ARTICLE **Communicated by Joel Zylberberg**

Nonlinear Decoding of Natural Images From Large-Scale Primate Retinal Ganglion Recordings

Young Joon Kim
yjkimnada@gmail.com
Columbia University, New York, NY 10027, U.S.A.

Nora Brackbill
nbrack@stanford.edu
Stanford University, Stanford, CA 94305, U.S.A.

Eleanor Batty
erb2180@columbia.edu
JinHyung Lee
jl4303@columbia.edu
Catalin Mitelut
mitelutco@gmail.com
William Tong
wlt2115@columbia.edu
Columbia University, New York, NY 10027, U.S.A.

E. J. Chichilnisky
ej@stanford.edu
Stanford University, Stanford, CA U.S.A.

Liam Paninski
liam@stat.columbia.edu
Columbia University, New York, NY 10027, U.S.A.

Decoding sensory stimuli from neural activity can provide insight into how the nervous system might interpret the physical environment, and facilitates the development of brain-machine interfaces. Nevertheless, the neural decoding problem remains a significant open challenge. Here, we present an efficient nonlinear decoding approach for inferring natural scene stimuli from the spiking activities of retinal ganglion cells (RGCs). Our approach uses neural networks to improve on existing decoders in both accuracy and scalability. Trained and validated on real retinal spike data from more than 1000 simultaneously recorded macaque RGC units, the decoder demonstrates the necessity of nonlinear computations for accurate decoding of the fine structures of visual stimuli. Specifically, high-pass spatial features of natural images can only be decoded using

Neural Computation **33, 1719–1750** (2021)
https://doi.org/10.1162/neco_a_01395

nonlinear techniques, while low-pass features can be extracted equally well by linear and nonlinear methods. Together, these results advance the state of the art in decoding natural stimuli from large populations of neurons.

1 Introduction

What is the relationship between stimuli and neural activity? While this critical neural coding problem has often been approached from the perspective of developing and testing encoding models, the inverse task of decoding—the mapping from neural signals to stimuli—can provide insight into understanding neural coding. Furthermore, efficient decoding is crucial for the development of brain-computer interfaces and neuroprosthetic devices (Cheng, Greenberg, & Borton, 2017; Cottaris & Elfar, 2009; Jarosiewicz et al., 2015; Liu et al., 2000; Moxon & Foffani, 2015; Nirenberg & Pandarinath, 2012; Schwemmer et al., 2018; Warland, Reinagel, & Meister, 1997; Weiland et al., 2004; Bialek, de Ruyter van Steveninck, Rieke, & Warland, 1997).

The retina has long provided a useful test bed for decoding methods, since mapping retinal ganglion cell (RGC) responses into a decoded image provides a direct visualization of decoding model performance. Most approaches to decoding images from retinal ganglion cells (RGCs) have depended on linear methods due to their interpretability and computational efficiency (Brackbill et al., 2020; Marre et al., 2015; Warland et al., 1997). Although linear methods successfully decoded spatially uniform white noise stimuli (Warland et al., 1997) and the coarse structure of natural scene stimuli from RGC population responses (Brackbill et al., 2020), they largely fail to recover final visual details of naturalistic images.

More recent decoders incorporate nonlinear methods for more accurate decoding of complex visual stimuli. Some have leveraged optimal Bayesian decoding for white noise stimuli but exhibited limited scalability to large neural populations (Pillow et al., 2008). Others have attempted to incorporate key prior information for natural scene image structures and perform computationally expensive approximations to Bayesian inference (Naselaris, Prenger, Kay, Oliver, & Gallant, 2009; Nishimoto et al., 2011). Unfortunately, computational complexity and difficulties in formulating an accurate prior for natural scenery have hindered these methods. Other studies have constructed decoders that explicitly model the correlations between spike trains of different cells, for example, by using the relative timings of first spikes as the measure of neural response (Portelli et al., 2016). Parallel endeavors into decoding calcium imaging recordings from the visual cortex have produced coarse reconstructions of naturalistic stimuli through both linear and nonlinear approaches (Ellis & Michaelides, 2018; Garasto, Bharath, & Schultz, 2018; Garasto, Nicola, Bharath, & Schultz, 2019; Yoshida & Ohki, 2020).

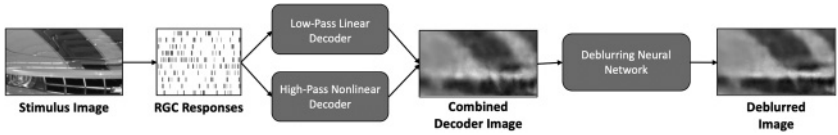

Figure 1: Outline of the decoding method. RGC responses to image stimuli are passed through both linear and nonlinear decoders to decode the low-pass and high-pass components of the original stimuli, respectively, before the combined decoded images are deblurred and denoised by a separate deblurring neural network.

In parallel, some recent decoders have relied on neural networks as efficient Bayesian inference approximators. However, established neural network decoders have either only been validated on artificial spike data sets (McCann, Hayhoe, & Geisler, 2011; Parthasarathy et al., 2017; Zhang, Jia et al., 2020) or on limited real-world data sets with modest numbers of simultaneously recorded cells (Botella-Soler et al., 2018; Ryu et al., 2011; Zhang, Jia et al., 2020). No nonlinear decoder has been developed and evaluated with the ultimate goal of efficiently decoding natural scenes from large populations (e.g., thousands) of neurons. Because the crux of the neural coding problem is to understand how the brain encodes and decodes naturalistic stimuli in through large neuronal populations, it is crucial to address this gap.

Therefore in this work we developed a multistage decoding approach that exhibits improved accuracy over linear methods and greater efficiency over existing nonlinear methods, and applied this decoder to decode natural images from large-scale multielectrode recordings from the primate retina.

2 Results

2.1 Overview. All decoding results were obtained on retinal data sets consisting of macaque RGC spike responses to natural scene images (Brackbill et al., 2020). Two identically prepared data sets, each containing responses to 10,000 images, were used for independent validation of our decoding methods. The electrophysiological recordings were spike sorted using YASS (Yet Another Spike Sorter; Lee et al., 2020) to identify 2094 and 1897 natural scene RGC units for the two data sets. We also recorded the responses to white noise visual stimulation and estimated receptive fields to classify these units into retinal ganglion cell types, to allow for analyses of cell-type specific natural scene decoding. (See section 3 for full details.)

Our decoding approach addresses accuracy and scalability by segmenting the decoding task into three subtasks (see Figures 1 and 2 and Table 1):

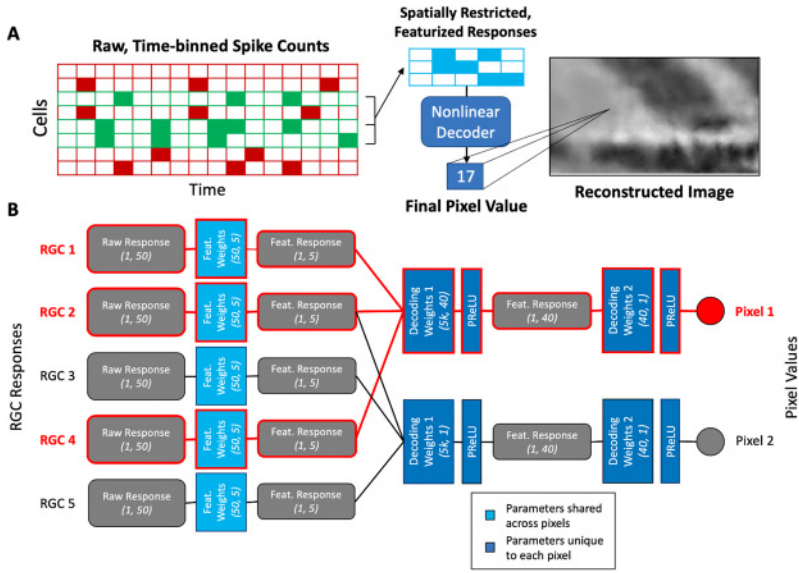

Figure 2: Outline of the nonlinear decoder. (A) The first part of the nonlinear decoder featurizes the RGC units' time-binned spike responses (50-dimensional vector for each RGC) to a lower dimension ($f = 5$). Afterward, each pixel's $k = 25$ most relevant units' featurized vectors are gathered and passed through a spatially restricted neural network, where each pixel is assigned its own nonlinear decoder to produce the final pixel value. (B) A miniaturized schematic of the spatially restricted neural network. Parameters that are shared across pixels versus those that are unique to each pixel are color-coded in different shades of blue. Furthermore, all the input values and weights that feed into a single pixel value are outlined in red to indicate the spatially restricted nature of the network. The vector dimensions of the weights and inputs are written in italicized parentheses; k represents the number of top units per pixel chosen for decoding.

- We use linear ridge regression to map the spike-sorted, time-binned RGC spikes to "low-pass," gaussian-smoothed versions of the target images. The smoothing filter size approximates the receptive fields of ON and OFF midget RGCs, the cell types with the highest densities in the primate retina.
- A spatially restricted neural network decoder is trained to capture the nonlinear relationship between the RGC spikes and the "high-pass" images, which are the residuals between the true and the low-pass images from the first step. The high-pass and low-pass outputs are summed to produce combined decoded images (see Figure 2).
- A deblurring network is trained and applied to improve the combined decoder outputs by enforcing natural image priors.

Table 1: Pixel-Wise Test Correlations of All Decoder Outputs (99% Confidence Interval Values in Parentheses).

	Versus True LP		Versus True HP		Versus True
LP ridge (2-bin)	0.975 (0.00016)			LP Ridge: 2-bin	0.887 (0.00062)
		HP NN	0.360 (0.0032)	**HP NN + LP ridge (2-bin)**	**0.901 (0.00059)**
				Combined-deblurred	**0.912 (0.00055)**
				Ridge-deblurred	**0.903 (0.00057)**
Whole ridge	0.963 (0.00021)	HP Ridge	0.282 (0.0028)	Whole RIDGE	0.890 (0.00061)
LP NN	0.960 (0.00033)			Whole NN	0.874 (0.00076)
LP ridge (50-bin)	0.979 (0.00015)				
LP LASSO	0.978 (0.00015)				

Notes: The best results are in bold.
The 2-bin and 50-bin LP ridge labels represent the two linear ridge decoders trained on the low-pass images. The whole ridge decoder is the 2-bin ridge decoder trained on the true whole images themselves, while the HP ridge decoder is the same decoder trained on the high-pass images only. The LP, HP, and whole NN labels denote the spatially restricted neural network decoder trained on low-pass, high-pass, and whole images, respectively. LP LASSO represents the 2-bin LASSO regression decoder trained on low-pass images. Finally, the combined-deblurred images are the deblurred versions of the sum of the HP NN and LP Ridge (2-bin) decoded images, while the ridge-deblurred images are the deblurred versions of the whole ridge decoder outputs. These final three—combined-deblurred, ridge-deblurred, and HP NN + LP Ridge (2-bin)—are in bold because they produced best results. The second, fourth, and sixth columns represent pixel-wise test correlations of each decoder's output versus the true low-pass, high-pass, and whole images, respectively.

