# Manifold Transform by Recurrent Cortical Circuit Enhances Robust Encoding of Familiar Stimuli

Weifan Wang[1,4], Xueyan Niu[1,2], Liyuan Liang[1,3], Tai-Sing Lee[1*,3,5],

**1** Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, PA, USA
**2** Center for Neural Science, New York University, New York City, NY, USA
**3** Neuroscience Institute, Carnegie Mellon University, Pittsburgh, PA, USA
**4** Department of Biomedical Engineering, Carnegie Mellon University, Pittsburgh, PA, USA
**5** Computer Science Department, Carnegie Mellon University, Pittsburgh, PA, USA

* taislee@cmu.edu

## Abstract

A ubiquitous phenomenon observed along the ventral stream of the primate hierarchical visual system is the suppression of neural responses to familiar stimuli at the population level. The observation of the suppression of the neural response in the early visual cortex (V1 and V2) to familiar stimuli of size that are multiple times larger in size than the receptive fields of individual neurons reflects the plausible development of recurrent circuits for encoding these global stimuli. In this work, we investigated the neural mechanisms of familiarity suppression and showed that an excitatory recurrent neural circuit, consisting of neurons with small and local receptive fields, can develop to encode specific global familiar stimuli robustly as a result of familiarity training. This Hebbian learning based model attributes the observed familiarity suppression effect to the sparsification of the population neural code for the familiar stimuli due to the formation of image-specific local excitatory circuits and competitive normalization among neurons, leading to the paradoxical neural response suppression to the familiar stimuli at the population level. We explored the computational implications of the proposed circuit by relating it to the sparse manifold transform. The recurrent circuit, by linking spatially co-occurring visual features together, compresses the dimensions of irrelevant variations of a familiar image in the neural response manifold relative to the dimensions for discriminating different familiar stimuli. The computation can be considered as a globally non-linear but locally linear manifold transform that orthogonalizes the slow modes of network dynamics relative to the subspace of irrelevant stimulus variations, resulting in increased robustness of the global stimulus representation against noises and other irrelevant perturbations. These results provide testable predictions for neurophysiological experiments.

## Author summary

In this research, we explored how the brain can become more efficient at processing familiar visual information. When we repeatedly see something, our brain's response to it changes. In response to familiar stimuli, neurons across the different visual areas of the mammalian visual system become more selective and their overall activities

decrease. We developed a computational model to investigate why this happens and what functional advantages these mechanisms might provide. We discovered that familiarity leads to the development of a more efficient and robust neural representation of what we see. It allows us to rapidly and robustly recognize our friend's face despite changes in lighting conditions, view angle, or facial expression. Our model showed that through repeated exposure, the brain's neural circuits, even in the early stages of visual processing, rapidly adapt and organize themselves to focus on important and consistent features in our visual environment while becoming less sensitive to irrelevant variations, and distractions.

# 1 Introduction

Familiarity suppression refers to a phenomenon observed in the inferotemporal cortex (ITC) [1, 2, 3, 4, 5, 6, 7] and more recently in early visual cortex [8] that repeated exposure to a set of familiar visual stimuli will lead to the suppression of neural responses to these stimuli, particularly in the later part of the temporal responses. There is evidence in the inferotemporal cortex that familiarity training leads to the sparsification of population neural representation to the familiar stimuli, as neurons' responses to their preferred familiar stimuli were found to be enhanced, while their responses to non-preferred familiar stimuli were suppressed, resulting in a sharpening of the stimulus selectivity tuning curves of the neurons [6, 7].

In the early visual cortex, Huang et al. [8] showed that neurons with localized receptive fields became sensitive to the global context of familiar images. Based on timing, it can be inferred that this sensitivity is mediated by the recurrent circuits within V2 rather than feedback from higher visual areas. Similar effects have also been observed in V1 as well, but with a shorter delay with stimulus onset, significantly earlier than the familiarity effects in IT. These observations suggest a rapid plasticity mechanism in the early visual cortex modifying the recurrent circuit within each visual area along the visual hierarchical system to encode global or semi-global familiar image context. These findings, together with the involvement of BCM-like learning rule in the familiarity effect [9], suggest that neurons in the early visual cortex, with local receptive fields, can rapidly learn recurrent excitatory circuits to encode global visual images.

However, the computational rationales and the neural mechanisms of this rapid plasticity are not well understood. Proposals on the behavioral and perceptual benefits of the familiarity effects in IT have been focused on image discrimination, reduced saliency to familiar entities, and novelty detection [6, 7, 10, 11]. Recent studies found that representations of natural movies in the primary visual cortex change with repeated exposures [12], which inspires us to investigate the relation between representational changes and familiarity training. The objective of this study is to link the familiarity effects to manifold learning, which seeks to find the correct geometric relations between visual concepts and their variants generated by different smooth transformations, such as different levels of occlusion, view angle, etc. [13, 14, 15, 16] (here the term "visual concepts" refers to images with different semantic meanings, in contrast to the general term "images" that refers to both visual concepts and their variants).

In this paper, we develop a V1-based neural circuit model based on BCM Hebbian learning and other standard V1 circuitry elements that can account for the familiarity training effects. This is a universal canonical circuit motif that can be generalized to V2, V4, and IT. We analyzed this circuit to show that familiarity training of a certain global image stimulus, henceforth called a global visual concept, transforms the neural representation manifold in such a way that unimportant variations of the same concept are ignored while discrimination of different visual concepts is maintained. We demonstrated that this manifold transform provides a more robust encoding of familiar

images or concepts so that they are more immune to noises. We found that the recurrent cortical circuit implements the manifold transform by locally linear computations that rotate the slow patterns of the local network dynamic to a direction in the input space that is orthogonal to the variation of the same concept. This novel perspective on cortical recurrent circuits provides insights into the functional rationales underlying the familiarity learning observed in the visual cortex.

# 2 Results

## 2.1 Recurrent neural circuit model of visual cortex shaped by BCM learning

Familiarity training effects have been reported in macaque ITC and V2. Similar effects have also been observed by us in macaque V1 (unpublished observations), as well as by others in mouse V1 [11, 17, 18]. Since V1 circuits are relatively well understood and modeled, we decided to use the V1 model for this study. We constructed a neural circuit model of the primary visual cortex to demonstrate that the plastic horizontal connections, driven by the BCM rule [9, 19, 20], can reproduce familiarity effects. This basic associative learning mechanism is likely generalizable to model familiarity effects in other visual and non-visual cortical areas.

The network model (Fig 1A) is a firing-rate-based recurrent neural network with $N_h$ hypercolumns (with $N_r$ rows and $N_c$ columns). Each hypercolumn comprises $N_d$ excitatory neurons with receptive fields (RF) derived from sparse coding [21, 22]. We have $N_e = N_r \times N_c \times N_d$ excitatory neurons and the same number ($N_i$) of inhibitory neurons in the network. Each excitatory neuron $k$ receives projection from its excitatory neighborhood ($NE(k)$) with a spatial extend of $R_e$ (Fig 1B). Each inhibitory neuron $k$ receives projections from the excitatory neurons of the same feature channel located in its inhibitory neighborhood ($NI(k)$) with range $R_i$ (Fig 1B) and projects back to all excitatory neurons in the network, mediating iso-orientation (or iso-feature) suppression. For neurons not at the border, $|NE(k)| = N_d \times (2R_e + 1)^2$, and $|NI(k)| = (2R_i + 1)^2$. In addition, excitatory neurons within the same hypercolumn project to an inhibitory neuron, which uniformly inhibits these excitatory neurons in return, as a form of divisive normalization [23, 24].

The dynamics of the excitatory population and inhibitory population are given, respectively as:

$$\tau_e \frac{dr_k^e}{dt} = -r_k^e + \sigma(\sum\nolimits_l W_{kl}^{ee}\, r_l^e - \sum\nolimits_{l'} W_{kl'}^{ei}\, r_{l'}^i + \alpha_k) \tag{1}$$

$$\tau_i \frac{dr_k^i}{dt} = -r_k^i + \sigma(\sum\nolimits_l W_{kl}^{ie}\, r_k^e) \tag{2}$$

where $r_k^e$, $r_k^i$ are the firing rates of the $k^{th}$ excitatory neuron and inhibitory neuron, respectively. $W_{kl}^{ee}$ is the E-E connection from excitatory neuron $l$ to excitatory neuron $k$; similarly $W_{kl}^{ie}$ and $W_{ij}^{ei}$ are the E-I and I-E connection; $\alpha_k$ is the input to the excitatory neuron $k$ obtained via convolutional sparse coding [22]. $\sigma(.) = \lfloor . \rfloor_+^2$ is the squared relu activation function (Fig 1C), as introduced in [24]. $\tau_e$ and $\tau_i$ are the time constants of excitatory and inhibitory neurons, respectively.

