Biological
Cybernetics

ORIGINAL ARTICLE

Efficient multi-scale representation of visual objects using a biologically plausible spike-latency code and winner-take-all inhibition

Melani Sanchez-Garcia[1] · Tushar Chauhan[2,3] · Benoit R. Cottereau[3,4] · Michael Beyeler[1,5]

Abstract

Deep neural networks have surpassed human performance in key visual challenges such as object recognition, but require a large amount of energy, computation, and memory. In contrast, spiking neural networks (SNNs) have the potential to improve both the efficiency and biological plausibility of object recognition systems. Here we present a SNN model that uses spike-latency coding and winner-take-all inhibition (WTA-I) to efficiently represent visual stimuli using multi-scale parallel processing. Mimicking neuronal response properties in early visual cortex, images were preprocessed with three different spatial frequency (SF) channels, before they were fed to a layer of spiking neurons whose synaptic weights were updated using spike-timing-dependent-plasticity. We investigate how the quality of the represented objects changes under different SF bands and WTA-I schemes. We demonstrate that a network of 200 spiking neurons tuned to three SFs can efficiently represent objects with as little as 15 spikes per neuron. Studying how core object recognition may be implemented using biologically plausible learning rules in SNNs may not only further our understanding of the brain, but also lead to novel and efficient artificial vision systems.

Keywords Spiking neural networks · Spike-timing-dependent-plasticity · Multi-scale processing · Spike-latency code · Winner-take-all inhibition

1 Introduction

Deep convolutional neural network (DCNNs) have been extremely successful in a wide range of computer vision applications, rivaling or exceeding human benchmark performance in key visual challenges such as object and face recognition (He et al. 2015; Sun et al. 2015; Jiang et al. 2022) or scene categorization (Stivaktakis et al. 2019). However, state-of-the-art DCNNs require too much energy, computation, and memory to be deployed on most computing devices

and embedded systems (Goel et al. 2020). In contrast, the brain is masterful at representing real-world objects with a cascade of reflexive, largely feedforward computations (DiCarlo et al. 2012) that rapidly unfold over time (Ales et al. 2013; Cichy et al. 2016) and rely on an extremely sparse, efficient neural code (for a recent review see Beyeler et al, 2019). For example, in macaques, faces are processed in localized patches along the Superior Temporal Sulcus (STS), where cells detect distinct constellations of face parts (e.g., eyes, noses, mouths), and whole faces can be recognized from

Melani Sanchez-Garcia and Tushar Chauhan are co-first authors.
Benoit R. Cottereau and Michael Beyeler are co-last authors.

✉ Melani Sanchez-Garcia
  mesangar@ucsb.edu

  Tushar Chauhan
  tchauhan@mit.edu

  Benoit R. Cottereau
  benoit.cottereau@cnrs.fr

  Michael Beyeler
  mbeyeler@ucsb.edu

[1] Department of Computer Science, University of California, Santa Barbara, CA, USA

[2] The Picower Institute for Learning and Memory, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Boston, MA, USA

[3] CerCo CNRS UMR5549, Université de Toulouse III-Paul Sabatier, Toulouse, France

[4] IPAL, CNRS IRL 2955, Singapore, Singapore

[5] Department of Psychological & Brain Sciences, University of California, Santa Barbara, CA, USA

a linear combination of neural responses within these face patches (Chang and Tsao 2017; Majaj et al. 2015).

In recent years, spiking neural networks (SNNs) have emerged as a promising approach to improving the efficiency and biological plausibility of neural networks such as DCNNs, due to their potential for low power consumption, fast inference, event-driven processing, and asynchronous operation (Gerstner and Kistler 2002; Stuijt et al. 2021). To facilitate learning in such networks, new learning algorithms based on varying degrees of biological plausibility have also been developed recently. For instance, spike-timing-dependent plasticity (STDP) is an unsupervised learning rule that is observed in biological systems (Bi and Poo 1998; Caporale and Dan 2008; Falez et al. 2019) and that can be used to extract the most notable spike patterns (Feldman 2012; Brzosko et al. 2019; Hao et al. 2020) by adjusting the efficacy of synaptic connections based on the relative timing of presynaptic and postsynaptic spikes. Studying how object recognition may be implemented using biologically plausible learning rules in SNNs may not only further our understanding of the brain, but also lead to the development of energy efficient systems, implementable on neuromorphic hardware.

Here we present a SNN model that uses spike-latency coding (Chauhan et al. 2018, 2021) and winner-take-all inhibition (WTA-I) (Maass 2000) to efficiently represent visual stimuli using multi-scale parallel processing. Part of this work (Sanchez-Garcia et al. 2022) was previously presented at the CVPR'22 NeuroVision workshop[1]. Given an input image, stimuli were preprocessed with parallel spatial frequency (SF) channels mimicking the sensitivity of neurons in early visual cortex (De Valois et al. 1982a). The resulting combination of the SF channels was then fed to a layer of spiking neurons whose synaptic weights were updated using STDP (Gütig et al. 2003b). We show that STDP can learn efficient object representations from the MNIST (LeCun 1998), FASHION-MNIST (Xiao et al. 2017), CIFAR10 (Krizhevsky and Hinton 2009), and ORL (Samaria and Harter 1994) datasets. In addition, we investigate how the quality of the represented objects changes under different SF bands and WTA-I schemes. Remarkably, our network is able to represent objects with as little as 200 neurons and 15 spikes per neuron.

The rest of the paper is organized as follows: Sect. 2 briefly introduces some of the most recent related works. Section 3 explains the main framework and the model equations. Next, we report the results of a computational study in which we explored the quality of the represented objects and the spar-

sity trade-off for the different networks schemes (see Sect. 4). Finally, a brief Discussion summarizes the main results and gives some perspectives in Sect. 5.