The division of visual decoding into low-pass and high-pass decoding subtasks allowed us to leverage linear regression, which is simple and quick, for obtaining the target images' global features, while having the neural network decoder focus its statistical power on the addition of finer visual details. As discussed below, this strategy yielded better results than applying the neural network decoder to either the low-pass or the whole test images (see Table 1).

2.2 Linear Decoding Efficiently Decodes Low-Pass Spatial Features. We used two penalized linear regression approaches, ridge and LASSO regression (Friedman, Hastie, & Tibshirani, 2001), for linearly decoding the

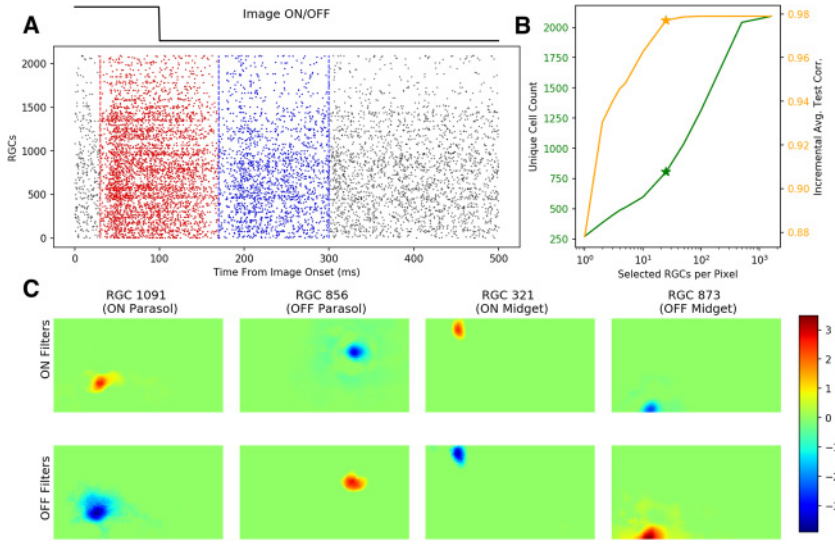

Figure 3: LASSO regression establishes a sparse mapping between RGC units and pixels. (A) Schematic of the ON (red; 30–170 ms) and OFF (blue; 170–300 ms) responses derived from RGC spikes. Each RGC's ON and OFF filter weights were multiplied to the summed spike counts within these windows. The spikes in these bins represent the cells' responses to stimuli onsets and offsets, respectively. The raster density (each dot represents a spike from a single RGC unit on a single trial) indicates that most of the RGC units' spikes were found in these two bins, which came slightly after the stimuli onsets and offsets themselves, as shown by the top line. (B) Total unique selected RGC unit count (green) and mean pixel-wise test correlations of partial LASSO decoded images (orange) as functions of the number of units chosen per pixel. For each pixel, {1, 2, 3, 4, 5, 10, 25, 50, 100, 500, 1000, 1600}, top units were chosen. Asterisks mark the top 25 units per pixel (805 unique units and 0.978 test correlation), the hyperparameter setting chosen for the nonlinear decoder below. (C) Representative ON and OFF spatial weights estimated by LASSO regression for four RGC units. Overall, LASSO regression successfully established a sparse mapping between RGC units and individual pixels by zeroing each cells' uninformative spatial weights, which comprise the majorities of the ON and OFF filters.

low-pass images. Both decoders considered only the neural responses during the image onset (30–170 ms) and offset (170–300 ms) time frames (see Figure 3A). While using the spikes from just the onset time bin produced reconstructions that were nearly as accurate as two bins, spikes from both bins were included to maximize accuracy with a minimal increase to computational workload (see Figure 12). For reference, LASSO regression is a form of linear regression whose regularization method enforces sparsity such that

the uninformative input variables are assigned zero weights while the informative inputs are assigned nonzero weights (Friedman et al., 2001). In the process, LASSO successfully identified each RGC unit's relevant linear spatial weights for both the image onset and offset time bins while zeroing out the insignificant spatial weights (see Figure 3C).

The LASSO spatial filters were roughly similar in appearance to the corresponding RGC unit receptive fields calculated from spike-triggered averages of white noise recordings (data not shown; see Brackbill et al., 2020). These linear filters eventually allowed for a sparse mapping between RGC units and image pixels so that only the most informative units for each pixel would be used as inputs for the nonlinear decoder (Botella-Soler et al., 2018). Partial LASSO-based decoding using smaller subsets of informative units demonstrated that these few hundred units were responsible for most of the decoding accuracy observed (see Figure 3B). Ultimately, 25 top units per pixel, corresponding to 805 total unique RGC units and a mean low-pass test correlation of 0.978 ($\pm$0.0002; this and all following error bars correspond to 99% CI values), were chosen. Choosing fewer than 25 informative RGC units per pixel resulted in lower LASSO regression test correlations, while choosing more units per pixel increased computational load without concomitant improvements in test correlation.

Consistent with previous findings (Brackbill et al., 2020), both linear decoders successfully decoded the global features of the stimuli by accurately modeling the low-pass images (see Figure 4). When evaluated by mean pixel-wise correlation against the true low-pass images, the decoded outputs from the ridge and LASSO decoders registered test correlations of 0.975 ($\pm$0.0002) and 0.978 ($\pm$0.0002), respectively (see Figure 4 and Table 1).[1] Increasing the temporal resolution of linear decoding beyond the two onset and offset time bins did not yield significant improvements in accuracy.

How different are decoding results if the linear decoder is instead applied to the true whole images rather than the low-pass images or if a nonlinear decoder is used for the low-pass targets? Notably, a ridge regression decoder trained on true images exhibited performance no better than the low-pass-specific linear decoders. Specifically, it registered a test correlation of 0.963 ($\pm$0.0002) versus true low-pass images and 0.890 ($\pm$0.0006) versus true images, suggesting that linear decoding can recover only low-pass details regardless of whether the decoding target contains high-pass details (see Table 1). The ridge low-pass decoded images registered a test correlation of 0.887 ($\pm$0.0006) against the whole test images. On the other hand, applying our neural network decoder to the low-pass targets

[1] Note that these correlation values are much higher than the subsequent correlation values in this manuscript as these low-pass decoded images were evaluated against the true low-pass images, which are much easier decoding targets than the true whole images themselves.

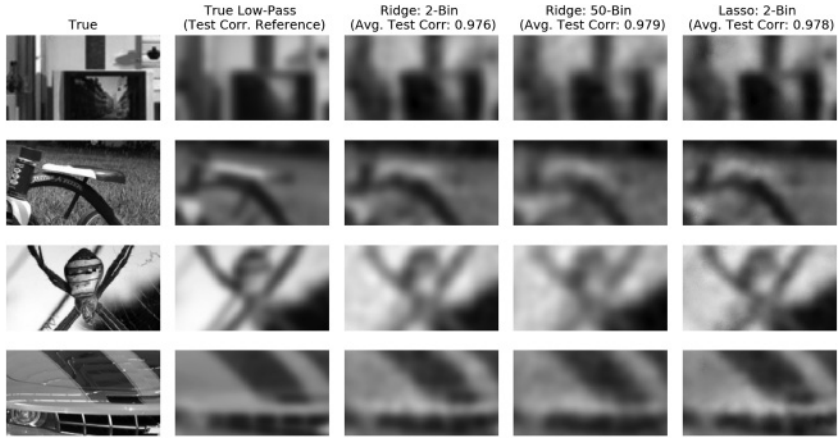

Figure 4: Linear decoding efficiently decodes low-pass spatial features. Representative true and true low-pass images along with their decoded low-pass counterparts produced via ridge (2-time-bin and 50-time-bin) and LASSO regression. Mean pixel-wise test correlations (evaluated against the true low-pass images, not the true images) are indicated within the top labels. The 50-bin decoder considers spike counts from the entire 500 ms stimulus window organized into 10 ms bins; this decoder achieved similar accuracy as the 2-bin decoder. All three linear regression techniques produce highly accurate decoding of the true low-pass images, suggesting that linear methods are sufficient for extracting the global features of natural scene image stimuli.

demonstrates that linear decoding is slightly more accurate (likely due to slight overfitting by the neural network) and vastly more efficient for low-pass decoding, as the former exhibited a lower test correlation of 0.960 ($\pm$0.0003) versus the low-pass targets (see Table 1). In sum, linear decoding is both the most accurate and appropriate approach for extracting the global features of natural scenes.

2.3 Nonlinear Methods Improve Decoding of High-Pass Details and Use Spike Temporal Correlations. Despite the high accuracy of low-pass linear decoding, the low-pass images and their decoded counterparts are (by construction) lacking the finer spatial details of the original stimuli. Therefore, we turned our attention next to decoding the spatially high-pass images formed as the differences of the low-pass and original images. Again, we compared linear and nonlinear decoders; unlike in the low-pass setting, we found that nonlinear decoders were able to extract significantly more information about the high-pass images than linear decoders. Specifically, a neural network decoder that used the nonzero LASSO regression weights to select its inputs (see Figure 3B) achieved a test correlation of

0.360 (±0.003) when evaluated against the high-pass stimuli, compared to ridge regression's test correlation of 0.282 (±0.003; see Figure 5B). While the high-pass reconstructions exhibited a greater spread in quality compared to their low-pass counterparts, nonlinear decoding consistently outperformed linear decoding even for the stimuli that both decoders struggled to decode (see Figure 13).

Moreover, the combined decoder output (summing the linearly decoded low-pass and nonlinearly decoded high-pass images) consistently produced higher test correlations compared to a simple linear decoder. Relative to the true images, ridge regression (for the whole images) and combined decoding yielded mean correlations of 0.890 (±0.0006) and 0.901 (±0.0006), respectively (see Figure 5A). In comparison, the linear low-pass decoded images alone yielded 0.887 (±0.0006). In other words, linear decoding of the whole image is almost no better than simply aiming for the low-pass image, and nonlinear decoding is necessary to recover significantly more detail beyond the low-pass target. Additionally, a neural network decoder that targets the whole true images falls short of the combined decoder with a mean test correlation of 0.874 (±0.0008) versus true images (see Table 1). In conjunction with the previous section's finding that the neural network decoder is not as successful with low-pass decoding as linear decoders, these results further justify our approach to reserve nonlinear decoding for the high-pass and linear decoding for the low-pass targets.

We then sought to analyze what characteristics of the RGC spike responses allowed for the superior performance of the combined decoding method. Previous studies have reported that nonlinear decoding better incorporates spike train temporal structure, which leads to its improvement over linear methods (Botella-Soler et al., 2018; Field & Chichilnisky, 2007; Passaglia & Troy, 2004). However, these studies were conducted with simplified random or white noise stimuli, and it is unclear how these findings translate to natural scene decoding. Thus, we hoped to shed light on how spike train correlations, both cross-neuronal and temporal, contribute to linear and nonlinear decoding. In previous literature, the former have been referred to as "noise correlations" and the latter as "history correlations" (Botella-Soler et al., 2018).