Here, we are pursuing a minimal circuit mechanism that could reproduce the familiarity effect in neural circuits. Hence, we start by only considering excitatory plasticity while assuming that familiarity learning did not fundamentally change the canonical inhibitory mechanisms. Hence, only $W^{ee}$ is subject to BCM learning, while $W^{ei}$ and $W^{ie}$ are fixed. The initial value of the E-E connection is set to $w_{ee}/|NE(k)|$, within the excitatory neighborhood. For a single inhibitory neuron, the E to I

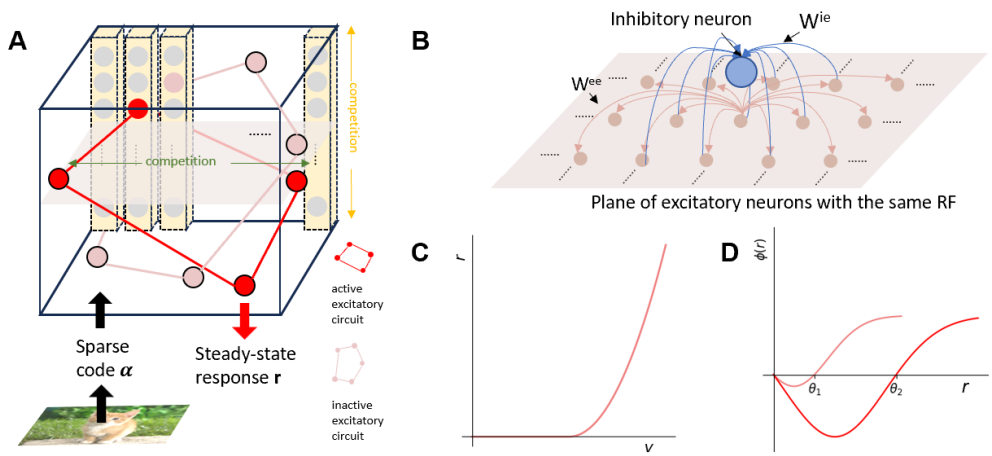


**Fig 1.** Recurrent circuit model of the visual cortex. (A) Illustration of the network structure. An excitatory sub-circuit is formed between excitatory neurons across different hypercolumns for encoding an image. Competitions exist among neurons within the same feature channel across hypercolumns and among neurons of different feature channels in the same hypercolumn. The input to the network or the neurons' initial responses $\boldsymbol{\alpha}$ are computed as the sparse codes of the image [21], and the output $\boldsymbol{r}$ is the steady-state population response of the excitatory-inhibitory network. (B) Illustration of the excitatory neighborhood (for a single feature channel) and inhibitory neighborhood. (C) The activation function of a neuron that maps membrane potential $v$ to neuronal firing rate $r$. (D) The graph illustrates the BCM learning rule, showing the synaptic potentiation $\phi$ of the post-synaptic neuron depends on its firing rate and the BCM threshold. The synaptic weight increases when the firing rate is above the BCM threshold and decreases otherwise. BCM rules for two thresholds $\theta_1$,$\theta_2$ are shown.

connection to this neuron is uniformly set to a fixed value $w_{ie}/|NI(k)|$ within its inhibitory neighbors. The I to E connection from this inhibitory neuron to all excitatory neurons is set to the fixed value $1/N_i$. $w_{ie}$ is set so that normalization within the hypercolumn and iso-orientation (iso-feature) surround suppression across the hypercolumns are strong enough to ensure the stability of the network.

The BCM learning rule [20] is written as:

$$\tau_w \frac{dW_{kl}^{ee}}{dt} = r_l^e \, r_k^e \, \frac{r_k^e - \theta_k}{\theta_k}; \tag{3}$$

where $\tau_w$ is the synaptic time scale determining the speed of learning. The BCM threshold $\theta_k$ for excitatory neuron $k$ is computed by taking the exponential moving average of the neuron's squared firing rate. The synaptic connection strength increases only if the neuronal firing rate exceeds the BCM threshold, and decreases if the firing rate is below the threshold. As the threshold traces the neuron's activity, for neurons after a period of high activity, the threshold increases (red curve in Fig 1D), and for neurons after a period of low activity, the threshold decreases (pink curve in Fig 1D). This negative feedback mechanism prevents unbounded synaptic increase and ensures the stability of the learning. We also added synaptic scaling to preserve the total strength of pre-synaptic connections to an excitatory neuron $i$ throughout the learning [25]:

$$\sum\nolimits_l W_{kl}^{ee} = w_{ee} \tag{4}$$

## 2.2 Familiarity suppression and tuning curve sharpening in the model

We verified that the model could indeed reproduce the effect of familiarity suppression. Fig 2B shows the averaged peri-stimulus firing rates of all excitatory neurons across a set of stimuli before and after familiarity training. Neural activity in the network peaks right after stimulus presentation. Surround suppression kicks in, leading to the response decay. This is followed by some small oscillations due to the interaction of the surrounding excitatory and inhibitory neurons. Finally, the activity settles into a steady state. After familiarity training, the averaged population responses to trained stimuli are profoundly suppressed (Fig 2B, yellow area), consistent with the findings of the neurophysiological experiment in the macaque visual cortex (Fig 2A).

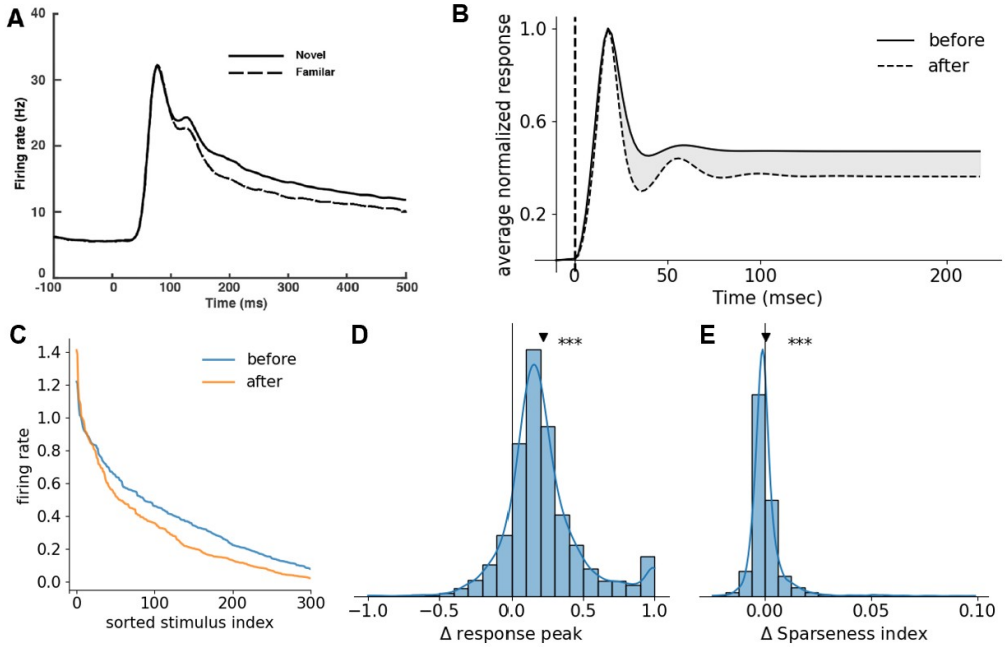


**Fig 2.** Tuning curve sharpening and familiarity suppression. (A) Experimental observation showing neural responses of familiar images are suppressed relative to novel images (Huang et al. 2018). (B) The average neural population response to a set of images is suppressed after the neurons are subjected to familiarity training. The grey area represents the suppression caused by familiarity training. This is equivalent to comparing the neural population responses to a set of trained images against a set of untrained novel images in the neurophysiological experiment, in which neurons could not be tracked and compared reliably across days. (C) Example neuron's tuning curve to 500 images before and after familiarity training. After training, the tuning curve becomes sharper: the response to the most preferred familiarized stimulus is enhanced, and responses to non-preferred familiarized stimuli are suppressed for this neuron, resulting in increased sparsity of the neuronal tuning. (D) The population histogram of change in the tuning peak shows a significant increase. (E) The population histogram of change in sparsity shows a small but statistically significant increase due to the long right tail. *** indicates $p < 0.05$ for the alternative hypothesis that the mean is greater than 0.

The trained network exhibited significant sharpening of the stimulus tuning curves of

the excitatory neurons to the familiar stimuli. Fig 2C shows sorted tuning curves of an example neuron before and after training, where the index was ordered by pre-trained response (in descending order). Familiarity training suppressed its response to most stimuli while only increasing the response to the strongly preferred stimuli, resulting in a sharper tuning curve. We quantified the relative change in the sparseness index (see Section 4.3) of the stimulus tuning curves for all neurons by the ratio ((after - before) / (after + before)). The overall effect on the tuning sparseness across the population is small but still significant, attributing to the long tail right tail of the distribution (Fig 2E), where about 3% of neurons show large changes (2.5%-10%) in the tuning sparseness before and after training. Meanwhile, the peak firing rates of individual neurons are increased significantly (Fig 2D). These results suggest that the familiarity suppression effect can result from the sparsification of the neural representation in response to familiar stimuli. The increase in tuning peak reflects the formation of a local circuit based on Hebbian learning to encode a particular global stimulus. When a familiar stimulus is presented, the neurons involved in this local circuit will boost each other's response, leading to an increase in the firing rates of neurons involved in the circuit. On the other hand, the enhanced responses of neurons involved in the circuit will suppress all the other neurons not involved in the circuit in that neighborhood due to the surround suppression and within-hypercolumn normalization, resulting in a net decrease in the averaged responses in the general population.

The connectivity strength in the spatial cortical network is typically modeled distance-dependently [24, 26, 27, 28], while for simplicity the fixed inhibition strength in our network is uniformly distributed in space. We experimented a form of spatial dependent inhibition $w_{ie}$ following a Gaussian profile (up to $\sigma = 1$) (Fig S1A) but found them inconsequential to familiarity suppression and the sharpening of tuning curves (Fig S1B-C). Having strong enough inhibition, however, is crucial. We varied $w_{ie}$ to assess the impact of inhibition strength. We found that the network would be unstable when the inhibition strength is too weak. However, the main effects (familiarity suppression and tuning curve sharpening) can be observed as long as the network is in a stable regime, with the effect decreasing as $w_{ie}$ moves away from the critical value where bifurcation happens (Fig S2A-B).

## 2.3 Manifold transform in the recurrent cortical circuit

In this section and the following, we investigate the computational rationales underlying the learning of the circuit associated with the familiarity effects. Computationally, our model is an attractor network; the input to the network is the sparse code $\boldsymbol{\alpha}$, and the network output is the steady-state response $\boldsymbol{r}$, i.e., the representation of the stimuli. A neural representation manifold is defined as a set of representations of a set of stimuli. [29, 30]. The recurrent circuit performs a manifold transform operation that is dynamically modulated through learning, which maps the input representation manifold to the output representation manifold. We are interested in how learning modulates the manifold transform and what computational benefit familiarity training could bring.

Inspired by the recently proposed sparse manifold transform framework [15, 31], we hypothesize that the manifold transform mediated by the recurrent network would increase the robustness of the representation of visual concepts, which compensates for the low reliability of the sparse code. In the following paragraph, We first discuss the problem of the neural representation caused by the sparse coding algorithm, and how the manifold transform makes the neural representation manifold better reflect the true geometry of the natural images manifold. Then, we propose a more general manifold transform that can be implemented by recurrent cortical circuits.

