Modeling spatial contrast sensitivity in responses of primate retinal ganglion cells to natural movies

Shashwat Sridhar [1,2] ✉, **Michaela Vystrčilová** [3], **Mohammad H. Khani** [1,2,‡], **Dimokratis Karamanlis** [1,2,§], **Helene M. Schreyer** [1,2,‡], **Varsha Ramakrishna** [1,2,4], **Steffen Krüppel** [1,2], **Sören J. Zapp** [1,2], **Matthias Mietsch** [5,6], **Alexander Ecker** [3,7], **Tim Gollisch** [1,2,8] ✉

[1]University Medical Center Göttingen, Department of Ophthalmology, Göttingen, Germany; [2]Bernstein Center for Computational Neuroscience Göttingen, Göttingen, Germany; [3]University of Göttingen, Institute of Computer Science and Campus Institute Data Science, Göttingen, Germany; [4]International Max Planck Research School for Neurosciences, Göttingen, Germany; [5]German Primate Center, Laboratory Animal Science Unit, Göttingen, Germany; [6]German Center for Cardiovascular Research, Partner Site Göttingen, Göttingen, Germany; [7]Max Planck Institute for Dynamics and Self-Organization, Göttingen, Germany; [8]Cluster of Excellence "Multiscale Bioimaging: from Molecular Machines to Networks of Excitable Cells" (MBExC), University of Göttingen, Göttingen, Germany

✉ **For correspondence:** shashwat.sridhar@med.uni-goettingen.de; tim.gollisch@med.uni-goettingen.de

Present address: ‡Institute of Molecular and Clinical Ophthalmology Basel, Basel, Switzerland

Present address: §University of Geneva, Department of Basic Neurosciences, Geneva, Switzerland

Funding: This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project numbers 432680300 (SFB 1456, project B05) and 515774656 – and by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement number 101041669). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Retinal ganglion cells, the output neurons of the vertebrate retina, often display nonlinear summation of visual signals over their receptive fields. This creates sensitivity to spatial contrast, letting the cells respond to spatially structured visual stimuli, such as a contrast-reversing grating, even when no net change in overall illumination of the receptive field occurs. Yet, computational models of ganglion cell responses are often based on linear receptive fields. Nonlinear extensions, on the other hand, such as subunit models, which separate receptive fields into smaller, nonlinearly combined subfields, are often cumbersome to fit to experimental data, in particular when natural stimuli are considered. Previous work in the salamander retina has shown that sensitivity to spatial contrast in response to flashed images can be partly captured by a model that combines signals from the mean and variance of luminance signals inside the receptive field. Here, we extend this spatial contrast model for application to spatiotemporal stimulation and explore its performance on spiking responses that we recorded from retinas of marmosets under artificial and natural movies. We show how the model can be fitted to experimental data and that it outperforms common models with linear spatial integration, in particular for parasol ganglion cells. Finally, we use the

model framework to infer the cells' spatial scale of nonlinear spatial integration and contrast sensitivity. Our work shows that the spatial contrast model provides a simple approach to capturing aspects of nonlinear spatial integration with only few free parameters, which can be used to assess the cells' functional properties under natural stimulation and which provides a simple-to-obtain benchmark for comparison with more detailed nonlinear encoding models.

Introduction

The ganglion cells of the vertebrate retina encode visual stimuli that occur inside their receptive fields. When they are activated by an appropriate stimulus, the cells report this by generating spikes for transmission along their axons, the fibers of the optic nerve. The most basic principle of this encoding is that ON ganglion cells are activated by an increase of light intensity in their receptive field centers, OFF ganglion cells by a decrease of light intensity, and ON-OFF cells by light-intensity changes of either sign (Hartline, 1938). In computational models of retinal ganglion cell activity, this is often captured by treating the receptive field as a (linear) filter of the incoming light-intensity pattern. Applying this filter amounts to computing a weighted sum over space of the change in illumination, that is, the temporal contrast, with weights given by the profile of the receptive field over space. The obtained filter activation can then be transformed into an actual response, such as a firing rate, in a way that takes the preferred contrast polarity of the ganglion cell — ON, OFF, or ON-OFF — into account. In its most basic form, this procedure corresponds to the commonly used linear–nonlinear (LN) model (Chichilnisky, 2001), where the receptive-field filtering represents the model's linear stage and the subsequent transformation into the cell's response the model's nonlinear stage.

Yet, despite the success of the LN model for describing retinal ganglion cells, as well as of its many variants that maintain the linear receptive field, it has long been known that many ganglion cells display nonlinear receptive fields (Enroth-Cugell and Robson, 1966; Zapp et al., 2022). Typically, this is demonstrated by responses to stimulation with contrast-reversing spatial gratings, showing that gratings can evoke sizeable responses even when the contrast reversals do not activate the linear receptive field because light-intensity increments and decrements symmetrically balance each other. The nonlinearities of the receptive field then lead to a characteristic frequency doubling in the response, as contrast reversals in both directions can activate the cell about equally strongly (Enroth-Cugell and Robson, 1966).

The receptive-field nonlinearities are, for some ganglion-cell types, considered a crucial component to be included in computational models for accurately capturing responses to natural stimuli (Gollisch, 2013; Heitman et al., 2016; Karamanlis et al., 2022; Schwartz & Rieke, 2011). A typical approach is to separate the receptive field into multiple smaller spatial filters, so-called subunits, and to then nonlinearly combine the resulting filter signals (Victor & Shapley, 1979). Several methods for inferring subunits from experimental data have been suggested (e.g., Krüppel et al., 2023; Liu et al., 2017; Maheswaranathan et al., 2018; Shah et al., 2020), which, in addition to serving as the basis for modeling, may also provide insights into retinal circuitry, as the subunits are thought to correspond to the receptive fields of presynaptic bipolar cells (Borghuis et al., 2001; Demb et al., 2001;

Liu et al., 2017). Yet, subunit inference remains a formidable challenge, requiring long recordings and computationally demanding analyses. Moreover, it is not clear how much the actual precise subunit locations matter for response predictions (Freedland & Rieke, 2022).

As a simpler alternative to subunit models, one may therefore consider capturing nonlinear spatial integration by including information about higher-order statistics of light intensities inside the receptive field. Indeed, prior work on modeling responses from salamander and mouse retinal ganglion cells to flashed natural images has shown that a substantial improvement over the LN model can be obtained by simply providing a measure of spatial contrast (given by the pixel-wise light-intensity variance weighted by the receptive field) as an additional input channel (Liu et al., 2022). Here, we extend this spatial contrast model to also include temporal integration, thus allowing application to spatiotemporal stimulation. Moreover, we apply the model to data recorded from the primate (marmoset) retina recorded in response to natural movies and demonstrate that the spatial contrast model can outperform the LN model in a cell-type-dependent and stimulus-dependent manner.

Methods & Materials

Ethics statement

All experimental procedures were performed in strict accordance with national and institutional guidelines. Electrophysiological recordings from marmoset retinas were performed on tissue obtained from animals used by other researchers, as approved by the institutional animal care committee of the German Primate Center and by the responsible regional government office (Niedersächsisches Landesamt für Verbraucherschutz und Lebensmittelsicherheit, permit number 33.19-42502-04-17/2496).

Electrophysiological recordings

We used retinas from adult common marmosets (*Callithrix jacchus*) of either sex. The eyes were enucleated and the cornea, lens, and vitreous humor were carefully removed immediately after the sacrifice of the animal. Eyecups were transported to the electrophysiology lab in a light-sealed container filled with freshly oxygenated (95% O_2 and 5% CO_2) Ames' medium (Sigma-Aldrich), supplemented with 6 mM D-glucose and buffered with 22 mM $NaHCO_3$ to maintain a pH of 7.4. Retinas were then dark-adapted for 1–2 hours in a light-sealed container while submerged in continually replenished oxygenated Ames' medium. Care was taken to avoid oxygen depletion and to reduce the amount of light incident onto the retinal tissue to a minimum.

Each retina was first hemisected along the superior-inferior axis. For recording, a smaller piece, a few millimeters wide, was then cut off, carefully separated from the sclera, choroid, and pigment epithelium, and gently flattened onto a transparent dialysis membrane held taut across a hollow, plastic support, with the photoreceptor-side of the tissue towards the membrane. The support, along with the membrane and retinal piece, was slid into a plastic bath chamber, designed to snugly fit the support and glued onto a planar microelectrode array (MEA; MultiChannel Systems) in such a way that the inverted support would position the retinal tissue on top of the recording electrodes with ganglion cell-side down. MEAs had 252 electrodes with 10 µm electrode diameter and

60 µm electrode spacing and were coated prior to experiments by a thin layer of poly-D-lysine, which helped keep the retinal tissue in place. The support with the membrane was carefully removed, leaving the retina piece on the MEA, and replaced with a support without a membrane, but with a small horizontal glass plate a few millimeters above the retinal tissue. This minimized refractive effects of the stimulus light when the bath chamber was filled with perfusion solution up to the lower surface of the glass plate. The MEA was then transferred to the recording setup and connected with an amplifier and analog-to-digital converter (MultiChannel Systems). Once mounted, the retina was perfused with the oxygenated Ames' medium at a rate of ~400–450 ml/h while being held at a temperature of 32–34°C with the help of an inline heater and a heating element under the array. The remaining retina pieces were stored in a light-sealed container at the same range of temperatures under a constant perfusion of the Ames' medium for later experiments.

Visual stimulation and basic cell characterizations

Visual stimuli were rendered by a custom-made software, written in C++ and OpenGL. The stimuli were displayed on a gamma-corrected monochrome grayscale OLED monitor (eMagin) at a refresh rate of 85 Hz and a resolution of 800x600 pixels. A telecentric lens (Edmund Optics) was used to project the stimuli onto the retina, resulting in a pixel size of 7.5 µm x 7.5 µm on the retinal surface.

Spatiotemporal white-noise stimulus

We stimulated each retina with a spatiotemporal white-noise stimulus, which consisted of a sequence of frames of black and white squares (100% Michelson contrast) arranged in a uniform checkerboard pattern, where each square was 30 µm x 30 µm on the surface of the retina. This resulted in a stimulus resolution of 150x200 stimulus squares. The frames of the stimulus were updated at the monitor refresh rate of 85 Hz. The brightness of each square was measured in contrast values of $+1$ or -1 around a measured mean light level of 5.5 mW/m^2.