On a separate data set of 150 test images, each repeated 10 times, we created two modified neural responses to remove the two types of spike train correlations. As before, we binned each cell's spike counts into 10 ms bins so that for a single presented image, each cell exhibited a 50-bin response. Then, to remove cross-neuronal correlations, we swapped each cell's 50-bin response to an image randomly across the 10 repeat trials. Since each cell's response was independently swapped of the other cells' responses, correlations between RGCs within a trial were removed. Meanwhile, to remove history correlations, the individual spike counts within each cell's 50-bin response were randomly and independently exchanged with those from the other repeat trials.

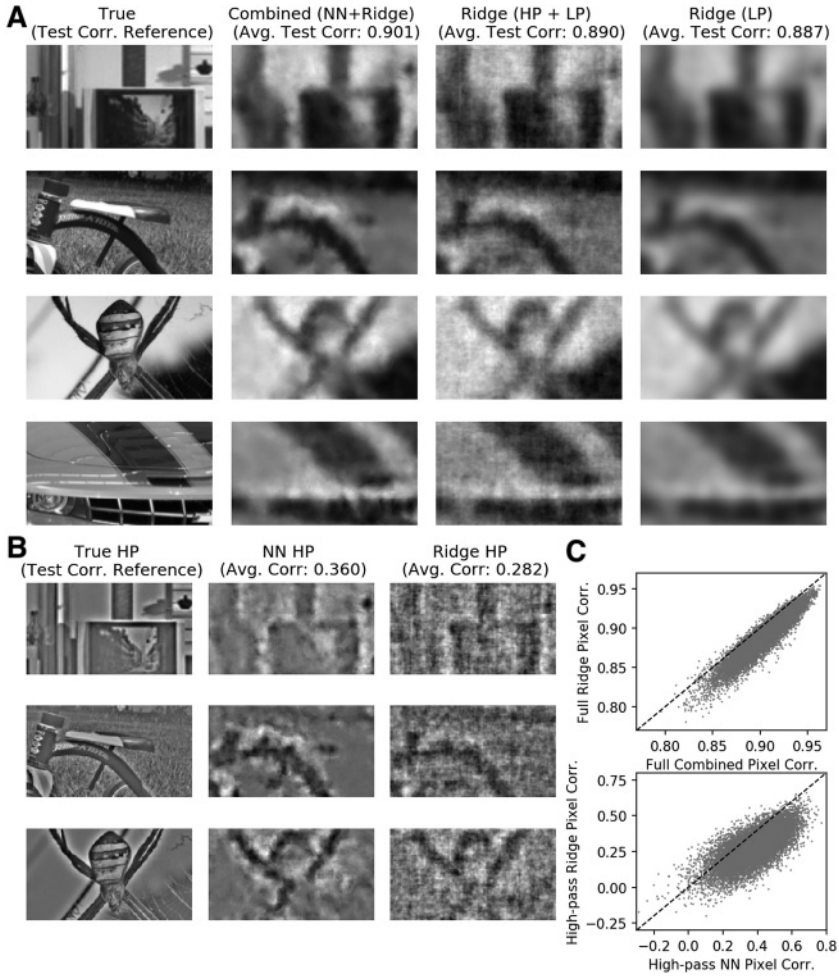

Figure 5: Nonlinear decoding extracts high-pass features more accurately than linear decoding. (A) Representative true images with their linearly decoded and combined decoder outputs; note that the linear decoder here decodes the true images (not just the true low-pass images) and was included for overall comparison. The correlation values here compare the decoded outputs against the true images. (B) Representative high-pass images with corresponding nonlinear and linear decoded versions. The correlation values here compare the high-pass decoded outputs against the true high-pass images. (C) Pixel-wise test correlation comparisons of linear and nonlinear decoding performance for the true and high-pass images. Linear decoding, either for the whole or low-pass images, is distinctly insufficient, and nonlinear methods are necessary for accurate decoding.

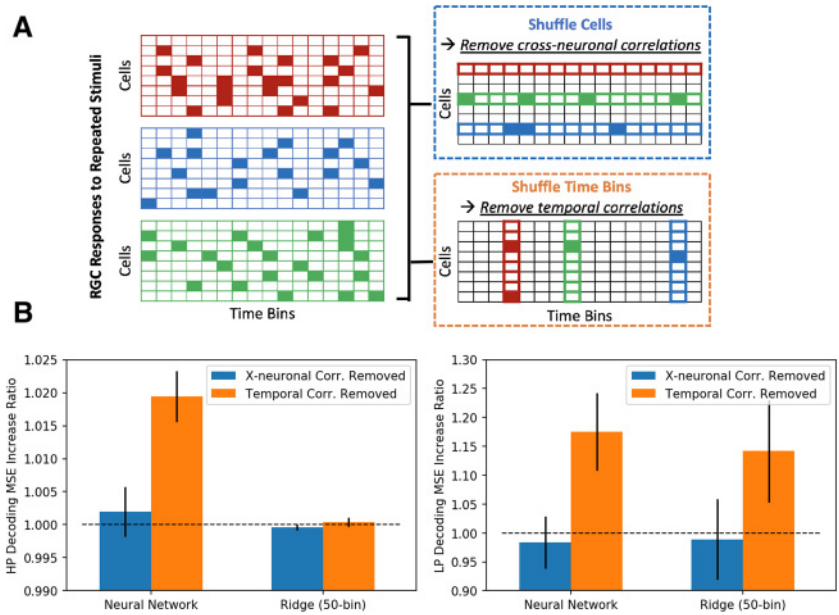

Figure 6: Spike temporal correlations are useful for high-pass nonlinear decoding and low-pass decoding. (A) Schematic of the shuffling of time bins and units' responses across repeated stimuli trials. (B) Ratio increases in MSE for neural network and linear decoders for high-pass and low-pass images before and after removing spike train correlations. While temporal correlations are important for both decoders in low-pass decoding, only the neural network decoder is reliant on temporal correlations in high-pass decoding. Cross-neuronal correlations are not crucial for both decoders in either decoding scheme.

For high-pass decoding, the neural network decoder exhibited a 1.9% (±0.4) increase in pixel-wise MSE when temporal correlations were removed, while the ridge decoder experienced a 0.04% (±0.07) increase in MSE (see Figure 6B); that is, nonlinear high-pass decoding is dependent on temporal correlations while linear high-pass decoding is not. Removing cross-neuronal correlations yielded no significant changes in either decoder, consistent with Brackbill et al. (2020). Meanwhile, for low-pass decoding, both decoders were equally and significantly affected by removing temporal correlations, as indicated by the 17.5% (±6.7) and 14.2% (±8.9) increases in MSE for the neural network and linear decoders, respectively (see Figure 6B). For the above comparisons, the ridge linear decoder for 50 time bins was used to maintain the same temporal resolution as the neural network decoder. In short, spike temporal correlations are important, specifically for the low-pass linear and all nonlinear decoders for optimal performance,

while cross-neuronal correlations are not influential in any decoding setup analyzed here (Botella-Soler et al., 2018).

2.4 OFF Midget RGC Units Drive Improvements in High-Pass Decoding when Using Nonlinear Methods. Next, we sought to investigate the differential contributions of each major RGC type toward visual decoding. Previous work has revealed that in the context of linear decoding, midget cells convey more high-frequency visual information, while parasol cells tend to encode more low-frequency information, consistent with the differences in density and receptive field size of these cell classes (Brackbill et al., 2020). Here we focused on the ON/OFF parasol/midget cells, the four numerically dominant RGC types, and their roles in linear versus nonlinear decoding. We classified the RGCs recorded during natural scene stimulation by first identifying units recorded during white noise stimulation and then using a conservative matching scheme that ensured one-to-one matching between recorded units in the two conditions. In total, 1033 units were matched, within which there were 72 ON parasol, 87 OFF parasol, 175 ON midget, and 195 OFF midget units (see section 4).

We performed standard ridge regression decoding for whole and low-pass images using spikes from the above four cell types and compared these decoded outputs to those derived from all 2094 RGC units, which include those not belonging to the four main types (see Figure 7). Consistent with previous results (Brackbill et al., 2020), midget decoding recovers more high-frequency visual information than parasol decoding, while ON and OFF units yield decoded images of similar quality. Meanwhile, differences between parasol and midget cell decoding are reduced for low-pass filtered images, as this task is not asking either cell population to decode high-frequency visual information.

We then investigated cell type contributions in the context of high-pass decoding (see Figure 8). Specifically, we investigated which cell type contributed most to the advantage of nonlinear over linear high-pass decoding and thus explained the improved performance of our decoding scheme. The advantages of nonlinear decoding were most prominent for midget and OFF units, with mean increases in test correlation of 7.1% and 6.8%, respectively (see Figure 8B). Parasol and ON units, meanwhile, saw a statistically insignificant change in test correlation. More finely grained analyses showed that only the OFF midget units enjoyed a statistically significant increase of 6.5% in mean test correlation in high-pass decoding. While ON midget units did indeed contribute meaningfully to high-pass decoding (as shown by their relatively high test correlations), they enjoyed no improvements with nonlinear over linear decoding. Therefore, one can conclude that the improvements in decoding for midget and OFF units via nonlinear methods can both be primarily attributed to the OFF midget subpopulation, which are also better encoders of high-pass details than their parasol counterparts. Previous studies have indeed indicated that midget units may