**The objective of manifold transform** Studies on manifold learning [13, 15, 32] have long noticed that the primary sensory areas, such as V1, with neurons selective to local features, such as the sparse code dictionary used in our model, cannot represent the geometry inherited in images. Smooth transitions in images are often represented by abrupt changes in the population neural codes, particularly when they are sparse codes. Consider the smooth manifold in the high-dimensional space of local visual features (Fig 3A), as discussed in [15], the learned spare code dictionaries are discrete samples on the manifold $\mathcal{M}$, and the sparse code $\boldsymbol{\alpha}_i$ of an image stimulus $\boldsymbol{x}_i$ is a composition of multiple delta functions on the manifold domain (Fig 3A). The underlying transformation of the stimulus, such as those induced by various degrees of noise and occlusion, or the changing view angle, is a collection of flows in $\mathcal{M}$. Because the learned dictionaries are typically unordered, the flow on the manifold leads to a drastic change in the sparse code (Fig 3A).

Here, we consider the problem setup where the image stimuli $\boldsymbol{x}$ comprise visual concepts $\boldsymbol{c}$ as well as their variants generated by the smooth transformation, denoted by $\{\boldsymbol{c}'\}$. We define the variants manifold associated with a concept to be $\mathcal{C}_i = \{\boldsymbol{x}_i \mid \boldsymbol{x}_i \in \boldsymbol{c}_i \cup \{\boldsymbol{c}_i'\}\}$, and the concept manifold to be a smooth manifold underlying all the visual concepts. Our goal is to find an encoder $\phi_{\boldsymbol{W}}$, parameterized by $\boldsymbol{W}$, that takes sparse code $\boldsymbol{\alpha}_i$ as input and captures the true geometric relation between stimuli in the manifold of its representation (denoted by $\boldsymbol{\beta}$). To do that, the learned representation should satisfy the following property compared to the sparse code: Locally around each stimulus $\boldsymbol{x}_i \in \mathcal{C}_i$, $\boldsymbol{x}_i$ should be closer to neighbor stimuli $\boldsymbol{x}_{n(i)}$ that are in the same variants manifold $\mathcal{C}_i$ relative to stimuli $\boldsymbol{x}_j$ that are in other variants manifolds $\mathcal{C}_j$. In other words, the manifold transform should compress all the variants manifolds relative to the concept manifold (Fig 3B).

The desired representation can be obtained by maximizing the similarity of representations within the variants manifolds of each concept while maintaining the difference between the representations of the different concepts. Various objectives have been proposed in the manifold learning and self-supervised learning literature to address this trade-off [13, 15, 16, 33, 34, 35]. Here, we considered an objective that maximizes the variant similarity with a soft constraint by optimizing the parameter $\boldsymbol{W}$ of the encoder.

$$\min_{\boldsymbol{W}} \sum_i \frac{\|\partial \boldsymbol{\beta}/\partial \theta\|^2_{\theta=\theta_i}}{\frac{1}{N-1}\sum_j^{N-1} \|\partial \boldsymbol{\beta}/\partial \gamma^j\|^2_{\gamma^j=\gamma_i^j}}, \quad \text{s.t.} \quad \boldsymbol{\beta} = \phi_{\boldsymbol{W}}(\boldsymbol{\alpha}), \tag{5}$$

The objective operates on the derivatives of the $\theta$ function and the $\gamma_j$ function on the variants manifold and concept manifold, respectively. The former controls the smooth change from a concept $\boldsymbol{x}_i$ to its neighborhood stimuli $\boldsymbol{x}_{n(i)}$ in a variants manifold, and the latter controls the smooth change from a concept $\boldsymbol{x}_i$ to other concepts $\boldsymbol{x}_j$ (Fig 3B). This objective seeks to minimize the change of $\boldsymbol{x}_i$ to $\boldsymbol{x}_{n(i)}$ relative to the averaged change of $\boldsymbol{x}_i$ to different $\boldsymbol{x}_j$, thereby ignoring the variations within the same concept while preserving the variation between the distinct concepts.

**Locally linear manifold transform by the recurrent cortical circuit** The traditional manifold transform algorithms [14, 15] typically use a single linear transform for all the visual concepts. However, organizing the representational geometry of all the concepts and their variations with this linear approach could be challenging. A more flexible way might be to let the manifold transform manipulate the local variations around different points on the manifold with distinct linear operations, thereby enabling concept-specific transforms while maintaining global nonlinearity. The recurrent circuit is inherently suitable for accomplishing this type of globally-nonlinear and locally-linear manifold transform. The nonlinear dynamic of a recurrent circuit can be approximated by a linear dynamical system determined by a Jacobian matrix (see Section 4.4) locally

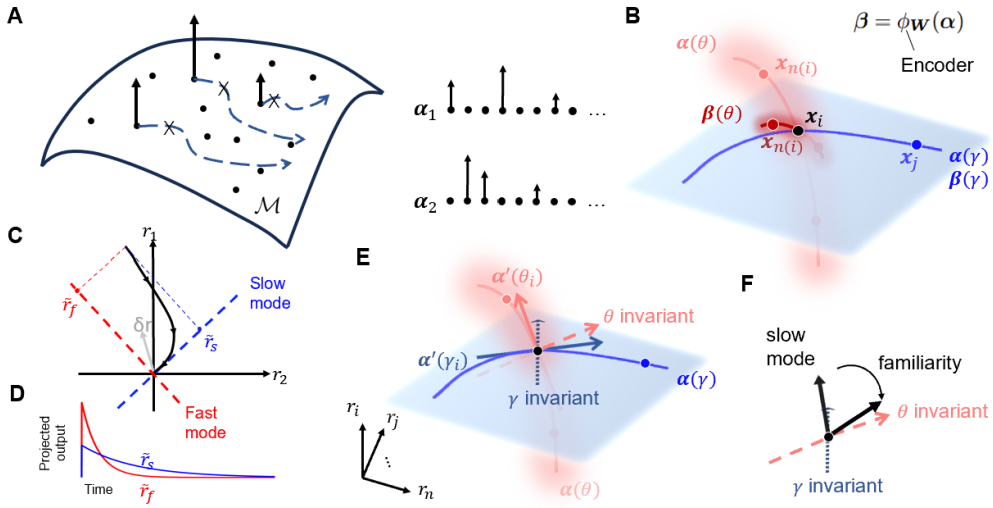


**Fig 3.** Manifold transform in the recurrent cortical circuit. (A) A conceptual model demonstrating how sparse code disrupts inherited geometric structures in images. $\mathcal{M}$ is the putative feature manifold; black dots represent sparse dictionaries that are landmarks on the manifold; black arrows are delta functions on the manifold domains constituting the sparse code of an image; blue arrows show the direction of flows; $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$ are two example sparse codes of two images along the flow. (B) Illustration of manifold transform. The blue area represents the smooth manifold of global concepts (concept manifold), and the blue curve shows a function on the concept manifold, determined by parameter $\gamma$, where stimulus $\boldsymbol{x}_i$ (black dot) smoothly changed to the non-neighbor stimulus $\boldsymbol{x}_j$ (blue dot) in the representation space. The red area represents the manifold of diverse variations of the concept (variants manifold), and the red curve is a function on the manifold determined by parameter $\theta$, along which $\boldsymbol{x}_i$ changes to the neighbor stimulus $\boldsymbol{x}_{n(i)}$ (red dots) in the representation space. Assuming no changes in the concept manifold (for visual simplicity), the learned encoder $\phi_{\boldsymbol{W}}$ compresses the variants manifold. (C) Illustration of dynamical mode decomposition. The black curve is the decaying trajectory of the neuronal response to the small perturbation $\delta r$ around the fixed point (origin). The red and blue dashed line denotes fast and slow dynamical modes, respectively. $\tilde{r}_s$ and $\tilde{r}_f$ indicate projected neuronal response onto the slow and fast modes, respectively. (D) shows the temporal evolution of $\tilde{r}_s$ and $\tilde{r}_f$, which are disentangled, exponential decay with fast decaying speeds ($\tilde{r}_f$, red curve) and slow decaying speed ($\tilde{r}_s$, slow curve). (E) Illustration of $\theta$ invariant direction and $\gamma$ invariant direction. $\boldsymbol{\alpha}'(\theta_i)$: derivative of $\boldsymbol{\alpha}$ w.r.t. $\theta$; $\boldsymbol{\alpha}'(\gamma)$: derivative of $\boldsymbol{\alpha}$ w.r.t. $\gamma$. $\theta$ invariant direction and $\gamma$ invariant direction are orthogonal to $\boldsymbol{\alpha}'(\theta_i)$ and $\boldsymbol{\alpha}'(\gamma)$. (F) Training should render slow modes of the network dynamics more aligned to $\theta$ invariant direction than to $\gamma$ invariant direction, leading to neural responses that are more similar to the neighbor stimuli than other stimuli.

around each fixed point [36, 37, 38]. As a result of this property, the manifold transform performed by the recurrent cortical circuit can be seen as a composition of multiple locally linear transforms around each stimulus in the representation space. Therefore, the recurrent circuit could potentially offer greater flexibility to the manifold transform.

