Efficient coding, channel capacity and the emergence of retinal mosaics

Anonymous Author(s)
Affiliation
Address
email

Abstract

Among the most striking features of retinal organization is the grouping of its output neurons, the retinal ganglion cells, into a diversity of functional types. Each of these types exhibits a mosaic-like organization of receptive fields that tiles the retina and visual space. Previous work has shown that many features of retinal ganglion cell organization, including the existence of ON and OFF cell types, the structure of spatial receptive fields, and their relative arrangement, can be predicted on the basis of efficient coding theory. This theory posits that the nervous system is organized to maximize information in its encoding of stimuli while minimizing metabolic costs. Here, we use efficient coding theory to present a comprehensive account of mosaic organization in the case of natural videos as the retinal channel capacity—the number of neurons available for encoding—is varied. Using a simple model of efficient coding, we show that, beginning with spatially localized temporal smoothing filters, mosaic density increases with channel capacity up to a series of critical points at which new mosaics focused on increasingly high temporal frequencies emerge. In addition, we show both experimentally and theoretically that a transition from mosaic alignment to anti-alignment is observed not only with increasing output noise as previously reported, but also with decreasing input noise. Together, these results offer a unified perspective on the relationship between retinal mosaics, efficient coding, and channel capacity that may help to explain the stunning functional diversity of retinal mosaics.

1 Introduction

The retina is one of the most intensely studied of all neural circuits, yet we still lack a computational understanding of its organization in relation to its function. At a structural level, the retina forms a three-layer circuit, with its primary feedforward pathway consisting of light-sensitive photoreceptors that project to bipolar cells that project in turn to retinal ganglion cells (RGCs), the axons of which form the optic nerve [1]. At a functional level, retinal ganglion cells are observed to be divided into 30-50 functional types, depending on species, with each cell responsive to a localized area of visual space (its receptive field (RF)) and the collection of receptive fields of each type tiling visual space to form distinct functional "mosaics" [2, 3]. At a computational level, each mosaic thus represents the extraction of a specific type of information from the visual scene, with different mosaics responding to light increments or decrements (ON and OFF cells), high or low spatial and temporal frequencies, color, motion, and a host of other variables. Yet, while much is known about the response properties of each type, the computational principles that drive RGC diversity remain unclear.

For more than sixty years, efficient coding theory has proven one of the most powerful ideas for understanding retinal organization. As originally proposed by Barlow [4], efficient coding posits that the nervous system attempts to encode the sensory world by minimizing redundancy subject to biological costs and constraints. As more commonly formulated, it seeks to maximize the mutual

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

information between sensory data and neural representations, with the most common cost in the retinal case being the energetic cost of action potentials transmitted by the ganglion cells. Despite its simplicity, this principle, has proven highly effective, having predicted the center-surround structure of receptive fields [5], the frequency response profile of contrast sensitivity [6], the structure of retinal mosaics [7, 8], the role of nonlinear rectification [9], different spatiotemporal kernels [10], and inter-mosaic arrangements [11, 12].

While previous studies have largely focused on either the spatial or the temporal aspects of efficient coding, we optimize for the first time a model of retinal efficient coding of natural spatiotemporal movies. This optimization revealed factors that determine the number of functional types. Specifically, we found that increasing the number of modeled ganglion cells in the optimization (analogous to the channel capacity in information theory) increased the number of distinct functional types that result from the optimization. Each of these functionally distinct types forms its own mosaic of receptive fields that approximately tiles space. In the case of linear encoding, we show analytically how a tradeoff between information gains from increasing mosaic density and information gains from adding new, specialized cell types leads to the addition of new mosaics capturing higher temporal frequency information. We then show that the same phenomenon occurs in the case of a nonlinear model trained to maximize mutual information subject to firing rate costs. Finally, we investigate the relative arrangement of these mosaics, showing that while they exhibit the same phase transition identified for purely spatial encoding [11], different ON-OFF mosaic pairs exhibit different transition profiles as a function of input noise.

Related work: As indicated above, this study builds on a long line of work using efficient coding principles to understand retinal processing. In addition, it is related to work examining encoding of natural videos [13, 14, 15] and prediction in the spacetime context. However, the most closely related work to this one is that of [10], which also considered efficient coding of natural videos and considered the tradeoffs involved in multiple mosaics. Our treatment here differs from that work in several key aspects: First, the linear analytical model of [10] focused on least squares reconstruction error in a convolutional neural framework with grids of regularly spaced receptive fields. In other words, that study predetermined the number of cell types and optimized their relative spacing. Here, we follow previous efficient coding models [5, 6, 7, 8] in maximizing mutual information and do not assume an a prior mosaic arrangement, a particular cell spacing, or a particular number of cell types. Second, while the computational model of [10] optimized strides for a pair of rectangular arrays of ganglion cells, we individually optimize receptive field locations and shapes, allowing us to study changes in optimal receptive field size and density as new, partial mosaics (up to five pairs) begin to form. Third, while [10] used zero-padding of natural videos to bias learned temporal filters toward those of observed ganglion cells, we link the form of our temporal receptive fields to biophysical limits on the filtering properties of bipolar cells, producing temporal filters with the delay properties observed in experimental data. Finally, while [10] only considered a single noise source in their model, we consider noise in both photoreceptor responses (input noise) and ganglion cell responses (output noise), which allows us to investigate the mosaic alignment transitions that are unique to this work.

