_{i-1}$.

In our experiments we use two image datasets, ImageNet [9] and the CelebA face dataset [21]. We apply preprocessing steps described previously in [18] to each image: **1)** Convert to gray scale, **2)** rescale to 256x256, **3)** crop the middle 128x128 region. From Imagenet we use 930k random images for training, 50K for validation, and a 10k held-out set for testing. We use ImageNet in all but one of our experiments - context-decoding. For the latter, we use the CelebA face dataset [21] with 160k images for training, 30k for validation, and a 10k held-out set for testing.

We evaluate all the models in our results using two separate metrics, pixelwise MSE and multiscale structural-similarity (SSIM) [36]. These metrics quantify pixel similarity and image structural similarity. Although each metric alone has known shortcomings, as a combination, they provide an objective evaluation of image reconstruction that is interpretable and well-understood.

Decoding models focused on minimizing MSE largely ignore higher level image information. In recent years, more complex perceptual similarity metrics (which minimize differences between high-level image features extracted by a pre-trained convolutional network) have gained traction in the deep learning community [20, 14]. In section 3.2 we experiment with a perceptual loss that draws on these ideas to encourage the CAE to extract finer detail from decoded images.

3 Results

3.1 ImageNet decoding

As expected [33], the linear decoder reconstructed blurry, noisy versions of the original natural images from the neural responses, a result that is attributable to the noisy responses from the RGCs down-sampling the input images. The CAE trained on the linear decoded images resulted in substantially improved reconstructions, perceptually and quantitatively (Figure 3). CAE decoding outperformed linear decoding both on average and for the vast majority of images, by both the MSE and $1-SSIM$ measures. Qualitatively, the improvements made by the CAE generally show increased sharpening of edges, adjustment of contrast, and smoothing within object boundaries that reduced overall noise. Similar improvement in decoding could not be replicated by utilizing static nonlinearities to transform the linear decoded output to the original images. We used a 6th degree polynomial fitted to approximate the relation between linearly decoded and original image pixel intensities, and then evaluated this nonlinear decoding on held out data. This approach produced a small improvement in reconstruction: 3.25% reduction in MSE compared to 27.06% for the CAE. This reveals that the improvement in performance with the CAE involves nonlinear image enhancement beyond simple remapping of pixel intensities.

3.2 Phase Scrambled Training

A possible explanation for the improved performance of the CAE compared to linear decoding is that it more fully exploits phase structure that is characteristic of natural images [3], perhaps by incorporating priors on phase structure that are not captured by linear decoding. To test this possibility, we trained both linear and CAE decoders on phase-scrambled natural images. The CAE input was produced by the linear decoder trained on the same image type as that CAE. Observed responses of RGCs to these stimuli followed approximately the same marginal distribution as responses to the original natural images. We then compared the performance of these linear and CAE decoders to the performance of the original decoders, on the original natural images (Figure 4). The linear decoders exhibited similar decoding performance when trained on the original and phase-scrambled images, while the CAE exhibited substantially higher performance when trained on real images. These findings are consistent with the idea that the CAE is able to capture prior information on image phase structure not captured by linear decoding.