Here, we consider $\phi_{\boldsymbol{W}}$ to be the recurrent cortical circuit, and its steady-state response $\boldsymbol{r}$ to be the output of manifold transform ($\boldsymbol{\beta}$ in Eq 5). To see how the recurrent circuit performs locally linear manifold transform, we consider the tuning

derivative defined as $\boldsymbol{r}' = \partial \boldsymbol{r}/\partial\theta$. The tuning derivative evaluated at $\theta_i$ can be decoupled into disentangled components that fall along each dynamical mode, obtained by eigendecomposition of the Jacobian matrix of that fixed point (see Section 4.4). The dynamical modes are a new set of bases along which the projected local dynamic around a fixed point is decoupled into independent, exponential evolution with different time scales (Fig 3C, D). The basis with a shorter time scale (faster evolution) is the fast mode, whereas the basis with a longer time scale (slower evolution) is called the slow mode. The decomposition of tuning derivative at $\theta_i$ w.r.t. dynamical modes is written as

$$\frac{\partial \boldsymbol{r}(\theta)}{\partial\theta}\Big|_{\theta=\theta_i} = \sum_j^N \tau_j \boldsymbol{v}_j^R (\boldsymbol{v}_j^L)^T \Lambda' \boldsymbol{\alpha}'(\theta_i), \tag{6}$$

where $\tau_j$ represents the dynamical timescale of the $j^{th}$ mode, derived from the eigenvalue of the Jacobian: $\tau_j = -1/Re(\lambda_j)$. The vector $\boldsymbol{v}_j^L$ denotes the $j^{th}$ dynamical mode, which is also the left eigenvector of the Jacobian. The vector $\boldsymbol{v}_j^R$ is the corresponding right eigenvector. $\Lambda'$ is the sensitivity matrix (see Section 4.4), and $\boldsymbol{\alpha}'(\theta_i)$ is the input derivatives at $\theta_i$. Eq 6 shows that the derivative along the representation manifold is the superposition of input derivatives processed by different modes, weighted by their respective time scales. The dynamical mode ($\boldsymbol{v}_j^L$) filters the input derivatives by projection, and the filtered information is projected back to the neural space using right eigenvectors ($\boldsymbol{v}_j^R$) as the basis. Therefore, the filtering performed by the dynamical modes with slow time scales (large $\tau$) determines what information in the input would be extracted or discarded.

Taking advantage of the orthogonality between left and right eigenvectors, a simpler linear relationship between input $\boldsymbol{\alpha}(\theta)$ and output $\boldsymbol{r}$ can be obtained by projecting the representation onto the slow dynamical modes:

$$L^T \boldsymbol{r}'(\theta_i) = \boldsymbol{T}\tilde{L}^T \boldsymbol{\alpha}'(\theta_i). \tag{7}$$

The column of $L$ is $\boldsymbol{v}_i^L$ and the column of $\tilde{L}$ is $\Lambda'\boldsymbol{v}_i^L$. The matrix $\boldsymbol{T} = \delta_{ij}\tau_i$, containing the time scales of each mode in the diagonal. Eq 7 shows that the slow dynamical modes constitute an efficient linear readout for the geometry of the representation manifold, as the representation projection onto each slow mode could retain the geometry-relevant information from the corresponding input projection through linear scaling. In contrast, the arbitrary projection of the representation is the superposition of input projections onto different dynamical modes, thus mixing the geometry-relevant and -irrelevant information.

The key aspect of the local linear transform lies in how slow modes filter inputs. We define the $\theta$-invariant direction to be orthogonal to the input derivatives $\boldsymbol{\alpha}'(\theta_i)$, along which change in $\theta$ is accompanied by little change in neural input activity $\boldsymbol{\alpha}$. $\gamma$-invariant direction is defined similarly with respect to change along the concept dimensions $\gamma$. (Fig 3E). Eq 6 and Eq 7 imply that the alignment of the network's slow modes with the invariant direction modulates the similarity of the stimulus representations. Minimization of the objective (Eq 5) requires the slow modes to align more to the $\theta$-invariant directions than the $\gamma^j$-invariant directions (Fig 3F), so that the representation $\boldsymbol{\beta}$ changes more slowly w.r.t. $\theta$ than w.r.t $\gamma^j$ around each fixed point. Consequently, the manifold's global geometry is reshaped through local optimization of stimulus representation similarities.

## 2.4 Simulation to relate familiarity-trained recurrent cortical circuit and manifold transform

Manifold transform, achieved by optimizing the objective function in Eq 5, by design, seeks to compress the variants manifold, i.e., manifold formed by variations of a visual

concept, relative to the concept manifold, i.e., the smooth manifold underlying all visual concepts. The variation of a visual concept can come in a variety of forms, such as noises and occlusions or changes in images associated with views. Manifold transform seeks to ignore these variations within a concept while maintaining sensitivity to the distinction between the different concepts.

Would familiarity training induce the same effect on a recurrent circuit, compressing the irrelevant variations such as noises and occlusion for each visual concept? To test this prediction, we perform a simulated neurophysiological experiment on spatial and temporal associative learning. The experimental paradigm is as follows: the network, simulating a monkey performing a fixation task, will be presented with a pair of stimuli. The first image is a noise-corrupted version of the second image called the target image. The network is trained for five epochs. Each epoch involves 150 steps of BCM learning (Fig 4A). The noises are salt and pepper occlusion noises at four levels: 0%, 10%, 30%, and 50% (Fig 4A). Corruption of target image $l$ with noise level $n$ results in a conditional distribution of stimuli $p(\boldsymbol{x} \mid n, l)$. We draw 10 image samples denoted by $\boldsymbol{x}_{n,l}$ from each conditional distribution. Here, the visual concepts are the target images, and the noise level for each image concept controls the variations of the input for that image concept, parameterized by the $\theta$ in the previous section.

Familiarity training in this experiment is designed to induce temporal association of images and spatial association of visual features within each image. Given the noisy image is followed by a clear version of the image, this is a form of temporal association learning between the input images, reminiscent of Slow Feature Analysis [32]. The repeated presentation of the target image during training will strengthen the connection between the neurons coding the spatially occurring events in a familiar image based on the Hebbian learning mechanism. If familiarity learning in a recurrent network is consistent with the manifold transform behaviorally, the learned representation in the recurrent circuit should ignore the variations induced by noise (i.e., the $\theta$ dimension) and maintain the discriminability among the different target images (i.e., the $\gamma$ dimension) in the manifold, resulting in greater robustness against the variation due to noise.

After each epoch of training, we probed the neural activity and analyzed the neural representation manifold formed by fixed points of the excitatory population, denoted by $\boldsymbol{r}^{\boldsymbol{e}}$. The familiarity suppression and sharpening of stimulus tuning and population tuning are also observed in the familiarity association experiment (Fig S3A-C). The evolution of the population activity of excitatory neurons in response to different noise levels forms distinct trajectories in the neural representation space, which start at the same initial resting state of the network, then diverge and settle into different fixed points corresponding to the distinct input images (Fig 4B). Fixed points of different noise levels are organized along certain directions in the neural representation space (Fig 4B), forming an image-to-noise axis along which the noise level changes gradually. We will show that familiarity training would compress this axis to increase the noise-robustness of the representation.

Next, we investigated the geometry of the neural representation manifold. We found that the neural representation lies on a lower dimensional space as compared to that of the input sparse code (Fig S3G). The neural representation manifold formed by all the target-noise image pairs has approximately a cone shape, as shown in figure 4C. Specifically, The 10 samples from each conditional distribution of stimuli given the noise level and visual concept form different clusters. The target image is located at one end of the manifold, and increasing the noise level drives the cluster mean farther away from the target image, accompanied by an increase in the variance of the distribution (noise variance)(Fig S3D). Such a geometry arises because the higher noise level reduces the similarity between different noisy image samples. In the meantime, as the noise level increased, the distances between cluster means in the same noise level (signal variance)

decreased. For 100% noises, all the "images" will converge to the same cluster, reflecting the reduced signal contents in the corrupted images. Therefore, if only the cluster means are considered, the manifold's shape is also a cone but has the opposite direction compared to the first cone (Fig S3E).

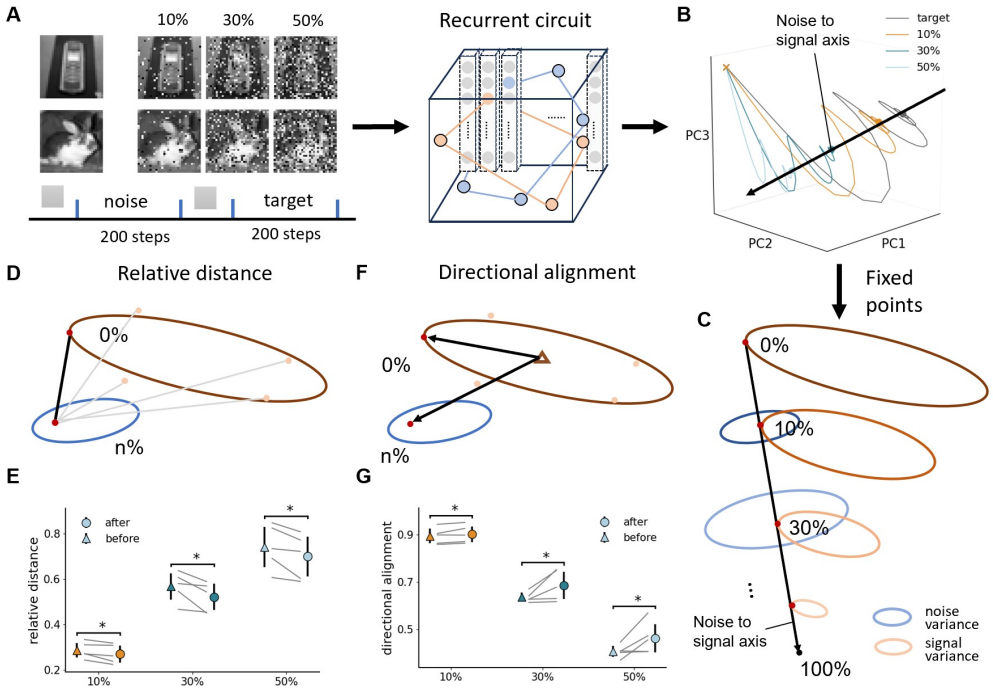


**Fig 4.** Manifold transformation in the familiarity association experiment. (A) Upper: Example images corrupted with different levels of noise. Lower: familiarity training for one noisy-target image pair. (B) Trajectories of different noise levels correspond to an example image in the model. The trajectory is averaged across noisy image samples. The black arrow indicates the direction along which the noise level changes (denoted as the image-to-noise axis). Cross: trial start; Dots: trial end. (C) Cartoon illustration of the cone formed by decreasing signal variance and the cone formed by increasing noise variance of an image. (D) Illustration of the relative distance. The black line between two red dots represents the Euclidean distance between the noisy image and the target image, which is normalized by the sum of distances from this noisy image to all other clear images (represented by the gray lines) (E) Relative distance significantly decreases after training for all noise levels. Triangle marker: before training, circle marker: after training (epoch5 for the simulation). Grey straight lines are for individual target images. *: $p < 0.05$. Error bar represents std across target images. (F) Illustration of the directional alignment. The triangle denotes the reference point formed by averaging all the target images. Cosine similarity is calculated between the two vectors pointing from the reference point to the noisy image and its target. (G), Similar to E but for the directional alignment, which significantly increases after training for all noise levels