2 Model

The model we develop is an extension of [11], a retinal model for efficient coding of natural images, which is based on a mutual information maximization objective proposed in [7]. The retinal model takes D-pixel patches of natural images $\mathbf{x} \in \mathbb{R}^D$ corrupted by input noise $\mathbf{n}_x \sim \mathcal{N}(0, \mathbf{C}_{n_{\text{in}}})$, which are filtered with unit-norm linear kernels $\{\mathbf{w}_j \in \mathbb{R}^D \mid \|\mathbf{w}_j\| = 1\}_{j=1,\cdots,J}$ representing J retinal ganglion cells. The resulting signals $y_j = \mathbf{w}_j^\top(\mathbf{x} + \mathbf{n}_{\text{in}})$ are then fed through softplus nonlinearities $\eta(y) = \log\left(1 + e^{\beta y}\right)/\beta$ with gain γ_j and threshold θ_j and finally further corrupted by additive output noise $\mathbf{n}_{\text{out}} \sim \mathcal{N}(0, \mathbf{C}_{n_{\text{out}}})$, to produce firing rates r_j:

$$r_j \quad = \quad \gamma_j \cdot \eta(y_j - \theta_j) + n_{\text{out},j}. \tag{1}$$

The parameters $\mathbf{w}_j$, γ_j, and θ_j are learned to maximize the mutual information between the inputs $\mathbf{x}$ and the outputs $\mathbf{r}$, under a mean firing rate constraint:

$$\text{maximize} \quad \log \frac{\det\left(\mathbf{G}\mathbf{W}^\top\left(\mathbf{C}_\mathbf{x}+\mathbf{C}_{n_\text{in}}\right)\mathbf{W}\mathbf{G}+\mathbf{C}_{n_\text{out}}\right)}{\det\left(\mathbf{G}\mathbf{W}^\top\mathbf{C}_{n_\text{in}}\mathbf{W}\mathbf{G}+\mathbf{C}_{n_\text{out}}\right)} \tag{2}$$

$$\text{subject to} \quad \mathbb{E}[r_j]=1. \tag{3}$$

Here $\mathbf{C}_\mathbf{x}$ is the covariance matrix of the input distribution, $\mathbf{W}\in\mathbb{R}^{D\times J}$ contains the filters $\mathbf{w}_j$ as its columns, the gain matrix $\mathbf{G}=\operatorname{diag}(\gamma_j\mathbf{1}_{r_j>0})$, and the noise covariances are $\mathbf{C}_{n_\text{in}}=\sigma_\text{in}^2\mathbb{1}_{D\times D}$ and $\mathbf{C}_{n_\text{out}}=\sigma_\text{out}^2\mathbb{1}_{J\times J}$. This objective is equivalent to the formulation in [7], which assumes normally distributed inputs and locally linear responses in order to approximate the mutual information in a closed form.

Here, we extend this model to time-varying inputs $\mathbf{x}(t)\in\mathbb{R}^D$ representing natural videos (Figure 1A), which are convolved with linear spatiotemporal kernels $\{\mathbf{w}_j(t)\}_{j=1,\cdots,J}$:

$$y_j(t) \;=\; \mathbf{w}_j(t)*\mathbf{x}(t) \;=\; \int_{-\infty}^{\infty}\mathbf{w}_j(\tau)^\top\mathbf{x}(t-\tau)d\tau. \tag{4}$$

We additionally assume that the convolution kernels are separable in time and space:

$$\mathbf{w}_j(t) \;=\; \phi_j(t)\mathbf{w}_j, \qquad \phi_j(t)\in\mathbb{R}, \quad \|\mathbf{w}_j\|=1, \quad \int_{-\infty}^{\infty}\phi(t)^2dt=1, \tag{5}$$

and the temporal kernels are unit-norm impulse responses taking the following parametric form:

$$\phi_j(t) \;\propto\; \begin{cases} \alpha_j t^n e^{-t/\tau_j}-\alpha_j' t^n e^{-t/\tau_j'} & \text{if delay } t\geq 0 \\ 0 & \text{otherwise} \end{cases}, \tag{6}$$

where α_j, α_j', $\tau_j>0$, $\tau_j'>0$ are learnable parameters and $n\in\mathbb{N}$ is fixed. Previous work assumed an unconstrained form for these filters, adding zero-padding before and after the model's image inputs to produce the characteristic shape of the temporal filters in primate midget and parasol cells [10], but this zero-padding represents a biologically implausible constraint, and the results fail to correctly reproduce the observed delay in retinal responses [16, 3, 17]. Rather, optimizing (2) with unconstrained temporal filters produces a filter bank uniformly tiling time (Supplementary Figure 1).

By contrast, (6) is motivated by the arguments of [18], which showed that the optimal minimum-phase temporal filters of retinal bipolar cells, the inputs to the ganglion cells, take the form

$$\phi(t>0) \;\propto\; e^{-t/\tau}[\sin\omega t-\omega t\cos\omega t] \;\approx\; e^{-t/\tau}\frac{(\omega t)^3}{3} \tag{7}$$

when $\omega\tau\ll 1$. Thus, we model ganglion cell temporal filters as a linear combination of these forms. In practice, we take only two filters and use $n=6$ rather than $n=3$, since these have been shown to perform well in capturing observed retinal responses [17]. The results produced by more filters or different exponents are qualitatively unchanged (Supplementary Figure 2). Likewise, for training on video data, we use discrete temporal filters and convolutions with $\sum_{t=0}^{T-1}\phi_j[t]^2=1$. Finally, while unconstrained spatial kernels $\mathbf{w}_j$ converge to characteristic center-surround shapes under optimization of (2) (Figure 1B), for computational efficiency and stability, we parameterized these filters using a radially-symmetric difference of Gaussians

$$w_j(r)\propto e^{-a_j r^2}-c_j e^{-b_j r^2}, \qquad b_j>a_j>0, \quad 0<c_j<1, \tag{8}$$

where r measures the spatial distance to the center of the receptive field. The parameters a_j, b_j, c_j determining the spatial kernel shape, as well as the center location, are different for each neuron j, allowing individual neurons to have different sizes and center-surround strengths. The result of optimizing (2) using these forms is a set of spatial and temporal kernels (Figure 1C-D) that replicate experimentally-observed shapes and spatial receptive fields tiling.