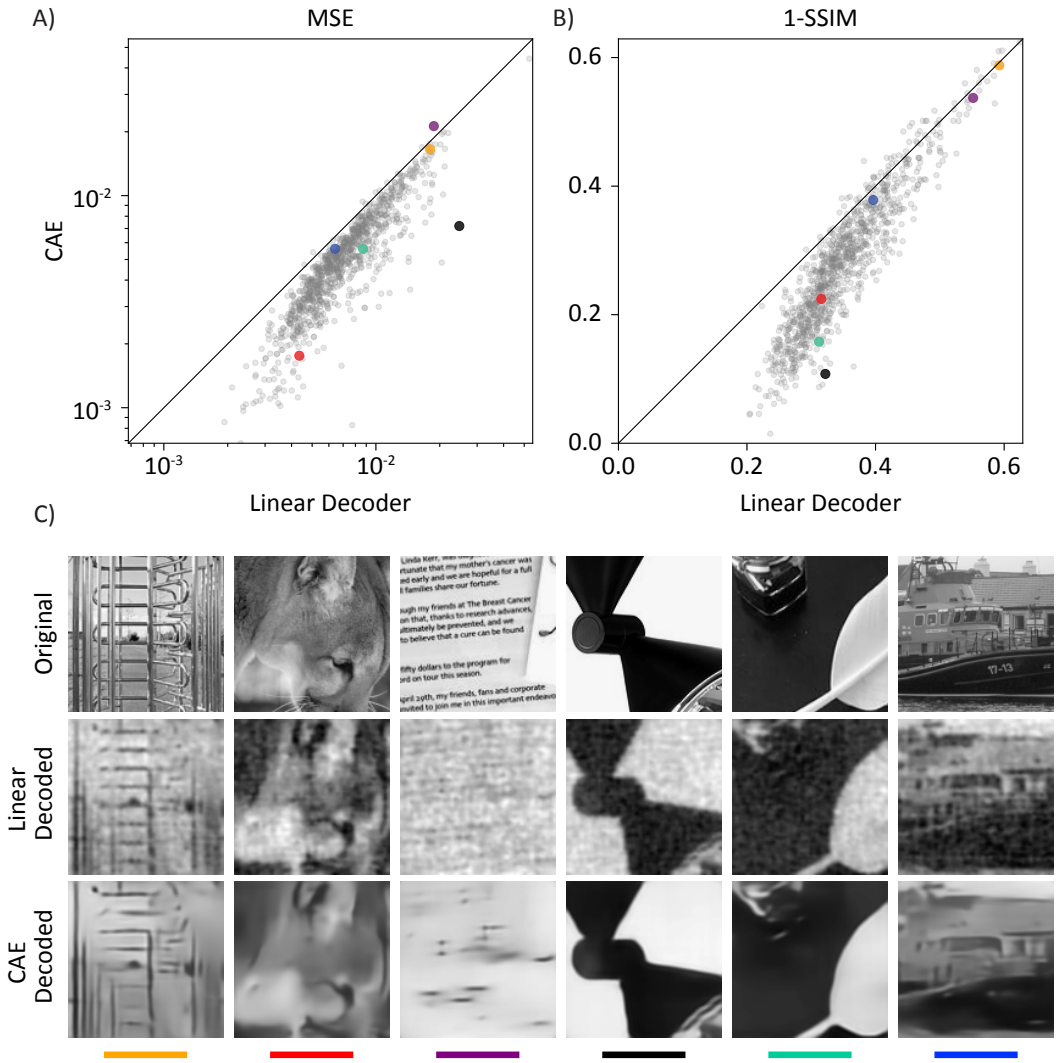

Figure 3: Comparison of linear and CAE decoding. A) MSE on a log-log plot for the ImageNet 10k example test set comparing the (Linear + CAE) model trained on ImageNet (only 1k subsampled examples are plotted here for visualization purposes). B) 1-SSIM version of the same figure. C) Example images from the test set show the original, linear decoded, CAE enhanced versions. The average (MSE, 1-SSIM) for LD over the full test set was $(0.0077, 0.35)$ and the corresponding averages for CAE were $(0.0056, 0.27)$.

3.3 Context Dependent Training

The above results suggest that the CAE is capturing important natural image priors. However, it remains unclear whether these priors are sufficient to decode specific classes of natural images as accurately as decoding models that are tuned to incorporate class-specific priors. We explored this in the context of human faces by fully re-training a class-specific CAE using the CelebA face dataset. Both linear and CAE models were trained from scratch (random initialization) using only this dataset. As with the phase scrambled comparisons, the CAE input is produced by the linear decoder trained on the same image type. We then compare these differently trained linear decoders and CAEs on a test set of CelebA faces. For the linear decoders, we see a 30% improvement in average test MSE and a 21 % improvement in 1-SSIM when training on CelebA as compared to training on ImageNet (Figure 5 A and C). We find that the differences in MSE and 1-SSIM between the differently trained CAEs are smaller (10% improvement in MSE and a 6% improvement in 1-SSIM) (Figure 5 B and