2 Related work

Significant efforts have been expended in recent years to demonstrate the efficacy of SNNs with STDP in object recognition applications (Vigneron and Martinet 2020; Liu et al. 2020; Fu and Dong 2021). Previous studies have used STDP to extract visual features of low or intermediate complexity from images and without supervision. Yu et al. (2013) proposed a novel SNN with a supervised learning rule and temporal coding scheme to generate temporal spike patterns, which could be used to classify a subset of handwritten digits found in the MNIST database. Liu and Yue (2016) combined Gabor filter banks with rank-order coding and STDP to push the MNIST classification rate to 82%. Beyeler et al. (2013) achieved 92% on MNIST using a Calcium-based STDP learning rule, which was later surpassed by Diehl and Cook (2015) using standard STDP and lateral inhibition. Masquelier and Thorpe (2007) used the STDP rule in an asynchronous feedforward SNN that mimics the ventral visual pathway and showed the emergence of selectivity to intermediate-complexity visual features when the network was presented with natural images.

More recent articles designed a deep SNN, comprising several convolutional and pooling layers trainable with either standard STDP (Kheradpisheh et al. 2018) or reward-based STDP (Mozafari et al. 2019). Bing et al. (2019) used a supervised reward-modulated STDP learning rule to train two SNN-based sub-controllers on obstacle avoidance tasks. Zhou and Li (2022) proposed a SNN with STDP learning and first-spike coding to extract object features from Gabor filters and even-driven convolutions.

Studying how object recognition may be implemented using biologically plausible learning rules in SNNs may not only further our understanding of the brain, but also lead to new efficient artificial vision systems.

Numerous studies in visual neuroscience demonstrated the existence of multiple channels, or multiple receptive field (RF) sizes, in early visual cortex and their implications for the processing of the spatial frequency (SF) content of images during object recognition (Kauffmann et al. 2014; Ginsburg 1986; Field 1987; Tolhurst et al. 1992; Hughes et al. 1996). Because RFs of neuronal populations in the visual pathway vary in size, the responses of different subsets of neurons would constitute a neural representation at some particular scale, allowing us to represent visual scenes as a combination of SF channels (Campbell 1973).

Selectivity for SF is one of the fundamental and most thoroughly studied properties of visual neurons (Henriksson
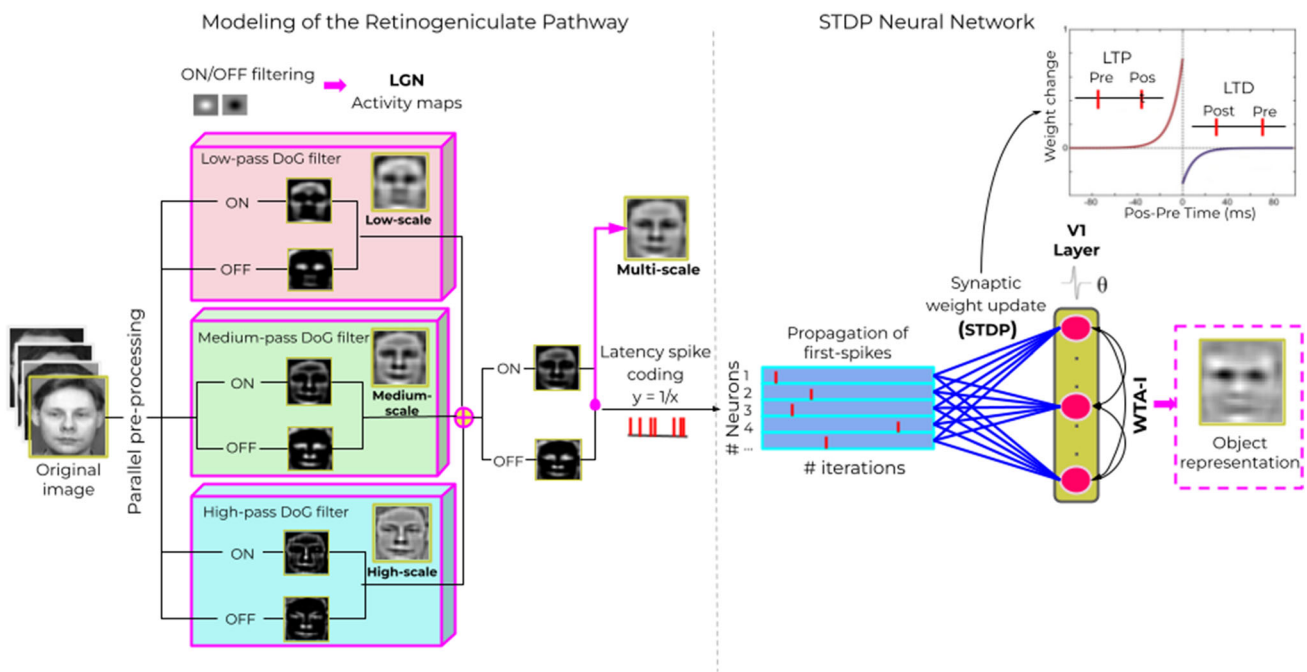

Fig. 1 Multi-scale network, illustrated using images from the ORL dataset (Samaria and Harter 1994). Images were convolved with ON and OFF center/surround kernels to simulate LGN responses. To simulate the multiple SF channels of the visual system, we used a pre-processing scheme where LGN maps were obtained from spatial filters at low, medium and high spatial frequencies (further illustrated in Fig. 2). The three LGN responses were added, converted to spike latencies, and fed to one layer spiking neural network (SNN) of firing-rate neurons with plas-

tic synapses implementing spike-timing-dependent-plasticity (STDP) and winner-take-all inhibition (WTA-I). The propagated LGN spikes contributed to an increase in the membrane potential of V1 neurons until one of the V1 membrane potentials reached threshold, resulting in a postsynaptic spike and inhibition of all other V1 neurons until the next iteration. Objects were reconstructed by taking a linear combination of spiking activity across the V1 population

et al. 2008; Shapley and Lennie 1985; De Valois et al. 1982b). The primary visual system processes low-level and high-level stimulus properties using inputs from the retina via the lateral geniculate nucleus (LGN). In the earliest stages of the visual pathway, the processing of different stimulus attributes occurs in a parallel fashion. This means that images are filtered by parallel, SF-selective channels (Enroth-Cugell and Robson 1966), which may converge in V1 (Nassi and Callaway 2009). The visual information from the LGN passes through V1 and multiple strategies might be used to transfer parallel input into multiple output streams.