The contrast values for each square were independently and randomly generated from a binary distribution. The duration of this stimulus in each recording was 60 min. The stimulus was divided into trials, each consisting of a 150 s-long non-repeating part, which was used for model training, and a 30 s-long repeating part, which was the same across trials and used for model testing.

Estimation of receptive fields and stimulus filters. We estimated the receptive field of each cell using the responses to the training set of the spatiotemporal white-noise stimulus and computing the spike-triggered average (STA). We computed the spatiotemporal STA from the binned responses (85 Hz) over the past 30 stimulus frames (corresponding to 353 ms). Responses to the first 29 frames in each trial were excluded in this computation.

From the spatiotemporal STA, we extracted a spatial and a temporal filter in the following way. For each STA, we selected pixels whose absolute peak intensity over time exceeded six times the robust standard deviation of the absolute peak intensities of all pixels around their median. The robust standard deviation is defined as 1.4826 times the median absolute deviation and is considered a consistent estimator of the standard

deviation of a normal distribution. The temporal filter was taken as the average of the time course of the selected pixels in the STA. The spatial component was computed by projecting the three-dimensional STA onto the temporal filter. The receptive field center was estimated by determining the pixel in the STA with the largest variance along the temporal dimension. For the spatial filter we took the spatial component of the STA, cropped to a region of 40x40 pixels (1200 µm x 1200 µm) around the receptive field center. The spatial and the temporal filters were normalized by their respective Euclidean norms.

When computing the spatial filter for modeling responses to the naturalistic movie stimulus, we followed the same procedure, but first upsampled the spatial dimensions of the STA from 200x150 squares of the checkerboard layout to 800x600 pixels, the resolution of the movie stimulus, before computing the spatial and temporal filters. The upsampling was done without interpolation. The resulting spatial filter after cropping around the receptive field center then had a resolution of 160x160 pixels.

For modeling, we fitted a 2-dimensional elliptical Gaussian function to each spatial filter and used the fit as a regularized spatial filter. The fitting was done using the `curve_fit` function of Python's `scipy` library. Pixels beyond a 3-σ boundary around the Gaussian center were set to zero. From the Gaussian fit, we also estimated the receptive-field size as the diameter of a circle that has the same area as within the 1.5-σ ellipse of the fit.

Autocorrelograms. To aid cell-type classification, we computed the autocorrelogram for each cell, based on its responses to the training segments of spatiotemporal white noise. We used a bin size of 0.5 ms and retained 150 bins, corresponding to a range of temporal lags of 0–75 ms. Each autocorrelogram was normalized to unit sum.

Naturalistic movie stimulus

To create a naturalistic movie stimulus, we utilized a short, publicly accessible science-fiction film released by the Blender Foundation under the CCA 3.0 license (Tears of Steel, accessible at https://mango.blender.org/). The movie was chosen for its diverse array of textures and light intensities. For experiments, the movie frames, provided in the Portable Network Graphics (PNG) format, were sequentially projected onto the retina. Excluding frames from the movie's introductory and credit sequences, as well as extended periods of homogeneous black, left us with 14,544 movie frames, equivalent to approximately 10 minutes of the movie. Of these, 1,434 continuous frames were reserved as a test stimulus, with the remaining 13,110 movie frames constituting the training set. The test segment was chosen through visual inspection, focusing on scenes featuring a balanced mix of fast and slow moving objects and diverse textures. To reduce memory requirements, we cropped each movie frame from a resolution of 1920x800 pixels to 1000x800 pixels, by removing 460 pixels from each end along the horizontal axis.

The original movie's frame rate was 24 Hz, which differed from the projector's refresh rate of 85 Hz. For presentation to the retina, we therefore upsampled the cropped frames by a factor of 3.54 by repeating individual frames three to four times, randomly selecting for each frame the presentation duration of three or four projector frames with probabilities of 0.46 and 0.54, respectively. Each cropped frame was converted from the RGB color space to grayscale using a weighted sum of R, G, and B values to compute linear luminance as $0.2126\boldsymbol{R} + 0.7152\boldsymbol{G} + 0.0722\boldsymbol{B}$.

To generate a unique, long stimulus sequence from a short movie and to increase the diversity of stimulus patterns within the receptive fields of individual ganglion cells, we

manipulated the stimulus in the following ways. First, we repeated the grayscale frames of the training set as many times as needed to yield a training set size of 255,000 frames, corresponding to 10 trials of 300 s each. Next, we added gaze-shift-like global motion signals to the stimulus by shifting the center of each frame independently using simulated eye movements (see below). Occasionally, due to the overlayed movement, the shifted frame exceeded the bounds of the cropped movie frame, leading to the replacement of missing pixels with gray pixels at the edges of the presentation region, but on the retina this did not overlap with the region from which ganglion cells were recorded. Lastly, to add further variability, 40% of the training segments were flipped vertically. The test stimulus consisted of 10 repeats of the grayscale test frames overlayed with simulated eye movements, but not flipped.

The final stimulus was projected onto the retina at a resolution of 800x600 pixels, with each pixel covering 7.5 µm x 7.5 µm on the surface of the retina, and a mean light level of 76% of the white-noise stimulus in order to accommodate for the range of pixel intensities and the limited maximal intensity of the projector. The root–mean–square contrast of the movie was 45%.

For analysis, we converted the luminance intensity values of each pixel in the stimulus to Weber contrast values by subtracting its mean intensity over time and subsequently normalizing by the same mean intensity.

Simulated eye movements. We introduced global motion to the movie stimulus by spatially shifting the center position of each frame, mimicking saccades, fixations, and fixational eye movements, consistent with the statistics of active vision in marmosets (Mitchell et al., 2014). Saccades were characterized by a combination of amplitude, direction, and duration. Saccade amplitudes were randomly sampled from an exponential distribution with a scale parameter of 200 µm. The directions of saccades were randomly drawn from a uniform distribution between 0 and 360 degrees. Given that the displayed movie's frame rate was constrained by the monitor's refresh rate, saccade duration was sampled from three possible values (23.5 ms, 35.2 ms, 47.1 ms), corresponding to 2, 3, and 4 frame updates, with probabilities of 0.35, 0.4, and 0.25, respectively. During these periods the frame's center position was linearly shifted from the original position to the target of the saccade.

Fixation durations were randomly sampled from an exponential distribution with a refractory period. This type of distribution was originally proposed as a model for fixation times in humans (Harris et al., 1988), and has since been used by others to model the statistics of fixation duration in other animal models (Rucci et al., 2000, Angueyra et al., 2022). We here applied a scale parameter of 200 ms and a refractory period of 100 ms as used previously for primate-retina recordings (Angueyra et al., 2022). Additionally, to simulate fixational eye movements, the center of each cropped frame was jittered in the x and y directions for every frame update during periods of fixation at 85 Hz. The amplitude of the jitter was sampled independently for each direction from a normal distribution with a mean of zero pixels and a standard deviation of two pixels.

Simulated eye movement traces were generated in 10 s chunks for both the training and the test segments. The fixation center was reset to the center of the monitor at the beginning of every chunk, preventing the fixation center from drifting out of the cropped frame. Additionally, chunks where the fixation center drifted more than 200 monitor pix-

els in any direction were rejected and regenerated.

Data processing

The recorded extracellular signals were amplified, bandpass-filtered between 300 Hz and 5 kHz, and stored digitally at a sampling rate of 25 kHz. Spikes were extracted using a modified version of the Kilosort2 spike-sorting algorithm (Pachitariu et al., 2023, modified code available at https://github.com/dimokaramanlis/KiloSortMEA). The resulting clusters were manually curated using the software package Phy2 (https://github.com/cortex-lab/phy). Only well-isolated clusters with minimal refractory period violations were chosen for further analysis, and each such cluster was deemed to correspond to a single cell. For the purpose of modeling, spike times of each cell were binned at the resolution of the projector refresh rate (85 Hz).

Cell selection

Since models cannot be expected to yield good predictions if responses are unreliable across trials, we selected cells based on the reliability of their responses to the test stimuli. We evaluated reliability separately for the white-noise stimulus and the naturalistic movie and required cells to pass reliability criteria for both stimuli.

We used two measures of reliability. The first was the fraction of explainable variance $\mathrm{FEV} = (\sigma^2_{total} - \sigma^2_{noise})/\sigma^2_{total}$ (Cadena et al., 2019; Heitman et al., 2016), used as a measure of the variability in the data that cannot be attributed to noise and thus should be predictable. Here, the total variance σ^2_{total} is the variance of spike counts per time bin across all stimuli and trials, and the noise variance σ^2_{noise} is the spike-count variance computed across trials for a given time bin, averaged across time.

The second measure was the coefficient of determination (R^2) between the binned firing-rate responses to odd versus even trials of the test set. Here, we used a symmetrized version of the coefficient of determination (Karamanlis and Gollisch, 2021), where the responses to either the odd or the even trials are taken as the prediction of the other and the average of both versions is used.

We selected, based on visual inspection, conservative thresholds of 0.15 for FEV and 0.0 for R^2 to include cells for analysis. In total, 180 recorded cells passed our reliability criterion.

Cell-type classification

We used a semi-automated approach to identify cell types in our data set, using a custom-made graphical user interface. To start, we constructed a feature vector for each reliable cell, which consisted of the elements of the temporal filter, the receptive-field size, and the scores corresponding to the first ten components from a principal component analysis (PCA) of the Z-scored autocorrelograms of all cells. After Z-scoring each entry in the feature vector independently across cells, we reduced the dimensionality of the feature vectors using principal component analysis, retaining as many components as necessary to explain 90% of the variance (26 components). Finally, we used KMeans++ on the reduced-dimensionality feature vectors to determine 6–8 initial clusters to which the cells could belong. These clusters were manually curated to ensure each cluster had similarly-sized receptive fields that tiled the retinal surface and showed consistency in the shapes of the temporal filters and autocorrelograms. We used functions from the

Python library `scikit-learn` to run the PCA and KMeans++ algorithms.

We identified two clusters as corresponding to OFF midget cells (n = 51) and OFF parasol cells (n = 31), according to receptive field size and temporal filters. For analysis, we selected three additional clusters with fairly consistent features and containing multiple cells and grouped these into a single class of Large OFF cells (n = 15). This multi-cluster class is not considered to correspond to a single cell type. Cells not included in one of these groups were discarded as they either displayed unreliable receptive field measurements or came from clusters with only few samples. This also concerned ON cells, of which only few reliable examples were found in our data set.

