Maximum *a posteriori* natural scene reconstruction from retinal ganglion cells with deep denoiser priors

Eric Gene Wu
Stanford University
wueric@stanford.edu

Alexander Sher
University of California, Santa Cruz

Alan M. Litke
University of California, Santa Cruz

Eero P. Simoncelli
Center for Neural Science, and
Courant Inst. of Mathematical Sciences,
New York University
Flatiron Institute, Simons Foundation
eero.simoncelli@nyu.edu

E.J. Chichilnisky
Stanford University
ej@stanford.edu

Abstract

A fraction of the visual information arriving at the retina is transmitted to the brain by signals in the optic nerve, and the brain must rely solely on these signals to make inferences about the visual world. Previous work has probed the visual information contained in retinal signals by reconstructing images from retinal activity using linear regression and nonlinear regression with neural networks. Maximum *a posteriori* (MAP) reconstruction offers a more general and principled approach. We develop a novel method for approximate MAP reconstruction by combining a generalized linear model of light responses in retinal neurons and their dependence on spike history and spikes of neighboring cells, with an image prior implicitly embedded in a deep convolutional neural network trained for image denoising. We use this method to reconstruct natural images from *ex vivo* simultaneously-recorded spikes of hundreds of ganglion cells uniformly sampling a region of the retina. The method produces reconstructions that match or exceed the state-of-the-art in perceptual similarity and exhibit additional fine detail, while using substantially fewer model parameters than previous approaches. The use of more rudimentary encoding models (a linear-nonlinear-Poisson cascade) or image priors (a 1/F spectral model) significantly reduces reconstruction performance, indicating the essential role of both components in achieving high-quality reconstructed images from the retinal signal.

1 Introduction

A torrent of visual information arrives at each of our eyes, but only a small portion of it is transmitted to the brain via the optic nerve, which is comprised of the axons of the retinal ganglion cells (RGCs). Elucidating the nature of this encoded information, and the inference process the brain uses to interpret it, is fundamental to understanding biological vision. Image reconstruction provides a method of visualizing the information encoded in RGC signals, evaluating it using standard image quality metrics, and reasoning about how the brain might interpret it [1, 2, 3, 4]. The fidelity and quality of reconstructed images also provides a useful objective function for optimizing the design of electrical stimulation patterns delivered by devices implanted to restore vision [5, 6].

The simplest and most well-studied image reconstruction method is linear regression [1, 3, 7]. Optimal reconstruction kernels are learned for each RGC using least-squares regression of recorded responses to many visual images, and the reconstruction of a new incident image is computed with the sum of the filters weighted by the response of each cell. The quality of linearly reconstructed images can be enhanced by applying an autoencoder neural network to leverage natural image priors [8], or by using deep neural networks to non-linearly recover additional high spatial frequency image components [4]. Neural networks can also be directly trained (supervised) for reconstruction, but this is data-intensive, and to date has limited their use to simulated data, or low-dimensional stimuli and small numbers of cells [9]. These regression approaches leave substantial room for improvement and interpretation. A Bayesian formulation, in which encoding model and prior probabilities are made explicit and are separately fitted, could provide a more flexible and interpretable solution, and could potentially improve the fidelity of reconstructed images.

Here we present a method for approximate maximum *a posteriori* (MAP) image reconstruction from RGC spikes, that combines a retinal encoding model that accurately captures retinal responses [10] with state-of-the-art image priors that are implicitly embedded in deep denoising networks [11, 12, 13, 14, 15, 16]. By separating the effects of image prior and retinal spiking response likelihood, our method offers two primary advantages over existing methods for reconstruction: (1) any pre-trained or closed-form natural image prior can be used, and the effects of different priors can be compared; and (2) any model of RGC encoding that provides an explicit likelihood can be used, and the method can quantify the relative importance of different model components in representing the visual signal, including spike train history, cell-to-cell correlations and output nonlinearities. We apply our method to reconstruct static flashed natural images from responses of several hundred macaque RGCs of identified types recorded with large-scale electrode arrays. We compare our method directly to published state-of-the-art linear and neural network regression methods (Section 4.1). The new method matches or significantly outperforms previous methods, producing sharper, more naturalistic reconstructions (Figure 2), and similar or greater perceptual similarity to ground truth (Figure 3, Tables 1 and 2). However, our method also produces some reconstructions with distinctive spurious image structure, as would be expected when RGC signals are noisy and image priors dominate the reconstruction process. Finally, comparisons to more conventional encoding models and image priors reveal that both aspects of the approach are important for the most accurate reconstructions.

2 Retinal data and stimuli

Extracellular recordings from RGCs in the peripheral macaque retina were performed *ex vivo* using a 512-electrode system [17] as described previously [3]. Retinas were obtained from terminally anesthetized macaque monkeys used by other laboratories, in accordance with Institutional Animal Care and Use Committee requirements. Spikes from individual RGCs were identified with the YASS [18] spike-sorter. A 30-minute spatiotemporal white noise stimulus [19] was used to compute spatio-temporal receptive fields, and to identify cells of distinct types. Analysis focused on the four major RGC types of the primate retina (ON parasol, OFF parasol, ON midget, and OFF midget) [20, 21], totaling roughly 700 cells per recording. The receptive fields [22] of all four cell types formed regular mosaics, uniformly covering a region of visual space (Figure 1).

Natural images were presented to the retina as described previously [3]. Images from the ImageNet database [23] were converted to grayscale and cropped to 256x160. Each pixel measured approximately 11×11 μm at the retina. Each stimulus image was displayed for 100 ms, followed by a 400 ms uniform gray display, allowing each image presentation to be treated as an independent trial. This trial design does not fully mimic natural vision because it does not account for eye movements [24], and because the temporal component of the stimulus is known to the reconstruction algorithm. Two retinas from different animals were used, with 19,000 and 10,000 image/response pairs, respectively. Details are summarized in Tables 4 and 5 in the Appendix.