3 Methods

3.1 Network architecture

The network architecture of our model is shown in Fig. 1. Inspired by Chauhan et al. (2018), our network consisted of an input layer corresponding to a simplified model of the LGN, followed by a layer of spiking neurons whose synaptic weights were updated using STDP. The LGN layer consisted of simulated firing-rate neurons with center-surround RFs, implemented using DoG filters which simulate the compu-

tations performed by the retinal ganglion cells and the LGN (Enroth-Cugell and Robson (1966); Derrington and Lennie (1982); further illustrated in Fig. 2). Based on Chauhan et al. (2018), the RF sizes were chosen to reflect the size of representative LGN center-surround cells. It is well known that the SFs of these neurons can differ by about a factor of 3. Some cells are therefore tuned to high SFs, while others are tuned to low SFs (Derrington et al. 1979). Here, we used the three following sizes of center-surround RFs: 0.375°/0.75° for low SF, 0.25°/0.5° for medium SF and 0.125°/0.25° for high SF (see Solomon et al, 2002). These values corresponded to the widths of the Gaussian used for the DoG filter.

The SF curves for the LGN images were thus fitted using a DoG model defined as follows:

$$\text{LGN}_{\text{ON}} = \frac{1}{2\pi\sigma_{\text{center}}^2} e^{-\frac{\hat{x}^2}{2\sigma_{\text{center}}^2}} - \frac{1}{2\pi\sigma_{\text{surround}}^2} e^{-\frac{\hat{x}^2}{2\sigma_{\text{surround}}^2}} \quad (1)$$

$$\text{LGN}_{\text{OFF}} = -\frac{1}{2\pi\sigma_{\text{center}}^2} e^{-\frac{\hat{x}^2}{2\sigma_{\text{center}}^2}} + \frac{1}{2\pi\sigma_{\text{surround}}^2} e^{-\frac{\hat{x}^2}{2\sigma_{\text{surround}}^2}} \quad (2)$$

where LGN_{ON} and LGN_{OFF} were the LGN maps, $\hat{x}$ was the input image, and σ_{center} and σ_{surround} were the center-
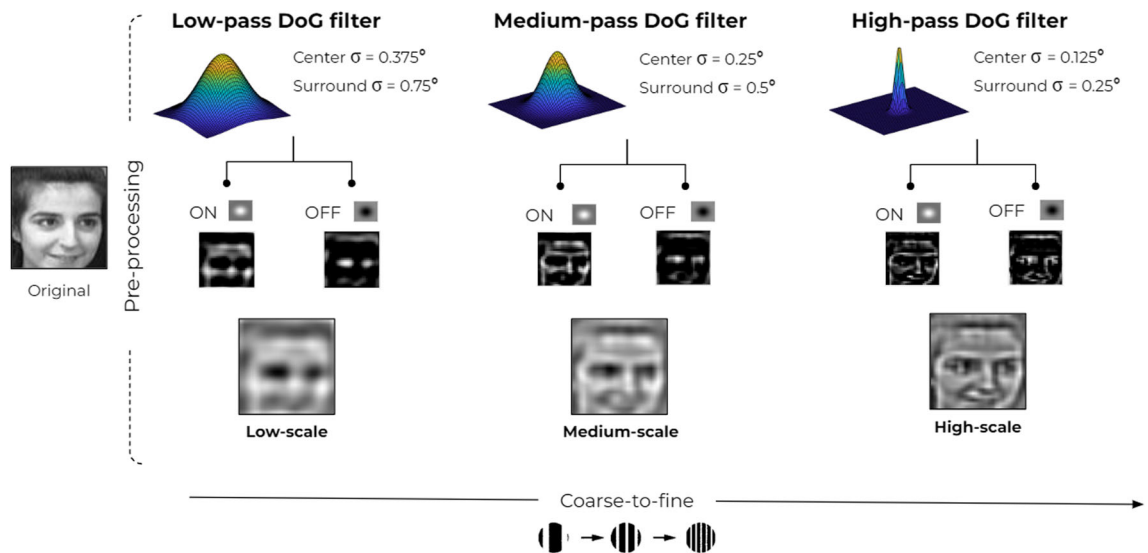
Fig. 2 LGN preprocessing. To simulate the computations performed by the retinal ganglion cells and the LGN, the images were convolved with ON and OFF center-surround kernels (Chauhan et al. 2018). Specifically, we chose three sizes based on an earlier study (Chauhan et al. 2018): 0.375°/0.75° for low SF, 0.25°/0.5° for medium SF and 0.125°/0.25° for high SF (Solomon et al. 2002). The resulting images processed with these filters correspond to low-scale, medium-scale and high-scale LGN maps, respectively

surround standard deviations used for the SF scales. The outputs of these filters, respectively, led to low-scale, medium-scale and high-scale images which were subsequently added together and converted into spikes using an intensity-to-latency conversion (Delorme and Thorpe 2001). These spikes were transmitted to the V1 layer, which was composed of integrate-and-fire neurons fully connected to the outputs of the LGN (see Fig. 1). In addition to this multi-scale architecture, we also developed an approach based on lateral scales, which is detailed in Appendix 7.

3.2 Neuron model

The membrane potential $E_n(t)$ of the n-th V1 neuron at time t within the iteration was given by:

$$E_n(t) = \begin{cases} \sum_{m \in LGN} w_{mn} \cdot H(t - t_m), & t < \min_t \left\{ t \mid \max_{n \in V1} E_n(t) \geq \theta \right\} \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

where t_m was the spike time of the m-th LGN neuron, H was the Heaviside or unit step function, and θ was the threshold of the V1 neurons (assumed to be a constant shared by the entire population). The expression $\min\{t \mid \max E_n(t) \geq \theta\}$ denoted the timing of the first spike in the V1 layer. Membrane potentials were calculated up to this point in time, after which a WTA-I scheme (Maass 2000) was triggered and all membrane potentials were reset to zero. In this scheme, the most frequently firing neuron exerted the strongest inhibition on its competitors and thereby stopped them from firing until the end of the iteration.