Modeling

Linear-nonlinear (LN) model

We employed a classical LN model to predict the responses of individual ganglion cells to white noise and natural stimuli. The linear component of the model involved a filtering operation of the stimulus with the cell's temporal and spatial filters, and the nonlinear part incorporated the activation function describing the relationship between the linear filter output and the cell's firing response.

We assumed that the spatiotemporal filter of a cell can be separated into a temporal and a spatial filter and obtained each, as described above, from the STA of the cell under spatiotemporal white noise. For filtering, we first convolved the stimulus X_{ijk} pixel-wise with the cell's temporal filter T_k (where i, j, and k index the two spatial dimensions as well as time) to obtain the convolved stimulus S_{ijk}:

$$S_{ijk} = \sum_{k'} X_{i,j,k-k'} \cdot T_{k'}. \tag{1}$$

We then applied the regularized spatial filter R_{ij} (taken as the Gaussian fit to the spatial component of the receptive field) by computing the scalar product of the pixel-wise elements of the stimulus and the filter to obtain the linear activation L_k^{LN} of the LN model at time bin k:

$$L_k^{LN} = \sum_{i,j} R_{ij} \cdot S_{ijk}. \tag{2}$$

Finally, the filtered signal is transformed by the nonlinear function $g^s(\cdot)$ to obtain the model's expected spike-count response f_k^{LN}:

$$f_k^{LN} = g^s\left(L_k^{LN}\right). \tag{3}$$

We here take the output nonlinearity g to be a softplus function, parameterized as

$$g(x) = a \cdot \log\left[1 + \exp(b \cdot x + c)\right] \tag{4}$$

The parameters a, b, and c are learned during model fitting. The details of model fitting are described below.

Spatial contrast (SC) model

We extended the SC model, originally proposed for flashed natural images (Liu et al., 2022), to spatiotemporal stimuli. In our version of the model, the temporal component of the stimulus is initially accounted for by convolving it with the cell's temporal filter as in Eq. 1. From the convolved stimulus, we computed the mean-luminance signal $\mathrm{I}_{\mathrm{mean}}$

and the local spatial contrast LSC within the receptive field of the cell and combined them linearly. Here, $\mathrm{I}_{\mathrm{mean},k}$ is a linearly filtered stimulus similar to Eq. 2, defined here as a weighted average of the (temporally convolved) stimulus S_{ijk} for every time point k, with weights coming from the regularized spatial filter R_{ij}:

$$\mathrm{I}_{\mathrm{mean},k} = \frac{\sum_{ij} R_{ij} \cdot S_{ij}}{\sum_{ij} R_{ij}}. \tag{5}$$

The LSC is analogously obtained as a weighted standard deviation of the pixel intensities from the (temporally convolved) stimulus S_{ij}:

$$\mathrm{LSC}_{,k} = \sqrt{\frac{\sum_{ij} R_{ij} \cdot \left(S_{ijk} - \mathrm{I}_{\mathrm{mean},k}\right)^2}{\sum_{ij} R_{ij}}}. \tag{6}$$

Note that this definition of LSC slightly differs from the original one suggested in Liu et al., 2022, where LSC was simply defined as a standard deviation of the spatially filtered stimulus. While for large receptive fields relative to pixel size, as was the case in Liu et al., 2022, the two definitions are nearly equivalent, Eq. 6 is superior for smaller receptive fields, as it is not sensitive to variance of pixel intensities induced by the spatial filtering.

To obtain the model response from $\mathrm{I}_{\mathrm{mean}}$ and LSC, a free parameter w was used to determine the relative importance of spatial contrast information for each cell. Finally, an activation function, again taken as a parameterized softplus function, modeled the relationship between the linearly combined signal and the expected spike-count response of the cell, similar to the LN model:

$$f_k^{SC} = g\left(\mathrm{I}_{\mathrm{mean},k} + w \cdot \mathrm{LSC}_{,k}\right). \tag{7}$$

Together, the SC model has four free parameters – a, b, and c from Eq. 4 and the weight w. Optimization of the parameters is described below.

Model fitting and evaluation

For both models described above, the responses of cells were binned by the update times of the stimulus (85 Hz). Model parameters were optimized by a maximum-likelihood fit under the assumption that spike counts per bin are determined according to an independent Poisson process with the expectation value given by the model response for that bin, f_k^{LN} (Eq. 3) and f_k^{SC} (Eq. 7) for the LN model and the SC model, respectively. The probability under each model of observing the complete binned spike train with spike counts n_k is given by the product of the probabilities for individual bins:

$$P(\{n_1, n_2, \ldots\}) = \prod_k \frac{f_k^{n_k} e^{-f_k}}{n_k!} \tag{8}$$

where f_k is either f_k^{LN} or f_k^{SC}. This yields the negative log-likelihood function (up to a constant)

$$\mathcal{L}\left[P(N)|\Theta\right] = -\sum_k n_k \ln f_k(\Theta) + \sum_k f_k(\Theta), \tag{9}$$

where Θ represents the parameters of the model to be optimized, that is the a, b, and c from the softplus nonlinearity, and, in the case of the SC model, the weight w. The minimization of Eq. 9 was carried out for both models by employing the `minimize` function

provided in Python's `scipy` library, using the `L-BFGS-B` algorithm, which we found to be the most reliable for our problem.

We Z-scored the activation signal obtained from the spatiotemporal filtering in the LN model as well as the $\mathrm{I}_{\mathrm{mean}}$ and the LSC signals for the SC model by their respective means and standard deviations. The same means and standard deviations were used to Z-score the corresponding signals from the test set. Such a normalization allowed us to use a common set of initial parameters for optimization of each model across all cells. For the LN model, we used the following initial values for the model parameters: $a = max(n_k)$, $b = 1.0$, $c = -2.0$, where n_k are the binned spike count responses to the training set of the stimulus. For the SC model, the following initial values were used: $a = max(n_k)$, $b = 1.0$, $c = -2.0$, $w = 0.0$. Since the firing response of a cell cannot be negative, a lower bound of zero was set for the parameter a in both models.

Model predictions were evaluated against the average response of each cell across trials of the test set using the Pearson correlation coefficient r.

Spatial smoothing and optimal scale of spatial integration

We used a modified version of the SC model to determine the scale at which ganglion cells integrate spatial nonlinearities. In this version of the model, the temporally convolved stimulus (S_{ijk}, Eq. 1) was smoothed before extracting the $\mathrm{LSC}_{,k}$ from it. The computation of $\mathrm{I}_{\mathrm{mean},k}$ remained unaffected. The smoothing was performed by convolving S_{ijk} with a circular Gaussian filter, and the scale of spatial smoothing was defined as three standard deviations of the filter in units of pixels. This definition of scale was chosen so as to be in line with our definition of receptive-field size via the area within the 1.5-σ contour of a Gaussian fit. To perform the smoothing, we used the function `gaussian_filter` provided by the Python library `scipy`, with the radius of the applied region of the filter taken to be three standard deviations of the filter.

For both white noise and naturalistic movie, we tested 20 different values of standard deviation of the smoothing filter, ranging from 0.8 to 12.0 pixels (in steps of 0.4 pixels from 0.8 to 6.4 and then 7.2, 8.0, 8.8, 10.4, and 12.0 pixels). In each case, model parameters were optimized independently. The change in model performance relative to the case without stimulus smoothing was measured for each cell as the ratio between the Pearson's *r* with smoothing and the Pearson's *r* without smoothing. The optimal scale of smoothing for a cell was taken to be the spatial scale at which the relative model improvement for the cell was largest.

Results

Spatial-contrast-driven nonlinear integration in retinal ganglion cells

To investigate how spatial contrast inside the receptive fields of primate retinal ganglion cells affects stimulus encoding under natural movies, we recorded the spiking activity from marmoset ganglion cells in response to various stimuli with microelectrode arrays. To be able to assess the stimulus inside a cell's receptive field, we used reverse-correlation analysis under stimulation with spatiotemporal white noise to determine, for each recorded cell, its spatial receptive field and temporal filter (Fig. 1A,B). This also allowed us to distinguish between different types of ganglion cells, such as parasol and midget cells (see *Methods*).