3 Mosaic formation as a function of channel capacity

3.1 Linear mosaics in the continuum limit

We begin by analyzing a linear version of the model (2) in the continuum limit, developing intuitions for the effects of the number of neurons on mosaic formation. We assume a single gain γ for all cells,

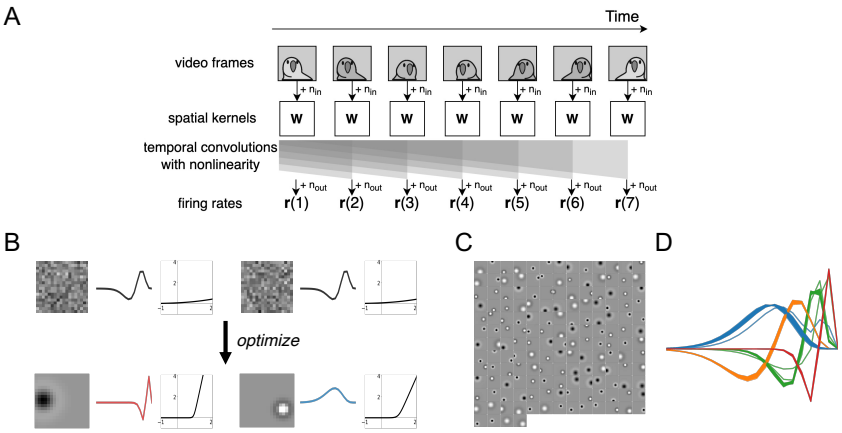

Figure 1: ON and OFF RF mosaics and their temporal RFs are predicted by efficient coding of natural videos. **(A)** Model architecture: frames of natural videos $\mathbf{x}(t)$, plus the input noise $\mathbf{n}_{\text{in}}$, are linearly filtered with the spatial kernels $\mathbf{w}_j$ and then passed through one-dimensional temporal convolutions followed by a nonlinearity, resulting firing rates $r_j(t)$ for each of J neurons. **(B)** Examples of initial and optimized spatial filters, temporal filters, and nonlinearities: **(left)** *fast OFF* kernel, **(right)** *slow ON* kernel. **(C)** Unconstrained spatial filters ($J = 160$) learned center-surround shapes, about half of which are ON RFs. **(D)** Learned temporal filters ($J = 160$) using the parameterization (6) converged to four distinct clusters.

no bias ($\theta = 0$), and a linear transfer function $\eta(y) = y$. Again, we follow [5, 7, 8, 11] in optimizing mutual information, though we obtain results in qualitative agreement with [10], which optimized reconstruction error. We assume an infinite retina on which receptive fields form mosaics described by a Bravais lattice R with principal vectors $(\mathbf{a}_1, \mathbf{a}_2)$ and reciprocal vectors $(\mathbf{b}_1, \mathbf{b}_2)$ in a dual lattice G with $\mathbf{a}_i \cdot \mathbf{b}_j = 2\pi\delta_{ij}$. Put another way, if the original mosaic in image space is periodic under shifts by elements of R, its Fourier representation is redundant under shifts in G. Under these conditions, we can then write the log determinants in (2) as integrals and optimize over the *unnormalized* filter $v \equiv \gamma w$ subject to a power constraint (Appendix A):

$$\max_v \int_{G_0} \frac{d^2\mathbf{k}}{(2\pi)^2} \left[\log \frac{\sum_{\mathbf{g}\in G} |v(\mathbf{k}+\mathbf{g})|^2 (C_x(\mathbf{k}+\mathbf{g}) + \sigma_{\text{in}}^2) + \sigma_{\text{out}}^2}{\sum_{\mathbf{g}\in G} |v(\mathbf{k}+\mathbf{g})|^2 \sigma_{\text{in}}^2 + \sigma_{\text{out}}^2} - \nu \sum_{\mathbf{g}\in G} |v(\mathbf{k}+\mathbf{g})|^2 (C_x(\mathbf{k}+\mathbf{g}) + \sigma_{\text{in}}^2) \right], \tag{9}$$

where $C_x(\mathbf{k})$ is the Fourier transform of the stationary image covariance $C_x(\mathbf{z}-\mathbf{z}')$. Here, the integral over $\mathbf{k}$ is restricted to G_0, the unit cell of the reciprocal lattice, and the sums are over reciprocal lattice cells, the result of aliasing due to the repeated spatial structure in the mosaic. In [5], the range $[-\pi, \pi]$ is used, corresponding to a one-dimensional lattice and units of mosaic spacing $\Delta z = 1$.

Now, performing the optimization in (9) results in a spatial kernel with the spectral form (full details in Appendix A)

$$|v(k)|^2 = \frac{\sigma_{\text{out}}^2}{\sigma_{\text{in}}^2} \left[\frac{1}{2} \frac{C_x(k)}{C_x(k) + \sigma_{\text{in}}^2} \left(\sqrt{1 + \frac{4\sigma_{\text{in}}^2}{\nu\, C_x(k)}} + 1 \right) - 1 \right]_+, \quad k \in G_0, \tag{10}$$

where $k = \|\mathbf{k}\|$ and ν is chosen to enforce the constraint on total power. This is exactly the solution found in [5], linking it (in the linear case) to the model of [7, 8]. Note, however, that (10) is only nonzero within the unit cell of the reciprocal lattice, indicating that receptive field spacing sets an upper limit on the passband of the resulting filters [10]. As we argue below, it is this restriction that critically determines the encoding tradeoffs that result in new mosaic formation as new receptive fields are added.