3.3 Spike-latency code

Following Chauhan et al. (2018), we converted the LGN activity maps to first-spike relative latencies using a simple inverse operation: $y = 1/x$, where x was the LGN input and y was the assigned spike-time latency. Any monotonically decreasing function would lead to equivalent results (i.e., where the most active units fire first, while units with lower activity fire later or not at all) (see (Masquelier and Thorpe 2007)). In this way, we ensured that the most active units fired first, while units with lower activity fired later or not at all.

3.4 Spike-timing-dependent-plasticity

The weights of plastic synapses connecting LGN and V1 were updated using multiplicative STDP, which is an unsupervised learning rule that modifies synaptic strength, w, as a function of the relative timing of pre- and postsynaptic spikes, Δt (Gütig et al. 2003b). LTP ($\Delta t > 0$) and LTD ($\Delta t \leq 0$) were driven by their respective learning rates α^+ and α^-, leading to a weight change (Δw):

$$\Delta w = \begin{cases} -\alpha^- \cdot w^{\mu^-} \cdot K(\Delta t, \tau_-), \Delta t \leq 0 \\ \alpha^+ \cdot (1 - w)^{\mu^+} \cdot K(\Delta t, \tau_+), \Delta t > 0, \end{cases} \quad (4)$$

where $\alpha^+ = 5 \times 10^{-3}$ and $\alpha^- = 3.75 \times 10^{-3}$, $K(\Delta t, \tau) = e^{-|\Delta t|/\tau}$ was a temporal windowing filter, and $\mu^+ = 0.65$ and $\mu^- = 0.05$ were constants $\in [0, 1]$ that defined the non-linearity of the LTP and LTD process, respectively. STDP has
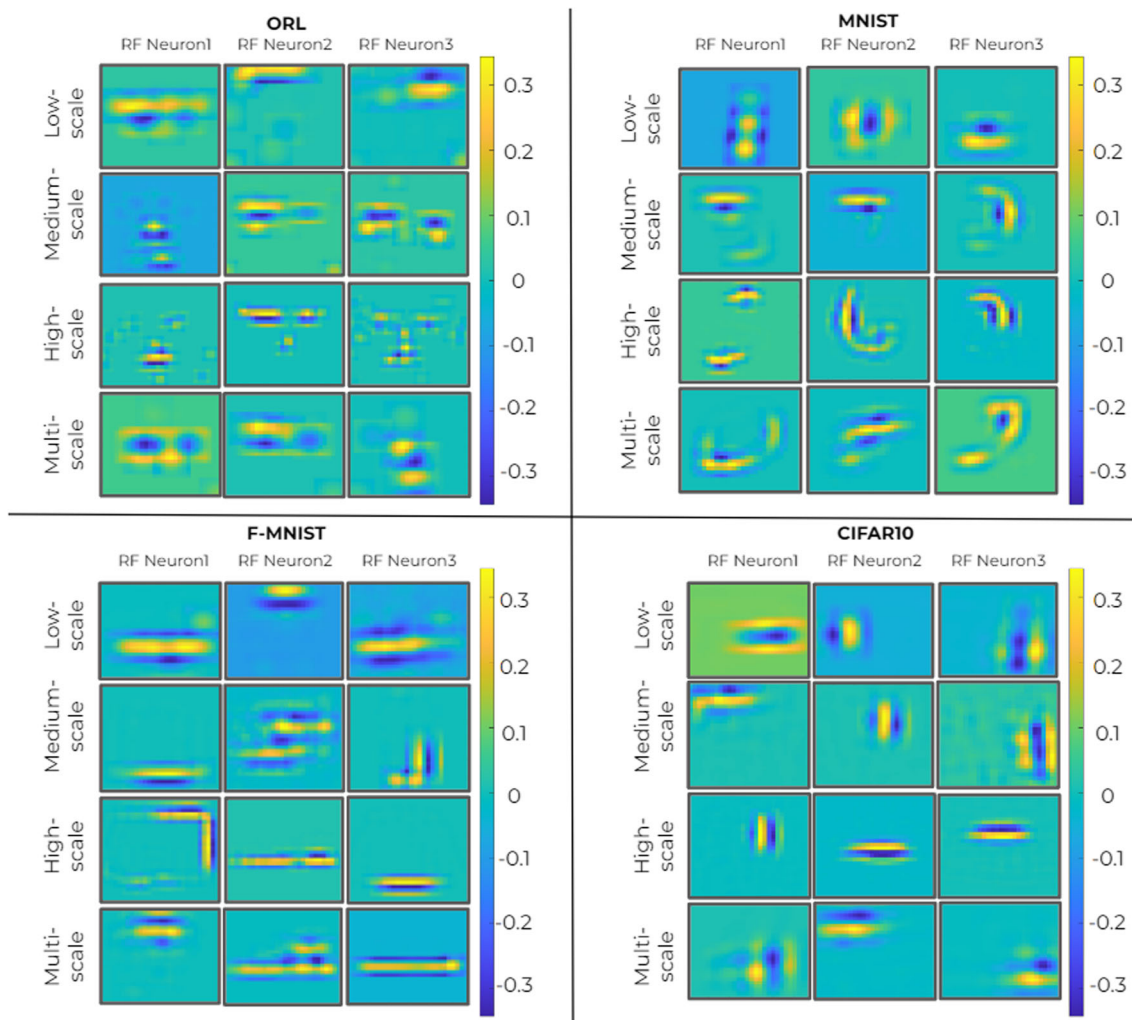
Fig. 3 Example RFs of three representative neurons (columns in each panel) of the simulated population for low-scale, medium-scale, high-scale and multi-scale networks (rows). With STDP, neurons progressively learned features corresponding to prototypical patterns that were both salient and frequent

the effect of concentrating high synaptic weights on afferents that systematically fire early, thereby decreasing postsynaptic spike latencies for these connections.