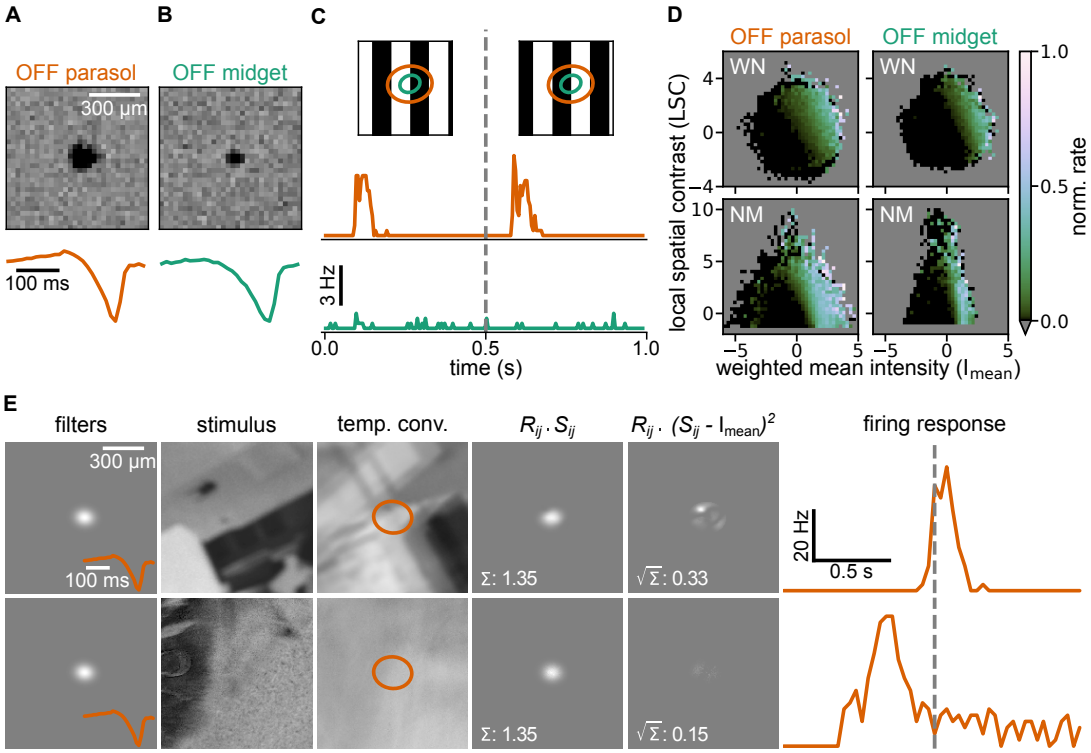

Figure 1. Spatial receptive-field nonlinearities of marmoset retinal ganglion cells. A: Spatial receptive field (top) and temporal filter (bottom) extracted from the spike-triggered average (STA) of an OFF parasol cell. **B**: Same as **A**, but for an OFF midget cell. **C**: Firing rates of the two sample cells in response to reversing gratings (60 µm spatial period), shown on top, with ellipses denoting the 3-σ contour of Gaussian fits to the receptive fields. **D**: Firing rate histograms for the two sample cells from the training segments for white-noise (WN, top) and naturalistic (NM, bottom) stimulation, depending on the mean luminance signal $\mathrm{I}_{\mathrm{mean}}$ and the local spatial contrast LSC. Both $\mathrm{I}_{\mathrm{mean}}$ and LSC have been Z-scored for this analysis. The firing rates have been normalized to the maximum entry in each histogram. Bins without any entries are depicted as gray. **E**: Illustration of the stimulus processing to obtain $\mathrm{I}_{\mathrm{mean}}$ and LSC by applying the filters of the OFF Parasol cell from **A** (1st column) to the stimulus (2nd column) to obtain a temporally convolved stimulus (3rd column), from which the weighted luminance signal (4th column) and the weighted spatial contrast (5th column) are extracted. The two rows correspond to two examples. The plots of the luminance and spatial contrast signals display the results of the pixel-wise computation as indicated by the formulas at the top, with the final values given in the lower left of the plots. The corresponding measured firing rates of the sample OFF parasol cell are shown on the right, with the dashed line corresponding to the time of the analyzed stimulus frame.

Different types of ganglion cells have different sensitivity to spatial structure inside their receptive fields, owing to different degrees of nonlinear spatial integration they perform. For example, compared to the relatively linear midget cells, the more nonlinear parasol cells typically display more pronounced and frequency-doubled responses under contrast-reversing gratings, as we also found in our recordings (Fig. 1C).

To understand how this difference manifests itself in responses to richer stimuli with more spatiotemporal complexity, we stimulated the retina with spatiotemporal white noise as well as with a naturalistic movie. The white-noise stimulus consisted of a flickering checker-board pattern, and the naturalistic movie of segments from a publicly available movie, which were modified by shifting the movie frames according to a model of gaze shifts, including saccades and fixational eye movements.

To assess how the spatial structure of the stimulus shapes the response of a ganglion cell to such stimuli beyond the simple, linear activation of its receptive field, we quantified the effective luminance as well as the spatial structure inside the receptive field at each time point of the stimulus. To do so, we first filtered the stimulus by convolving it with the temporal filter of the cell, as obtained from the reverse-correlation analysis, to take care of the cell's temporal stimulus integration. This yielded, for every point in time, an effective stimulus S_{ij}, with spatial coordinates i and j. We then measured the mean luminance signal, $\mathrm{I_{mean}}$, as the weighted average of pixel intensity values (given by their contrast relative to the mean stimulus intensity) in the receptive field of the cell, where the weights are given by the regularized spatial filter R_{ij}, here defined as a Gaussian fit to the spatial component of the spike-triggered average under white noise. $\mathrm{I_{mean}}$ is a measure of the activation of the linear part of the receptive field and is commonly used as the input for computational models that do not account for receptive-field nonlinearities.

In addition, we also measured the local spatial contrast, LSC, in each frame of the temporally filtered stimulus by computing the weighted standard deviation of the pixel intensity values within the receptive field of the cell, where the weights are again given by the regularized spatial filter, R_{ij} (see *Methods*, Eq. 6). The LSC provides a measure of the effective spatial structure in the stimulus that the cell was exposed to at a given point in time.

After computing the $\mathrm{I_{mean}}$ and the LSC for the spatiotemporal white-noise and naturalistic movie stimuli for a given cell, we analyzed how the cell's activity depended on the two measures by compiling two-dimensional histograms of the average firing rate in response to different combinations of $\mathrm{I_{mean}}$ and LSC (Fig. 1D). For the sample parasol cell, the colored regions in these plots, which correspond to regions of high firing rate, have a slanted orientation, indicating higher firing rate towards the upper right corner of the plot, where both $\mathrm{I_{mean}}$ and the LSC are large. If LSC had no effect on the firing rate, regions of increased firing rate should not depend on the value along the y-axis and rather be fully vertically oriented. Thus, the plots reveal a dependence of the firing rate on the LSC for both stimuli beyond the effect of the $\mathrm{I_{mean}}$. For the midget cell, this tendency is less pronounced, suggesting that the LSC has less influence over the cell's firing.

To illustrate how the LSC can shape the firing rate of the parasol cell, Fig. 1E shows examples from two different segments of the natural movie. Each row displays the computation of $\mathrm{I_{mean}}$ and LSC at a given point in time from the corresponding temporally convolved movie frame, which here resulted in nearly identical $\mathrm{I_{mean}}$, but substantially different LSC values. Comparing the corresponding firing rates (Fig.1E, right), shows that

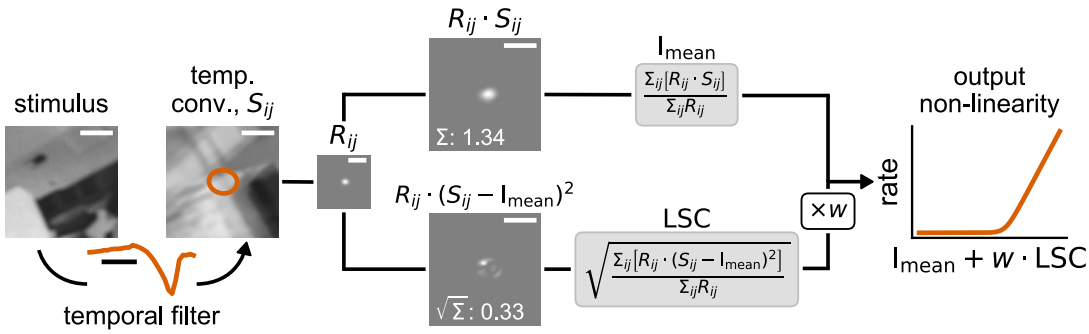

Figure 2. Schematic of the spatial contrast (SC) model. For a given cell, the incoming stimulus is first convolved with the cell's temporal filter to yield the effective stimulus frame S_{ij}. From this, the mean luminance signal $\mathrm{I_{mean}}$ and the spatial contrast signal LSC are computed by using the the cell's regularlized spatial filter R_{ij} as weights. The $\mathrm{I_{mean}}$ and LSC are linearly combined with a weight parameter w used to scale the LSC contribution, and the combined signal is transformed by an output nonlinearity to yield the model's firing rate response. Scale bars: spatial: 300 µm, temporal: 100 ms.

the larger LSC value (top row) yielded a much larger firing rate at the time point of interest (dashed line). Note that the earlier firing-rate peak visible for the example of the smaller LSC value (bottom row) corresponds to earlier, and thus different, $\mathrm{I_{mean}}$ and LSC values. These examples corroborate that spatial contrast plays a crucial role in shaping the firing response of certain ganglion cells and suggest that the LSC can be used to capture these effects in computational models.

Predictive model with spatial-contrast sensitivity

In order to incorporate the LSC into a predictive model for responses to spatiotemporal stimuli, we extended the spatial contrast (SC) model that had previously been used to predict response to flashed natural images in the salamander retina (Liu et al., 2022). The SC model itself is an extension of the commonly used linear–nonlinear (LN) model (Chichilnisky, 2001). In the linear part of the LN model, the stimulus is filtered with the receptive field of the cell to yield the mean luminosity in the stimulus weighted by the receptive field, equivalent to computing the cell's $\mathrm{I_{mean}}$ signal. The nonlinear part of the LN model then transforms the weighted mean luminosity into the model's response (e.g., firing rate or spike probability).

In the SC model, in addition to the $\mathrm{I_{mean}}$ signal, the LSC is measured for the stimulus, and the two are linearly combined (see *Methods*, Eq. 7) using the weight parameter w, which determines the importance of the local spatial contrast relative to the mean luminance signal. This enables the SC model to scale the dependence on LSC independently for each cell. The linearly combined signal is then transformed into the model's response by the output nonlinearity, much like in the LN model. For both models, we here apply a softplus rectifying nonlinearity that is parameterized with three free parameters (see *Methods*, Eq. 4).

In our spatiotemporal version of the SC model, the calculation of $\mathrm{I_{mean}}$ and LSC is preceded by a temporal convolution of the stimulus with the cell's temporal stimulus filter (see *Methods*, Eq. 1). The full model is summarized in the schematic in Fig. 2. Together, implementing the SC model for a given cell requires estimating the cell's spatial

receptive field and temporal filter, which can both be obtained by reverse correlation under spatiotemporal white-noise stimulation, as well as optimizing four free parameters, three for the softplus nonlinearity and one for the weight w of the LSC. We optimize the free parameters simultaneously, using negative log-likelihood minimization under the assumption of a Poisson spiking process.

Cell-type-specific comparison of LN and SC model

We next tested whether the SC model indeed improves model predictions as compared to the LN model for white noise as well as for naturalistic movies. To do this in a cell-type-specific manner, we first performed a cluster analysis of the recorded cells, based on their receptive field sizes, their temporal filters, and the autocorrelograms of their spike trains. For the final analysis, we focused on three classes of cells: OFF midget cells, OFF parasol cells, and Large OFF cells (Fig. 3B). OFF midget and OFF parasol cells were obtained as well-separated clusters, each with consistent autocorrelograms and temporal filters and with receptive fields of similar sizes, which approximately tile the retinal surface (despite gaps in the mosaics from cells there were not recorded). The Large OFF cells in our analysis are a collection of cells from several clusters with somewhat more heterogeneous properties and likely belong to several cell types, which are nonetheless pooled together here, as the boundaries between the contributing clusters were less clear than for midget and parasol cells. We did not consider ON cells for further analyses as our recordings and spike sorting yielded too few reliable ON cells.