We then quantitatively measured whether the trained network transformed the representation manifold in a way consistent with the manifold transform specified in Eq 5. We defined the relative distance as a discrete approximation of the objective (Eq 5). The relative distance between the neural representation of a noisy image and the target image is their Euclidean distance, normalized by the sum of the distances between that specific noisy image and all the clear images except its target (Fig 4D). The directional

alignment is another metric to measure the noise robustness, which is the cosine similarity between the vector from a reference point to the representation of the noisy images and the vector from the reference point to the representation of the target image, where the reference point is the mean of all the target image representations (Fig 4F). We found that for all noise levels, the directional alignment increases and the relative distance decreases, comparing before and after training (Fig 4E, G). Consistent with this compression effect, we observed a decrease in the dimensionality of representation after training (Fig S3G). These results show that the familiarity-trained network behaves in a way consistent with the manifold transform, ignoring the variation specifically arising from the noises, thereby enhancing the noise-robustness of natural image representations.

### 2.5 Familiarity training learns manifold transform by aligning the slow modes of network dynamics to the $\theta$-invariant dimension.

We performed numerical experiments to evaluate the predictions of our analysis on the locally linear manifold transform induced in the recurrent circuit by familiarity training. We used numerical methods to perform linearization analysis around each fixed point of the network. The dynamical modes around each fixed point are all decaying, suggesting that all the fixed points are stable attractors. The decaying modes can be categorized into several groups according to the decaying time constants. The first group of decaying modes has time constants greater and gradually approaches $\tau_e$ (the time constant of excitatory neurons, see Eq 1); The second group has $\tau = \tau_e$, but the modes are all zero vector, thus will not contribute to the dynamic; The third group has time constants smaller than $\tau_e$ and larger than $\tau_i$ (the time constant of inhibitory neurons, see Eq 2), and the last two groups have $\tau = \tau_i$, decaying with the same speed as the inhibitory neurons (Fig 5A). Eigenvectors in groups 1 and 3 have exclusive selectivity for excitatory neurons. Hence, they are relevant to the manifold formed by excitatory population responses. In contrast, eigenvectors in group 4 constitute mainly inhibitory loading (Fig 5B). These observations suggest that the group 1 dynamical mode is most relevant to the transformation of the neural manifold formed by excitatory population response.

In this experiment, target images represent distinct visual concepts, where variations within each concept are governed by the noise level, parameterized by $\theta$. We refer to the $\theta$-invariant direction as the direction insensitive to noise variation, whereas $\gamma^j$-invariant direction as the direction insensitive to the changes in concepts. The alignment of the top 100 slow modes of the trained network with noise-invariant directions and concept-invariant directions is examined. We used the distance between two adjacent noise levels (noise distance $\mathcal{D}_N$) and the distance between concepts (concept distance $\mathcal{D}_C$) to approximate the two tuning derivatives (see Section 4.7 and Fig S4A). We found that the slow decaying dynamic modes aligned more with the noise-invariant direction than the concept-invariant directions, evidenced by the decreased normalized noise distances (ratio of noise and concept distance, $\mathcal{D}_{N/C}$) in the projected input ($\tilde{L}^T\boldsymbol{\alpha}$) compared to the inputs ($\boldsymbol{\alpha}$) (Fig 5C). Projecting response onto the slow modes ($L^T\boldsymbol{r}$) further decreases the normalized noise distance (Fig 5C), confirming that slow modes constitute an efficient subspace for the noise invariance.

We then investigated the effect of training by comparing the trained network with the pre-trained network. For subspace spanned by the top 100 slow modes, training decreases the normalized noise distance in input projection and response projection around each fixed point (Fig 5D-E). The results suggest that familiarity training facilitates manifold transformation by continuously improving the local alignment of

slow dynamical modes with noise-invariant direction, making the slow subspace more robust against noise. We also observed the decrease in the normalized noise distance in the original neural representations ($\boldsymbol{r^e}$) (Fig 5F), which is consistent with the finding of relative distances in Fig 4E.

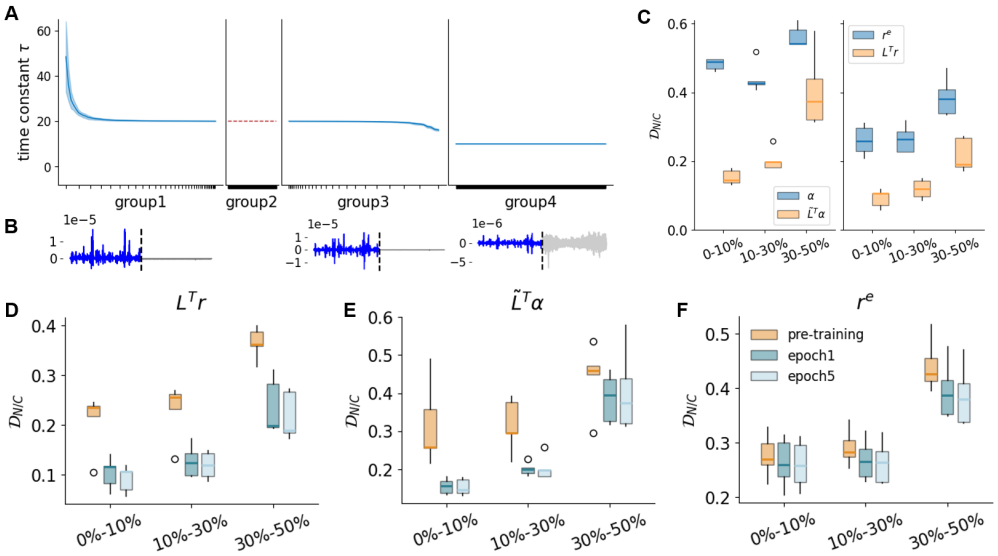


**Fig 5.** Familiarity training aligns slow dynamical modes with $\theta$-invariant direction. (A) Groups of dynamical modes that are specified by decaying the time constant $\tau$. First panel: Group 1, with $\tau > \tau_e$; Second panel: Group2, with $\tau = \tau_e$ and all eigenvectors are zeros vectors. Third panel: Group 3, with $\tau_i < \tau < \tau_e$. Last panel: Group 4 with $\tau = \tau_i$. The blue line represents the mean, and the shaded area represents s.e.m. across visual concepts, noise levels, and training epochs. (B) Averaged eigenvectors of groups 1, 2, and 4 correspond to excitatory neurons (colored in blue) and inhibitory neurons (colored in gray). Eigenvectors in group 1 and group 3 are irrelevant to inhibitory neurons, while eigenvectors in group 4 primarily represent loadings to inhibitory neurons. (C) Normalized noise distance $\mathcal{D}_{N/C}$ in the epoch5 trained network decreases in the projection space. Left panel: comparing feedforward input ($\boldsymbol{\alpha}$) and input projection ($\tilde{L}^L\boldsymbol{\alpha}$); Right panel: comparing the original representation ($\boldsymbol{r^e}$) and response projection ($L^T\boldsymbol{r}$). Boxplots show distributions of visual concepts. (D-F) Familiarity training continuously reduces normalized noise distance $\mathcal{D}_{N/C}$ for input projection, response projection, and original neural response. Boxplots show distributions of visual concepts.

Furthermore, we asked whether the observed effects are consistent across different project subspace dimensions, i.e., the number of slow modes we used to construct the projection subspaces. We found that the first group of decaying mode consistently aligned with the noise-invariant directions more so than the concept-invariant direction (Fig S4D), and noise invariance in the slow modes subspace spanned by different numbers of group 1 modes is consistently better than the original response (Fig S4E). The effects of familiarity training are presented in a wide range of subspace dimensions. We quantified the training effect by the Pearson correlation coefficient between training epochs and normalized noise distance. Fig S4F-G shows that the training effect of the input and output projection exhibits a decreasing trend, with saturation at higher subspace dimensions. Therefore, the observed effects are consistent with different projection subspace dimensions. The result also emphasizes the critical role of the first quarter of the slow modes in compressing the irrelevant variation, as further adding

more slow modes yields diminishing effects.

Taken together, these results demonstrate that recurrent networks align the slow dynamical mode specifically to the $\theta$-invariant direction, and familiarity training continuously enhances the alignment. These properties of the network dynamics facilitate the formation of a representation manifold geometry that improves the representation of familiar stimuli and their robustness against noise.

# 3 Discussion

In this work, we showed that a neural circuit model with standard V1 motifs of hypercolumns, local sparse codes, and local inhibitory and excitatory mechanisms can rapidly learn to encode global image concepts by modifying the excitatory connections in the near-surround using a BCM Hebbian learning mechanism. BCM rule has been used to explain the development of the receptive field's orientation selectivity [20]. The BCM rule is also consistent with the plasticity rule inferred from earlier physiological studies on familiarity effects in IT [9]. We developed a V1 circuit model with the BCM rule to successfully account for the familiarity suppression effects observed in the early visual cortex [8]. The model further predicts the sharpening of the tuning curves for the familiar stimuli in the early visual cortical neurons. Huang et al. [8] did not observe the tuning curve "sharpening." However, in that study, only 25 familiar stimuli were trained and tested, which likely did not include the preferred stimulus of the neurons tested. In our simulation, we found that the tuning sharpening effect can be observed when 500 random natural images were used for familiarity training. Given that the familiarity effect is absent when the stimulus is limited to the receptive fields and the onset of the effect in V1 and V2 is 30 ms earlier than in IT [8], the development of sensitivity to global familiar images, 4-6 times larger than the size of the receptive fields of individual neurons, can be attributed to the formation of local recurrent circuits. These recurrent circuits exist in each visual area, processing familiar images in each area with progressively more global context down the ventral visual hierarchy.