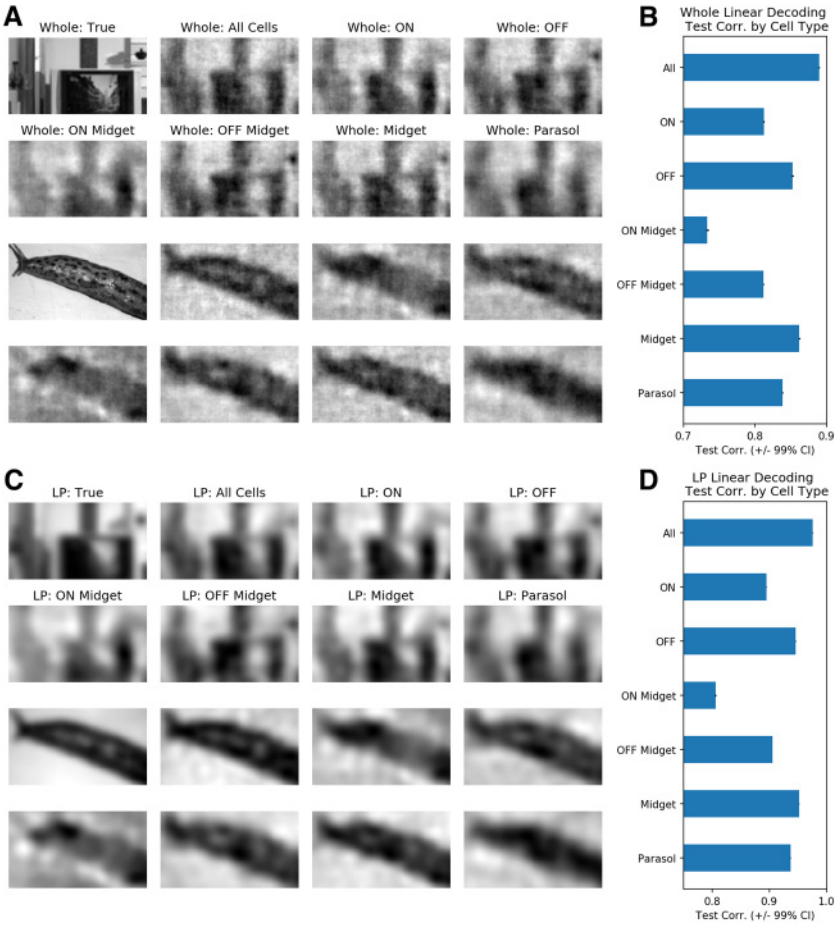

Figure 7: All major RGC types meaningfully contribute to low-pass linear decoding. (A) Representative whole images with their corresponding linearly decoded outputs using all, ON, OFF, ON Midget, OFF midget, midget, and parasol units, respectively. (B) Whole test correlations as functions of RGC type used for linear decoding. (C) Representative low-pass images with their corresponding linearly decoded outputs using all, ON, OFF, ON Midget, OFF midget, midget, and parasol units, respectively. (D) Low-pass test correlations as functions of RGC type used for linear decoding. Overall, all RGC types contribute meaningfully to low-pass, linear decoding.

encode more high-frequency visual information and that OFF midget units, in particular, exhibit nonlinear encoding properties (Brackbill et al., 2020; Chichilnisky & Kalmar, 2002; Freeman et al., 2015).

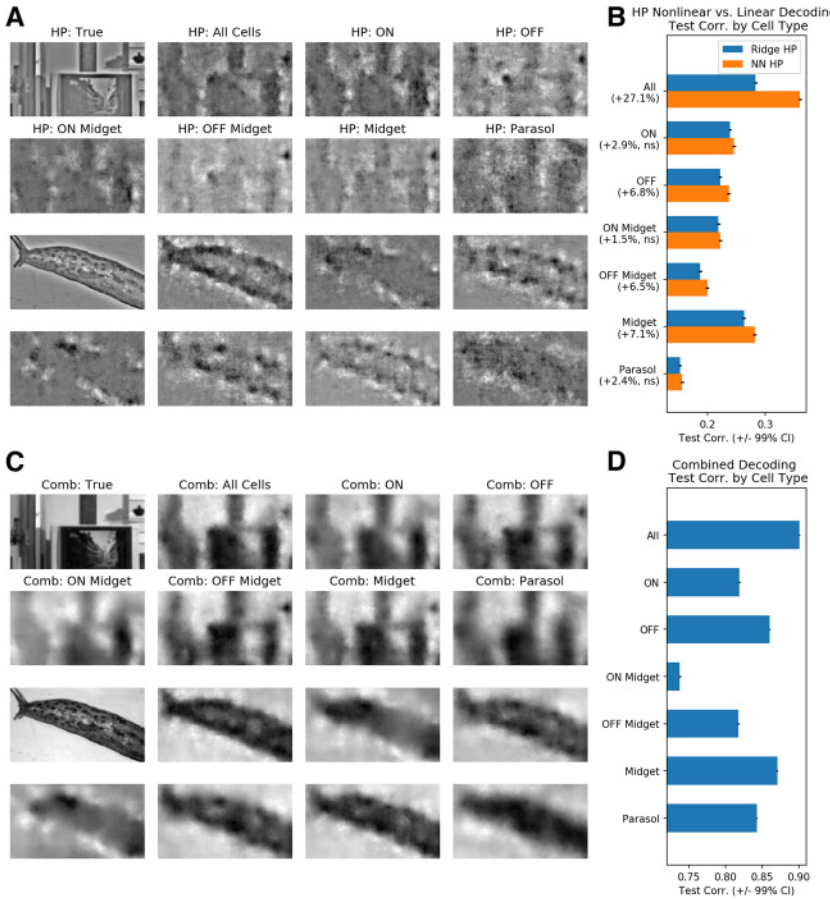

Figure 8: Midget and OFF units contribute most to high-pass, nonlinear decoding. (A) Representative high-pass images with their corresponding nonlinear decoded versions using all, ON, OFF, ON Midget, OFF midget, midget, and parasol units, respectively. (B) Comparison of test correlations between linear and nonlinear high-pass decoding versus cell type. (C) Representative true images with their corresponding combined decoder outputs using all, ON, OFF, ON Midget, OFF midget, midget, and parasol units, respectively. (D) Comparison of test correlations for the combined decoded images per cell type. Nonlinear decoding most significantly improves midget and OFF cell high-pass and combined decoding but does not bring any significant benefit to parasol and ON cell decoding of high-pass details.

2.5 A Final "Deblurring" Neural Network Further Improves Accuracy, but Only in Conjunction with Nonlinear High-Pass Decoding. Despite the success of the neural network decoder in extracting more spatial detail than the linear decoder, the combined decoder output still exhibited the blurriness near edges that is characteristic of low-pass image decoding. Therefore we trained a final convolutional "deblurring" network and found that this network was indeed qualitatively able to sharpen object edges present in the decoder output images (see Figure 9A; see Parthasarathy et al., 2017, for a related approach applied to simulated data). Quantitatively, the test pixel-wise correlation improved from 0.890 ($\pm$0.0006) and 0.901 ($\pm$0.0006) in the linear and combined decoder images, respectively, to 0.912 ($\pm$0.0006) in the combined-deblurred images (see Figure 9B and Table 1). Comparison by SSIM, a more perceptually oriented measure (Wang, Bovik, Sheikh, & Simoncelli, 2004), also revealed similar advantages in deblurring in combination with nonlinear decoding over other methods (see Figure 9C). In short, this final addition to the decoding scheme brought both subjective and objective improvements to the quality of the final decoder outputs.

The deblurring network is trained to map noisy, blurry decoded images back to the original true natural image—and therefore implicitly takes advantage of statistical regularities in natural images. (See Parthasarathy et al., 2017, for further discussion on this point.) Hypothetically, applying the deblurring network to linear decoder outputs could be sufficient for improved decoding. We therefore investigated the necessity of nonlinear decoding in the context of the deblurring network. Retraining and applying the deblurring network on the simple ridge decoder outputs (with the result denoted "ridge-deblurred" images) produced a final mean pixel-wise test correlation of 0.903 ($\pm$0.0006), which is lower than that of the combined-deblurred images (see Figure 9 and Table 1). Comparison by SSIM also yielded identical findings. We note that the deblurring network brought significant perceptual image quality improvements with or without the nonlinear decoder, as can be seen in the sample pipeline outputs. However, applying the deblurring network on the ridge decoder outputs did not fully remove the grainy, salt-and-pepper noise that is the product of the noisy linear attempt toward recovering the high-pass details (see Figure 5). This noise is not seen in the full pipeline (combined-deblurred) outputs, suggesting that one of the nonlinear decoder's unique roles is to remove noise during high-pass decoding that neither the linear decoder nor the deblurring network can accomplish. Thus, to obtain maximal results, the nonlinear decoder must be included alongside all the other components.

3 Discussion

The approach we have presented combines recent innovations in image restoration with prior knowledge of neuronal receptive fields to yield a

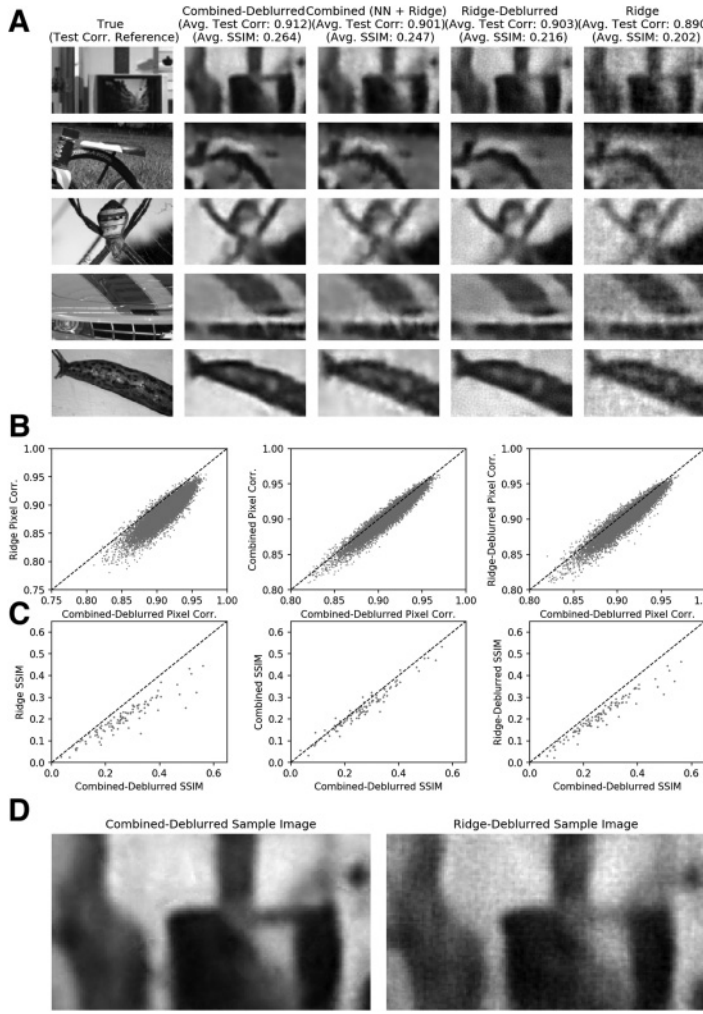

Figure 9: Neural network deblurring further improves nonlinear decoding quality. (A) Representative true images and their corresponding combined-deblurred, combined, ridge-deblurred, and ridge decoder outputs. Comparisons of pixel-wise test correlation (B) and SSIM (C) of the combined-deblurred versus ridge, combined, and ridge-deblurred decoder outputs, respectively. The combined-deblurred images had the highest mean SSIM at 0.265 (±0.018, 90% CI). The ridge-deblurred images had an SSIM of 0.216 (±0.015), which is lower than that of the combined-deblurred images. (D) Sample outputs from the combined-deblurred and ridge-deblurred pipelines. The deblurring network, specifically in combination with nonlinear decoding, brings quantitative and qualitative improvements to the decoded images. See Figure 10 for a similar analysis on a second data set.

decoder that is both more accurate and scalable than the previous state of the art. A comparison of linear and nonlinear decoding reveals that linear methods are just as effective as nonlinear approaches for low-pass decoding, while nonlinear methods are necessary for accurate decoding of high-pass image details. The nonlinear decoder was able to take advantage of spike temporal correlations in high-pass decoding while the linear decoder was not; both decoders used temporal correlations in low-pass decoding. Furthermore, much of the advantage that nonlinear decoding brings can be attributed to the fact that OFF midget units best encode high-pass visual details in a manner that is more nonlinear than the other RGC types, which aligns with previous findings about the nonlinear encoding properties of this RGC sub-class (Freeman et al., 2015).