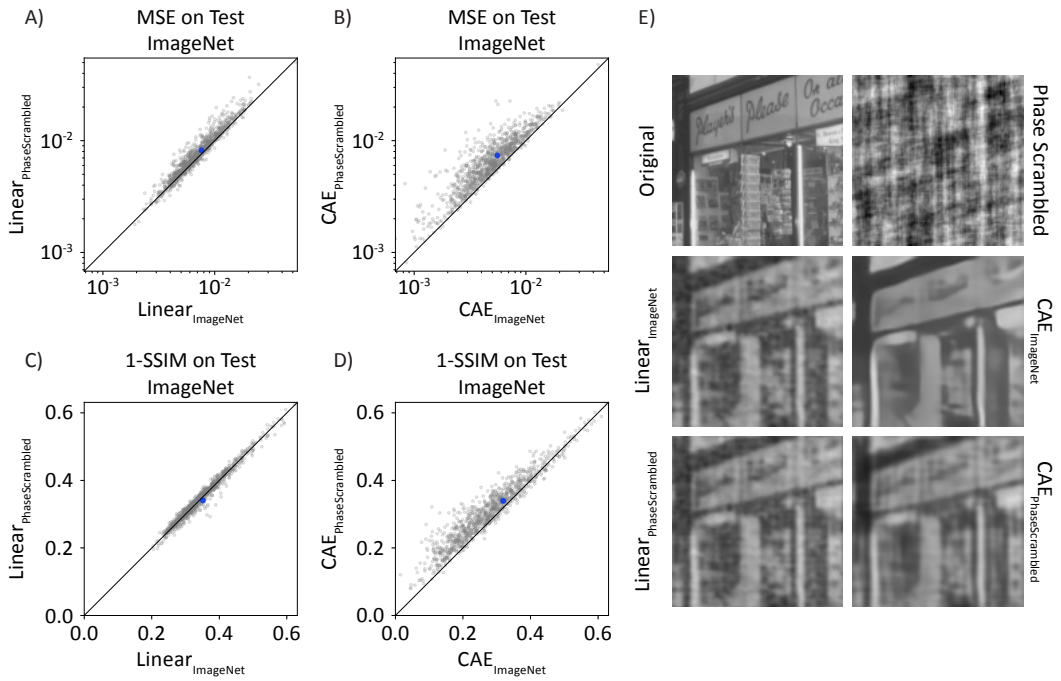

Figure 4: Comparison of phase scrambled and ImageNet trained models. A) MSE on log-log plot comparing the performance of the linear decoder fit on natural images to the linear decoder fit on phase scrambled images. The subscript of each model indicates the dataset on which it was trained. The reported MSE values are based on performance on the natural image test set (1k subsampled examples shown). B) Similar plot to A but comparing the CAE fit on natural images to the CAE fit on phase scrambled images. C) 1-SSIM version of A. D) 1-SSIM version of B. E) One example test natural image (represented by blue dot in A-D) showing the reconstructions from all 4 models and the phase scrambled version. The average (MSE, 1-SSIM) for $Linear_{PhaseScrambled}$ over the full ImageNet test set was $(0.0085, 0.35)$ and the corresponding averages for $CAE_{PhaseScrambled}$ were $(0.0074, 0.31)$.

D). The example in Figure 5E suggests that while the CAE trained on CelebA does capture some face-specific statistics better than the one trained on ImageNet, the latter is able to utilize the implicit general natural image prior to enhance the faces fairly well, with the exception of sometimes missing details of certain facial features (e.g. eye definition). This result suggests that given enough training data, the CAE trained on a large selection of natural images can be used to obtain reasonable results within a highly structured specific context.