The Hebbian-based cortical recurrent network in each visual area resembles a Hopfield network, with relatively local connections, capable of encoding associative memory. The connections in V1 and V2 extend outside the receptive fields but are as global as a Hopfield net. In addition to Hebbian excitatory horizontal connections, this cortical network is characterized by an additional strong surround iso-orientation (or iso-feature) inhibition and competitive local normalization, which is important for maintaining the stability of the network and for generating familiarity suppression and tuning curve sharpening effects. A neuron strongly activated by a specific familiar image, boosted by the excitatory horizontal circuit for encoding that image, would suppress other less-responsive neurons in the local neighborhood, leading to the familiarity suppression effect at the population level. Here, the super-linear activation function is used to model the neurons [24], but this specific mechanism is not necessary as familiarity effects can also be observed when a rectified linear activation function (ReLU) is used. Nonetheless, the supra-linear activation mechanism enhances the sparsity of the population responses as it more strongly suppresses low activity and amplifies high activity.

While recurrent circuits have been considered as a form of associative memory, we advanced the idea that the recurrent cortical circuit is performing a manifold transform, and training would induce modifications in this manifold transform. In general, a recurrent circuit is performing a globally nonlinear but locally linear manifold transform, in contrast to a single linear transform for all the points on the manifold used in traditional manifold learning [14, 15]. It is due to the nature of recurrent dynamics that in the vicinity of attractors, the behavior of the system is approximately linear, where

the tuning derivatives can be untangled and decomposed into different components that fall along different dynamic modes operating on distinct timescales. In the locally linear manifold transforms, the slow mode of the network dynamic should be rotated during the training process to become more orthogonal to the dimensions of irrelevant variations of the stimulus concepts, so that their impact on the discriminative dimensions can be minimized.

We empirically connected familiarity training and manifold transform by showing that recurrent circuits trained with familiarity training perform the proposed manifold transform for different image concepts in the familiarity association experiment. We found that familiarity training decreases the distances in the noise direction relative to the distances in the stimulus direction locally around each fixed point, which is the expected behavior of manifold transform, resulting in a more robust representation of the global image concepts. We further confirmed that such behaviors were mediated by orthogonalizing the slow modes with the direction of irrelevant variation, as predicted by the theory. Thus, the nonlinear computation of the recurrent circuits shaped by familiarity training can be considered to be "thinking nonlinear globally and acting linearly locally" to achieve flexible manifold transformation around each remembered concept. These findings suggest that familiarity learning simultaneously sparsify the neuronal tuning (which leads to the suppression of population activities) and rotates the slow dynamical mode around each attractor (which results in the manifold transform). A natural question arises: why do these two phenomena co-occur in the learning process? One possibility is that, for the flexibility of each local manifold transform to rotate the slow modes against the specific irrelevant ($\theta$) directions around each concept, the representations or memories should have less interference with each other. Thus, the sparsification of the representation is favored [39]. Nonetheless, future research should address this question by theoretically analyzing the causal relation between the sparsification and mode rotation during the learning medicated by the BCM rule.

This study has additional limitations that should be addressed in future research. Firstly, the architecture of our model does not follow the detailed biological neural circuit structure with different types of neurons and connectivity. Rather, we incorporate the standard circuit mechanisms of the macaque's primary visual cortex, such as surround excitation/inhibition, normalization, etc., to construct a generalizable circuit model and assume such architecture design is canonical and applicable to the familiarity effects observed in V2, V4, and IT in principle. Future research should investigate how incorporating more biologically realistic neural circuit structures, including diverse neuron types and specific connectivity patterns, might influence the model's explanatory power for the familiarity effects and capacity for the manifold transform. Another important remaining question is the broader applicability of the proposed manifold transform mechanism in different continuous transformations. While our current work demonstrates the effectiveness of this approach in handling noise corruption, its adaptation to other transformations, like rotation, translation, contrast modulation, and spatial frequency alterations, as well as to dynamic stimuli that vary in the temporal dimension, remains to be examined. These generalizations might require some variations at the circuit implementation level but should validate the generality of our proposed mechanism and provide insights into the fundamental principles of neural information processing of sensory stimuli.

# 4 Methods

## 4.1 Parameter setting and assumptions

The network we used is a standard V1 network with sparse feedforward encoding. The feedforward response is computed using $N_k = 64$ convolutional filter is obtained using the convolutional sparse coding algorithm that jointly minimizes the reconstruction loss and $L1$ norm of the per-hypercolumn activation [22]. The strategy improved the coding efficiency and was shown to enhance the characterization of the receptive field of neurons in the primary visual cortex [40]

Most of the parameters used in the model are rather standard. The time constant of excitatory neurons ($\tau_e = 20$) double that of inhibitory neurons ($\tau_i = 10$) [41]. The local extent of the surround inhibition and excitation is also relatively well-known based on the spatial extent of the surround inhibition [42], here we set the radius for mutual facilitation among excitatory neurons to be $R_e = 2$, and the radius for iso-orientation suppression is $R_i = 1$. Also, in Fig S1, we showed that the exact profile of the inhibition strength (uniform or Gaussian) is inconsequential in our model. The initial weight of plastic excitatory-excitatory connections is $w_{ee} = 5$. The absolute strength of the inhibitory surround $w_{ie}$ is the main free parameter that was tuned for the model stability. We performed a parameter sweep (Fig S2) to understand the impact of this parameter on the main effects (sharpening of tuning curves and familiar suppression). In the main experiment, we typically set $w_{ie} = 30$. We set the synaptic time constant of BCM learning to be $\tau_w = 2e9$, and the time constant of the exponential moving average for the BCM threshold to be $2e7$.

## 4.2 Familiarity training for the familiarity effect

We trained the network on 500 natural images from CIFAR100. The size of the image is $32 \times 32$. Each image was convolved by 64 filters of size $9 \times 9$ derived from the convolutional sparse coding [22], resulting in outputs of size $8 \times 8$, which is the feedforward input to excitatory neurons in the network. We set the network size $N_r = N_c = 8$ accordingly. Each image was shown to the network for 150 ms in one epoch of training by forward-simulating Euler method with $dt = 1$ ms. The network was trained for 5 epochs. The network was reset to the initial state ($\boldsymbol{r^e} = \mathbf{0}; \boldsymbol{r^i} = \mathbf{0}$) between each image to mitigate inter-stimulus interference. During training, we scale the feedforward input by 30 to increase the training speed. The initial BCM threshold for each neuron is determined by averaging its responses across all time steps and stimuli in the pre-trained network.

## 4.3 Quantification of familiarity effects (Fig 2, S1 Fig, S2 Fig, S3 Fig)

The familiarity suppression index is calculated as follows:

$$\text{suppression index} = \frac{r_i^{after} - r_i^{before}}{r_i^{after} + r_i^{before}}, \tag{8}$$

where $r_i^{after}$, $r_i^{after}$ denote the steady-state response of neuron $i$ after and before familiarity training. The neuron response was min-max normalized before computing the familiarity suppression index.

The sparsity index of a neuron tuning curve is introduced in [43], which is expressed as:

$$S = \frac{n-1}{n}\left(1 - \frac{(\sum_j r_j/n)^2}{\sum_j (r_j^2/n)}\right) \tag{9}$$

where $r_j$ indicate the neuronal response to $j^{th}$ stimulus. This index measures the tail length of the neuronal tuning curve; a high value indicates the tuning curve has a long tail, with only a few neurons having non-zero firing rates [44].

## 4.4 Derivation of locally linear manifold transform in the recurrent circuit (Fig 3)

Locally linear manifold transform is determined by dynamical modes derived from the Jacobian matrix of the linearized system. To obtain the Jacobian matrix and the dynamical modes, we first rewrite the network dynamic (Eq 1&2) into the matrix form:

$$\tilde{\boldsymbol{T}}\frac{\partial \boldsymbol{r}}{\partial t} = -\boldsymbol{r} + \sigma(\boldsymbol{W}\boldsymbol{r} + \boldsymbol{\alpha}(\theta)), \tag{10}$$

where $\boldsymbol{r} = [\boldsymbol{r_e}; \boldsymbol{r_i}]$, $\tilde{T}_{ij} = \delta_{ij}\tau_i$, and $W = [W_{ee}, W_{ei}; W_{ie}, \mathbf{0}]$ (We use "," to separate column, ";" to separate row). $\boldsymbol{\alpha}(\theta)$ denotes the sparse code input of $\theta$. To perform linearization on Eq 10, we expand the network around the fixed point $\boldsymbol{r}(\theta)$ using Taylor series:

$$\dot{\boldsymbol{r}(\theta) + \delta\boldsymbol{r}} = F(\boldsymbol{r}(\theta)) + F'(\boldsymbol{r}(\theta))\delta\boldsymbol{r} + \frac{1}{2}\delta\boldsymbol{r}^T F''(\boldsymbol{r}(\theta))\delta\boldsymbol{r} + \cdots$$

Since $\boldsymbol{r}(\theta)$ is the fixed points, $F(\boldsymbol{r}(\theta)) = 0$, and the second order term vanished assume small $\delta\boldsymbol{r}$, we can obtain the linearized dynamic of the network around $\boldsymbol{r}(\theta)$ by substituting back Eq 10:

$$\dot{\delta\boldsymbol{r}} \approx (\Lambda'\boldsymbol{W} - \boldsymbol{T}^{-1})\delta\boldsymbol{r}, \tag{11}$$

where $\Lambda'_{ij} = \delta_{ij}\tau_i^{-1}\sigma'(\boldsymbol{W}\boldsymbol{r}(\theta)) + \boldsymbol{\alpha}(\theta))$. This matrix quantifies the sensitivity of each neuron's firing rate to small changes in its membrane potential.