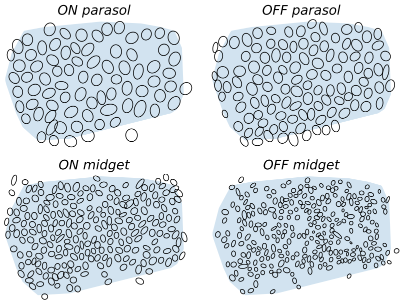

Figure 1: Receptive field mosaics from one retina for the four major RGC types (ON parasol, OFF parasol, ON midget, OFF midget) used in the image reconstructions. Image quality metrics were computed over the shaded blue region, to exclude areas that were insufficiently covered by receptive fields of recorded RGCs.

3 MAP image reconstruction from RGC spikes

MAP reconstruction estimates the stimulus image $\mathbf{x}$ from observed RGC spike trains $\mathbf{s}$ by minimizing the negative log of the posterior, $-\log p(\mathbf{x} \mid \mathbf{s})$, which can be expressed using Bayes' Rule as:

$$\hat{\mathbf{x}}(\mathbf{s}) = \arg\min_{\mathbf{x}} \left\{ -\log p(\mathbf{s} \mid \mathbf{x}) - \log p(\mathbf{x}) \right\}. \tag{1}$$

Both terms in equation (1) have intuitive interpretations in the context of reconstruction from RGC spikes. The first, $-\log p(\mathbf{s} \mid \mathbf{x})$, is the negative log likelihood (NLL) of an encoding model describing the probabilistic spiking of RGCs given a stimulus. The parameters of this model can be learned from experimental data. The second is the negative log prior of the stimulus image $\mathbf{x}$ and can be learned from natural images independently of retinal responses. Because encoding models with varying levels of fidelity and detail can be mixed and matched with priors of varying sophistication, the MAP approach allows us to probe the distinct roles of these two components in image reconstruction [25, 26].

3.1 RGC encoding models

Encoding models for each RGC must be learned from the experimental data before performing MAP reconstruction. Two types of encoding models were fitted to the data: (1) a linear-nonlinear-Poisson (LNP) cascade model with an exponential nonlinearity, the most commonly used model of RGC responses to visual stimuli [7]; and (2) a generalized linear model (GLM) that augments the LNP model with a feedback loop and cross-connections between neighboring cells, and which can accurately capture fine spike timing structure and cell-cell correlations [10]. Encoding models were fitted on the entire training partition, and regularization hyperparameters were tuned by evaluating the test partition NLL for a small subset of RGCs of each type.

Linear-nonlinear-Poisson (LNP) encoding model The LNP model is the *de facto* standard model for describing the probabilistic spiking of RGCs in response to visual stimuli [7]. The model parameters for a single RGC consist of a linear spatial filter $\mathbf{m}$, and a scalar bias b. In the model, a scalar generator signal $\mathbf{m}^T\mathbf{x} + b$ is passed through a nonlinearity to compute a spike rate λ. The RGC spike count in a 150 ms interval is modeled as a draw from a Poisson distribution with rate λ. Assuming an exponential nonlinearity, the encoding NLL for a single RGC is

$$-\log p(s \mid \mathbf{x}) = \exp\{\mathbf{m}^T\mathbf{x} + b\} - s(\mathbf{m}^T\mathbf{x} + b), \tag{2}$$

which is convex in $\mathbf{m}$ and b. In practice, to ensure that the spatial filters were spatially compact and corresponded approximately with the receptive fields obtained using reverse correlation with white noise, the MAP objective was augmented with two regularization terms: an L1 sparsity-inducing penalty, and an L2 penalty enforcing similarity to the receptive field. The complete objective function with both regularization terms is described in A.4. The parameter optimization was solved separately for each cell using FISTA [27].

MAP reconstruction requires the joint encoding NLL of the observed spikes from every RGC given the image. Since the LNP model assumes that the RGC responses are statistically independent, this is simply the sum of the single-cell NLLs, which is convex in the image $\mathbf{x}$:

$$-\log p(\mathbf{s} \mid \mathbf{x}) = \sum_{i \in \text{Cells}} \exp\{\mathbf{m_i}^T \mathbf{x} + b_i\} - (\mathbf{m_i}^T \mathbf{x} + b_i). \tag{3}$$

Generalized linear encoding models The generalized linear model (GLM) is an augmentation of the LNP model that incorporates the effects of spiking history and cell-cell correlations on neural response [10]. In the GLM, each RGC (indexed by i) is parameterized by a spatio-temporal stimulus filter $k_i[x, y, t]$, a feedback (spike history) filter $f_i[t]$, neighboring cell coupling filters $c_i^{(j)}[t]$, and a bias b_i. The GLM was fitted to spike counts measured within 1 ms time bins, approximately matched to the refractory period of the cells [28, 29]. To limit the number of parameters and improve computational efficiency, the stimulus filter was assumed to be space-time separable $k_i[x, y, t] = \mathbf{m}_i[x, y]h_i[t]$. Letting $*$ denote time-domain convolution, $w[t]$ denote the separable time component of the stimulus, and $\{i\}^C$ denote the set of coupled neighboring cells, the generator signal $g[t]$ is

$$g_i[t] = (\mathbf{m}_i^T \mathbf{x})(h_i * w)[t-1] + (s_i * f_i)[t-1] + \sum_{j \in \{i\}^C} (s_j * c_i^{(j)})[t-1] + b_i. \tag{4}$$

Using a sigmoidal nonlinearity and Bernoulli spiking, the encoding NLL used to fit a single cell (see A.5.2 for complete derivation) is

$$-\log p(s_i[N:T] \mid \mathbf{x}, s_i[0:N], \mathbf{s}_{\{i\}^C}[0:T]) = \sum_{t=N}^{T-1} \Big\{ \log(1 + \exp\{g_i[t]\}) - s_i[t]g_i[t] \Big\}. \tag{5}$$

To simplify the GLM, the filters h_i, f_i, and $c_i^{(j)}$ were each represented as weighted sums over a set of cosine "bump" functions [10]. As with the LNP model, L1 and L2 regularization terms were added to constrain the spatial filters, and an additional $L_{1,2}$ group sparsity penalty for the coupling filters was added to eliminate spurious cell-cell correlations. The complete objective function is described in detail in A.5.3. Model parameters were found by alternating between spatial and temporal filter convex minimization steps for each RGC, using FISTA [27, 30] for each step.