These results differ from previous findings (using non-natural stimuli) that linear decoders are unaffected by spike temporal correlations (Botella-Soler et al., 2018; Passaglia & Troy, 2004) as, evidently, the low-pass linear decoder is just as reliant on such correlations as the nonlinear decoder for low-pass decoding. On the other hand, they also seem to support prior work indicating that nonlinear decoders are able to extract temporally coded information that linear decoders cannot (Field & Chichilnisky, 2007; Passaglia & Troy, 2004). Indeed, previous studies have noted that retinal cells can encode some characteristics of visual stimuli linearly and others nonlinearly (Gollisch, 2013; Passaglia & Troy, 2004; Schreyer & Gollisch, 2020; Schwartz & Rieke, 2011), which corresponds with our findings that temporally encoded low-pass stimuli information can be recovered linearly while temporally encoded high-pass information cannot. The above may help explain why linear and neural network decoders perform equally well for low-pass images but exhibit significantly different efficacies for high-pass details. We note that different experimental and recording conditions may yield alternative conclusions on the role of correlations in RGC population behavior. For instance, it has been suggested that different luminance conditions can affect the degree to which RGC populations rely on spike train correlations to encode visual information (Ruda, Zylberberg, & Field, 2020). Our recorded data set exhibited relatively low trial-to-trial variability, and this may have influenced the spike train correlation results. Indeed, for data sets with greater trial-to-trial noise, such as in low-light settings, different findings could have been made.

Nevertheless, several key questions remain. While our nonlinear decoder demonstrated state-of-the-art performance in decoding the high-pass images, the neural networks still missed many spatial details from the true image. Although it is unclear how much of these missing details can theoretically be decoded from spikes from the peripheral retina, we suspect that improvements in nonlinear decoding methods are possible. It is entirely possible that our spatially restricted parameterization of the nonlinear decoding may result in loss of information during the dimentionality-reduction process even though close analysis of architecture

choice on decoding performance does not suggest so (see Figure 14). We found that nonlinear high-pass decoding performance did not improve beyond 20 to 25 unique RGCs per pixel and actually decreased when using more than two nonlinear layers. Nevertheless, we do not rule out the possibility of other architecture choices producing better decoding results.

The deblurring of the combined decoder outputs is a challenging problem that current image restoration methods in computer vision likely cannot fully capture. Specifically, this step represents an unknown combination of superresolution, deblurring, denoising, and inpainting. With ongoing advances in image restoration networks that can handle more complex blur kernels and noise, it is likely that further improvements in performance are possible (Kupyn, Martyniuk, Wu, & Wang, 2019; Ledig et al., 2017; Maeda, 2020; Wang et al., 2018; Wang, Chen, & Hoi, 2020; Zhang, Zuo, & Zhang, 2019; Zhang, Zuo, Gu, & Zhang, 2017; Zhang, Tian et al., 2020; Zhou & Susstrunk, 2019).

Finally, while our decoding approach helped shed some light on the importance of nonlinear spike temporal correlations and OFF midget cell signals on accurate, high-pass decoding, the specific mechanisms of visual decoding have yet to be fully investigated. Indeed, many other sources of nonlinearity, including nonlinear spatial interactions within RGCs or nonlinear interactions between RGCs or RGC types, are all factors that could help justify nonlinear decoding that we did not explore (Gollisch, 2013; Odermatt, Nikolaev, & Lagnado, 2012; Pitkow & Meister, 2012; Schreyer & Gollisch, 2020; Schwartz & Rieke, 2011; Turner, Schwartz, & Rieke, 2018; Turner & Rieke, 2016). For example, it has been suggested that nonlinear interactions between jointly activated, neighboring ON and OFF cells may signal edges in natural scenes (Brackbill et al., 2020). We hope to investigate these issues further in future work.

4 Materials and Methods

The nonlinear decoder and deblurring network codes can be found at https://github.com/yjkimnada/ns_decoding.

4.1 RGC Data Sets. See Brackbill et al. (2020) for full experimental procedures. Briefly, retinas were obtained from terminally anesthetized macaques used by other researchers in accordance with animal ethics guidelines (see the Ethics Statement). After the eyes were enucleated, only the eye cup was placed in a bicarbonate-buffered Ames' solution. In a dark setting, retinal patches, roughly 3 mm in diameter, were placed with the RGC side facing down on a planar array of 512 extracellular microelectrodes covering a 1.8 mm-by-0.9 mm region. For the duration of the recording, the ex vivo preparation was perfused with Ames' solution (30–34C°, pH 7.4) bubbled with 95% O_2, 5% CO_2 and the raw voltage traces were bandpass

filtered, amplified, and digitized at 20 kHz (Chichilnisky & Kalmar, 2002; Field et al., 2010; Frechette et al., 2005; Litke et al., 2004).

In total, 10,000 natural scene images were displayed, with each image being displayed for 100 ms before and after 400 ms intervals of a blank, gray screen. For training, 9900 images were chosen and the remaining 100 for testing. The recorded neural spikes were spike-sorted using the YASS spike sorter to obtain the spiking activities of 2094 RGC units (Lee et al., 2020), which is significantly more units than previous decoders were trained to decode (Botella-Soler et al., 2018; Brackbill et al., 2020; Ryu et al., 2011; Zhang, Jia et al., 2020). Due to spike sorting errors, some of these 2094 units may be either oversplit (partial-cell) or overmerged (multicell). Nevertheless, oversplit and overmerged units can still provide decoding information (Deng, Liu, Kay, K., Frank, & Eden, 2015), and we therefore chose to include all spike-sorted units in the analyses here in an effort to maximize decoding accuracy. In the LASSO regression analysis (described below), we perform feature selection to choose the most informative subset of units, reducing the selected population roughly by a factor of two. Finally, to incorporate temporal spike train information, the binary spike responses were time-binned into 10 ms bins (50 bins per displayed image). A second retinal data set prepared in an identical manner was used to validate our decoding method and accompanying findings (see Figure 10).

While the displayed images were 160-by-256 in pixel dimensions, we restricted the images to a center portion of size 80-by-144 that corresponded to the placement of the multielectrode array. To facilitate low-pass and high-pass decoding, each of the train and test images was blurred with a gaussian blur of $\sigma = 4$ pixels and radius 3σ to produce the low-pass images. The filter size approximates the average size of the midget RGC. The high-pass images were subsequently produced by subtracting the low-pass images from their corresponding whole images.

4.2 RGC Unit Matching and Classification. To begin, we obtained spatiotemporal spike-triggered averages (STAs) of the RGC units from their responses to a separate white noise stimulus movie and classified them based on their relative spatial receptive field sizes and the first principal component of their temporal STAs (Chichilnisky & Kalmar, 2002). Afterward, both MSE and cosine similarity between electrical spike waveforms were used to identify each white noise RGC unit's best natural scene unit match and vice versa. Specifically, for each identified white noise unit, we chose the natural scene unit with the closest electrical spike waveform using both measures and kept only the white noise units that had the same top natural scene candidate found by both metrics. Then we performed the same procedure on all natural scene units, keeping only the units that had the same top white noise match using both metrics. Finally, we kept only the white noise-natural scene RGC unit pairs where each member of the pair chose each other as the top match via both MSE and cosine similarity. This

ensured one-to-one matching and that no white noise or natural scene RGC was represented more than once in the final matched pairs. In total, 1033 RGC units were matched in this one-to-one fashion, within which there were 72 ON parasol, 87 OFF parasol, 175 ON midget, and 195 OFF midget units. Several other cell types, such as small bistratified and ON/OFF large RGC units, were also found in smaller numbers. We also confirmed that the top 25 units chosen per pixel by LASSO, which comprise the 805 unique units feeding into the nonlinear decoder, also represented the four main RGC classes proportionally.

We chose a very conservative matching strategy to ensure one-to-one representation and maximize the confidence in the classification of the natural scene units. Naturally, such a matching scheme produced many unmatched natural scene units and a smaller number of unmatched white noise units. On average, the unmatched natural scene units had similar firing rates to the matched units while having smaller maximum channel spike waveform peak-to-peak magnitudes. While it is likely that a relaxation of matching requirements would yield more matched pairs, we confirmed that our matching strategy still resulted in full coverage of the stimulus area by each of the four RGC types (see Figure 11).

4.3 Low-Pass Linear Decoding. To perform efficient linear decoding on a large neural spike matrix without overfitting, for each RGC, we summed spikes within the 30 to 170 ms and 170 to 300 ms time bins, which correspond to the image onset and offset response windows. Thus, with n, t, x indexing the RGC units, training images, and pixels, respectively, the RGC spikes were organized into matrix $X \in \mathbb{R}^{t \times 2n}$ and the training images into $Y \in \mathbb{R}^{t \times x}$. To initially solve the linear equation $Y = X\beta$, the weights were inferred through the expression $\hat{\beta} = (X^T X + \lambda I)^{-1} X^T Y$, in which the regularization parameter $\lambda = 4833$ was selected via three-fold cross-validation on the training set (Friedman et al., 2001). Although we reduced the number of per-image time bins from 50 to 2, we confirmed that performing ridge regression on the augmented $\bar{X} = \mathbb{R}^{t \times mn}$ with m indexing the 50 time bins yielded essentially identical low-pass decoding performance, as discussed in the section 2.

Additionally, to perform pixel-specific feature selection for high-pass decoding, we performed LASSO regression (Friedman et al., 2001), which was proven to successfully select for relevant units, on the same neural bin matrix X from above (Botella-Soler et al., 2018). Due to the enormity of the neural bin matrix, Celer, a recently developed accelerated L1 solver, was used to individually set each pixel's L1 regularization parameter, as decoding each pixel represents an independent regression subtask (Massias, Gramfort, & Salmon, 2018).

4.4 High-Pass Nonlinear Decoding. To maximize high-pass decoding efficacy with the nonlinear decoder, the augmented $\tilde{X} = \mathbb{R}^{t \times mn}$ was chosen

as the training neural bin matrix. As noted, nonlinear methods, including kernel ridge regression and feedforward neural networks, have been successfully applied to decode both the locations of black disks on white backgrounds (Botella-Soler et al., 2018) and natural scene images (Zhang, Jia et al., 2020). Notably the former study used L1 sparsification of the neural response matrix so that only a handful of RGC responses contributed to each pixel before applying kernel ridge regression. We borrow this idea of using L1 regression to create a sparse mapping between RGC units and pixels before applying our own neural network decoding, as explained below. However, the successful applications of feedforward decoding networks above crucially depended on the fact that they used a small number of RGCs (91 RGCs with 5460 input values and 90 RGCs with 90 input values, respectively). For reference, constructing a feedforward network for our spike data of 2094 RGC units and 104,700 inputs would yield an infeasibly large number of parameters in the first feedforward layer alone. Similarly, kernel ridge regression, which is more time-consuming than a feedforward network, would be even more impractical for large neural data sets.