In this implementation, computation speed greatly increased by making the windowing filter K infinitely wide, which is equivalent to assuming $\tau_\pm \to \infty$ or $K = 1$ (Gütig et al. 2003a). A ratio $\alpha^+/\alpha^- = 4/3$ was chosen based on previous experiments that demonstrated network stability (Masquelier and Thorpe 2007). Also, Chauhan et al. (2018) showed that the results were robust to variations of this ratio. The threshold of the V1 neurons was fixed through trial and error at $\theta = 20$. This value was unmodified for all experiments.

Initial weight values were sampled from a random uniform distribution between 0 and 1. After each iteration, the synaptic weights for the first V1 neuron to fire were updated using STDP (Eq. 4), and the membrane potentials of all the

other neurons in the V1 population were reset to zero. The STDP rule was active only during the training phase.

3.5 Winner-take-all inhibition

We used a hard WTA-I scheme such that, if any V1 neuron fired during a certain iteration, it simultaneously prevented other neurons from firing until the next sample (Maass 2000). This scheme computes a function WTA-I_n: $\mathbb{R}^n \to \{0, 1\}^n$ whose output $\langle y_1, \ldots, y_n \rangle = \text{WTA-}I_n(x_1, \ldots, x_n)$ satisfied:

$$y_i = \begin{cases} 1, & \text{if } x_i > x_j \quad \text{for all} \quad j \neq i \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

For a given set of n different inputs $x_1, \ldots, x_n$, a hard WTA-I scheme would thus yield a single output y_i with value 1 (corresponding to the neuron that received the largest
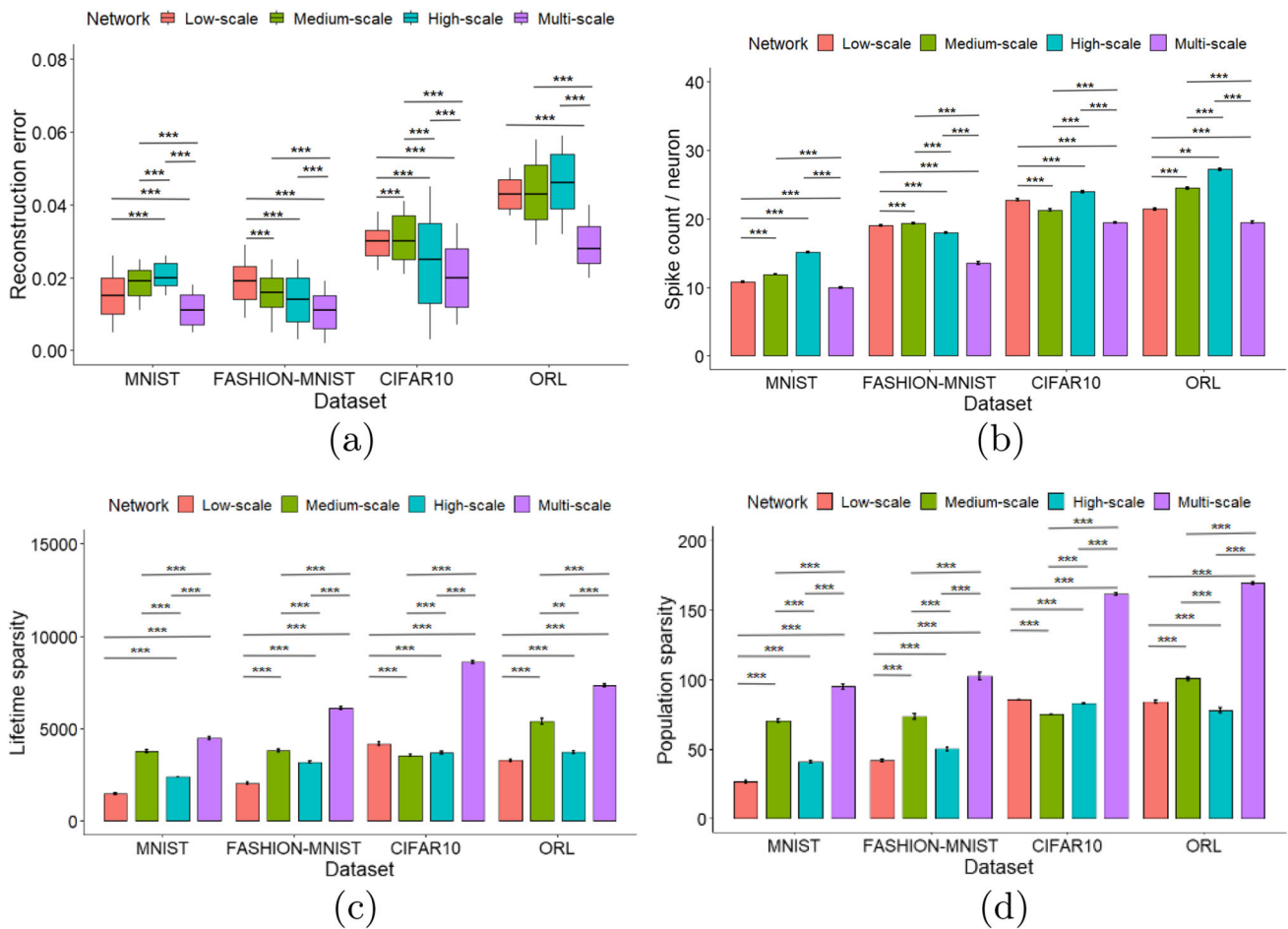
Fig. 4 Multi-scale network. (**a**) Reconstruction error (MSE) of test set. (**b**) Spike count per neuron: number of spikes fired by an active neuron. (**c**) Lifetime sparsity: active stimuli during the lifetime of a neuron. (**d**) Population sparsity: neurons active at any point in time. Mean responses and standard deviation grouped by type of network (low-scale, medium-scale, high-scale and multi-scale). Error bars have been averaged across neurons for lifetime sparsity and averaged across images for population sparsity. $*** = p < .001; ** = p < .01; * = p < .05; ns = p > .05$. All t tests paired samples, two-tailed

input x_i), whereas all other neurons would be silent. Sanchez-Garcia et al. (2022) showed that a hard WTA-I scheme was essential for enforcing competition among neurons, which led to sparser object representations and lower reconstruction error compared to softer WTA-I schemes.