The joint encoding NLL over all of the cells used for image reconstruction (see A.5.4 for derivation) is again convex in the image $\mathbf{x}$:

$$-\log p(\mathbf{s}[N:T] \mid \mathbf{x}, \mathbf{s}[0:N]) = -\sum_{t=N}^{T-1} \sum_{i \in \text{cells}} \log p(s_i[t] \mid \mathbf{x}, s_i[0:t], \mathbf{s}_{\{i\}^C}[0:t]). \tag{6}$$

3.2 MAP with Gaussian 1/F priors

The Gaussian 1/F prior is the among the simplest and most commonly used image priors [31], and is the basis for many classical image processing algorithms. The 1/F prior assumes that pixels of the image are drawn from a stationary jointly Gaussian distribution (and thus that the spatial covariance matrix is diagonalized by the 2D Fourier basis) and that spectral power (variance of each spatial frequency component) falls off in inverse proportion to the square of the frequency F^2. Discarding terms that do not depend on the image $\mathbf{x}$, the negative log prior can be written as $-\log p(\mathbf{x}) = \sum_k |a_k(x)|^2/f_k^2$ where $a_k(x)$ is the amplitude of the k^{th} Fourier coefficient at frequency f_k. MAP image reconstruction using the 1/F prior can be performed using standard unconstrained convex minimization methods as both the negative log prior and the RGC encoding NLLs described in (3) and (6) are convex.

3.3 Approximate MAP with denoising convolutional neural network (dCNN) priors

Modern denoising convolutional neural networks can represent powerful image priors, but these priors are implicit [12, 32]: they are not expressed in closed form, and their values cannot be computed explicitly, making exact MAP inference difficult. The "plug-and-play" methodology provides an approximate iterative procedure for using such denoisers in MAP estimation problems [11], by