3.4 Perceptual Loss Training

Next we experimented with training the decoder using a non-MSE-based loss. Multiple "perceptual losses" have been proposed in the image processing literature [42]. We design our loss function L_{MSE} as a weighed sum of pixel-wise MSE P_{MSE} and content-wise MSE V_{MSE}, with weighing parameters λ_1, λ_2:

$$L_{MSE} = \lambda_1 P_{MSE} + \lambda_2 V_{MSE}. \tag{1}$$

Here V_{MSE} measures the MSE distance between feature activation maps $\phi_{i,j}$ obtained from a pre-trained VGG16 [32] convolutional network. We feed the output of the CAE G through the (i, j)th layers of the VGG network to obtain $\phi_{i,j}(G(I^{CAE}))$ and feed the original stimuli image I^{ST} through the VGG to obtain $\phi_{i,j}(I^{ST})$. Our final perceptual component is:

$$V_{MSE} = \frac{1}{W_{i,j} H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} (\phi_{i,j}(G(I^{CAE}))_{x,y} - \phi_{i,j}(I^{ST})_{x,y})^2 \tag{2}$$

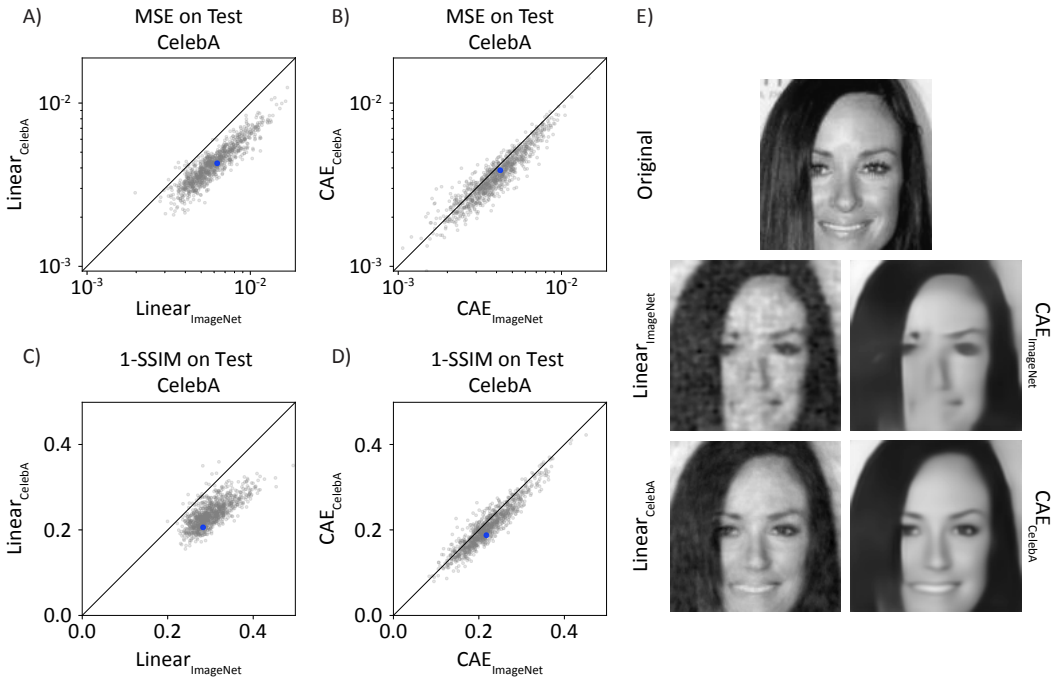

Figure 5: Comparison of CelebA and ImageNet trained models. A) MSE on log-log plot comparing the performance of the linear decoder fit on CelebA to the linear decoder fit on ImageNet. The subscript of each model indicates the dataset on which it was trained. The reported MSE values are based on performance on the natural image test set (1k subsampled examples shown). B) Similar plot to A but comparing the CAE fit on CelebA to the CAE fit on ImageNet. C) 1-SSIM version of A. D) 1-SSIM version of B. E) One example test natural image (represented by blue dot in A-D) showing the reconstructions from all 4 models. The average (MSE, 1-SSIM) for $Linear_{CelebA}$ over the full test set was $(0.0044, 0.24)$ and the corresponding averages for $Linear_{ImageNet}$ were $(0.0063, 0.30)$). The average (MSE, 1-SSIM) for CAE_{CelebA} over the full test set was $(0.0038, 0.21)$ and the corresponding averages for $CAE_{ImageNet}$ were $(0.0042, 0.22)$.