Therefore, we constructed a spatially restricted network based on the fact that each RGC's receptive field encodes a small subset of the pixels and, conversely, each pixel is represented by a small number of RGCs. Specifically, each unit's image-specific response m-vector is featurized to a reduced f-vector so that each unit is assigned its own featurization mapping that is preserved across all pixels. Afterward, for each pixel, the featurized response vectors of the k most relevant units are gathered into a fk-vector and further processed by nonlinear layers to produce a final pixel intensity value. The k relevant units are derived from the L1 weight matrix $\beta \in \mathbb{R}^{2n \times x}$ from above. Within each pixel's weight vector $\beta_x \in \mathbb{R}^{2n \times 1}$ and an individual unit's pixel-specific weights ($\beta_{n,x} \in \mathbb{R}^{2 \times 1}$), we calculate the L1-norm $\lambda_{x,n} = |\beta_{n,x}|_1$ and select the units corresponding to the k largest norms for each pixel. The resulting high-pass decoded images are added to the low-pass decoded images to produce the combined decoder output. Note that while the RGC featurization weights are shared across all pixels, each pixel has its own optimized set of nonlinear decoding weights (see Figure 2).

The hyperparameters $f = 5, k = 25$ were chosen from an exhaustive grid search spanning $f \in \{5, 10, 15, 20\} k \in \{5, 10, 15, 20, 25\}$ so that the values at which no further performance gains were observed were selected. The neural network itself was trained with a variant of the traditional stochastic gradient descent (SGD) optimizer that includes a momentum term to speed up training (Qian, 1999) (momentum hyperparameter of 0.9, learning rate of 0.1, and weight regularization of 5.0×10^{-6} used for training the network over 32 epochs).

4.5 Deblurring Network. To further improve the quality of the decoded images, we sought to borrow image restoration techniques from the ever-growing domain of neural network–based deblurring. Specifically,

a deblurring network leveraging natural image priors would take in the combined decoder outputs and produce sharpened versions of the inputs. However, these networks usually come with high requirements for training data set size; using only the 100 decoded images corresponding to the originally held out test images would be insufficient.

As a result, we sought to virtually augment our decoder training data set of 9,900 spikes-image pairs for use as training examples in the deblurring scheme. The 9900 training spikes-image pairs were subdivided into 10 subsets of 990 pairs. Then each subset was held out and decoded (both linearly and nonlinearly) with the other 9 subsets used as the decoders' training examples. Rotating and repeating through each of the 10 subsets allowed for all 9900 training examples to be transformed into test-quality decoder outputs, which could be used to train the deblurring network. (To be clear, 100 of the original 10,000 spikes-images pairs were held out for final evaluation of the deblurring network, with no data leakage between these 100 test pairs and the 9900 training pairs obtained through the above data set augmentation.) An existing alternative method would be to craft and use a generative model for artificial neural spikes corresponding to any arbitrary input image (Parthasarathy et al., 2017; Zhang, Jia et al., 2020). However, the search for a solution for the encoding problem is still a topic of active investigation in neuroscience; our method circumvents this need for a forward generative model.

With a sufficiently large set of decoder outputs, we could adopt well-established neural network methods for image deblurring and super-resolution (Kupyn et al., 2019; Ledig et al., 2017; Maeda, 2020; X. Wang et al., 2018; Wang et al., 2020; Zhang et al., 2019; Zhang et al., 2017; Zhang, Tian et al., 2020; Zhou & Susstrunk, 2019). Specifically, we chose the convolutional generator of DeblurGANv2, an improvement of the widely adopted DeblurGAN with superior deblurring capabilities (Kupyn et al., 2019). After performing a grid search of the generator ResNet block number hyperparameter ranging $\{1, 2, \ldots, 7, 8\}$, the 6-block generator was chosen for training under the Adam optimizer (Kingma & Ba, 2017) for 32 epochs at an initial learning rate of 1×10^{-5} that was reduced by half every 8 epochs.

We do not expect that the decoded images will be near-perfect replicas of the original image. Recordings here were taken from the peripheral retina, where spatial acuity is lower; as a result, one would expect the neural decoding of the stimuli to miss some of the fine details of the original image. Therefore, while the original DeblurGANv2 paper includes pixel-wise L1 loss, a VGG discriminator-based content/perceptual loss, and an additional adversarial loss during training, we excluded the final adversarial loss term due to the fact that the deblurred images of the decoder would not be perfect (or near-perfect) look-alikes of the raw stimuli images. Instead, we focus on improving the perceptual qualities of the output image, including edge sharpness and contrast, for more facile visual identification. We use both pixel-wise L1 loss and L1 loss between the features extracted from the true images and from the reconstructions in the third convolutional layer

of the pretrained VGG-19 network before the corresponding pooling layer (Johnson, Alahi, & Fei-Fei, 2016; Wang et al., 2018).

Appendix: Supplemental Information

A.1 Validation of Decoding Methods on Second RGC Data Set (Figure 10).

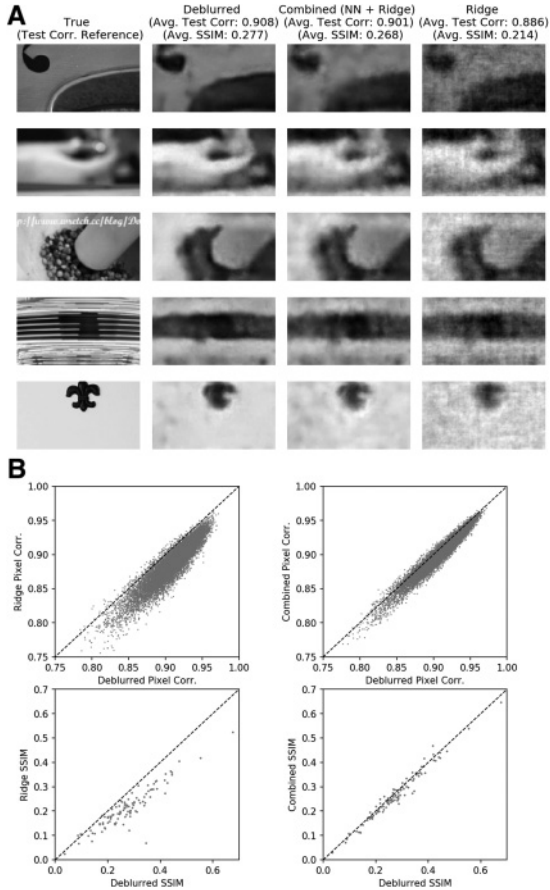

Figure 10: Decoding method results corroborated on a second RGC data set. (A) Representative outputs from the decoding algorithm compared to those from a simple linear decoder. (B) Comparison of pixel-wise test correlations and SSIM values between deblurred and linear decoder outputs and against combined decoder outputs, respectively. The second data set consisted of the responses of 1987 RGC units to 10,000 images, prepared in an identical manner as the first data set. The superiority of nonlinear decoding with deblurring is apparent.

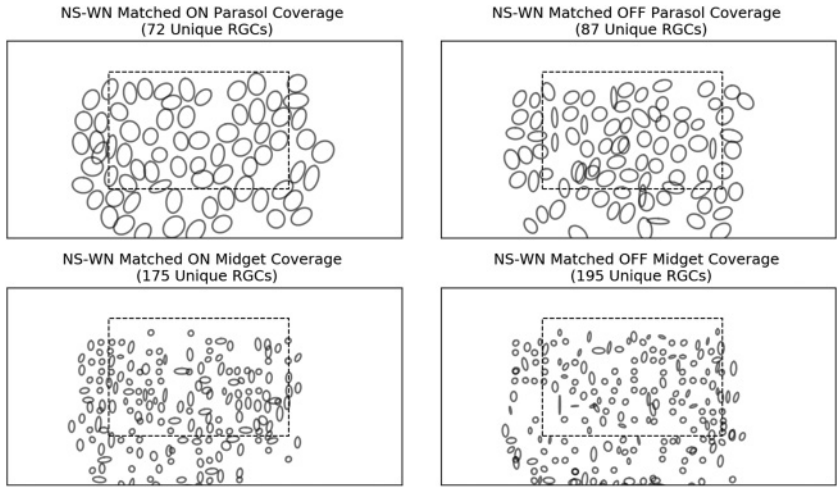

Figure 11: Coverage of image area by matched RGC cells. All four cell types, ON/OFF parasol/midget, sufficiently cover the image area (marked in dashed rectangle) with the receptive fields of their constituent white noise-natural scene matched units.

A.2 Matching of White Noise and Natural Scene RGC Units (Figure 11). Because hundreds of white noise and more than a thousand natural scene RGC units were discarded during the matching process, these unmatched units were analyzed to see whether they exhibited any distinguishing properties from the matched units. Comparing the mean firing rates of the matched and unmatched units revealed no clear differences: 10.53 Hz versus 11.46 Hz for matched and unmatched natural scene units and 6.56 Hz versus 7.03 Hz for matched and unmatched white noise units. However, the mean maximum channel peak-to-peak values (PTPs) were markedly different between matched and unmatched units within both experimental settings: 22.06 versus. 10.21 for matched and unmatched natural scene units and 24.93 versus 18.48 for matched and unmatched white noise units.

Nonmatching of units is likely caused by several factors. To begin, MSE and cosine similarity are not perfect measures of template similarity. Many close candidates were quite similar in shape to the reference templates but had either a slightly different amplitude or peaks and troughs at different temporal locations. Indeed it is possible that using a more flexible similarity metric would recover more matching units. Meanwhile, it is also likely that some of the unmatched units in either experimental setting are simply inactive units. Specifically, it could be the case that some units are inactive during white noise stimulation but more active for natural scene input, and

Figure 12: Comparison of ridge regression decoding using both onset (30–170 ms) and offset (170–300 ms) time bins versus using only one time bin. (A) Ridge decoding on one versus two time bins for whole image and (B) low-pass image reconstruction. Using the onset but not the offset time bin gives reconstructions of nearly the same quality as using both time bins.

vice versa. Finally, difficulties with spike-sorting smaller units could also lead to mismatches. Nevertheless, despite the above issues, we were able to recover full coverage of the stimulus region for each cell type, as shown in Figure 11.