3.6 Stimulus reconstruction

The activity map ξ_j of the i-th V1 neuron was estimated as follows:

$$\xi_j \approx \sum_{j \in LGN} w_{ij} \psi_j, \tag{6}$$

where ψ_j was the RF of the j-th LGN afferent, and w_{ij} was the weight of the synapse connecting the j-th afferent to the i-th V1 neuron.

Stimuli k were then linearly reconstructed from the V1 population activity:

$$OR_k = \sum_{j \in V1} r_{kj} \xi_j, \tag{7}$$

where r_{kj} was the response of the j-th V1 neuron to the k-th image and ξ_j was its activity map. Reconstruction error for an image k was calculated as the pixel-wise mean square error (MSE) between the LGN (LGN_k) and the V1 activity maps OR_k.

Table 1 Global results for type of networks

Dataset	Network	RE (MSE)	RE (SSIM)	SC	LS	PS
ORL	Low-scale	4.54e−2±8.40e−3	0.70±0.19	21.4±1.08	*3282.5±1525.4*	84.0±11.97
	Medium-scale	4.54e−2±8.43e−3	0.70±0.12	24.5±1.48	5404.5±704.3	100.8±18.32
	High-scale	4.30e−2±3.99e−3	0.73±0.14	27.2±1.72	3732.7±559.5	*78.0±9.24*
	Multi-scale	*2.91e−2±6.07e−3*	*0.76±0.09*	*19.5±1.78*	7320±847.0	169.4±11.7
MNIST	Low-scale	1.81e−2±4.53e−3	0.83±0.07	10.8±0.77	*1483.7±695.0*	*26.6±11.9*
	Medium-scale	2.07e−2±1.29e−2	0.79±0.06	11.9±0.78	3788.0±103.2	70.6±5.14
	High-scale	1.51e−2±4.84e−3	0.86±0.04	15.1±0.81	2386.7±295.9	40.9±3.31
	Multi-scale	*1.16e−2±3.77e−3*	*0.89±0.05*	*10.0±0.79*	4500.1±782.7	95.0±19.9
FASHION-MNIST	Low-scale	1.49e−2±5.99e−3	0.62±0.10	19.0±1.07	*2037.5±735.4*	*42.1±19.8*
	Medium-scale	1.37e−2±6.87e−3	0.62±0.05	19.3±1.39	3822.9±493.8	73.5±11.12
	High-scale	1.90e−2±6.17e−3	0.60±0.13	18.0±1.38	3201.8±591.0	50.1±10.39
	Multi-scale	*9.34e−3±5.15e−3*	*0.69±0.12*	*13.6±1.76*	6105.0±907.9	102.5±25.2
CIFAR10	Low-scale	3.10e−2±6.66e−3	0.73±0.16	22.8±1.72	4179.9±795.7	85.5±3.29
	Medium-scale	*2.13e−2±3.43e−3*	*0.82±0.09*	21.3±1.40	*3542.4±693.9*	*75.0±11.18*
	High-scale	3.07e−2±6.61e−3	0.75±0.15	24.0±1.75	3692.9±1006.7	83.2±8.32
	Multi-scale	2.15e−2±8.22e−3	0.80±0.13	*19.5±1.06*	8599.5±830.7	161.5±10.8

Comparison of mean responses and standard deviation grouped by type of network and dataset. Best performing networks for each dataset and error metric are highlighted in italics

3.7 Sparsity

We computed a sparsity metric for the population activity in the network schemes according to the definition of sparsity by Vinje and Gallant (2000). On average, we measured how many neurons were activated by any given stimulus (population sparsity) and for all active neurons, how many stimuli any given neuron responded to (lifetime sparsity), as can be seen in Eq. 8).

$$\text{sparsity} = \left(1 - \frac{1}{N} \frac{(\sum_{n=1} r_i)^2}{\sum_{n=1} r_i^2}\right) \Big/ \left(1 - \frac{1}{N}\right), \quad (8)$$

For population sparsity, r_i was the response of the i-th neuron to a particular stimulus, and N was the number of model neurons. For lifetime sparsity, r_i was the response of a neuron to the i-th stimulus, and N was the number of stimuli. Population sparsity was averaged across stimuli, and lifetime sparsity was averaged across neurons (Beyeler et al. 2016). We also calculated the average number of spikes per stimulus.

3.8 Dataset

To demonstrate the generality of our approach, we assessed the ability of our SNN network to represent visual stimuli from the MNIST (LeCun 1998), FASHION-MNIST (Xiao et al. 2017), CIFAR10 (Krizhevsky and Hinton 2009) and ORL (Samaria and Harter 1994) datasets. MNIST is a dataset

of handwritten digits and consists of 60,000 training patterns and 10,000 test patterns. FASHION-MNIST is a dataset of Zalando article images consisting of a training set of 60,000 examples and a test set of 10,000 examples. Each example of both, MNIST and FASHION-MNIST, is a 28 × 28 grayscale image, associated with a label from 10 classes. The CIFAR10 database consists of 60,000 32 × 32 color images in 10 classes, with 6000 images per class. There are 50,000 training images and 10,000 test images. The ORL database of faces contains 400 images from 40 distinct subjects. The size of each image is 92 × 112 pixels, with 256 gray levels per pixel.

We enlarged images from the CIFAR10 and ORL database using data augmentation with different orientations of the original images to match the data size with MNIST and FASHION-MNIST datasets.