The generalization of this formulation to the spacetime case is straightforward. Given a spacetime stationary image spectrum $C_x(\mathbf{z}-\mathbf{z}', t-t')$ and radially-symmetric, causal filter $w(\mathbf{z}, t)$ in the continuum limit, calculating the information present in firing rates $r_j(t)$ across both neurons and time

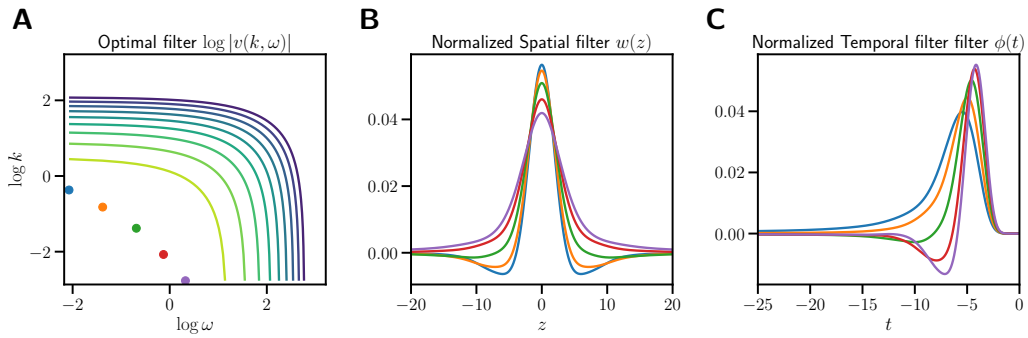

Figure 2: Optimal filters in the linear spacetime case. **(A)** Spectrum of the optimal linear spacetime filter in $d = 1$ spatial dimension. Contour lines indicate constant (log) power. **(B)** Spatial filters at representative temporal frequencies. Each filter represents a vertical section at the correspondingly colored dot in **A**. **(C)** Temporal filters at representative spatial frequencies. Each filter represents a horizontal section at the correspondingly colored dot in **A**.

requires computing log determinants of matrices with entries of the form

$$\begin{aligned} F_{ijtt'} &= \int d\mathbf{z}d\mathbf{z}'d\tau d\tau'\gamma^2 w(\mathbf{z}_i - \mathbf{z}, t-\tau)C_x(\mathbf{z}-\mathbf{z}', \tau-\tau')w(\mathbf{z}_j - \mathbf{z}', t'-\tau') \\ &= \int \frac{d^2\mathbf{k}}{(2\pi)^2}\frac{d\omega}{2\pi}e^{i\mathbf{k}\cdot(\mathbf{z}_i-\mathbf{z}_j)+i\omega(t-t')}|v(k,\omega)|^2 C_x(k,\omega). \end{aligned} \tag{11}$$

Again, such matrices can be diagonalized in the Fourier basis, with the result that we again obtain (9) and (10) under the substitutions $v(k) \to v(k,\omega)$, $C_x(k) \to C_x(k,\omega)$. Figure 2 depicts the corresponding spatial and minimum-phase temporal filters for this solution in $d = 1$ spatial dimension. As we will argue below, division of this single filter into multiple specialized cell types as mosaic density increases results in the emergence of multiple cell types roughly corresponding to the dots in Figure 2A. A detailed derivation of these results is presented in Appendix A.

3.2 Multiple mosaics

Of course, the formulation above assumes only a single filter type, spatially translated by the lattice vectors that define the mosaic. More generally, for multiple filter types, calculating the determinants in (2) involves not only terms of the form (11) but also cross-terms involving $v(k,\omega)$ from different mosaics. Here, we argue that such cross-terms vanish at the optimum, yielding an information objective that is equivalent to a sum of individual mosaic terms like (9). This is because (Appendix A), our power constraint in (9) remains unchanged under an orthogonal transformation of the filters, while information is subadditive when cross-terms between filters are nonzero. Thus, we can always increase information by choosing a filter basis orthogonal with respect to both $\mathbf{C}_x$ and $\mathbf{C}_{\text{in}}$. This logic is similar to the argument of [10], where it was reconstruction performance that was equivalent under transformations of the filters but costs were superadditive.

However, this reasoning does not fully fix the choice of independent filters, since there are multiple ways to make cross-terms in the determinants vanish. For example, in the discrete (pixel, frame) formulation of the problem, the generalized eigenvectors of $(\mathbf{C}_x, \mathbf{C}_{\text{in}})$ represent a special solution (the one that fully diagonalizes both matrices). Yet these filters need not match the optimal form (10). But there is an alternative solution that does respect the optimal filter shapes: choosing spectrally disjoint filters with $v^*(k,\omega)v'(k,\omega) = 0$ for different mosaics defined by v and v'. In [10], the authors considered a form of this with spatial bandwidth partitioning ($v^*(k)v'(k) = 0$), but as we show below, this special case encounters a limit beyond the first two mosaics, when the spatial passband spectra of new filters substantially overlap. Rather, the more general solution is that mosaic filters are band-limited in both space and time, arranging themselves to tile minimally overlapping regions with highest power in the (k,ω) plane (Appendix A).

3.3 Effects of channel capacity

How, then, does this picture change as the channel capacity of the system is increased by adding additional neurons? For a finite retina of size L^d in d dimensions, J receptive fields lead to a density $\rho \sim J/L^d \sim \mathrm{vol}(G_0)$ and a mosaic spacing $\Delta z \sim \rho^{-1/d}$. Likewise, under changes in density, the principal vectors scale as $(\mathbf{a}_1, \mathbf{a}_2) = (\hat{\mathbf{a}}_1, \hat{\mathbf{a}}_2)\Delta z$ and the reciprocal vectors of the dual lattice as $(\mathbf{b}_1, \mathbf{b}_2) = (\hat{\mathbf{b}}_1, \hat{\mathbf{b}}_2)\frac{1}{\Delta z}$. In other words, increasing cell number leads to increasing density, which leads to increases in the size of the unit cell G_0 in spatial frequency space. Thus, in the infinite retina limit, we consider the optimal arrangement of mosaics as density and spatial bandwidth grow.