A.3 Linear Decoding Using Spikes from Only Either Onset or Offset Time Bins (Figure 12). While our standard ridge regression decoder used spikes from both the onset (30–170 ms) and offset (170–300 ms) time bins, we investigated how much the onset and offset spikes, respectively, contributed to linear decoding. Whole and low-pass image reconstructions using just the onset spikes resulted in test correlations of 0.885 ($\pm$0.0006) and 0.968 ($\pm$0.000) versus true whole and low-pass images, respectively.

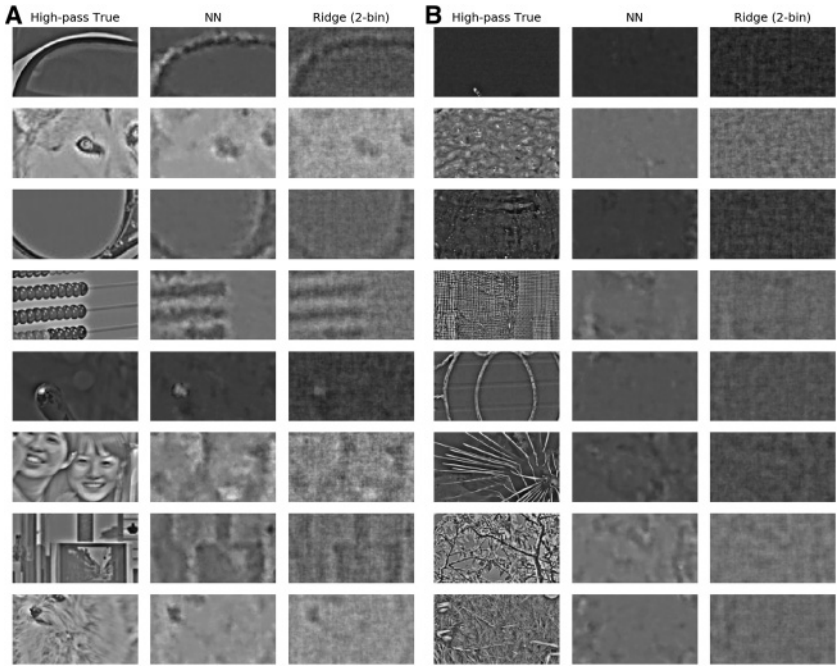

Figure 13: The eight best (A) and worst (B) high-pass images and their linear and nonlinear reconstructions. Both decoders performed best for stimuli with clear edges marked by high contrast and performed worst for low-contrast stimuli with either too few or too many edges.

Meanwhile, using just offset spikes yielded 0.0837 (±0.0008) and 0.936 (±0.0004) against true whole and low-pass images. For reference, using both time bins results in 0.890 (±0.0006) and 0.975 (±0.0002) against the same targets, respectively. In short, while it is possible to use just the onset spikes to give reconstructions that were nearly as good as those produced from two time bins, using both onset and offset spikes produced the best ridge decoding results.

A.4 High-Pass Decoding Benefits from Stimuli with High Contrast Edges (Figure 13). Because high-pass decoding via both the linear and nonlinear decoders exhibited greater spread in the quality of reconstructions compared to low-pass decoding, we investigated which high-pass images yielded the best and worst reconstructions, respectively. The 20 best high-pass images (measured by nonlinear decoder outputs' test correlation versus their respective true high-pass images) resulted in mean image correlations of 0.519 (±0.029) and 0.380 (±0.031) via the nonlinear and linear

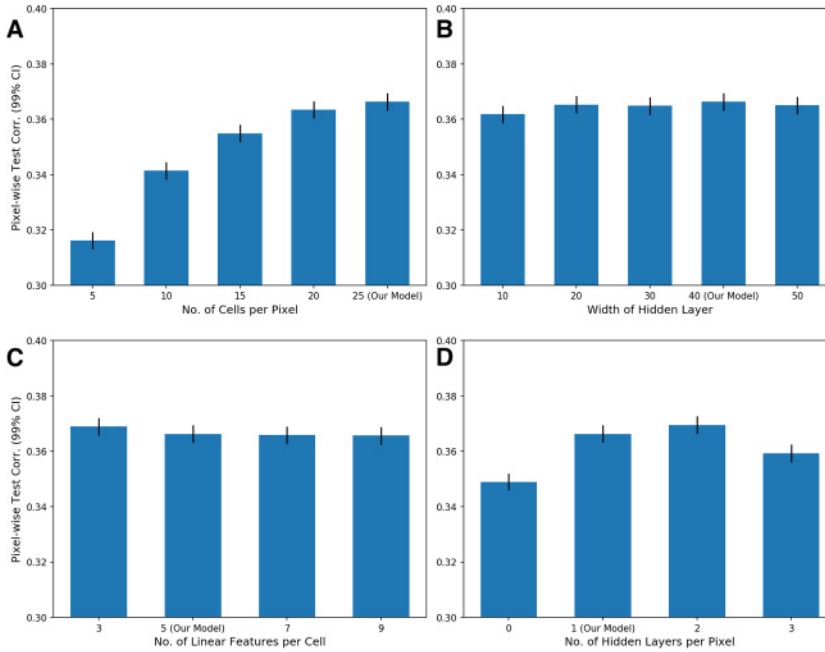

Figure 14: Comparison of pixel-wise high-pass test correlation for variants of the neural network decoder. Comparisons included architectures with (A) different cell numbers used per pixel, (B) width of the pixel-specific hidden layer, (C) linear features extracted per cell, and (D) number of pixel-specific hidden layers.

decoders, respectively. Meanwhile the 20 worst high-pass images produced mean image correlations of 0.232 ($\pm$0.030) and 0.152 ($\pm$0.029) via nonlinear and linear decoding, respectively. Note that these values represent correlations between individual images and not the pixel-wise test correlations used throughout the study.

A closer analysis of the best and worst high-pass images suggests that both linear and nonlinear high-pass decoding performed the best when the target stimuli had clear edges marked by high contrast. Meanwhile, both high-pass decoders struggled to reconstruct stimuli with too few or too many object edges combined with low color contrast. Nevertheless, as indicated by the above image correlations, nonlinear decoding still significantly outperforms linear decoding in both best and worst case scenarios.

A.5 Influence of Cell Number and Architecture on High-Pass Decoding (Figure 14). The high-pass neural network decoder takes in cell number used, number and width of pixel-specific layers, and number of linear

features per cell as its hyperparameters. Naturally, we sought to investigate how these values affect the performance of high-pass decoding. We remind readers that our chosen model takes in 25 cells per pixel, transforms each cell's 50-bin response into 5 linear features, and uses 1 hidden layer that is 40 units wide.

To begin, we compared performance when using 5, 10, 15, 20, and 25 cells per pixel, which corresponds to 512, 597, 676, 746, and 805 unique cells incorporated. Unsurprisingly, this was the hyperparameter with the greatest influence on high-pass decoding with significant gains in performance being observed up until at least 20 cells per pixel are used. We note that we could not go much below using 512 unique cells as our sparse cell-to-pixel feature selection process zeroes out each cell's influence on pixels outside of its receptive field.

Next, we noted that at least one nonlinear hidden layer per pixel was required for optimal decoding performance. However, using more than two hidden layers resulted in lower performance, suggesting that using more parameters than necessary results in either overfitting or difficulties in the optimization process. This conclusion is reinforced by the fact that increasing the number of linear features per cell and width of each pixel-specific hidden layer did not further improve decoding performance. Since our neural network decoder does not seem to improve indefinitely with increasing number of parameters and layers, our choice of a spatially restricted neural network decoder still stands as a reasonable one over much more massive, intractable fully connected architectures.

Ethics Statement

Eyes were removed from terminally anesthetized macaque monkeys (*Macaca mulatta*, *Macaca fascicularis*) used by other laboratories in the course of their experiments, in accordance with the Institutional Animal Care and Use Committee guidelines. All of the animals were handled according to approved institutional animal care and use committee (IACUC) protocols (28860) of the Stanford University. The protocol was approved by the Administrative Panel on Laboratory Animal Care of the Stanford University (Assurance Number: A3213-01).

Acknowledgments

We thank Eric Wu and Nishal Shah for helpful discussions.

References

Bialek, W., de Ruyter van Steveninck, R., Rieke, F., & Warland, D. (1997). *Spikes: Exploring the neural code*. Cambridge, MA: MIT Press.

Botella-Soler, V., Deny, S., Martius, G., Marre, O., & Tkačik, G. (2018). Nonlinear decoding of a complex movie from the mammalian retina. *PLOS Computational Biology*, *14*(5), e1006057. https://doi.org/10.1371/journal.pcbi.1006057.

Brackbill, N., Rhoades, C., Kling, A., Shah, N. P., Sher, A., Litke, A. M., & Chichilnisky, E. J. (2020). Reconstruction of natural images from responses of primate retinal ganglion cells. *Neuroscience*. https://doi.org/10.1101/2020.05.04.077693.

Cheng, D. L., Greenberg, P. B., & Borton, D. A. (2017). Advances in retinal prosthetic research: A systematic review of engineering and clinical characteristics of current prosthetic initiatives. *Current Eye Research*, *42*(3), 334–347. https://doi.org/10.1080/02713683.2016.1270326.

Chichilnisky, E. J., & Kalmar, R. S. (2002). Functional asymmetries in ON and OFF ganglion cells of primate retina. *Journal of Neuroscience*, *22*(7), 2737–2747. https://doi.org/10.1523/JNEUROSCI.22-07-02737.2002.

Cottaris, N. P., & Elfar, S. D. (2009). Assessing the efficacy of visual prostheses by decoding ms-LFPs: Application to retinal implants. *Journal of Neural Engineering*, *6*(2), 026007. https://doi.org/10.1088/1741-2560/6/2/026007.

Deng, X., Liu, D. F., Kay, K., Frank, L. M., & Eden, U. T. (2015). Clusterless decoding of position from multiunit activity using a marked point process filter. *Neural Computation*, *27*(7), 1438–1460. https://doi.org/10.1162/NECO_a_00744.

Ellis, R. J., & Michaelides, M. (2018). High-accuracy decoding of complex visual scenes from neuronal calcium responses. *Neuroscience*. https://doi.org/10.1101/271296.

Field, G. D., & Chichilnisky, E. J. (2007). Information processing in the primate retina: Circuitry and coding. *Annual Review of Neuroscience*, *30*(1), 1–30. https://doi.org/10.1146/annurev.neuro.30.051606.094252.

Field, Greg D., Gauthier, J. L., Sher, A., Greschner, M., Machado, T. A., Jepson, L. H., . . . Chichilnisky, E. J. (2010). Functional connectivity in the retina at the resolution of photoreceptors. *Nature*, *467*(7316), 673–677. https://doi.org/10.1038/nature09424.

Frechette, E. S., Sher, A., Grivich, M. I., Petrusca, D., Litke, A. M., & Chichilnisky, E. J. (2005). Fidelity of the ensemble code for visual motion in primate retina. *Journal of Neurophysiology*, *94*(1), 119–135. https://doi.org/10.1152/jn.01175.2004.