3.9 Statistical analysis

Data were analyzed using two-way ANOVA and post hoc test with Tukey's method to evaluate simultaneously the effect of the two grouping variables (Dataset and Networks/WTA-I schemes/V1 neurons) on the following response variables: reconstruction error (RE), spike count/neuron (SC), lifetime sparsity (LS), population sparsity (PS), and recognition time with $*** = p < .001$; $** = p < .01$; $* = p < .05$ and $ns = p \geq .05$. For the reconstruction error, we have used the mean squared error (MSE) which is the most widely used metric reference and the Structured Similarity

Fig. 5 Representative object representation (OR) examples using low-scale, medium-scale, high-scale and multi-scale networks (columns). The number below each image indicates the reconstruction error (MSE) for that particular image. The black frame highlights the image with the smallest error
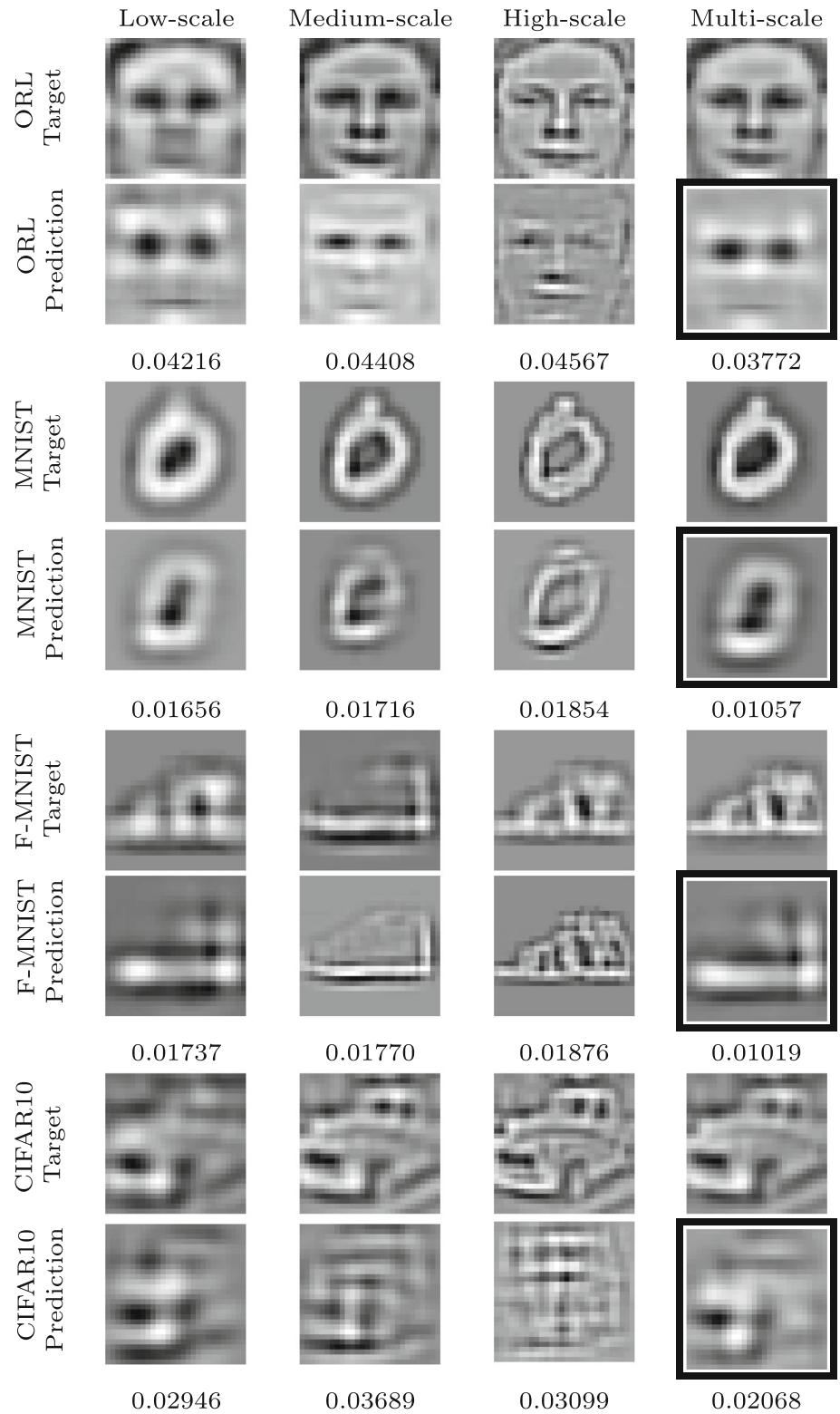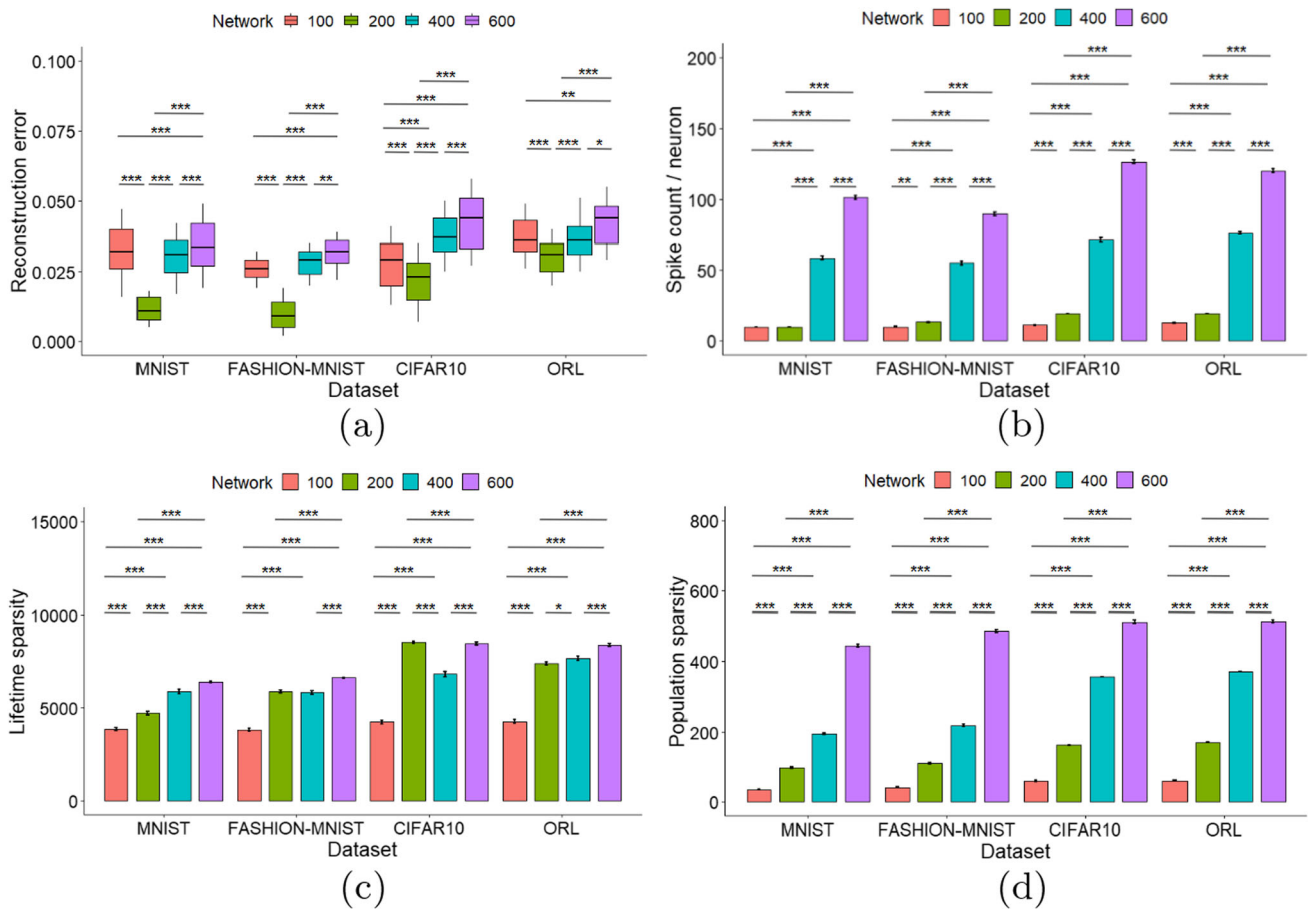
Fig. 6 V1 neurons. (a) Reconstruction error (MSE) of test set using different number of V1 neurons: 100, 200, 400 and 600. (b) Spike count per neuron: number of spikes fired by an active neuron. (c) Lifetime sparsity: active stimuli during the lifetime of a neuron. (d) Population sparsity: neurons active at any point in time. Mean responses and standard deviation grouped by type of network architecture (low-scale, medium-scale, high-scale and multi-scale). Error bars have been averaged across neurons for lifetime sparsity and averaged across images for population sparsity