Eq 11 defines a simple linear system that approximates the dynamic of the original system near the fixed point $\boldsymbol{r}(\theta)$. We define Jacobian $J(\theta) = \Lambda'\boldsymbol{W} - \boldsymbol{T}^{-1}$. Note that the dependency of the Jacobian on the stimulus variable will be omitted in the notation for simplicity. Dynamical modes of the linear system can be found by diagonalizing the Jacobian using eigendecomposition. After diagonalizing $J$, the dynamic can be written as:

$$\dot{\delta\boldsymbol{r}} = \sum_i^N \lambda_i \boldsymbol{v}_i^R (\boldsymbol{v}_i^L)^T \delta\boldsymbol{r},$$

where $\boldsymbol{v}_i^R$, $\boldsymbol{v}_i^L$ denotes $i^{th}$ right/left eigenvector, and $\lambda_i$ refers to the corresponding eigenvalue. Due to the orthogonality between left and right eigenvectors $\boldsymbol{v}_i^L \cdot \boldsymbol{v}_j^R = 0$, by projecting $\delta\boldsymbol{r}$ onto the left eigenvector, we efficiently 'diagonalize' the dynamics, such that it is decoupled into multiple independent processes:

$$(\boldsymbol{v}_i^L)^T \dot{\delta\boldsymbol{r}} = \lambda_i (\boldsymbol{v}_i^L)^T \delta\boldsymbol{r} \tag{12}$$

We define the dynamical mode as the left eigenvectors $\boldsymbol{v}^L$, and the dynamical time scale as $\tau = -1/Re(\lambda)$. Eq 12 shows that each dynamical mode describes a specific deviation pattern of population activity from the fixed point. The evolution of each pattern is independent and exponential, with the speed being controlled by the corresponding time scale $\tau$. A negative $\tau$ entails escaping from the fixed point, while a positive $\tau$ entails decaying back to the fixed point.

The tuning derivative is computed by differentiating the steady-state response $\boldsymbol{r}(\theta) = \sigma(\boldsymbol{W}\boldsymbol{r}(\theta) + \boldsymbol{\alpha}(\theta))$ with respect to $\theta$. After plugging in the eigendecomposition of

the Jacobian, we obtain Eq 6

$$\begin{aligned}\frac{\partial \boldsymbol{r}(\theta)}{\partial \theta}\Big|_{\theta=\theta_i} &= -J^{-1}(\theta_i)\,\Lambda'\,\boldsymbol{\alpha}'(\theta_i)\\ &= \sum_j^N -\frac{1}{\lambda_j}\boldsymbol{v}_j^R(\boldsymbol{v}_j^L)^T\Lambda'\boldsymbol{\alpha}'(\theta_i)\\ &= \sum_j^N \tau_j\boldsymbol{v}_j^R(\boldsymbol{v}_j^L)^T\Lambda'\boldsymbol{\alpha}'(\theta_i),\end{aligned}$$

Again applying the orthogonality of the left and right eigenvectors, projecting the tuning derivative on the dynamical mode obtains:

$$(\boldsymbol{v}_j^L)^T\,\frac{\partial \boldsymbol{r}(\theta)}{\partial \theta}\Big|_{\theta=\theta_i} = \tau_j(\boldsymbol{v}_j^L)^T\Lambda'\boldsymbol{\alpha}'(\theta_i).$$

Projecting onto multiple dynamical modes results in Eq 7

$$L^T\,\boldsymbol{r}'(\theta_i) = \boldsymbol{T}\tilde{L}^T\boldsymbol{\alpha}'(\theta_i),$$

where the matrix L has columns $\boldsymbol{v}_i^L$, while $\tilde{L}$ has columns $\Lambda'\boldsymbol{v}_i^L$. The diagonal matrix $\boldsymbol{T}$ contains the time scales $\tau_i$ of each mode on its diagonal.

## 4.5 Familiarity training for the familiarity association experiment

For the familiarity association experiment, we randomly selected five 32x32 images from the CIFAR-100 dataset to serve as the target images. The images were converted to the grey scale before adding the noise. To corrupt the target image $l$ by noise level $n\%$, we randomly chose $n\%$ pixels and sampled the pixel value from $\varepsilon \sim \mathcal{U}(0,1)$ to replace the original one. This results in a distribution of stimuli $\boldsymbol{x}_{n,l} \sim p(\boldsymbol{x}\,|\,n,l)$. We sampled 3 noise levels (10%, 30%, 50%) and 10 stimuli for each noise level and target image; hence, there are $5 \times 3 \times 10$ target-noise pairs in total. The feedforward input was obtained by convolution with the same sparse coding filter bank as in 4.2.

The network architecture was configured with dimensions $N_r = N_c = 8$ to match the image size. The network was trained for 5 epoches. During each training epoch, the noisy image was first presented to the network for 150 ms, then the clear image for 150 ms, through simulating the Euler method with $dt = 1$ ms. To ensure the specificity of training on the presented image, the network state was reset to $\boldsymbol{r^e} = \boldsymbol{0}; \boldsymbol{r^i} = \boldsymbol{0}$ prior to the presentation of each image. During training, we multiply the feedforward input by 30 to accelerate the training process. The initial BCM threshold for each neuron was established by calculating the mean of its responses across all time steps and stimuli in the pre-trained network.

## 4.6 Fixed points analysis in the familiarity association experiment (Fig 4, S3 Fig)

The fixed point was found by forward simulating the network until convergence in the probe test period. The trajectory was visualized by reducing the dimension to 3 via PCA. The fixed point manifold was visualized by embedding the manifold in 2D via Multidimensional Scaling (MDS) that preserves the pairwise Euclidean distances between fixed points of interest.

The relative distance is computed as:

$$d(\boldsymbol{x}_{n,l}, \boldsymbol{x}_{0,l}) = \frac{\|\boldsymbol{r}^{\boldsymbol{e}}(\boldsymbol{x}_{n,l}) - \boldsymbol{r}^{\boldsymbol{e}}(\boldsymbol{x}_{0,l})\|}{\sqrt{\sum_{m\neq l} \|\boldsymbol{r}^{\boldsymbol{e}}(\boldsymbol{x}_{n,l}) - \boldsymbol{r}^{\boldsymbol{e}}(\boldsymbol{x}_{0,m})\|^2 \,/\, (N-1)}} \tag{13}$$

where $N$ is the number of target images, and $\boldsymbol{r}^{\boldsymbol{e}}(\boldsymbol{x}_{n,l})$ refers to representation of $\boldsymbol{x}_{n,l}$ which was sampled from the conditional distribution $p(\boldsymbol{x} \,|\, n, l)$ (where $n$ denotes the noise level and $l$ denotes the target image). We sampled 10 stimuli for each $n$ and $l$, and the relative distance was averaged over those 10 stimuli.

To measure the directional alignment, we computed the cosine similarity given by

$$s(\boldsymbol{x}_{n,l}, \boldsymbol{x}_{0,l}) = \frac{\langle \boldsymbol{r}^{\boldsymbol{e}}(\boldsymbol{x}_{n,l})\rangle'^T \langle \boldsymbol{r}^{\boldsymbol{e}}(\boldsymbol{x}_{0,l})\rangle'}{\|\langle \boldsymbol{r}^{\boldsymbol{e}}(\boldsymbol{x}_{n,l})\rangle'\| \|\langle \boldsymbol{r}^{\boldsymbol{e}}(\boldsymbol{x}_{0,l})\rangle'\|} \tag{14}$$

where $\langle \boldsymbol{r}^{\boldsymbol{e}}(\boldsymbol{x}_{n,l})\rangle$ refers to the averaged representation of 10 stimuli sampled from $p(\boldsymbol{x} \,|\, n, l)$. $\langle \boldsymbol{r}^{e}(\boldsymbol{x}_{n,l})\rangle' = \langle \boldsymbol{r}^{\boldsymbol{e}}(\boldsymbol{x}_{n,l})\rangle - \langle \boldsymbol{r}^{\boldsymbol{e}}(\boldsymbol{x}_{0,l})\rangle_l$, which denotes the population vector of excitatory neurons that starts from the mean of target images.

## 4.7 Analysis of the dynamical modes in the familiarity-trained network (Fig 5, S4 Fig)

To analyze dynamical modes and their associated decaying time constants in our neural circuit model, we forward-simulated the system until it converges to find the fixed point and used the numerical method to obtain the eigenvalues and eigenvectors of the Jacobian matrix. We define $L$ to be the matrix of eigenmode, whose columns are top M slow left eigenvectors ($\boldsymbol{v}_i^L$) of the $J(\theta)$ (M ranges from 1 to 200). Similarly, $\tilde{L}$ is column concatenation of $\tilde{\boldsymbol{v}}_i^L = \Lambda' \boldsymbol{v}_i^L$. The decaying time constant of mode $i$ is: $\tau_i = -1/Re(\lambda_i)$, where $\lambda_i$ is the $i^{th}$ eigenvalue of $J$. Eigenvectors and eigenvalues were pooled across visual concepts, noise levels, and training epochs. The mean and s.e.m. of $\tau_i$ and the mean of $\tilde{\boldsymbol{v}}_i^L$ were plotted.