The primary effect of this increase in density on the optimization problem (9) is to raise the cutoff on the sum over spatial frequencies. That is, denser retinal mosaics offer higher spatial resolving power [10]. Moreover, this change in bandwidth changes the shape (much less so the size) of the optimal filters. Yet increases in high-frequency cutoff for a single filter type result in diminishing returns in terms of information. However, as detailed above, adding filters covering non-overlapping spectral windows allows mosaics to specialize to different regions of spatial and temporal frequency, offering gains in information as density (and thus total spatial frequency bandwidth) increases. As we show below, these results continue to hold in the nonlinear, discretized case.

4 Experiments

We analyzed the characteristics of the optimal spatiotemporal RFs obtained from the model (2, 3) trained on videos from the Chicago Motion Database [14]. We used square-shaped input patches $\mathbf{x}(t)$ of size $D = 18^2 = 324$ pixels unless otherwise specified. We applied a circular mask on the input patches following [11], which effectively constrained the spatial kernels within a circle bounded by the input square. For computational efficiency in large-scale experiments in Sections 4.2 and 4.3, we approximated the input data distribution as a multivariate Gaussian matched to the statistics of the natural image data set estimated using one million samples of video patches. Input patches consisted of either $T = 10$ or 20 frames sliced from 30-fps videos, and for computational efficiency, we used "valid"-mode convolutions with temporal kernels of the same size T, so each neuron yielded a single firing rate value r_j for each input. Models using larger convolutions, capturing information across both neurons and time, produced similar results. Model parameters for spatial kernels, temporal kernels, and the nonlinearities were jointly optimized using Adam [19] to maximize (2) subject to the mean firing rate constraint (3), enforced using the augmented Lagrangian method with the quadratic penalty $\rho = 1$ [20]. Further technical details of analysis and model training are in Appendix E.

4.1 Natural video statistics affect learned receptive fields

As previously noted, the power spectral density of natural videos is can be well approximated by a product of spatial and temporal power-law densities, implying an anticorrelation between high spatial and temporal frequency content [21]. To examine the effect of these statistics on the learned RFs, we divided the dataset into four progressively smaller subsets by the proportion of their temporal spectral content below 3 Hz, their *spectral attenuation*. Using values of 70%, 80%, and 90% then yielded a progression of data sets ranging from most videos to only the slowest videos (Figure 3A, B). As shown in Figure 3C, when the model was trained on these progressively slower data subsets, it produced only temporal smoothing filters, whereas the same model trained on all videos produced a variety of "fast" temporal filter types. We also note that these experiments used *unconstrianed* spatial kernels in place of (8), yet still converged on spatial RFs with typical center-surround structure as in [7, 12, 11].

4.2 Mosaics fill in order of temporal frequency

As the number of neurons available to the model increased, we observed the formation of new mosaics with new spectral properties (Figure 4). We characterized each neuron's learned filters in terms of their spectral centroid, defined as the center of mass of the Fourier (spatial) or Discrete Cosine (temporal) transform. Despite the fact that each model neuron was given its own spatial and temporal filter parameters (8, 6), neurons' learned filter shapes strongly clustered, forming mosaics with nearly uniform response properties (Figure 4A–C). Critically, the emergence of new mosaics shifted the spectral responses of previously established ones, with new mosaics compressing the

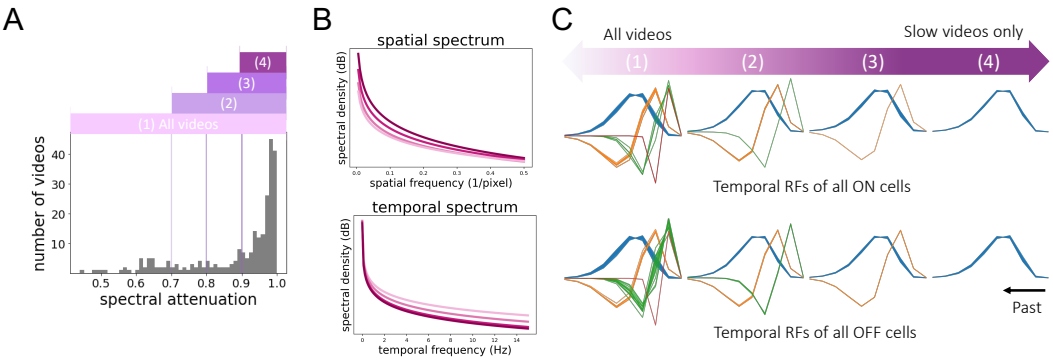

Figure 3: Statistics of natural videos affect learned receptive fields. **(A)** Histogram of spectral attenuation (fraction of power < 3 Hz) for each video clip from the Chicago Motion Database. A significant portion of the dataset exhibits predominaly low-frequency spectral content in time. Videos with spectral attenuation above 0.9, 0.8, and 0.7, are denoted (4), (3), and (2), respectively, while (1) refers to all videos in the dataset. **(B)** Spatial (top) and temporal (bottom) spectral density of the four subsets. **(C)** Temporal filters learned by training on each of the four subsets. Training on slow videos produced only smoothing kernels, while training on all videos produced a variety of temporal filters.

spectral windows of one another as they further specialized. Moreover, mosaic density increased with increasing neuron number, shifting the centroids of early mosaics toward increasingly higher spatial frequencies. This is also apparent in the forms of the typical learned filters and their power spectra: new filters selected for increasingly high-frequency content in the temporal domain (Figure 4D).