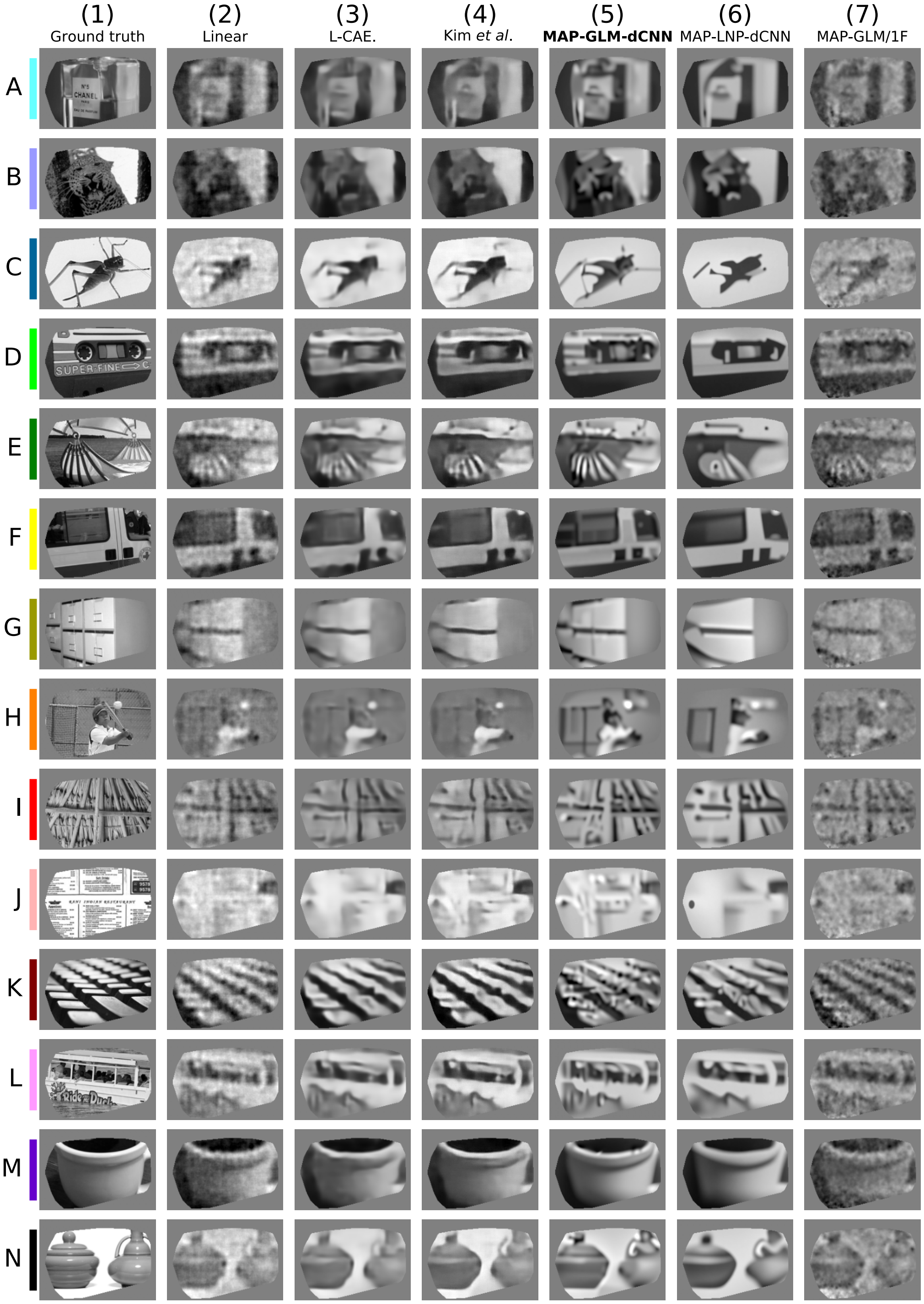

Figure 2: Selected stimulus images and reconstructions for the retinal mosaic of Figure 1. Images are spatially masked to only include regions covered by recorded RGCs. Colors on the left correspond to the colored markers in the scatterplots in Figure 3. Column contents: (1) Ground truth stimulus image; (2) Linear reconstruction; (3) Linear reconstruction followed by CNN autoencoder (L-CAE, [8]); (4) Linear reconstruction followed by CNN regression [4]; (5) Our method, approximate MAP with GLM and denoiser prior (MAP-GLM-dCNN); (6) Approximate MAP with LNP encoder and denoiser prior (MAP-LNP-dCNN); (7) Exact MAP with GLM encoder and 1/F prior (MAP-GLM-1F).

incorporating them into variable-splitting optimization methods such as half-quadratic splitting (HQS) [33] or alternating direction method of multipliers (ADMM) [34, 35]. Here, we adopt a method based on the HQS method presented in [15] to perform MAP reconstruction from RGC spikes. As in [15], we introduce an auxiliary variable $\mathbf{z}$, split the original problem in (1) into two sub-problems, incorporate a regularization parameter λ_p to control the prior term, and solve by alternating between two complementary optimization problems:

$$\mathbf{x}^{(k+1)} = \arg\min_{\mathbf{x}} \Big\{ -\log p(\mathbf{s} \mid \mathbf{x}) + \frac{\rho^{(k)}}{2} |\mathbf{x} - \mathbf{z}^{(k)}|_2^2 \Big\} \tag{7}$$

$$\mathbf{z}^{(k+1)} = \arg\min_{\mathbf{z}} \Big\{ -\lambda_p \log p(\mathbf{z}) + \frac{\rho^{(k)}}{2} |\mathbf{z} - \mathbf{x}^{(k+1)}|_2^2 \Big\}. \tag{8}$$

Since the problem in equation (8) has the same objective function as MAP Gaussian denoising with known noise variance $\lambda_p/\rho^{(k)}$, we solve it approximately with a single forward pass of a pre-trained unblind DRUNet denoiser network [15], resulting in Algorithm 1.

Algorithm 1 Approximate MAP reconstruction from RGC spikes (MAP-*encoding*-dCNN)

```
Hyperparameters: $\lambda_p$, schedule $\rho^{(1)}, \rho^{(2)}, ..., \rho^{(N)}$
Inputs: observed spike count vector $\mathbf{s} \in \mathbb{R}^C$
Initialize $\mathbf{z}^{(1)}$ to linear solution
for $k \in 1, 2, ...K$ do
    $\mathbf{x}^{(k+1)} \leftarrow \arg\min_{\mathbf{x}} \{ -\log p(\mathbf{s} \mid \mathbf{x}) + \frac{\rho^{(k)}}{2} |\mathbf{x} - \mathbf{z}^{(k)}|_2^2 \}$
    $\mathbf{z}^{(k+1)} \leftarrow$Unblind-DRUNet-Denoiser($\mathbf{x}^{(k+1)}; \sigma^2 = \frac{\lambda_p}{\rho^{(k)}}$)
end for
```

Unlike most applications of HQS, the encoding term in equation (7) is non-quadratic in $\mathbf{x}$ and hence (7) was solved iteratively (gradient descent with momentum and backtracking line search) rather than in closed form. $\mathbf{z}^{(1)}$ was initialized as the linear solution, though using random Gaussian initialization does not significantly affect the results (see Appendix A.6). Because convergence in the mathematical sense is not necessary for most imaging applications [36], $K = 25$ iterations were used in Algorithm 1. As in [15], $\rho^{(k)}$ was increased per-iteration on a log-spaced schedule. The hyperparameters λ_p and $[\rho^{(1)}, \rho^{(25)}]$ were determined by performing a grid search and evaluating reconstruction quality on an 80-image subset of the test partition. Variations in hyperparameters over a reasonable range ($\rho^{(1)} \in [10^{-2}, 10^{-1}]$, $\rho^{(25)} \in [30, 500]$, $\lambda_p \approx 0.1$) produced similar reconstruction quality, and optimal hyperparameters were similar across the two retinas and across LNP and GLM encoding models. Approximate MAP reconstructions using this algorithm are termed MAP-GLM-dCNN and MAP-LNP-dCNN for the GLM and LNP encoding models, respectively.

3.4 Benchmark: nonlinear regression with artificial neural networks

Current state-of-the-art methods for reconstruction of natural images from RGC spikes rely on an initial linear reconstruction step [1, 3], followed by *ad hoc* application of nonlinear neural networks. Specifically, Parthasarathy *et al.* [8] use a deep convolutional autoencoder (L-CAE) trained with MSE loss to apply natural image priors to linearly-reconstructed images. The authors trained and tested the model on simulated RGC spikes. Here, we used the published architecture and hyperparameters, but trained the model on experimentally measured retinal spike counts using backpropagation. Further details and characterization of the L-CAE on experimental data are provided in Appendix A.7. Another method was developed by Kim *et al.* [4], who partition the target image into high and low spatial frequency components. The low-frequency component is obtained via linear regression from the RGC spike counts, and a fully-connected neural network is used to non-linearly reconstruct the high-frequency component. The two components are summed and then passed through a final deblurring CNN. Both the high-frequency networks and the deblurring CNN are trained with backpropagation on RGC responses to natural images. This method achieves the most accurate reconstructions in the literature to date. Details of our implementation of the Kim *et al.* method are provided in Appendix A.8.