Both the white-noise stimulus and the naturalistic movie were presented to the retina in several trials consisting of non-repeating segments used for model training, alternating with a fixed stimulus segment, displayed repeatedly and used for model testing. For both models, we used the same spatial filter, which was obtained as a two-dimensional Gaussian fit to the spatial receptive field (see *Methods*). Model performances were evaluated for each cell on the average time-dependent firing rate across trials of the test segments by computing the Pearson correlation with the model prediction (Fig. 3A).

For spatiotemporal white noise, we found that the SC model yielded better model predictions across all cell classes (Fig. 3C). The improvement is most pronounced for the OFF parasol cells (average ratio of correlation values r_{SC}/r_{LN}: 1.16 ± 0.08) and Large OFF cells (r_{SC}/r_{LN}: 1.41 ± 0.32), whereas OFF midget cells displayed a more modest improvement (r_{SC}/r_{LN}: 1.09 ± 0.05). For naturalistic movies (Fig. 3E), the SC model outperformed the LN model for OFF parasol cells (r_{SC}/r_{LN}: 1.10 ± 0.07) and Large OFF cells (r_{SC}/r_{LN}: 1.32 ± 0.13). However, no change in the prediction performance was observed for OFF midget cells (r_{SC}/r_{LN}: 1.00 ± 0.02).

The finding that models of OFF midget cells profit least from including spatial contrast information is consistent with their weak responses to reversing gratings (Fig. 1C) and their smaller dependence on spatial contrast in firing rate histograms (Fig. 1D) as well as with the notion that midget cells are classically considered to have linear receptive field (Derrington et al., 1984). Note, though, that midget cells can also display substantial receptive-field nonlinearities, for example, as demonstrated in the peripheral macaque retina (Freedland and Rieke, 2022; Freeman et al., 2015).

The results suggest that cells with larger receptive fields have a greater improvement in model performance by including spatial contrast information. This is confirmed by plotting the ratio of model performances r_{SC}/r_{LN} against receptive field size (Fig. 3D,F,

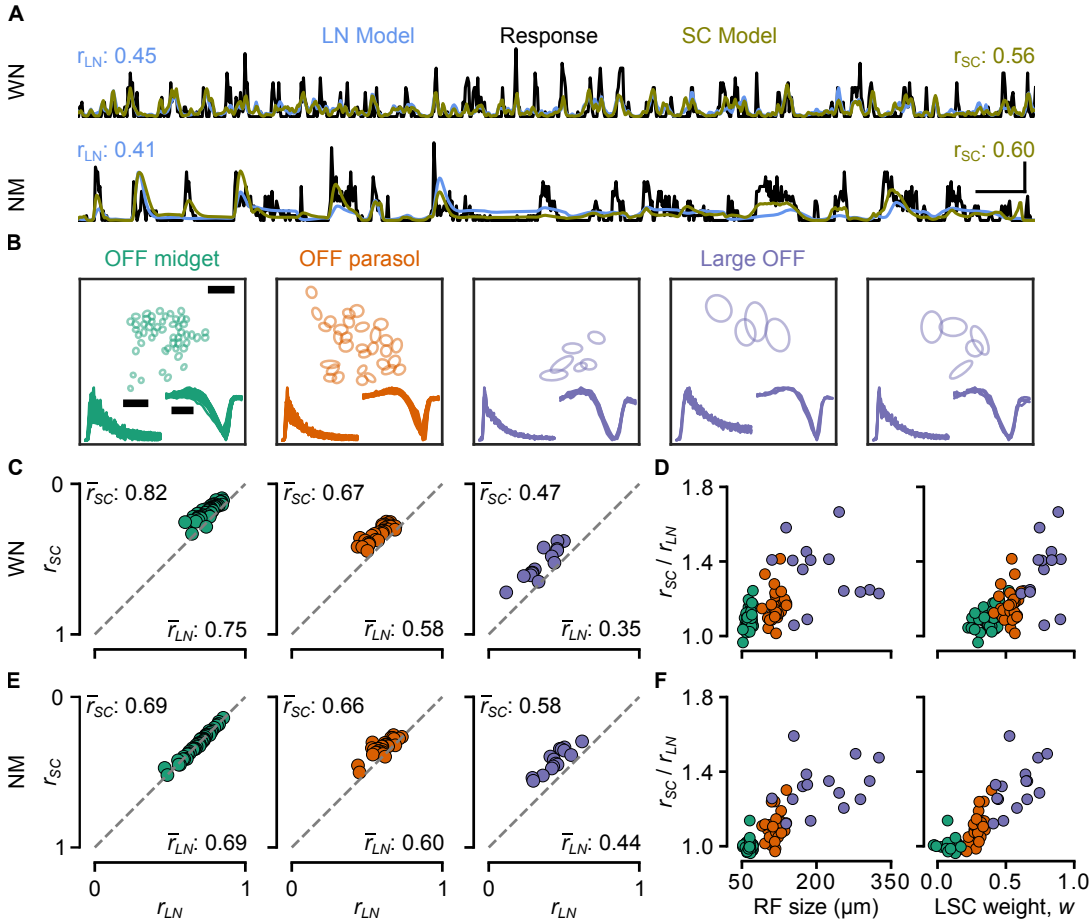

Figure 3. Spatial contrast model performance by cell type. A: Responses of a sample Large OFF cell to 10 s segments of test stimuli (gray) for white noise (top) and naturalistic movie (bottom), together with the predictions from the LN model (blue) and the SC model (olive). The numbers in the corners of each panel show correlation coefficients between the model prediction and the cell's response. Scale bars: x-axis: 0.5 s, y-axis: 50 Hz. **B**: Layouts of receptive fields (shown as 1.5-σ contours), spike-train autocorrelograms (lower left), and temporal filters (lower right) for all cells from one recording of the three analyzed cell classes. The Large OFF cells pool cells from three clusters, which are shown separately. Scale bars: spatial: 300 µm, temporal filters: 100 ms, autocorrelograms: 25 ms. **C**: Comparison of correlation coefficients for SC model (r_{SC}) and LN model (r_{LN}) for each cell under spatiotemporal white noise. The three panels correspond to the three cell classes in **B**. The average correlation coefficients $\bar{r}_{SC}$ and $\bar{r}_{LN}$ are shown next to the corresponding axes. **D**: Relative improvement in model prediction (r_{SC}/r_{LN}) under white noise for each cell versus receptive-field size (left) and the obtained weight w (right) for the LSC contribution. Colors according to cell type as in **B**. **E, F**: Same as **C** and **D**, but for the naturalistic movie.

left). Cell types with larger receptive fields display greater relative improvement in model performance. For midget cells under white-noise stimulation, this trend is also apparent for cells of the same type (Pearson's r: 0.61, p-val < 0.0001). That larger cells are more sensitive to spatial contrast makes intuitive sense, as their larger receptive fields are more likely to actually experience substantial spatial structure. For OFF midget cells under natural stimuli, for example, there may simply be too little spatial structure available at the high spatial frequencies that would fit into their small receptive fields. Thus, unlike for white-noise stimuli where signal power extends to high frequencies, midget cells might simply not experience sufficient spatial contrast to shape their responses beyond the activation of the linear receptive field.

We also checked how the improvement in model performance is related to the weight parameter w, which sets the relative contribution of the spatial-contrast information in determining the activation of the SC model. As expected, we found that larger weights are tightly associated with a larger relative improvement of the SC over the LN model (Fig. 3D,F, right).

To illustrate how the SC model can lead to improvements in model prediction over the LN model, we examined the responses and model firing rates of a sample OFF parasol cell for different episodes taken from the test data of the naturalistic movie (Fig. 4A). For each of the four representative cases (Fig. 4B), the LN model predicted a firing response lower than the ground truth, owing to the low average luminance in the receptive field. In fact, the temporally convolved stimulus, weighted by the cell's receptive field (Fig. 4B, 3rd column) displays bright as well as dark regions of similar extent, not unlike a grating stimulus. Thus, each of the four cases displayed substantial spatial structure inside the receptive field, leading to large values of the local spatial contrast. This allowed the SC model to predict higher firing rates in response to the displayed stimulus episodes, which better matched the response of the cell.

Identifying spatial scales of stimulus integration with the SC model

Visual input can contain information at a wide range of spatial frequencies, but not all of them are relevant for downstream sensory processing. Spatial contrast at very high spatial frequencies, corresponding to scales below the size of bipolar cell or even photoreceptor receptive fields, for example, is unlikely to matter even for a nonlinear ganglion cell. One may thus ask what the relevant spatial scales of integrating spatial contrast are for the analyzed marmoset retinal ganglion cells. This question can be targeted with the SC model as a tool by analyzing what the optimal spatial scale is to assess local spatial contrast. To do so, we used a modified version of the SC model, in which the LSC is computed from the temporally convolved stimulus after an additional spatial smoothing, here performed with a two-dimensional circular Gaussian filter (Fig. 5A). The smoothing eliminates spatial-contrast information on scales smaller than the spatial scale of the Gaussian filter but retains spatial-contrast information on larger scales. The computation of $\mathrm{I}_{\mathrm{mean}}$, on the other hand, remains unaffected.