For our experiments we chose the loss weight coefficients $\lambda_1 = 0.3$ and $\lambda_2 = 0.7$ by trying a range of weight combinations (90/10, 80/20, 70/30) as proposed by [42]. For the VGG layers, we chose the $(i = 2, j = 2)$th layers as recommended by [14]. We chose a weighted combination of P_{MSE} and V_{MSE} to counter the square patterns or artifacts introduced throughout the image by using the V_{MSE} alone.

In figure 6 we compare the CAE trained on this perceptual loss ("VGG-CAE"), against the LD and the CAE trained on MSE ("MSE-CAE"). While the VGG-CAE outputs decoded images that look perceptually different from the MSE-CAE decoded images (VGG-CAE output tends to be less smoothed), it is not immediately clear that VGG-CAE is significantly improving on MSE-CAE perceptually here. (In fact, VGG-CAE obtains both higher mean MSE and 1-SSIM on the test set here.) We discuss these results in more depth below.

4 Discussion

The work presented here develops a novel approximate Bayesian decoding technique that uses non-linear DNNs to decode images from simulated responses of retinal neurons. The approach substantially outperforms linear reconstruction techniques that have usually been used to decode neural responses to high-dimensional stimuli.

Perhaps the most successful previous applications of Bayesian neural decoding are in cases where the variable to be decoded is low-dimensional. The work of [6] stimulated much progress in hippocampus and motor cortex using Bayesian state-space approaches applied to low-dimensional (typically