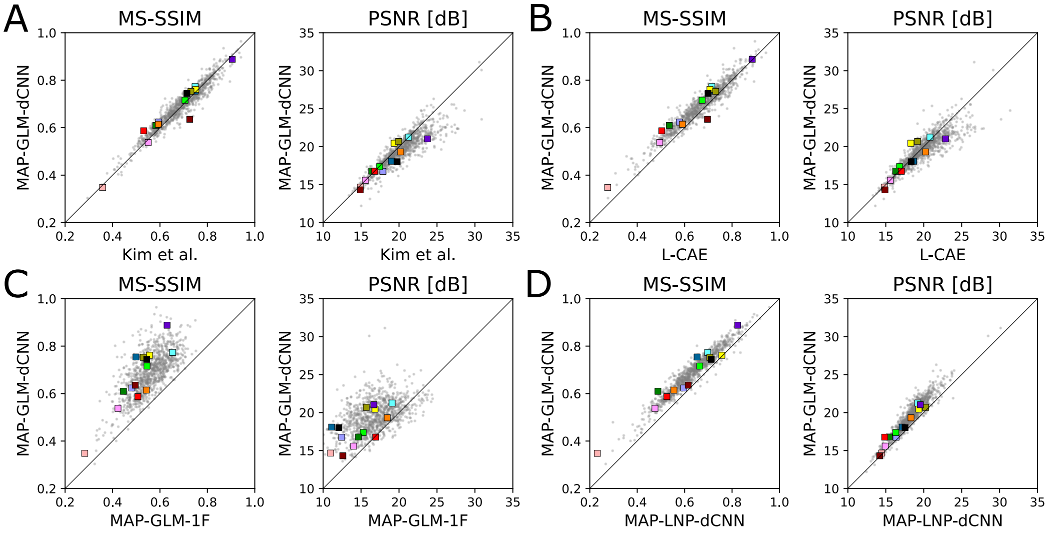

Figure 3: Comparisons of reconstruction quality of test data based on spikes of retina shown in Figure 1. Each panel shows values of two metrics – multi-scale structural similarity (MS-SSIM) [37], and PSNR – for all test images (individual points). Colored markers correspond to the example images in Figure 2. Vertical axis of all graphs corresponds to our method (MAP-GLM-dCNN), and horizontal axis corresponds to other methods: (A) MAP-GLM-dCNN vs. Kim *et al.* linear with neural network regression [4]. MAP-GLM-dCNN achieves comparable perceptual similarity but worse PSNR (B) MAP-GLM-dCNN vs. L-CAE [8]. MAP-GLM-dCNN achieves improved perceptual similarity with slightly worse PSNR. (C) MAP-GLM-dCNN vs. MAP-GLM-1F. Reconstructions using the weaker 1/F prior have worse perceptual similarity and PSNR. (D) MAP-GLM-dCNN vs. MAP-LNP-dCNN. Reconstructions using the simpler LNP encoding model have worse perceptual similarity and PSNR.

4 Results

4.1 Approximate MAP with GLM/dCNN matches or exceeds state-of-the-art results

To test whether our MAP-GLM-dCNN method outperforms state-of-the-art approaches, image reconstructions were generated from the test partitions of the datasets, and were compared both qualitatively and quantitatively. Example reconstructions are shown in Figure 2 for the L-CAE [8], for Kim *et al.* [4], and for our method. MAP-GLM-dCNN reconstructions are seen to be sharper than those of L-CAE, and contain additional image details (especially extended contours, as in rows C, E, G, H, I, and L). When compared to Kim *et al.*, MAP-GLM-dCNN tended to recover more content, particularly straight edges (rows E, G, H, I, and L), but sometimes exaggerated the contrast (rows A, H, and I). MAP-GLM-dCNN produced qualitatively different artifacts than the other methods. In particular, it sometimes hallucinated naturalistic structure not present in the stimulus images (rows J, K, and N), including striking irregularities in contours (rows D, K, L, and M).

Quantitative comparisons between MAP-GLM-dCNN and the two benchmark regression methods were also made. Scatter plots comparing MS-SSIM and PSNR on the test partition of one retina are shown in Figure 3A and 3B for Kim *et al.* and the L-CAE, respectively, and summary statistics over the test and heldout partitions for both retinas are presented in Tables 1 and 2. On an image-for-image basis, MAP-GLM-dCNN reconstructions have greater MS-SSIM than those of L-CAE (3B), demonstrating that the new method systematically achieves greater perceptual similarity to ground truth. The MAP-GLM-dCNN method resulted in comparable MS-SSIM perceptual similarity to the much more complicated Kim *et al.* method (3A). The PSNR of MAP-GLM-dCNN reconstructions was systematically worse than either benchmark. This is not surprising, as the MAP optimization procedure does not necessarily minimize MSE. These results held for both retinas (Tables 1 and 2).

Table 1: Average test and heldout MS-SSIM for each reconstruction method and retina.

	L-CAE		Kim *et al.*		**MAP-GLM-dCNN**		MAP-LNP-dCNN		MAP-GLM-1F	
	test	held	test	held	test	held	test	held	test	held
Retina 1	0.665	0.661	0.687	0.681	**0.689**	**0.688**	0.635	0.636	0.557	0.552
Retina 2	0.643	0.645	**0.675**	**0.675**	0.668	0.673	0.601	0.609	0.584	0.577

Table 2: Average test and heldout PSNR for each reconstruction method and retina.