Freeman, J., Field, G. D., Li, P. H., Greschner, M., Gunning, D. E., Mathieson, K., . . . Chichilnisky, E. (2015). Mapping nonlinear receptive field structure in primate retina at single cone resolution. *ELife*, *4*, e05241. https://doi.org/10.7554/eLife.05241.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*. New York: Springer.

Garasto, S., Bharath, A. A., & Schultz, S. R. (2018). *Visual reconstruction from 2-photon calcium imaging suggests linear readout properties of neurons in mouse primary visual cortex*. bioRxiv:300392. https://doi.org/10.1101/300392.

Garasto, S., Nicola, W., Bharath, A. A., & Schultz, S. R. (2019). Neural sampling strategies for visual stimulus reconstruction from two-photon imaging of mouse primary visual cortex. In *Proceedings of the 9th International IEEE/EMBS Conference on Neural Engineering* (pp. 566–570). Piscataway, NJ: IEEE. https://doi.org/10.1109/NER.2019.8716934.

Gollisch, T. (2013). Features and functions of nonlinear spatial integration by retinal ganglion cells. *Journal of Physiology–Paris*, *107*(5), 338–348. https://doi.org/10.1016/j.jphysparis.2012.12.001.

Jarosiewicz, B., Sarma, A. A., Bacher, D., Masse, N. Y., Simeral, J. D., Sorice, B., . . . Hochberg, L. R. (2015). Virtual typing by people with tetraplegia using a self-calibrating intracortical brain-computer interface. *Science Translational Medicine*, *7*(313), 313ra179-313ra179. https://doi.org/10.1126/scitranslmed.aac7328.

Johnson, J., Alahi, A., & Fei-Fei, L. (2016). *Perceptual losses for real-time style transfer and super-resolution*. arXiv:1603.08155 [Cs]. http://arxiv.org/abs/1603.08155.

Kingma, D. P., & Ba, J. (2017). *Adam: A method for stochastic optimization*. arXiv:1412.6980 [Cs]. http://arxiv.org/abs/1412.6980.

Kupyn, O., Martyniuk, T., Wu, J., & Wang, Z. (2019). *DeblurGAN-v2: Deblurring (orders-of-magnitude) faster and better*. arXiv:1908.03826 [Cs]. http://arxiv.org/abs/1908.03826.

Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., . . . Shi, W. (2017). *Photo-realistic single image super-resolution using a generative adversarial network*. arXiv:1609.04802 [Cs, Stat].

Lee, J., Mitelut, C., Shokri, H., Kinsella, I., Dethe, N., Wu, S., . . . Paninski, L. (2020). YASS: Yet another spike sorter applied to large-scale multi-electrode array recordings in primate retina. *Neuroscience*. https://doi.org/10.1101/2020.03.18.997924.

Litke, A. M., Bezayiff, N., Chichilnisky, E. J., Cunningham, W., Dabrowski, W., Grillo, A. A., . . . Sher, A. (2004). 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*, *51*(4), 1434–1440. https://doi.org/10.1109/TNS.2004.832706.

Liu, W., Vichienchom, K., Clements, M., DeMarco, S. C., Hughes, C., McGucken, E., Humayun, M. S., De Juan, E., Weiland, J. D., & Greenberg, R. (2000). A neuro-stimulus chip with telemetry unit for retinal prosthetic device. *IEEE Journal of Solid-State Circuits*, *35*(10), 1487–1497. https://doi.org/10.1109/4.871327.

Maeda, S. (2020). *Unpaired image super-resolution using pseudo-supervision*. arXiv:2002.11397 [Cs, Eess].

Marre, O., Botella-Soler, V., Simmons, K. D., Mora, T., Tkačik, G., & Berry, M. J. (2015). High accuracy decoding of dynamical motion from a large retinal population. *PLOS Computational Biology*, *11*(7), e1004304. https://doi.org/10.1371/journal.pcbi.1004304.

Massias, M., Gramfort, A., & Salmon, J. (2018). *Celer: A fast solver for the lasso with dual extrapolation*. arXiv:1802.07481 [Stat].

McCann, B. C., Hayhoe, M. M., & Geisler, W. S. (2011). Decoding natural signals from the peripheral retina. *Journal of Vision*, *11*(10), 1–11. https://doi.org/10.1167/11.10.19.

Moxon, K. A., & Foffani, G. (2015). Brain-machine interfaces beyond neuroprosthetics. *Neuron*, *86*(1), 55–67. https://doi.org/10.1016/j.neuron.2015.03.036.

Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., & Gallant, J. L. (2009). Bayesian reconstruction of natural images from human brain activity. *Neuron*, *63*(6), 902–915. https://doi.org/10.1016/j.neuron.2009.09.006.

Nirenberg, S., & Pandarinath, C. (2012). Retinal prosthetic strategy with the capacity to restore normal vision. In *Proceedings of the National Academy of Sciences*, *109*(37), 15012–15017. https://doi.org/10.1073/pnas.1207035109.

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, *21*(19), 1641–1646. https://doi.org/10.1016/j.cub.2011.08.031.

Odermatt, B., Nikolaev, A., & Lagnado, L. (2012). Encoding of luminance and contrast by linear and nonlinear synapses in the retina. *Neuron*, *73*(4), 758–773. https://doi.org/10.1016/j.neuron.2011.12.023.

Parthasarathy, N., Batty, E., Falcon, W., Rutten, T., Rajpal, M., Chichilnisky, E. J., & Paninski, L. (2017). Neural networks for efficient Bayesian decoding of natural images from retinal neurons [Preprint]. *Neuroscience*. https://doi.org/10.1101/153759.

Passaglia, C. L., & Troy, J. B. (2004). Information transmission rates of cat retinal ganglion cells. *Journal of Neurophysiology*, *91*(3), 1217–1229. https://doi.org/10.1152/jn.00796.2003.

Pillow, J. W., Shlens, J., Paninski, L., Sher, A., Litke, A. M., Chichilnisky, E. J., & Simoncelli, E. P. (2008). Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, *454*(7207), 995–999. https://doi.org/10.1038/nature07140.

Pitkow, X., & Meister, M. (2012). Decorrelation and efficient coding by retinal ganglion cells. *Nature Neuroscience*, *15*(4), 628–635. https://doi.org/10.1038/nn.3064.

Portelli, G., Barrett, J. M., Hilgen, G., Masquelier, T., Maccione, A., Di Marco, S., . . . Sernagor, E. (2016). Rank order coding: A retinal information decoding strategy revealed by large-scale multielectrode array retinal recordings. *Eneuro*, *3*(3). https://doi.org/10.1523/ENEURO.0134-15.2016.

Qian, N. (1999). On the momentum term in gradient descent learning algorithms. *Neural Networks*, *12*(1), 145–151. https://doi.org/10.1016/S0893-6080(98)00116-6.

Ruda, K., Zylberberg, J., & Field, G. D. (2020). Ignoring correlated activity causes a failure of retinal population codes. *Nature Communications*, *11*(1), 4605. https://doi.org/10.1038/s41467-020-18436-2.

Ryu, S. B., Ye, J. H., Goo, Y. S., Kim, C. H., & Kim, K. H. (2011). Decoding of temporal visual information from electrically evoked retinal ganglion cell activities in photoreceptor-degenerated retinas. *Investigative Opthalmology and Visual Science*, *52*(9), 6271. https://doi.org/10.1167/iovs.11-7597.

Schreyer, H. M., & Gollisch, T. (2020). *Nonlinearities in retinal bipolar cells shape the encoding of artificial and natural stimuli.* bioRxiv:2020.06.10.144576.

Schwartz, G., & Rieke, F. (2011). Nonlinear spatial encoding by retinal ganglion cells: When $1 + 1 \neq 2$. *Journal of General Physiology*, *138*(3), 283–290. https://doi.org/10.1085/jgp.201110629.

Schwemmer, M. A., Skomrock, N. D., Sederberg, P. B., Ting, J. E., Sharma, G., Bockbrader, M. A., & Friedenberg, D. A. (2018). Meeting brain–computer interface user performance expectations using a deep neural network decoding framework. *Nature Medicine*, *24*(11), 1669–1676. https://doi.org/10.1038/s41591-018-0171-y.

Turner, M. H., & Rieke, F. (2016). Synaptic rectification controls nonlinear spatial integration of natural visual inputs. *Neuron*, *90*(6), 1257–1271. https://doi.org/10.1016/j.neuron.2016.05.006.

Turner, M. H., Schwartz, G. W., & Rieke, F. (2018). Receptive field center-surround interactions mediate context-dependent spatial contrast encoding in the retina. *eLife*, *7*, e38841. https://doi.org/10.7554/eLife.38841.

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Loy, C. C., Qiao, Y., & Tang, X. (2018). *ESRGAN: Enhanced super-resolution generative adversarial networks*. arXiv:1809.00219 [Cs].

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612. https://doi.org/10.1109/TIP.2003.819861.

Wang, Z., Chen, J., & Hoi, S. C. H. (2020). *Deep learning for image super-resolution: A survey.* arXiv:1902.06068 [Cs].

Warland, D. K., Reinagel, P., & Meister, M. (1997). Decoding visual information from a population of retinal ganglion cells. *Journal of Neurophysiology*, *78*(5), 2336–2350. https://doi.org/10.1152/jn.1997.78.5.2336.

Weiland, J. D., Yanai, D., Mahadevappa, M., Williamson, R., Mech, B. V., Fujii, G. Y., . . . Humayun, M. S. (2004). Visual task performance in blind humans with retinal prosthetic implants. In *Proceedings of the 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (p.p. 4172–4173). Piscataway, NJ: IEEE.

Yoshida, T., & Ohki, K. (2020). Natural images are reliably represented by sparse and variable populations of neurons in visual cortex. *Nature Communications*, *11*(1), 872. https://doi.org/10.1038/s41467-020-14645-x.

Zhang, K., Zuo, W., Gu, S., & Zhang, L. (2017). Learning deep CNN denoiser prior for image restoration. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2808–2817). Piscataway, NJ: IEEE. https://doi.org/10.1109/CVPR.2017.300.

Zhang, K., Zuo, W., & Zhang, L. (2019). *Deep plug-and-play super-resolution for arbitrary blur kernels.* arXiv:1903.12529 [Cs].

Zhang, Y., Jia, S., Zheng, Y., Yu, Z., Tian, Y., Ma, S., . . . Liu, J. K. (2020). Reconstruction of natural visual scenes from neural spikes with deep neural networks. *Neural Networks*, *125*, 19–30. https://doi.org/10.1016/j.neunet.2020.01.033.

Zhang, Y., Tian, Y., Kong, Y., Zhong, B., & Fu, Y. (2020). Residual dense network for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, January. https://doi.org/10.1109/TPAMI.2020.2968521.

Zhou, R., & Susstrunk, S. (2019). Kernel modeling super-resolution on real low-resolution images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 2433–2443). https://doi.org/10.1109/ICCV.2019.00252.

Received September 15, 2020; accepted January 25, 2021.