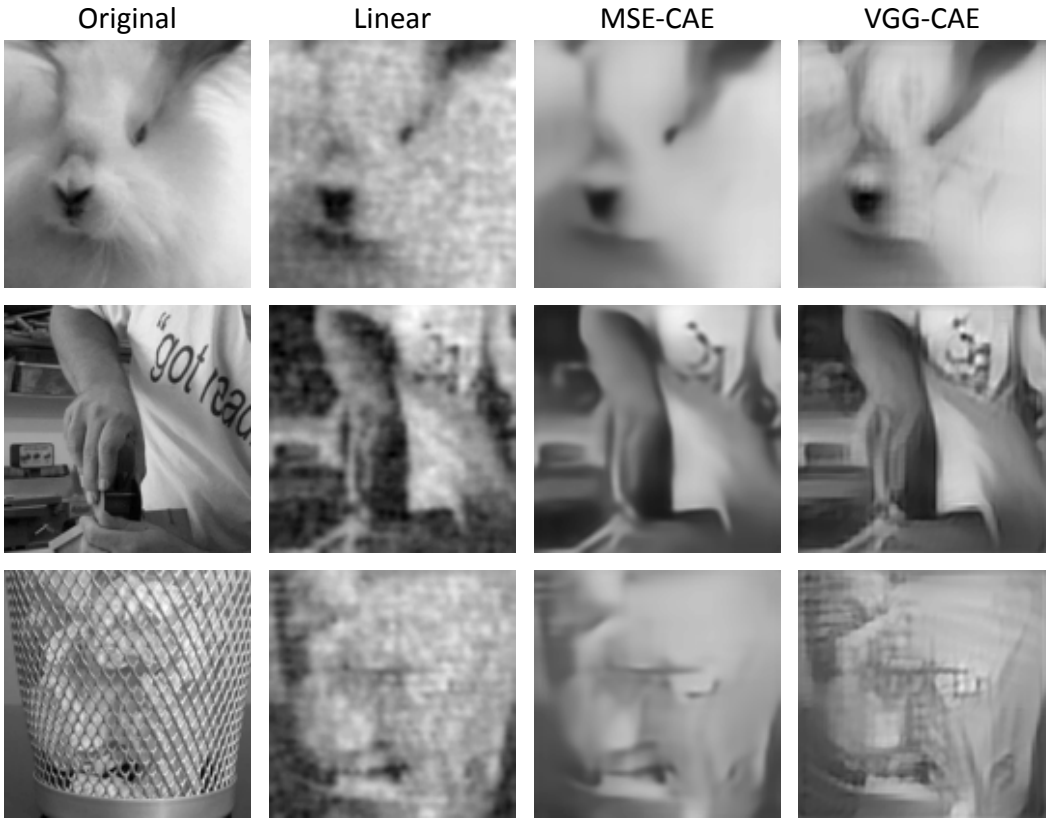

Figure 6: Perceptual loss CAE. Comparison of each model output on random test set images. MSE-CAE and VGG-CAE denote CAE trained on MSE and VGG-based perceptual loss, respectively.

two- or three-dimensional) position variables; see also [17] and [27] for further details. The low dimensionality of the state variable and simple Markovian priors leads to fast Bayesian computation in these models. At the same time, non-Bayesian approaches based on support vector regression [31] or recurrent neural networks [34] have also proven powerful in these applications.

Decoding information from the retina or early visual pathway requires efficient computations over objects of much larger dimensionality: images and movies. Several threads are worth noting here. First, some previous work has focused on decoding of flicker stimuli [37] or motion statistics [19, 23], both of which reduce to low-dimensional decoding problems. Other work has applied straightforward linear decoding methods [33, 10]. Finally, some work has tackled the challenging problem of decoding still images undergoing random perturbations due to eye movements [7, 1]. These studies developed approximate Bayesian decoders under simplified natural image priors, and it would be interesting in future work to examine potential extensions of our approach to those applications.

While our focus here has been on the decoding of spike counts from populations of neurons recorded with single-cell precision, the ideas developed here could also be applied in the context of decoding fMRI data. Our approach shares some conceptual similarity to previous work [24, 26] which used elegant encoding models combined with brute-force computation over a large discrete sample space to compute posteriors, and to other work [38] which used neural network methods similar to those developed in [41] to decode image features. Our approach, for example, could be extended to replace a brute-force discrete-sample decoder [24, 26] with a decoder that operates over the full high-dimensional continuous space of all images.

Many state-of-the-art models for in-painting and super-resolution image enhancement rely on generative adversarial networks (GANs). However, these models require very specific architecture tuning based on the exact problem structure. They have also been notably unstable during training due to problems of mode-collapse that are still a current topic of research [2]. Because our problem involves

some complex and unknown combination of denoising, super-resolution, and inpainting, we required a more robust model that could be tested with little hand-tuning. Furthermore, we have no parametric form for the noise in the linear decoded images, so standard pre-trained networks could not be applied directly. Based on previous work in [39], it seems that autoencoder architectures can robustly achieve reasonable results for these types of tasks without the issues that come with training complex deep learning models. Therefore, we chose the CAE architecture as a useful starting point. Nevertheless, GAN models have been shown to perform extremely well on problems such as patch in-painting or super-resolution with specifically tuned architectures, so we plan to explore these networks in future work.

Along with the architecture, the chosen loss function can have a large impact on the resulting decodings. While we did not extensively explore this aspect, preliminary analysis shows that the MSE loss and perceptual loss lead to noticeably different outputs, each with advantages and disadvantages. Loss functions for image-related neural network training are an active area of research, and we plan to further explore this area. Importantly, any retinal prosthetics application of this work would require decoding of visual scenes that is accurate by perceptual metrics rather than MSE.

We have shown improved reconstruction based on simulated data but an important next step is to move to more realistic simulated data, and then eventually real data. In addition, we have shown better CAE reconstruction only based on one perfect mosaic of the simulated neurons. In reality, these mosaics differ from retina to retina and there are gaps in the mosaic when we record from retinal neurons. Therefore, it will be important to investigate whether the CAE can learn to generalize over different mosaic patterns. Additional directions of future work include reconstruction of movies and color images.