	L-CAE		Kim *et al.*		**MAP-GLM-dCNN**		MAP-LNP-dCNN		MAP-GLM-1F	
	test	held	test	held	test	held	test	held	test	held
Retina 1	19.9	20.1	**20.2**	**20.5**	19.5	19.5	18.3	18.4	16.8	16.8
Retina 2	19.3	19.6	**19.8**	**20.1**	18.5	18.5	17.0	17.1	17.3	17.2

4.2 Deep denoiser prior substantially improves image quality over 1/F prior

To test the importance of the image prior, MAP-GLM-dCNN results were compared against reconstruction using the GLM encoding model with the classical 1/F Gaussian prior (MAP-GLM-1F). Example reconstructed images using the denoiser prior and 1/F prior are shown in columns 5 and 7, respectively, of Figure 2. Images reconstructed with the denoiser prior are less "grainy", and tend to have better-defined edges and smoother surfaces. The artifacts seen in the 1/F examples are expected, since this simple prior does not constrain phase [31], whose alignment is essential for generating sharp spatially-localized features. Scatterplots of image quality on the test partition using MS-SSIM and PSNR are shown for one retina in Figure 3C, and mean values for both retinas are summarized in Tables 1 and 2. Consistent with the visual appearance, PSNR and MS-SSIM were systematically higher when using the denoiser prior, in both retinas. Thus, using the more sophisticated denoiser image prior substantially increased the perceptual similarity of the reconstructions to ground truth.

4.3 GLM encoding model recovers additional image structure over LNP encoding model

To test the importance of the encoding model, we compared images reconstructed using the GLM and LNP encoding models, both using the same denoiser prior. Example images are shown in columns 5 and 6 of Figure 2. Images reconstructed using both models exhibit natural image structure like smooth surfaces and well-defined edges, but the GLM-reconstructed images tended to have more realistic-looking textures, whereas the LNP-reconstructed images tended to be overly simplified. Moreover, the GLM method recovered more high spatial frequency details (e.g., the legs of the insect in row C, the horizontal stripes on the tape cassette in row D, and the details on the hammock in row E, and the structure on the file cabinets in row G). The quality of image reconstructions for each image/response pair in the test partition for one retina were compared using MS-SSIM and PSNR in Figure 3D, and their mean values over the test and held out partitions for both retinas are summarized in Tables 1 and 2. In both retinas, images reconstructed using the GLM encoding model had systematically greater MS-SSIM scores, indicating greater perceptual similarity to ground truth, than those reconstructed using the LNP encoding model. This demonstrates that the choice of encoding model significantly affects reconstruction quality, and that the inclusion of temporal spike dependencies and cell-to-cell correlations in the more sophisticated GLM encoding model provides important constraints on the information encoded by the RGC spikes. This finding is consistent with previous work showing that decoding using the GLM (without priors) can access more information than simplified models lacking the cell-cell correlations or spiking history [10].

Table 3: Number of parameters trained on retinal data for each method for retina 1. MAP-GLM-dCNN and L-CAE have comparable numbers of parameters. Because of sparsity regularization in the GLM spatial encoding filters, more than half (55%) of the parameters in the GLM model are zero. The Kim *et al.* model contains nearly an order-of-magnitude more parameters than either MAP-GLM-dCNN or L-CAE.

	L-CAE	Kim *et al.*	**MAP-GLM-dCNN**	MAP-LNP-dCNN	MAP-GLM-1F
Trained params.	$3.07 \cdot 10^7$	$2.44 \cdot 10^8$	$2.88 \cdot 10^7$	$2.85 \cdot 10^7$	$2.88 \cdot 10^7$

5 Discussion

This paper presents a novel approximate MAP method for reconstructing natural images from the simultaneously recorded spikes of several hundred RGCs, using an accurate probabilistic model of retinal encoding and a natural image prior implicit in a pre-trained denoising neural network. The method matches or outperforms the current state-of-the-art in terms of recovering naturalistic image structure and/or the perceptual similarity of reconstructions to ground truth, while also being more principled and interpretable due to the explicit Bayesian separation of the encoding model and prior. The new approach uses substantially fewer parameters than previous state-of-the-art methods based on CNNs, and does not require training CNNs on retinal data (the prior is obtained from a network trained exclusively on image denoising). We showed that both encoding model and image prior contributed to the high-quality image reconstructions: removal of either substantially degraded performance. Thus, we expect that cell-cell correlations and temporal structure of spike trains, as well as image priors, will prove important in understanding how the retinal signal is used by the brain.

Several previous studies have used GLM encoding models for stimulus reconstruction from experimentally-recorded retinal signals, revealing the significance of cell-cell correlations for decoding temporal structure in white noise stimuli [10, 38], and the significance of the temporal structure of spike trains in tracking moving features [2]. By including a complex natural image prior into a Bayesian reconstruction method, the present work more efficiently exploits both the GLM and experimental data to produce state-of-the-art natural image reconstructions.

The enhanced reconstructions and interpretability obtained with our method could lead to improved function of retinal implants for restoring vision. Previous work [5] has suggested that electrical stimulation with a retinal implant can be guided by minimizing the expected MSE of linearly reconstructed images. This method ignores potentially important cell-cell correlations and fine temporal structure in RGC spike trains, and assumes that image priors captured by linear regression are sufficient for high performance. The method presented here offers an alternative approach to choosing simulation patterns to produce higher-fidelity artificial vision, while potentially being more robust than *ad hoc* neural network methods. However, achieving this in real time with minimal latency presents a substantial technical challenge.