We likewise analyzed the coverage factors of both individual mosaics and the entire collection, defined as the proportion of visual space covered by the learned receptive fields. More specifically, we defined the spatial radius of an RF as the distance from its center at which intensity dropped to 20% of its peak and used this area to compute a coverage factor, the ratio of total RF area to total visual space ($\pi/4$ of the square's area due to circular masking) to obtain the coverage factor. Since coverage factors depend not simply on neuron number but on RF density, they provide an alternative measure of the effective number of distinct mosaics learned by the model. As Figure 4E shows, coverage increases nearly linearly with neuron number, while coverage for newly formed mosaics increases linearly before leveling off. In other words, new mosaics initially increase coverage of visual space by adding new RFs, but marginal gains in coverage diminish as density increases. In all cases, the model dynamically adjusts the number of learned RF types and the proportion of neurons assigned to them as channel capacity increases.

4.3 Phase changes in mosaic arrangement

In addition to retinal organization at the level of mosaics, a pair of recent papers reported both experimental [12] and theoretical [11] evidence for an additional degree of freedom in optimizing information encoding: the relative arrangement of ON and OFF mosaics. Jun et al. studied this for the case of natural images in [11], demonstrating that the optimal configuration of ON and OFF mosaics is alignment (RFs co-located) at low output noise levels and anti-alignment (OFF RFs between ON RFs and vice-versa) under higher levels of retinal output noise. Moreover, this transition is abrupt, constituting phase change in optimal mosaic arrangement.

We thus asked whether learned spatiotemporal mosaics in the space-time optimization exhibited a similar phase transition and how the characteristics of this transition might vary for different mosaics. To do so, following [11], we repeatedly optimized a small ($J = 14$, 7 ON, 7 OFF) model for multiple learned filter types while systematically varying levels of input and output noise. In each case, one ON-OFF pair was fixed at the center of the space, while the locations of the others were allowed to vary. We used receptive field size $D = 8^2$ pixels for *Slow* and $D = 12^2$ for *FastA* and *FastB* to allow the size of spatial kernels to be similar to those of the previous experiments, and we imposed the additional constraint that the shape parameters a_j, b_j, and c_j in (6) be same across neurons.

Under these conditions, the six free pairs of RFs converged to either aligned (overlapping) or anti-aligned (alternating) positions along the edges of the circular visual space, allowing for a straight-