For each cell, we thus trained the modified SC model for various scales of the Gaussian filter (defined as three times the filter's standard deviation, in line with our definition of the size of a cell's receptive field). The resulting performance in predicting responses to the test stimulus was then compared with that of the original SC model without stimulus smoothing. For spatiotemporal white noise, a clear pattern emerged. Across all cell

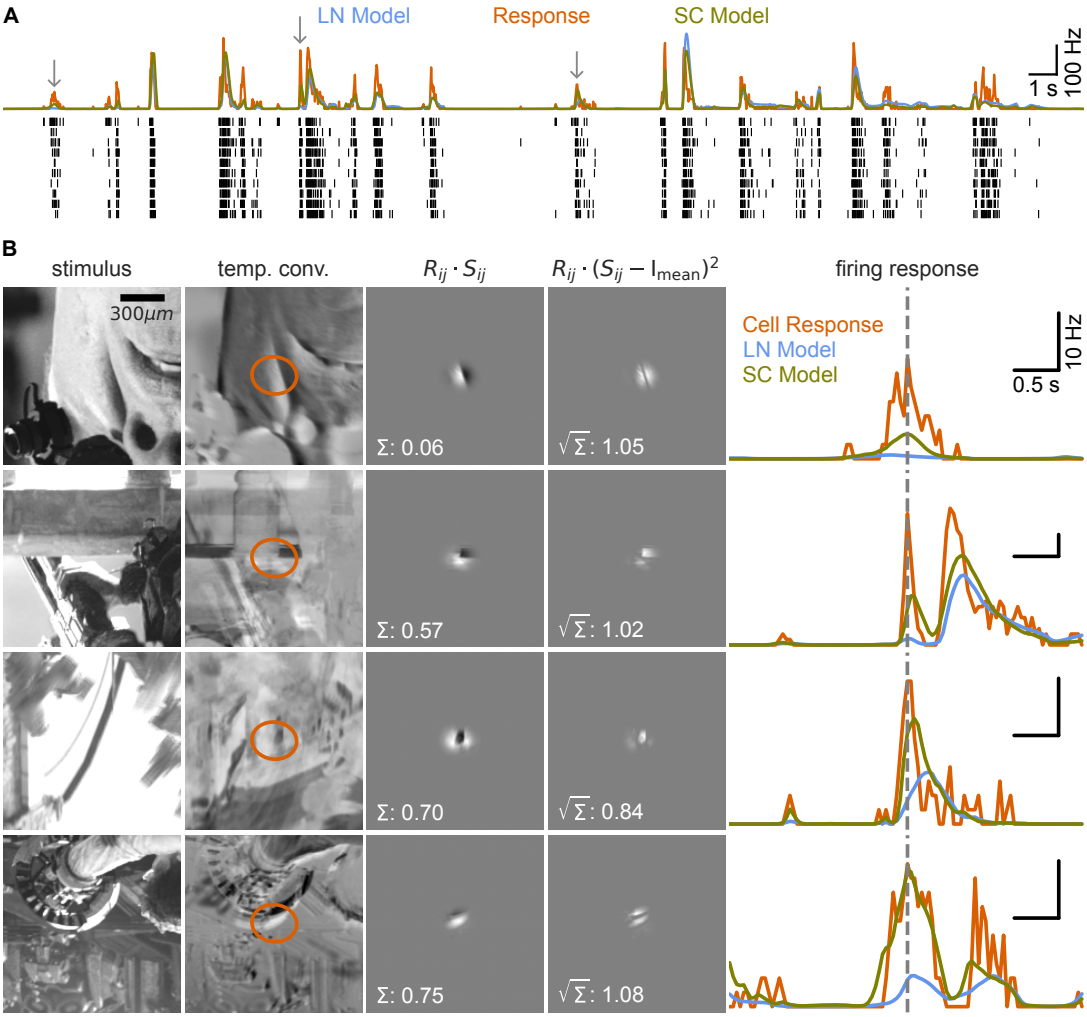

Figure 4. Examples of improved predictions by SC model versus LN model. A: Firing rate (top, orange) and spike raster (bottom) for the sample OFF parasol from Fig. 1 to a 20 s segment of the naturalistic movie test stimulus. Overlaid are response predictions for the LN model (blue) and the SC model (olive). Gray arrows indicate examples where the SC model markedly outperformed the LN model. **B**: Examples of episodes from the natural movie stimulus with substantially better response prediction by the SC model versus the LN model. First three rows correspond to the gray arrows in **A**. The first four columns are shown in the same way as in Fig. 1E. Right: 2 s snippets of responses and model predictions. Colors as in **A**. The dashed line corresponds to the time of the analyzed stimulus frame.

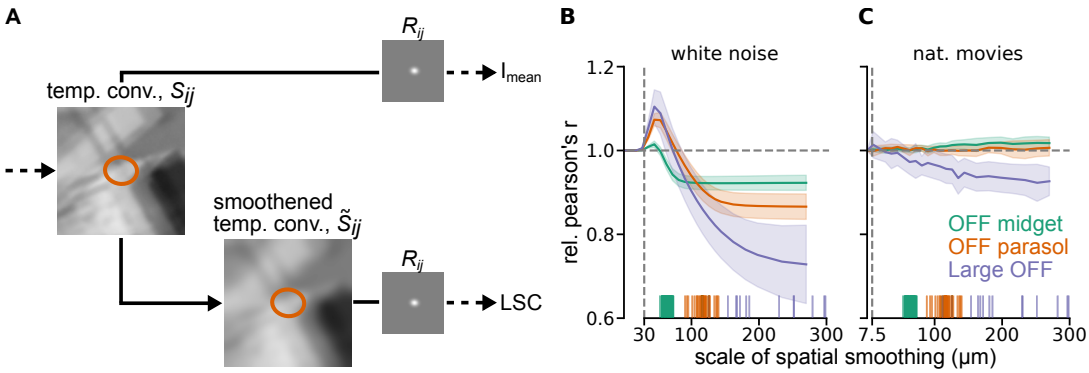

Figure 5. Analysis of spatial scale of contrast sensitivity. A: Modification of the schematic of the SC model from Fig. 2 to include a spatial stimulus smoothing step before the computation of LSC. **B**: Average SC model performance over cells of each class under white noise, depending on the scale of spatial smoothing, normalized for each cell to the model's performance without smoothing. Shaded bands represent three times the standard error. The dashed line marks the size of the pixels in the stimulus on the surface of the retina. Receptive-field sizes of individual cells are shown as a rug plot along the x-axis for reference. **C**: Same as in **B**, but for the naturalistic movie.

types, increasing the scale of spatial smoothing led, on average, to an initial increase in performance followed by a steady decrease (Fig. 5B). The relative performance peaked at a spatial scale of around 40–50 µm (mean ± std in units of µm: all cells: 47 ± 24; OFF midget cells: 46 ± 33, OFF parasol cells: 49 ± 5, Large OFF cells: 46 ± 5).

The magnitude of the model improvement obtained by assessing the local spatial contrast at its optimal scale can be read off from the peak size of the performance-improvement curves. The strength of this peak varied between cell types, with larger improvements for OFF parasol cells (relative Pearson's r: 1.08 ± 0.03, mean ± std) and Large OFF cells (relative Pearson's r: 1.11 ± 0.05) cells as compared to OFF midget cells (relative Pearson's r: 1.02 ± 0.02). This likely reflects the more nonlinear receptive fields of the parasol and Large OFF cells, whose models can therefore benefit more from properly assessing spatial contrast information.

For the naturalistic movie stimulus, the same analysis of different smoothing scales did not reveal optimal scales for any of the three cell classes (Fig. 5C). Increasing the scale of spatial smoothing did not systematically affect average model performance for OFF midget or OFF parasol cells and led to a continuous decrease in average model performance for Large OFF cells. Most likely, the lack of an optimal smoothing scale here comes from the fact that natural scenes – unlike white noise – are dominated by low spatial frequencies so that even without the stimulus smoothing, the LSC is largely determined by spatial structure at larger scales and that additional smoothing has no effect on the LSC or may even deteriorate the LSC signal. This indicates that, despite the many advantages of using naturalistic stimuli to analyze neuronal coding, specific aspects of stimulus processing may still be more easily investigated with artificial stimuli that emphasize the stimulus feature of interest, such as spatial contrast at high spatial frequency.

Discussion

Despite the ubiquity of spatial nonlinearities within the receptive fields of retinal ganglion cells (Zapp et al., 2022), current models of retinal stimulus encoding often start with a linear receptive field, implemented by a spatial filter applied to the incoming stimulus. The widespread use (and success) of these linear-receptive-field models is undoubtedly related to the relative ease with which their parameters, in particular the filters themselves, can be obtained from experimental data. Under Gaussian white noise stimulation, for example, the spike-triggered average can provide an excellent and easy-to-obtain estimate of the filter (Chichilnisky, 2001). And natural stimuli allow estimating the filter via likelihood estimation (Paninski, 2004; Pillow et al., 2008; Truccolo et al., 2005), supported by a good mathematical understanding of the parameter optimization challenge, or via information-theoretic tools (Sharpee et al., 2004).

This has led to several extensions of simple single-filter-based models, such as the LN models, by including additional model components that capture, for example, cell-intrinsic feedback underlying refractoriness, adaptation, and bursting (Keat et al., 2001; Pillow et al., 2005), influences by spiking activity in other neurons (Pillow et al., 2008), biophysical adaptation models (Ozuysal et al., 2018), and spike-timing dynamics (Aldworth et al., 2005; Gollisch, 2006). Including receptive-field nonlinearities, on the other hand, is typically done by splitting up the receptive field into spatial subunits which act as multiple, parallel filters and whose signals are nonlinearly combined, often by applying a threshold-linear rectification to the individual subunit signals before additive pooling. Yet, although such subunit models of retinal ganglion cells have been introduced almost half a century ago (Victor and Shapley, 1979), it has remained difficult to optimize subunit models with respect to data from actual, experimentally measured cells. Among the problems here are that the optimal number of subunits is not known beforehand, which makes parameterizing subunit models difficult, and that the small spatial scale of subunits requires finely structured visual stimuli to resolve their boundaries, which often results in inefficient stimulation and the need for long recording durations.

We therefore here turned to a simpler, phenomenological approach for including effects of nonlinear spatial integration into computational models, aiming to keep the advantage of conceptual model simplicity and relatively easy parameter optimization as in linear-receptive-field models. After demonstrating the spatial nonlinearity and sensitivity to spatial contrast (as defined by the luminance variance inside a cell's receptive field) in parasol cells of the primate retina (Fig. 1), we extended the previously developed spatial contrast (SC) model (Liu et al., 2022) for flashed images to spatiotemporal stimuli (Fig. 2). Evaluation on data that we recorded from the marmoset retina, we found that, compared to a linear-receptive-field LN model, the SC model yielded improved response predictions under artificial as well as naturalistic stimulation in a cell-type-specific fashion (Fig. 3), by predicting responses to stimulus episodes with high spatial contrast but little mean luminance signal (Fig. 4). Finally, we demonstrated that the SC model can be employed to identify the spatial scale of nonlinear stimulus integration by comparing model performance with different levels of spatial smoothing prior to the computation of the local spatial contrast (Fig. 5).