The present results have two implications for visual neuroscience. First, the results provide a framework for understanding how an altered neural code, such as the patterns of activity elicited in a retinal prosthesis, could influence perception of the visual image. With our approach, this can be assessed in the image domain directly (instead of the domain of spikes) by examining the quality of "optimal" reconstruction from electrical activity induced by the prosthesis. Second, the results provide a way to understand which aspects of natural scenes are effectively encoded in the natural output of the retina, again, as assessed in the image domain. Previous efforts toward these two goals have relied on linear reconstruction. The substantially higher performance of the CAE provides a more stringent assessment of prosthesis function, and suggests that the retina may convey visual images to the brain with higher fidelity than was previously appreciated.

References

[1] Alexander G Anderson, Bruno A Olshausen, Kavitha Ratnam, and Austin Roorda. A neural model of high-acuity vision in the presence of fixational eye movements. In *Signals, Systems and Computers, 2016 50th Asilomar Conference on*, pages 588–592. IEEE, 2016.

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.

[3] Elizabeth Arsenault, Ahmad Yoonessi, and Curtis Baker. Higher order texture statistics impair contrast boundary segmentation. *Journal of vision*, 11(10):14–14, 2011.

[4] Eleanor Batty, Josh Merel, Nora Brackbill, Alexander Heitman, Alexander Sher, Alan Litke, E.J. Chichilnisky, and Liam Paninski. Multilayer recurrent network models of primate retinal ganglion cell responses. *International Conference on Learning Representations*, 2017.

[5] Gijs Joost Brouwer and David J. Heeger. Decoding and reconstructing color from responses in human visual cortex. *The Journal of Neuroscience: the official journal of the Society for Neuroscience*, 29(44):13992–14003, 2009.

[6] Emery N Brown, Loren M Frank, Dengda Tang, Michael C Quirk, and Matthew A Wilson. A statistical paradigm for neural spike train decoding applied to position prediction from ensemble firing patterns of rat hippocampal place cells. *Journal of Neuroscience*, 18(18):7411–7425, 1998.

[7] Yoram Burak, Uri Rokni, Markus Meister, and Haim Sompolinsky. Bayesian model of dynamic image stabilization in the visual system. *Proceedings of the National Academy of Sciences*, 107(45):19525–19530, 2010.

[8] E.J. Chichilnisky. A simple white noise analysis of neuronal light responses. *Network: Computation in Neural Systems*, 12(2):199–213, 2001.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[10] Ariadna R. Diaz-Tahoces, Antonio Martinez-Alvarez, Alejandro Garcia-Moll, and Eduardo Fernandez. Towards the reconstruction of moving images by populations of retinal ganglion cells. In *6th International Work-Conference on the Interplay Between Natural and Artificial Computation, IWINAC*, volume 9107, 2015.

[11] ES Frechette, A Sher, MI Grivich, D Petrusca, AM Litke, and EJ Chichilnisky. Fidelity of the ensemble code for visual motion in primate retina. *Journal of neurophysiology*, 94(1):119–135, 2005.

[12] Lovedeep Gondara. Medical image denoising using convolutional denoising autoencoders. *arXiv pre-print 1608.04667*, 2016.

[13] Alexander Heitman, Nora Brackbill, Martin Greschner, Alexander Sher, Alan M Litke, and EJ Chichilnisky. Testing pseudo-linear models of responses to natural scenes in primate retina. *bioRxiv*, page 045336, 2016.

[14] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016.

[15] Yukiyasu Kamitani and Frank Tong. Decoding the visual and subjective ontents of the human brain. *Nature Neuroscience*, 8(5):679–685, 2005.

[16] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[17] Shinsuke Koyama, Lucia Castellanos Pérez-Bolde, Cosma Rohilla Shalizi, and Robert E Kass. Approximate methods for state-space models. *Journal of the American Statistical Association*, 105(489):170–180, 2010.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

[19] Edmund C Lalor, Yashar Ahmadian, and Liam Paninski. The relationship between optimal and biologically plausible decoding of stimulus velocity in the retina. *JOSA A*, 26(11):B25–B42, 2009.

[20] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802*, 2016.