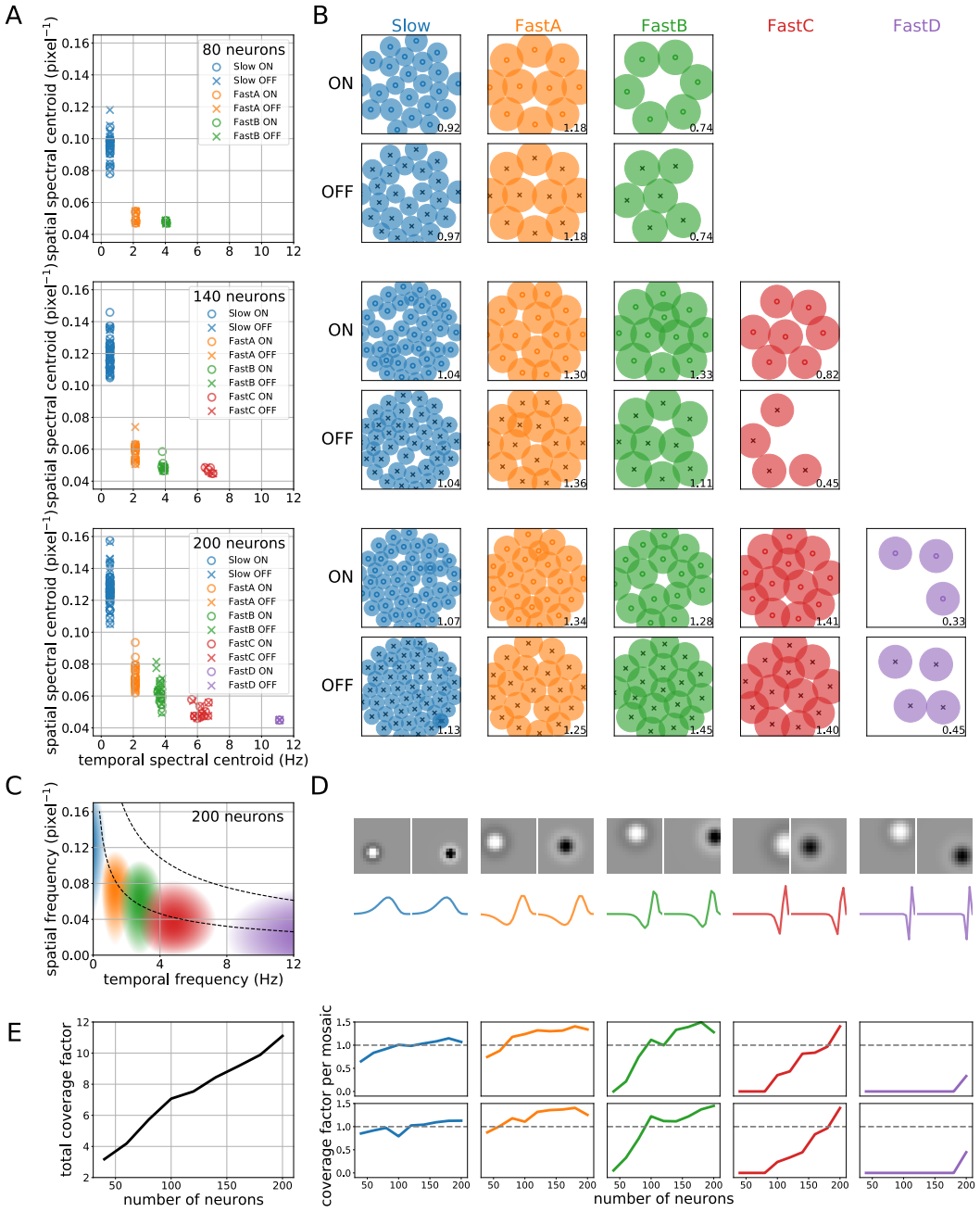

Figure 4: Emergence of new RF types with increasing neuron number. **(A)** Distribution of spatial and temporal spectral centroids for $J = 80, 140, 200$ neurons. ON and OFF RFs form distinct clusters corresponding to different learned filter types. **(B)** ON and OFF mosaics corresponding to each RF type. The number in the lower right of each plot is the coverage factor for the mosaic. **(C)** Power spectral density of a typical kernel in each mosaic for $J = 200$. As predicted, learned kernels filter over roughly nonoverlapping regions of spatiotemporal frequency. Contour lines represent isopower lines of the signal correlation $C_x(k, \omega)$. **(D)** Learned shapes of example spatial ON and OFF filters (top) and corresponding temporal filters (bottom) from each RF type for the $J = 200$ case. **(E)** Total (left) and per-mosaic (right) coverage factors as the number of neurons J increases from 40 to 200. New mosaics increase coverage linearly with the number of RFs, while nearly full mosaics see diminishing returns in coverage from density increases. See Supplementary Figures 4-5 for similar plots for all neuron numbers.

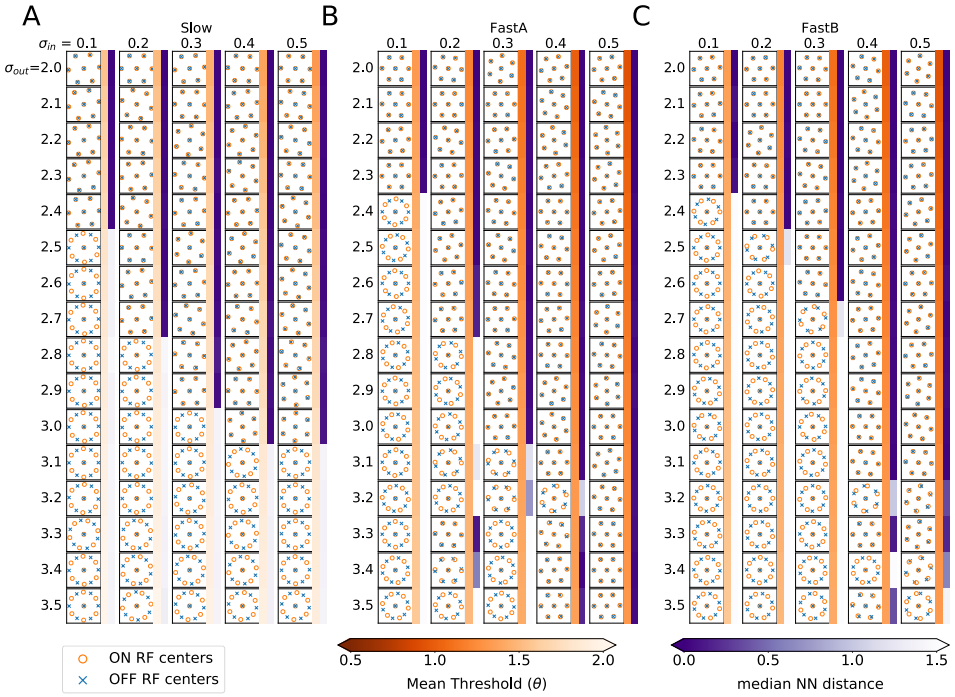

Figure 5: Learned spatiotemporal mosaics exhibit a phase transition as a function of input and output noise. **(A-C)** Spatial kernel centers for *Slow* **(A)**, *FastA* **(B)**, and *FastB* **(C)** as a function of σ_{in} and σ_{out}. In all three cases, the optimal configuration changes from aligned to anti-aligned when output noise increases or input noise decreases. Orange color bars denote learned firing thresholds, blue bars alignment as measured by median distance to the nearest RF center of the opposite type.

forward examination of the effect of input and output noises on mosaic arrangement. Figure 5A–C shows that the phase transition boundaries closely follow the pattern observed in [11]: increasing output noise shifts the optimal configuration from alignment to anti-alignment. Moreover, for each of the tested filters, increasing input noise *discourages* this transition. This effect also follows from the analysis presented in [11], since higher input noise increases coactivation of nearby pairs of RFs, requiring larger thresholds to render ON-OFF pairs approximately indpendent (Appendix B).

5 Discussion

Here, we have shown that efficient coding of natural movies results in the formation of multiple receptive field mosaics with complementary filtering properties [10]. In addition, we have shown how the number and characteristics of these mosaics depend crucially on the channel capacity of the retina as dictated by the number of available neurons. As new neurons become available, they are initially concentrated into denser RFs, improving the spatial bandwidth over which information is encoded, but as this strategy becomes less effective, new cell types encoding higher-frequency temporal features form. These new cell types capture information over distinct spatiotemporal frequency bands, and their formation leads to resultant shifts in the response properties of previous cell types (up in spatial frequency, down in temporal frequency). Moreover, these receptive field mosaics continue to exhibit the phase transition between alignment and anti-alignment between pairs of ON and OFF pairs studied in [11], suggesting that mosaic coordination may prove a general strategy for increasing coding efficiency in the retina. Moreover, despite the strong assumptions of our model—linear filtering, separable filters, firing rates instead of spikes—our results are in strong agreement with observed retinal data. Thus, these results further testify to the power of efficient coding principles in providing a principled framework for understanding nervous system organization.