Due to the complexity and unobservability of the feature space of the natural images, we used the distance between the nearby fixed points to approximate the derivative. Specifically, when the linearization was performed at the representation of stimulus $\boldsymbol{x}_{n,l}$ sampled from the conditional distribution $p(\boldsymbol{x} \,|\, n, l)$ (where $n$ denotes the noise level and $l$ denotes the visual concept), the noise distance is:

$$\mathcal{D}_N^{\psi} = \frac{1}{h} \|\psi(\boldsymbol{x}_{n,l+h}) - \psi(\boldsymbol{x}_{n,l})\|, \tag{15}$$

where $h$ denotes the increment of noise level to the adjacent noise level, with a unit of 10%. In total 10 noise pairs $(\boldsymbol{x}_{n,l}, \boldsymbol{x}_{n,l+h})$ were sampled from each conditional distribution (the linearization was always performed at $\boldsymbol{x}_{n,l}$), and the distance was averaged over noise pairs. The concept distance is

$$\mathcal{D}_C^{\psi} = \sqrt{\frac{1}{N-1} \sum_{i\neq n} \|\psi(\boldsymbol{x}_{n,l}) - \psi(\boldsymbol{x}_{i,l})\|^2}, \tag{16}$$

where $N$ is the number of visual concepts. Similarly, ten concept sets $(\boldsymbol{x}_{n,l},\, n = 1, \ldots, N)$ were sampled from each conditional distribution (the linearization was always performed at $\boldsymbol{x}_{n,l}$), and the distance was averaged over concept sets. The normalized distance is computed as the ratio of noise and concept distance $\mathcal{D}_{N/C}^{\psi} = \mathcal{D}_N^{\psi}/\mathcal{D}_C^{\psi}$, and the ratio was computed before averaging over noise pairs and concept sets. $\psi$ can be input $\boldsymbol{\alpha}$, representation $\boldsymbol{r}^{\boldsymbol{e}}$, input projection $\tilde{L}^T\boldsymbol{\alpha}$ and output projection $L^T\boldsymbol{r}$, where $L$ and $\tilde{L}$ are derived from the Jacobian of $\boldsymbol{x}_{n,l}$.

In Fig S4D-E, the relative difference between $\mathcal{D}_{N/C}$ in $\psi_1$ and $\psi_2$ were computed using the formula: $(\mathcal{D}^{\psi_1}_{N/C} - \mathcal{D}^{\psi_2}_{N/C}) \,/\, \mathcal{D}^{\psi_2}_{N/C}$. To compute the correlation in Fig S4F-G, $\mathcal{D}^{\tilde{L}^T\boldsymbol{\alpha}}_{N/C}$ and $\mathcal{D}^{L^T\boldsymbol{r}}_{N/C}$ were concatenated across 3 epochs (pre-training, epoch1, epoch5), and the Pearson correlation coefficient is then computed between the concatenated distances and the corresponding epochs.

# Supporting information

**S1 Text Supplementray figures S1-S4**. Additional results of numerical experiments

# Acknowledgments

This work was supported by NSF CISE RI 1816568 and NIH R01 EY030226-01A1. We thank Wenhao Zhang and Carl Olson for their insightful discussions.

# Data availability statement

All code written in support of this publication is publicly available at https://github.com/leelabcnbc/familiarity-circuit-V1.git. Simulation input files and generated data are available from https://doi.org/10.5281/zenodo.14729086

# References

1. Meyer T, Walker C, Cho RY, Olson CR. Image familiarization sharpens response dynamics of neurons in inferotemporal cortex. Nature neuroscience. 2014;17(10):1388–1394.

2. Fahy F, Riches I, Brown M. Neuronal activity related to visual recognition memory: long-term memory and the encoding of recency and familiarity information in the primate anterior and medial inferior temporal and rhinal cortex. Experimental Brain Research. 1993;96:457–472.

3. Xiang JZ, Brown M. Differential neuronal encoding of novelty, familiarity and recency in regions of the anterior temporal lobe. Neuropharmacology. 1998;37(4-5):657–676.

4. Mruczek RE, Sheinberg DL. Context familiarity enhances target processing by inferior temporal cortex neurons. Journal of Neuroscience. 2007;27(32):8533–8545.

5. Sobotka S, Ringo JL. Investigation of long term recognition and association memory in unit responses from inferotemporal cortex. Experimental Brain Research. 1993;96:28–38.

6. Freedman DJ, Riesenhuber M, Poggio T, Miller EK. Experience-dependent sharpening of visual shape selectivity in inferior temporal cortex. Cerebral Cortex. 2006;16(11):1631–1644.

7. Woloszyn L, Sheinberg DL. Effects of long-term visual experience on responses of distinct classes of single units in inferior temporal cortex. Neuron. 2012;74(1):193–205.

8. Huang G, Ramachandran S, Lee TS, Olson CR. Neural correlate of visual familiarity in macaque area V2. Journal of Neuroscience. 2018;38(42):8967–8975.

9. Lim S, McKee JL, Woloszyn L, Amit Y, Freedman DJ, Sheinberg DL, et al. Inferring learning rules from distributions of firing rates in cortical neurons. Nature neuroscience. 2015;18(12):1804–1810.

10. Jutras MJ, Buffalo EA. Recognition memory signals in the macaque hippocampus. Proceedings of the National Academy of Sciences. 2010;107(1):401–406.

11. Aitken K, Campagnola L, Garrett ME, Olsen SR, Mihalas S. Simple synaptic modulations implement diverse novelty computations. Cell Reports. 2024;43(5).

12. Deitch D, Rubin A, Ziv Y. Representational drift in the mouse visual cortex. Current biology. 2021;31(19):4327–4339.

13. Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. science. 2000;290(5500):2323–2326.

14. Saul LK, Roweis ST. Think globally, fit locally: unsupervised learning of low dimensional manifolds. Journal of Machine Learning Research. 2003;.

15. Chen Y, Paiton D, Olshausen B. The sparse manifold transform. Advances in neural information processing systems. 2018;31.

16. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: International conference on machine learning. PMLR; 2020. p. 1597–1607.

17. Cooke SF, Komorowski RW, Kaplan ES, Gavornik JP, Bear MF. Visual recognition memory, manifested as long-term habituation, requires synaptic plasticity in V1. Nature neuroscience. 2015;18(2):262–271.

18. Kaplan ES, Cooke SF, Komorowski RW, Chubykin AA, Thomazeau A, Khibnik LA, et al. Contrasting roles for parvalbumin-expressing inhibitory neurons in two forms of adult visual cortical plasticity. elife. 2016;5:e11450.

19. Cooper LN, Bear MF. The BCM theory of synapse modification at 30: interaction of theory with experiment. Nature Reviews Neuroscience. 2012;13(11):798–810.

20. Law CC, Cooper LN. Formation of receptive fields in realistic visual environments according to the Bienenstock, Cooper, and Munro (BCM) theory. Proceedings of the National Academy of Sciences. 1994;91(16):7797–7801.

21. Olshausen BA, Field DJ. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature. 1996;381(6583):607–609.

22. Szlam A, Kavukcuoglu K, LeCun Y. Convolutional matching pursuit and dictionary training. arXiv preprint arXiv:10100422. 2010;.

23. Carandini M, Heeger DJ. Normalization as a canonical neural computation. Nature reviews neuroscience. 2012;13(1):51–62.

24. Rubin DB, Van Hooser SD, Miller KD. The stabilized supralinear network: a unifying circuit motif underlying multi-input integration in sensory cortex. Neuron. 2015;85(2):402–417.

25. Toyoizumi T, Kaneko M, Stryker MP, Miller KD. Modeling the dynamic interaction of Hebbian and homeostatic plasticity. Neuron. 2014;84(2):497–510.

26. Liou Jy, Smith EH, Bateman LM, Bruce SL, McKhann GM, Goodman RR, et al. A model for focal seizure onset, propagation, evolution, and progression. Elife. 2020;9:e50927.

27. Hennequin G, Ahmadian Y, Rubin DB, Lengyel M, Miller KD. The dynamical regime of sensory cortex: stable dynamics around a single stimulus-tuned attractor account for patterns of noise variability. Neuron. 2018;98(4):846–860.

28. Mosheiff N, Ermentrout B, Huang C. Chaotic dynamics in spatially distributed neuronal networks generate population-wide shared variability. PLOS Computational Biology. 2023;19(1):e1010843.

29. Chung S, Abbott LF. Neural population geometry: An approach for understanding biological and artificial neural networks. Current opinion in neurobiology. 2021;70:137–144.

30. Kriegeskorte N, Wei XX. Neural tuning and representational geometry. Nature Reviews Neuroscience. 2021;22(11):703–718.

31. Chen Y, Yun Z, Ma Y, Olshausen B, LeCun Y. Minimalistic unsupervised learning with the sparse manifold transform. arXiv preprint arXiv:220915261. 2022;.

32. Wiskott L, Sejnowski TJ. Slow feature analysis: Unsupervised learning of invariances. Neural computation. 2002;14(4):715–770.

33. Oord Avd, Li Y, Vinyals O. Representation learning with contrastive predictive coding. arXiv preprint arXiv:180703748. 2018;.

34. Zbontar J, Jing L, Misra I, LeCun Y, Deny S. Barlow twins: Self-supervised learning via redundancy reduction. In: International conference on machine learning. PMLR; 2021. p. 12310–12320.

35. Bardes A, Ponce J, LeCun Y. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. arXiv preprint arXiv:210504906. 2021;.

36. Chadwick A, Khan AG, Poort J, Blot A, Hofer SB, Mrsic-Flogel TD, et al. Learning shapes cortical dynamics to enhance integration of relevant sensory input. Neuron. 2023;111(1):106–120.

37. Mante V, Sussillo D, Shenoy KV, Newsome WT. Context-dependent computation by recurrent dynamics in prefrontal cortex. nature. 2013;503(7474):78–84.

38. Fung CA, Wong KM, Wu S. A moving bump in a continuous manifold: a comprehensive study of the tracking dynamics of continuous attractor neural networks. Neural Computation. 2010;22(3):752–792.

39. Palm G. Neural associative memories and sparse coding. Neural Networks. 2013;37:165–171.

40. Wu Z, Rockwell H, Zhang Y, Tang S, Lee TS. Complexity and diversity in sparse code priors improve receptive field characterization of Macaque V1 neurons. PLoS computational biology. 2021;17(10):e1009528.

41. King PD, Zylberberg J, DeWeese MR. Inhibitory interneurons decorrelate excitatory cells to drive sparse code formation in a spiking model of V1. Journal of Neuroscience. 2013;33(13):5475–5485.

42. Cavanaugh JR, Bair W, Movshon JA. Nature and interaction of signals from the receptive field center and surround in macaque V1 neurons. Journal of neurophysiology. 2002;88(5):2530–2546.

43. Vinje WE, Gallant JL. Sparse coding and decorrelation in primary visual cortex during natural vision. Science. 2000;287(5456):1273–1276.

44. Rolls ET, Tovee MJ. Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. Journal of neurophysiology. 1995;73(2):713–726.