Though the present work is limited to reconstruction of flashed static natural images from RGC spikes, extensions of our approximate MAP reconstruction method could be used to probe how neurons encode visual information under more natural conditions. For example, a central problem is understanding how the visual system achieves high-acuity perception in the presence of “jitter” in eye position, even when fixated [24]. Previous computational efforts have probed this question, but have been largely limited to simulated data with simple encoding models and stimuli [39, 40, 41]. Combining the methods put forth here with modern algorithms for image deblurring and motion-correction [42, 15] could yield more powerful methods to decode images from jittered retinal inputs. A related problem is understanding how the retina encodes the information contained in complex naturalistic movies [2], including movement of objects within a scene and other non-rigid transformations over time. The dimensionality of such stimuli and the consequent data requirements are high, so the ability to capture stimulus priors using modern machine learning tools [43, 44] separately from the retinal data, as was done here, will be important for understanding reconstruction in more naturalistic visual contexts.

References

[1] D. K. Warland, P. Reinagel, and M. Meister, "Decoding Visual Information From a Population of Retinal Ganglion Cells," *Journal of Neurophysiology*, vol. 78, pp. 2336–2350, Nov. 1997.

[2] V. Botella-Soler, S. Deny, G. Martius, O. Marre, and G. Tkačik, "Nonlinear decoding of a complex movie from the mammalian retina," *PLOS Computational Biology*, vol. 14, p. e1006057, May 2018.

[3] N. Brackbill, C. Rhoades, A. Kling, N. P. Shah, A. Sher, A. M. Litke, and E. Chichilnisky, "Reconstruction of natural images from responses of primate retinal ganglion cells," *eLife*, vol. 9, p. e58516, Nov. 2020.

[4] Y. J. Kim, N. Brackbill, E. Batty, J. Lee, C. Mitelut, W. Tong, E. J. Chichilnisky, and L. Paninski, "Nonlinear Decoding of Natural Images From Large-Scale Primate Retinal Ganglion Recordings," *Neural Computation*, vol. 33, pp. 1719–1750, June 2021.

[5] N. P. Shah, S. Madugula, L. Grosberg, G. Mena, P. Tandon, P. Hottowy, A. Sher, A. Litke, S. Mitra, and E. Chichilnisky, "Optimization of Electrical Stimulation for a High-Fidelity Artificial Retina," in *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*, (San Francisco, CA, USA), pp. 714–718, IEEE, Mar. 2019.

[6] N. P. Shah and E. J. Chichilnisky, "Computational challenges and opportunities for a bi-directional artificial retina," *Journal of Neural Engineering*, vol. 17, p. 055002, Oct. 2020.

[7] F. Rieke, D. Warland, R. de Ruyter van Steveninck, and W. Bialek, *Spikes: Exploring the Neural Code*. Cambridge, MA, USA: MIT Press, 1999.

[8] N. Parthasarathy, E. Batty, W. Falcon, T. Rutten, M. Rajpal, E. Chichilnisky, and L. Paninski, "Neural Networks for Efficient Bayesian Decoding of Natural Images from Retinal Neurons," conference, Neuroscience, June 2017.

[9] Y. Zhang, S. Jia, Y. Zheng, Z. Yu, Y. Tian, S. Ma, T. Huang, and J. K. Liu, "Reconstruction of natural visual scenes from neural spikes with deep neural networks," *Neural Networks*, vol. 125, pp. 19–30, May 2020.

[10] J. W. Pillow, J. Shlens, L. Paninski, A. Sher, A. M. Litke, E. J. Chichilnisky, and E. P. Simoncelli, "Spatio-temporal correlations and visual signalling in a complete neuronal population," *Nature*, vol. 454, pp. 995–999, Aug. 2008.

[11] S. V. Venkatakrishnan, C. A. Bouman, and B. Wohlberg, "Plug-and-Play priors for model based reconstruction," in *2013 IEEE Global Conference on Signal and Information Processing*, (Austin, TX, USA), pp. 945–948, IEEE, Dec. 2013.

[12] Y. Romano, M. Elad, and P. Milanfar, "The Little Engine That Could: Regularization by Denoising (RED)," *SIAM Journal on Imaging Sciences*, vol. 10, pp. 1804–1844, Jan. 2017.

[13] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *Advances in Neural Information Processing Systems 32*, pp. 11918–11930, Curran Associates, Inc., 2019.

[14] Z. Kadkhodaie and E. P. Simoncelli, "Stochastic Solutions for Linear Inverse Problems using the Prior Implicit in a Denoiser," in *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, p. 13, 2021.

[15] K. Zhang, Y. Li, W. Zuo, L. Zhang, L. Van Gool, and R. Timofte, "Plug-and-Play Image Restoration with Deep Denoiser Prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021.

[16] B. Kawar, G. Vaksman, and M. Elad, "Stochastic image denoising by sampling from the posterior distribution," tech. rep., aug 2021.

[17] A. Litke, N. Bezayiff, E. Chichilnisky, W. Cunningham, W. Dabrowski, A. Grillo, M. Grivich, P. Grybos, P. Hottowy, S. Kachiguine, R. Kalmar, K. Mathieson, D. Petrusca, M. Rahman, and A. Sher, "What does the eye tell the brain?: Development of a system for the large-scale recording of retinal output activity," *IEEE Transactions on Nuclear Science*, vol. 51, pp. 1434–1440, Aug. 2004.

[18] J. Lee, C. Mitelut, H. Shokri, I. Kinsella, N. Dethe, S. Wu, K. Li, E. B. Reyes, D. Turcu, E. Batty, Y. J. Kim, N. Brackbill, A. Kling, G. Goetz, E. Chichilnisky, D. Carlson, and L. Paninski, "YASS:

Yet Another Spike Sorter applied to large-scale multi-electrode array recordings in primate retina," preprint, Neuroscience, Mar. 2020.

[19] E. J. Chichilnisky, "A simple white noise analysis of neuronal light responses," *Network: Computation in Neural Systems*, vol. 12, pp. 199–213, 2001.