Advantages and limits of the SC model

An important feature of the SC model, in particular in comparison to subunit models, is that matching the model to a given recorded cell mostly requires an estimate of the receptive field and temporal filter, which can be obtained from standard stimuli and analyses, such as reverse correlation under spatiotemporal white-noise stimulation. Beyond this, the model has few free parameters, the weight factor of the local spatial contrast as well as parameters that describe the nonlinearity. As receptive field measurements are among the typical standard characterizations that form part of any retinal recording, obtaining the SC model essentially amounts to fitting a handful of parameters, which is feasible even with little training data.

Obtaining the receptive field and the temporal filter from white-noise stimulation, however, disregards that adaptation to different stimulus statistics, such as those of natural stimuli, may change a cell's filtering characteristics. An interesting future extension might thus be to directly fit the full model, including its spatial and temporal filter, under the stimulus statistics of interest and compare with the filtering characteristics under white noise. It may even be feasible to allow for different spatial filters for computing the mean luminance signal and the local spatial contrast, though from a biological perspective it seems plausible that evaluation of luminance and spatial contrast occurs with the same spatial weighting.

Another current simplification is the assumption that spatial and temporal filtering are separable. This does not only reduce the number of parameters required for describing the filters (and therefore makes the model more robust to noise in their estimation) but also conceptually simplifies the assessment of the local spatial contrast. Without assuming space-time separability, for example, it is unclear how the spatiotemporal weighting in the computation of the standard deviation should occur, as temporal variations in illumination most likely contribute differently (or not at all) to nonlinear signal processing as compared to spatial structure. Similarly, the current approach neglects the receptive-field surround, which may – besides standard (linear) surround suppression – also contribute nonlinear components (Takeshita and Gollisch, 2014) or influence the nonlinearities of the receptive field center (Turner et al., 2018). An interesting extension could thus be to apply the SC model in the context of a difference-of-Gaussians receptive field (Rodieck, 1965), with local spatial contrast independently evaluated in the center and in the surround Gaussian. This may, however, require a different estimation of the receptive field, as spatiotemporal white noise often does not reveal proper receptive-field surrounds (Wienbar and Schwartz, 2018), consistent with the present data (cf. Fig. 1A).

Alternative approaches

Despite the difficulties associated with subunit models discussed above, a number of approaches have successfully tackled the challenge of identifying subunits from recordings of retinal ganglion cells. These include direct fits of cascaded LN models (Karamanlis et al., 2023; Maheswaranathan et al., 2018; Real et al., 2017), statistical inference techniques (Liu et al., 2017; Shah et al., 2020), tomographic reconstruction (Krüppel et al., 2023), and making use of anatomical data (Schwartz et al., 2012) or prior determination of photoreceptor locations (Freeman et al., 2015). Given that subunits are thought to correspond to presynaptic bipolar cells (Borghuis et al., 2001; Demb et al., 2001; Liu et al., 2017), these efforts of subunit inference also aim at providing insight into the layout of

bipolar cell receptive fields and bipolar-to-ganglion cell connectivity. Yet, how much the precise knowledge of subunit number and position matters for response predictions of computational models of ganglion cells has been debated (Freedland and Rieke, 2022). Moreover, obtaining sufficient recording data for the challenging task of subunit identification may often be impractical.

Subunit models are essentially a form of multi-filter models, which apply several parallel filters before combining their signals nonlinearly. Thus, in principle, other approaches to obtaining multi-filter models could tackle the challenge of nonlinear spatial integration. Standard techniques like spike-triggered covariance analysis (Bialek and de Ruyter van Steveninck, 2005; Samengo and Gollisch, 2013; Schwartz et al., 2006) or most-informative dimensions (Sharpee et al., 2004), however, do not easily provide functional, interpretable models (Gollisch, 2013; Kaardal et al., 2013). An alternative is a likelihood-based model fit, such as in the Nonlinear Input Model (McFarland et al., 2013), which has had success in capturing excitatory and suppressive components as well as ON–OFF interactions of retinal ganglion cells (Shi et al., 2019), but which has not yet been applied to nonlinear spatial integration within the receptive field center.

Conceptually, the closest alternative approach would be to fit a generalized quadratic model (GQM; Park et al., 2013; Park and Pillow, 2011; Rajan et al., 2013) to the data. In the GQM, a neuron's activation is described by a quadratic form (general second-order polynomial function) of the stimulus components, such as here the pixel-wise intensity values over space and time. By a nonlinear transformation (typically an exponential function) and a stochastic process (typically Poisson), the activation is then turned into an output spike train, and feedback from generated spikes can be integrated as well via a filtering process that feeds into the model activation. It is the use of a quadratic form of the stimulus components, rather than a linear function as is the case in the generalized linear model (GLM; Paninski, 2004; Pillow et al., 2008; Truccolo et al., 2005) or the LN model, which yields sensitivity to receptive-field nonlinearities, similar to our use of the local spatial contrast. In fact, the GQM can be viewed as a more general form of the SC model. This generality, on the other hand, makes fitting the GQM much harder as compared to the more constrained SC model, owing to the many free parameters in the quadratic form, in particular in the spatiotemporal scenario considered here.

Spatial scale of nonlinear integration

Using the SC model with a spatial stimulus smoothing prior to computing the local spatial contrast allowed us, at least for the case of white-noise stimulation, to determine the optimal spatial scale for assessing receptive-field nonlinearities in the model. Interestingly, all three analyzed cell classes yielded a similar optimal scale of around 40–50 µm. This matches typical scales of receptive fields and dendritic trees of presynaptic bipolar cells in the peripheral monkey retina. In macaque, for example, receptive field sizes of midget and diffuse bipolar cells, the putative main inputs of midget and parasol ganglion cells, range from 30–50 µm and 70–110 µm, respectively (Dacey et al., 2000). Anatomically determined dendritic tree sizes for diffuse bipolar cells, which may be expected to match their receptive fields, have been reported around 30–50 µm (Boycott and Wässle, 1991). For the marmoset retina, data are more sparse, but anatomical measurements in the retinal periphery suggest bipolar cell dendritic tree sizes in the range of 20–80 µm (Chan et al., 2001). Despite partly comparing across different species (macaque versus marmoset)

and across likely different retinal eccentricities as well as different measures (anatomical versus electrophysiological), these numbers indicate that the 40–50 µm spatial scale of nonlinear stimulus integration lies in the expected range, given that the spatial contrast sensitivity is thought to result from nonlinear signal transmission between bipolar and ganglion cells.

Artificial versus naturalistic stimuli

It may seem surprising that the analysis under natural stimuli did not yield an optimal scale of spatial nonlinearities. Rather, stimulus smoothing before computing the local spatial contrast generally had no or a detrimental effect. The likely reason is that natural scenes, such as those in the applied naturalistic movie, are dominated by low spatial frequencies, owing to the approximate $1/f^2$-scaling of the power spectrum of light intensities in natural images with spatial frequency f (Ruderman and Bialek, 1994). Thus, smoothing at small spatial scales does little to the stimulus, and any residual improvements in the assessment of local spatial contrast may be counterbalanced by influences of the stimulus from outside the receptive field.

The relative lack of high-spatial-frequency signals in the naturalistic movie as compared to the white-noise stimulus is also apparent in the reduced importance of local spatial contrast in the SC model under the naturalistic movie. Performance gains relative to the LN model and values of the parameter for weighting the local spatial contrast contribution were generally smaller than under white noise. This may simply reflect that responses are, on average, determined more by the luminance signals inside the receptive field than by spatial contrast, so that the model naturally profits from putting more weight on the former when determining its activation.

On the other hand, it also seems feasible that the retinal neurons adapt to the different stimulus statistics under naturalistic versus white-noise stimulation (Sharpee et al., 2006) and that change in the models between those two scenarios reflects an actual change in nonlinear spatial integration, as can be observed for other alterations of stimulus context (Grimes et al., 2014; Yu et al., 2022). Future investigations may aim at resolving this by adding appropriate probe stimuli to the artificial and naturalistic stimulation or by frequently switching between the two, as done to analyze other complex adaptation phenomena (Hosoya et al., 2005). Moreover, comparing model performance and the importance of spatial contrast information under different types of naturalistic stimuli may help understand how specific the observed effects are.

Acknowledgment

We thank Fred Rieke for advice on experiments with primate retina and Fernando Rozenblit for help with the stimulus generation.

Author contributions

Conceptualization: S.S., A.E., T.G.; Methodology: S.S., M.K., D.K., M.M., T.G.; Software: S.S., M.V.; Formal analysis: S.S.; Investigation: S.S., M.V., M.K., D.K., H.M.S., V.R., S.K., S.J.Z.; Resources: A.E., T.G.; Writing - original draft: S.S., T.G.; Writing - review & editing: S.S., M.V., M.K., D.K., H.M.S., V.R., S.K., S.J.Z., M.M., A.E., T.G.; Visualization: S.S.; Supervision: T.G.; Funding acquisition: A.E., T.G.

References

Aldworth, Z. N., Miller, J. P., Gedeon, T., Cummins, G. I., & Dimitrov, A. G. (2005). Dejittered spike-conditioned stimulus waveforms yield improved estimates of neuronal feature selectivity and spike-timing precision of sensory interneurons. *J Neurosci*, *25*, 5323–5332.

Angueyra, J. M., Baudin, J., Schwartz, G. W., & Rieke, F. (2022). Predicting and Manipulating Cone Responses to Naturalistic Inputs. *J Neurosci*, *42*, 1254–1274.

Bialek, W., & de Ruyter van Steveninck, R. R. (2005). Features and dimensions: Motion estimation in fly vision. *arXiv*, q–bio/0505003.

Borghuis, B. G., Marvin, J. S., Looger, L. L., & Demb, J. B. (2001). Two-photon imaging of nonlinear glutamate release dynamics at bipolar cell synapses in the mouse retina. *J Neurosci*, *33*, 10972–10985.

Boycott, B. B., & Wässle, H. (1991). Morphological classification of bipolar cells of the primate retina. *Eur J Neurosci*, *3*, 1069–1088.

Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., & Ecker, A. S. (2019). Deep convolutional models improve predictions of macaque V1 responses to natural images. *PLoS Comp Biol*, *15*, e1006897.

Chan, T. L., Martin, P. R., Clunas, N., & Grünert, U. (2001). Bipolar cell diversity in the primate retina: Morphologic and immunocytochemical analysis of a new world monkey, the marmoset *Callithrix jacchus*. *J Comp Neurol*, *437*, 219–239.