References

[1] Mrinalini Hoon, Haruhisa Okawa, Luca Della Santina, and Rachel OL Wong. Functional architecture of the retina: development and disease. *Progress in retinal and eye research*, 42:44–84, 2014.

[2] Joshua R Sanes and Richard H Masland. The types of retinal ganglion cells: current status and implications for neuronal classification. *Annual review of neuroscience*, 38:221–246, 2015.

[3] Tom Baden, Philipp Berens, Katrin Franke, Miroslav Román Rosón, Matthias Bethge, and Thomas Euler. The functional diversity of retinal ganglion cells in the mouse. *Nature*, 529(7586):345–350, 2016.

[4] Horace B Barlow. Possible principles underlying the transformation of sensory messages. *Sensory communication*, 1:217–234, 1961.

[5] Joseph J Atick and A Norman Redlich. Towards a theory of early visual processing. *Neural computation*, 2(3):308–320, 1990.

[6] Joseph J Atick and A Norman Redlich. What does the retina know about natural scenes? *Neural computation*, 4(2):196–210, 1992.

[7] Yan Karklin and Eero P Simoncelli. Efficient coding of natural images with a population of noisy linear-nonlinear neurons. In *Advances in neural information processing systems*, pages 999–1007, 2011.

[8] Eizaburo Doi, Jeffrey L Gauthier, Greg D Field, Jonathon Shlens, Alexander Sher, Martin Greschner, Timothy A Machado, Lauren H Jepson, Keith Mathieson, Deborah E Gunning, et al. Efficient coding of spatial information in the primate retina. *Journal of Neuroscience*, 32(46):16256–16264, 2012.

[9] Xaq Pitkow and Markus Meister. Decorrelation and efficient coding by retinal ganglion cells. *Nature neuroscience*, 15(4):628–635, 2012.

[10] Samuel Ocko, Jack Lindsey, Surya Ganguli, and Stephane Deny. The emergence of multiple retinal cell types through efficient coding of natural movies. In *Advances in Neural Information Processing Systems*, pages 9389–9400, 2018.

[11] Na Young Jun, Greg D Field, and John Pearson. Scene statistics and noise determine the relative arrangement of receptive field mosaics. *Proceedings of the National Academy of Sciences*, 118(39), 2021.

[12] Suva Roy, Na Young Jun, Emily L Davis, John Pearson, and Greg D Field. Inter-mosaic coordination of retinal receptive fields. *Nature*, 592(7854):409–413, 2021.

[13] Stephanie E Palmer, Olivier Marre, Michael J Berry, and William Bialek. Predictive information in a sensory population. *Proceedings of the National Academy of Sciences*, 112(22):6908–6913, 2015.

[14] Jared M. Salisbury and Stephanie E. Palmer. Optimal prediction in the retina and natural motion statistics. *Journal of Statistical Physics*, 162(5):1309–1323, January 2016.

[15] Audrey J Sederberg, Jason N MacLean, and Stephanie E Palmer. Learning to make external sensory stimulus predictions using internal correlations in populations of neurons. *Proceedings of the National Academy of Sciences*, 115(5):1105–1110, 2018.

[16] EJ Chichilnisky and Rachel S Kalmar. Functional asymmetries in on and off ganglion cells of primate retina. *Journal of Neuroscience*, 22(7):2737–2747, 2002.

[17] Sneha Ravi, Daniel Ahn, Martin Greschner, EJ Chichilnisky, and Greg D Field. Pathway-specific asymmetries between on and off visual signals. *Journal of Neuroscience*, 38(45):9728–9740, 2018.

[18] W. Bialek and W.G. Owen. Temporal filtering in retinal bipolar cells. elements of an optimal computation? *Biophysical Journal*, 58(5):1227–1233, 1990.

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

[20] Jorge Nocedal and Stephen Wright. Numerical optimization. *Springer Science*, 35(67-68):7, 1999.

[21] Dawei W Dong and Joseph J Atick. Statistics of natural time-varying images. *Network: Computation in Neural Systems*, 6(3):345, 1995.

[22] Michael S Lewicki. Efficient coding of natural sounds. *Nature neuroscience*, 5(4):356–363, 2002.

[23] Stephen W Kuffler. Discharge patterns and functional organization of mammalian retina. *Journal of neurophysiology*, 16(1):37–68, 1953.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We address this in the Discussion.
   (c) Did you discuss any potential negative societal impacts of your work? [No] The focus of this work is information-theoretic analyses on a simple computational model for the retina. As such, we do not anticipate any direct, negative societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Detailed treatments are Appendices A and B.
   (b) Did you include complete proofs of all theoretical results? [Yes] Detailed treatments are Appendices A and B.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The code and instructions to run the complete implementation of this work are included in the supplementary materials, with the necessary links to download the datasets for full replication.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] (and partially N/A) our work did not include data splits, and other, empirically chosen hyperparameters are detailed in the main text and the supplement.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] Our figures were either conceptual or theoretical plots or visualizing exemplar data, where error bars are not suitable.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We detail the compute usage together with other training details in the supplement.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We primarily use the Chicago Motion Database [14] and data derived from it.
   (b) Did you mention the license of the assets? [No] The license is not specified in the paper and the website `https://cmd.rcc.uchicago.edu`, but the website's about page says it is "intended for use in visual neuroscience experiments, as training data for machine vision applications, ...".
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We have included zipped source code in the supplement to preserve anonymity for review but will include a link to the GitHub repository in a camera-ready version.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] We did not use human data.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] Our data consisted of images of natural scenes.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] We did not use crowdsourcing or human subjects

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] We did not use crowdsourcing or human subjects

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] We did not use crowdsourcing or human subjects