[20] E. J. Chichilnisky and R. S. Kalmar, "Functional Asymmetries in ON and OFF Ganglion Cells of Primate Retina," *The Journal of Neuroscience*, vol. 22, pp. 2737–2747, Apr. 2002.

[21] D. Dacey, "The mosaic of midget ganglion cells in the human retina," *The Journal of Neuroscience*, vol. 13, pp. 5334–5355, Dec. 1993.

[22] J. L. Gauthier, G. D. Field, A. Sher, M. Greschner, J. Shlens, A. M. Litke, and E. J. Chichilnisky, "Receptive Fields in Primate Retina Are Coordinated to Sample Visual Space More Uniformly," *PLoS Biology*, vol. 7, p. e1000063, Apr. 2009.

[23] L. Fei-Fei, J. Deng, and K. Li, "ImageNet: Constructing a large-scale image database," *Journal of Vision*, vol. 9, pp. 1037–1037, Aug. 2009.

[24] M. Rucci and M. Poletti, "Control and Functions of Fixational Eye Movements," *Annual Review of Vision Science*, vol. 1, pp. 499–518, Nov. 2015.

[25] T. Naselaris, R. J. Prenger, K. N. Kay, M. Oliver, and J. L. Gallant, "Bayesian Reconstruction of Natural Images from Human Brain Activity," *Neuron*, vol. 63, pp. 902–915, Sept. 2009.

[26] L.-Q. Zhang, N. P. Cottaris, and D. H. Brainard, "An image reconstruction framework for characterizing initial visual encoding," *eLife*, vol. 11, p. e71132, Jan. 2022.

[27] A. Beck and M. Teboulle, "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems," *SIAM Journal on Imaging Sciences*, vol. 2, pp. 183–202, Jan. 2009.

[28] M. J. Berry and M. Meister, "Refractoriness and Neural Precision," *The Journal of Neuroscience*, vol. 18, pp. 2200–2211, Mar. 1998.

[29] V. J. Uzzell and E. J. Chichilnisky, "Precision of Spike Trains in Primate Retinal Ganglion Cells," *Journal of Neurophysiology*, vol. 92, pp. 780–789, Aug. 2004.

[30] J. Liu, S. Ji, and J. Ye, "Multi-Task Feature Learning Via Efficient 2,1-Norm Minimization," in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, p. 10, 2009.

[31] E. P. Simoncelli, "Statistical Modeling of Photographic Images," in *Handbook of Image and Video Processing*, pp. 431–441, Elsevier, 2005.

[32] Z. Kadkhodaie and E. P. Simoncelli, "Solving linear inverse problems using the prior implicit in a denoiser," *arXiv*, July 2020.

[33] D. Geman and Chengda Yang, "Nonlinear image recovery with half-quadratic regularization," *IEEE Transactions on Image Processing*, vol. 4, pp. 932–946, July 1995.

[34] S. H. Chan, X. Wang, and O. A. Elgendy, "Plug-and-Play ADMM for Image Restoration: Fixed-Point Convergence and Applications," *IEEE Transactions on Computational Imaging*, vol. 3, pp. 84–98, Mar. 2017.

[35] Y. Sun, Z. Wu, X. Xu, B. Wohlberg, and U. Kamilov, "Scalable Plug-and-Play ADMM With Convergence Guarantees," *IEEE Transactions on Computational Imaging*, vol. 7, pp. 849–863, 2021.

[36] S. Diamond, V. Sitzmann, F. Heide, and G. Wetzstein, "Unrolled Optimization with Deep Priors," *arXiv:1705.08041 [cs]*, Dec. 2018. arXiv: 1705.08041.

[37] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*, vol. 13, pp. 600–612, Apr. 2004.

[38] K. Ruda, J. Zylberberg, and G. D. Field, "Ignoring correlated activity causes a failure of retinal population codes," *Nature Communications*, vol. 11, p. 4605, Dec. 2020.

[39] X. Pitkow, H. Sompolinsky, and M. Meister, "A Neural Computation for Visual Acuity in the Presence of Eye Movements," *PLoS Biology*, vol. 5, p. e331, Dec. 2007.

[40] Y. Burak, U. Rokni, M. Meister, and H. Sompolinsky, "Bayesian model of dynamic image stabilization in the visual system," *Proceedings of the National Academy of Sciences*, vol. 107, pp. 19525–19530, Nov. 2010.

[41] A. G. Anderson, K. Ratnam, A. Roorda, and B. A. Olshausen, "High-acuity vision from retinal image motion," *Journal of Vision*, vol. 20, p. 34, July 2020.

[42] Y. Li, M. Tofighi, J. Geng, V. Monga, and Y. C. Eldar, "Efficient and Interpretable Deep Blind Image Deblurring Via Algorithm Unrolling," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 666–681, 2020.

[43] A. Brifman, Y. Romano, and M. Elad, "Unified Single-Image and Video Super-Resolution via Denoising Algorithms," *IEEE Transactions on Image Processing*, vol. 28, pp. 6063–6076, Dec. 2019.

[44] D. Y. Sheth*, S. Mohan*, J. L. Vincent, R. Manzorro, P. A. Crozier, M. M. Khapra, E. P. Simoncelli, and C. Fernandez-Granda, "Unsupervised deep video denoising," in *Int'l Conf. Computer Vision (ICCV)*, Oct 2021.

[45] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), pp. 8024–8035, Curran Associates, Inc., 2019.

[46] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *3rd International Conference on Learning Representations*, 2015. arXiv: 1412.6980.

[47] O. Kupyn, T. Martyniuk, J. Wu, and Z. Wang, "DeblurGAN-v2: Deblurring (Orders-of-Magnitude) Faster and Better," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, (Seoul, Korea (South)), pp. 8877–8886, IEEE, Oct. 2019.

[48] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *International Conference on Learning Representations 2015*, Apr. 2015. arXiv: 1409.1556.