[21] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[22] Xiao-Jiao Mao, Chunhua Shen, and Yu-Bin Yang. Image restoration using convolutional auto-encoders with symmetric skip connections. In *Advances in Neural Information Processing*, 2016.

[23] Olivier Marre, Vicente Botella-Soler, Kristina D Simmons, Thierry Mora, Gašper Tkačik, and Michael J Berry II. High accuracy decoding of dynamical motion from a large retinal population. *PLoS Comput Biol*, 11(7):e1004304, 2015.

[24] Thomas Naselaris, Ryan J. Prenger, Kendrick N. Kay, Michael Oliver, and Jack L. Gallant. Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63(9):902–915, 2009.

[25] Sheila Nirenberg and Chetan Pandarinath. Retinal prosthetic strategy with the capacity to restore normal vision. *PNAS*, 109(37), 2012.

[26] Shinji Nishimoto, An T. Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L. Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 21(19):1641–1646, 2011.

[27] Liam Paninski, Yashar Ahmadian, Daniel Gil Ferreira, Shinsuke Koyama, Kamiar Rahnama Rad, Michael Vidne, Joshua Vogelstein, and Wei Wu. A new look at state-space models for neural data. *Journal of computational neuroscience*, 29(1-2):107–126, 2010.

[28] Brian N. Pasley, Stephen V. David, Nima Mesgarani, Adeen Flinker, Shibab A. Shamma, Nathan E. Crone, Robert T. Knight, and Edward F. Chang. Reconstructing speech from human auditory cortex. *PLOS Biology*, 10(1), 2012.

[29] Alexandro D Ramirez, Yashar Ahmadian, Joseph Schumacher, David Schneider, Sarah M. N. Woolley, and Liam Paninski. Incorporating naturalistic correlation structure improves spectrogram reconstruction from neuronal activity in the songbird auditory midbrain. *Journal of Neuroscience*, 31(10):3828–3842, 2011.

[30] Fred Rieke, Davd Warland, Rob de Ruyter van Steveninck, and William Bialek. *Spikes: Exploring the Neural Code*. MIT Press, Cambridge, MA, USA, 1999.

[31] Lavi Shpigelman, Hagai Lalazar, and Eilon Vaadia. Kernel-arma for hand tracking and brain-machine interfacing during 3d motor control. In *Advances in neural information processing systems*, pages 1489–1496, 2009.

[32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[33] Garrett B. Stanley, Fei F. Li, and Yang Dan. Reconstruction of natural scenes from ensemble responses in the lateral geniculate nucleus. *Journal of Neuroscience*, 19(18):8036–8042, 1999.

[34] David Sussillo, Sergey D Stavisky, Jonathan C Kao, Stephen I Ryu, and Krishna V Shenoy. Making brain–machine interfaces robust to future neural variability. *Nature Communications*, 7, 2016.

[35] Zhangyang Wang, Yingzhen Yang, Zhaowen Wang, Shiyu Chang, Wen Han, Jianchao Yang, and Thomas S. Huang. Self-tuned deep super resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015.

[36] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[37] David K. Warland, Pamela Reinagel, and Markus Meister. Decoding visual information from a population of retinal ganglion cells. *Journal of neurophysiology*, 78(5):2336–2350, 1997.

[38] Haiguang Wen, Junxing Shi, Yizhen Zhang, Kun-Han Lu, and Zhongming Liu. Neural encoding and decoding with deep learning for dynamic natural vision. *arXiv pre-print 1608.03425*, 2016.

[39] Junyuan Xie, Linli Xu, and Enhong Chen. Image denoising and inpainting with deep neural networks. In *Advances in Neural Information Processing Systems*, pages 341–349, 2012.

[40] Kai Xu, Yueming Wnag, Shaomin Zhang, Ting Zhao, Yiwen Wang, Weidong Chen, and Xiaoxiang Zhang. Comparisons between linear and nonlinear methods for decoding motor cortical activities of monkey. In *Engineering in Medicine and Biology Society, EMBC, Annual International Conference of the IEEE*, 2011.

[41] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.

[42] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for neural networks for image processing. *arXiv preprint arXiv:1511.08861*, 2015.