Chichilnisky, E. J. (2001). A simple white noise analysis of neuronal light responses. *Netw Comput Neural Syst*, *12*, 199–213.

Dacey, D. M., Packer, O. S., Diller, D., L.and Brainard, Peterson, B., & Lee, B. (2000). Center surround receptive field structure of cone bipolar cells in primate retina. *Vis Res*, *40*, 1801–1811.

Demb, J. B., Zaghloul, K., Haarsma, L., & Sterling, P. (2001). Bipolar cells contribute to nonlinear spatial summation in the brisk-transient (Y) ganglion cell in mammalian retina. *J Neurosci*, *21*, 7447–7454.

Derrington, A. M., Krauskopf, J., & Lennie, P. (1984). Chromatic mechanisms in lateral geniculate nucleus of macaque. *J Physiol*, *357*, 241–265.

Enroth-Cugell, C., & Robson, J. G. (1966). The contrast sensitivity of retinal ganglion cells of the cat. *J Physiol*, *187*, 517–552.

Freedland, J., & Rieke, F. (2022). Systematic reduction of the dimensionality of natural scenes allows accurate predictions of retinal ganglion cell spike outputs. *Proc Natl Acad Sci U S A*, *119*, e2121744119.

Freeman, J., Field, G. D., Li, P. H., Greschner, M., Gunning, D. E., Mathieson, K., Sher, A., Litke, A. M., Paninski, L., Simoncelli, E. P., & Chichilnisky, E. J. (2015). Mapping nonlinear receptive field structure in primate retina at single cone resolution. *eLife*, *4*, e05241.

Gollisch, T. (2006). Estimating receptive fields in the presence of spike-time jitter. *Netw Comput Neural Syst*, *17*, 103–129.

Gollisch, T. (2013). Features and functions of nonlinear spatial integration by retinal ganglion cells. *J Physiol Paris*, *107*, 338–348.

Grimes, W. N., Schwartz, G. W., & Rieke, F. (2014). The synaptic and circuit mechanisms underlying a change in spatial encoding in the retina. *Neuron*, *82*, 460–473.

Harris, C. M., Hainline, L., Abramov, I., Lemerise, E., & Camenzuli, C. (1988). The distribution of fixation durations in infants and naive adults. *Vis Res*, *28*, 419–432.

Hartline, H. K. (1938). The response of single optic nerve fibers of the vertebrate eye to illumination of the retina. *Am J Physiol*, *121*, 400–415.

Heitman, A., Brackbill, N., Greschner, M., Sher, A., Litke, A. M., & Chichilnisky, E. J. (2016). Testing pseudo-linear models of responses to natural scenes in primate retina. *bioRxiv*, 045336v2.

Hosoya, T., Baccus, S. A., & Meister, M. (2005). Dynamic predictive coding by the retina. *Nature*, *436*, 71–77.

Kaardal, J., Fitzgerald, J. D., Berry II, M. J., & Sharpee, T. O. (2013). Identifying functional bases for multidimensional neural computations. *Neural Comput*, *25*, 1870–1890.

Karamanlis, D., Khani, M. H., Schreyer, H. M., Zapp, S. J., Mietsch, M., & Gollisch, T. (2023). Natural stimuli drive concerted nonlinear responses in populations of retinal ganglion cells. *bioRxiv*, 2023.01.10.523412.

Karamanlis, D., & Gollisch, T. (2021). Nonlinear spatial integration underlies the diversity of retinal ganglion cell responses to natural images. *J Neurosci*, *41*, 3479–3498.

Karamanlis, D., Schreyer, H. M., & Gollisch, T. (2022). Retinal encoding of natural scenes. *Annu Rev Vis Sci*, *8*, 171–193.

Keat, J., Reinagel, P., Reid, R. C., & Meister, M. (2001). Predicting every spike: A model for the responses of visual neurons. *Neuron*, *30*, 803–817.

Krüppel, S., Khani, M. H., Schreyer, H. M., Sridhar, S., Ramakrishna, V., Zapp, S. J., Mietsch, M., Karamanlis, D., & Gollisch, T. (2023). Applying super-resolution and tomography concepts to identify receptive field subunits in the retina. *bioRxiv*, 2023.11.27.568854.

Liu, J. K., Karamanlis, D., & Gollisch, T. (2022). Simple model for encoding natural images by retinal ganglion cells with nonlinear spatial integration. *PLoS Comput Biol*, *18*, e1009925.

Liu, J. K., Schreyer, H. M., Onken, A., Rozenblit, F., Khani, M. H., Krishnamoorthy, V., Panzeri, S., & Gollisch, T. (2017). Inference of neuronal functional circuitry with spike-triggered non-negative matrix factorization. *Nat Commun*, *8*, 149.

Maheswaranathan, N., Kastner, D. B., Baccus, S. A., & Ganguli, S. (2018). Inferring hidden structure in multilayered neural circuits. *PLoS Comput. Biol.*, *14*, e1006291.

McFarland, J. M., Cui, Y., & Butts, D. A. (2013). Inferring nonlinear neuronal computation based on physiologically plausible inputs. *PLoS Comput Biol*, *9*, e1003143.

Mitchell, J. F., Reynolds, J. H., & Miller, C. T. (2014). Active vision in marmosets: A model system for visual neuroscience. *J Neurosci*, *34*, 1183–1194.

Ozuysal, Y., Kastner, D. B., & Baccus, S. A. (2018). Adaptive feature detection from differential processing in parallel retinal pathways. *PLoS Comput Biol*, *14*, e1006560.

Pachitariu, M., Sridhar, S., & Stringer, C. (2023). Solving the spike sorting problem with Kilosort. *bioRxiv*, 2023.01.07.523036.

Paninski, L. (2004). Maximum likelihood estimation of cascade point-process neural encoding models. *Netw Comput Neural Syst*, *15*, 243–262.

Park, I. M., Archer, E., Priebe, N., & Pillow, J. W. (2013). Spectral methods for neural characterization using generalized quadratic models. *Adv Neural Inf Process Syst*, *26*, 2454–2462.

Park, I. M., & Pillow, J. W. (2011). Bayesian spike-triggered covariance. *Adv Neural Inf Process Syst*, *24*, 1692–1700.

Pillow, J. W., Paninski, L., Uzzell, V. J., Simoncelli, E. P., & Chichilnisky, E. J. (2005). Prediction and decoding of retinal ganglion cell responses with a probabilistic spiking model. *J Neurosci*, *25*, 11003–11013.

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*, 995–999.

Rajan, K., Marre, O., & Tkacik, G. (2013). Learning quadratic receptive fields from neural responses to natural stimuli. *Neural Comput*, *25*, 1661–1692.

Real, E., Asari, H., Gollisch, T., & Meister, M. (2017). Neural circuit inference from function to structure. *Curr Biol*, *27*, 189–198.

Rodieck, R. W. (1965). Quantitative analysis of cat retinal ganglion cell response to visual stimuli. *Vis Res*, *5*, 583–601.

Rucci, M., Edelman, G. M., & Wray, J. (2000). Modeling LGN responses during free-viewing: A possible role of microscopic eye movements in the refinement of cortical orientation selectivity. *J Neurosci*, *20*, 4708–4720.

Ruderman, D. L., & Bialek, W. (1994). Statistics of natural images: Scaling in the woods. *Phys Rev Lett*, *73*, 814–817.

Samengo, I., & Gollisch, T. (2013). Spike-triggered covariance: Geometric proof, symmetry properties, and extension beyond gaussian stimuli. *J Comput Neurosci*, *34*, 137–161.

Schwartz, G. W., & Rieke, F. (2011). Nonlinear spatial encoding by retinal ganglion cells: When $1 + 1 \neq 2$. *J Gen Physiol*, *138*, 283–290.

Schwartz, G. W., Okawa, H., Dunn, F. A., Morgan, J. L., Kerschensteiner, D., Wong, R. O., & Rieke, F. (2012). The spatial structure of a nonlinear receptive field. *Nat Neurosci*, *15*, 1572–1580.

Schwartz, O., Pillow, J. W., Rust, N. C., & Simoncelli, E. P. (2006). Spike-triggered neural characterization. *J Vis*, *6*, 484–507.

Shah, N. P., Brackbill, N., Rhoades, C., Kling, A., Goetz, G., Litke, A. M., Sher, A., Simoncelli, E., & Chichilnisky, E. J. (2020). Inference of nonlinear receptive field subunits with spike-triggered clustering. *eLife*, *9*, e45743.

Sharpee, T. O., Sugihara, H., Kurgansky, A. V., Rebrik, S. P., Stryker, M. P., & Miller, K. D. (2006). Adaptive filtering enhances information transmission in visual cortex. *Nature*, *439*, 936–942.

Sharpee, T. O., Rust, N. C., & Bialek, W. (2004). Analyzing neural responses to natural signals: Maximally informative dimensions. *Neural Comput*, *16*, 223–250.

Shi, Q., Gupta, P., Boukhvalova, A. K., Singer, J. H., & Butts, D. A. (2019). Functional characterization of retinal ganglion cells using tailored nonlinear modeling. *Sci Rep*, *9*, 8713.

Takeshita, D., & Gollisch, T. (2014). Nonlinear spatial integration in the receptive field surround of retinal ganglion cells. *J Neurosci*, *34*, 7548–7561.

Truccolo, W., Eden, U. T., Fellows, M. R., Donoghue, J. P., & Brown, E. N. (2005). A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *J Neurophysiol*, *93*, 1074–1089.

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.

Victor, J. D., & Shapley, R. M. (1979). The nonlinear pathway of Y ganglion cells in the cat retina. *J Gen Physiol*, *74*, 671–689.

Wienbar, S., & Schwartz, G. W. (2018). The dynamic receptive fields of retinal ganglion cells. *Prog Retin Eye Res*, *57*, 102–117.

Yu, Z., Turner, M. H., Baudin, J., & Rieke, F. (2022). Adaptation in cone photoreceptors contributes to an unexpected insensitivity of primate on parasol retinal ganglion cells to spatial structure in natural images. *eLife*, *11*, e70611.

Zapp, S. J., Nitsche, S., & Gollisch, T. (2022). Retinal receptive-field substructure: Scaffolding for coding and computation. *Trends Neurosci*, *45*, 430–445.