Indexing Method (SSIM) which compare the structural and feature similarity measures between reconstructed and original images on the basis of perception.

4 Results

4.1 Object representation using multi-scale network

The performance using a single-scale (i.e., low-scale, medium-scale, or high-scale networks) and multi-scale network is summarized in Fig. 4. The results show the reconstruction error, lifetime sparsity, population sparsity and spike count per neuron (mean ± standard deviation) achieved on the test sets for all databases (see Table 1). The reconstruction error for the four networks (low-scale, medium-scale, high-scale and multi-scale) is shown in Fig. 4a. We found similarity between the reconstruction errors of the three single net-

works (low-, medium- and high-scale) for all datasets, with some slight discrepancy in the more complex CIFAR10 and ORL datasets. Interestingly, the use of multi-scale manages to further reduce the reconstruction error, being the same trend for all datasets. We also performed a test to determine if the mean difference between networks are statically significant using two-tailed test with a significant level $\alpha = 0.05$. The analysis of the average reconstruction error reveals a significant difference between networks (low-/multi-scale, medium-/multi-scale and high-/multi-scale). Examples of object representations for all datasets can be found in Fig. 5.

Figure 4b shows the number of spikes per neuron needed for object representation. The number of spikes needed to represent an object decreased with the Multi-scale scheme compared to low-, medium- and high-scale networks. On the other hand, we found that the CIFAR10 and ORL dataset, which we considered two of the most complex of the four

Table 2 Global results for V1 neurons

Dataset	V1 neurons	RE (MSE)	RE (SSIM)	SC	LS	PS
ORL	100	3.73e−2±7.09e−3	0.76±0.04	*12.4±1.7*	*4092.1±1058.3*	*58.7±14.9*
	200	*2.91*e−2±*6.07*e−*3*	*0.76±0.09*	19.5±1.78	7320±847.0	169.4±11.7
	400	3.90e−2±7.90e−3	0.73±0.02	75.5±8.8	7769.2±1151.9	370.5±9.0
	600	4.17e−2±7.70e−3	0.73±0.08	120.4±12.5	8288.6±828.0	513.7±45.3
MNIST	100	2.62e−2±8.61e−3	0.87±0.08	10.3±0.8	*3667.8±718.3*	*35.2±9.3*
	200	*1.16*e−2±*3.77*e−*3*	*0.89±0.05*	*10.0±0.79*	4500.1±782.7	95±19.9
	400	2.94e−2±7.72e−3	0.86±0.05	59.7±13.1	5719.3±1568.7	206.2±9.3
	600	3.47e−2±8.29e−3	0.82±0.09	101.8±13.4	6379.4±388.3	445.4±38.7
FASHION-MNIST	100	2.56e−2±4.38e−3	0.67±0.012	*9.9±1.5*	*3893.7±694.5*	*43.1±10.2*
	200	*9.34*e−3±*5.15*e−*3*	*0.69±0.12*	13.6±1.76	6105.0±907.9	102.5±25.2
	400	2.78e−2±4.48e−3	0.68±0.16	53.4±11.9	6251.3±979.1	216.7±31.8
	600	3.09e−2±5.45e−3	0.66±0.11	87.8±13.3	6585.8±279.6	486.5±37.1
CIFAR10	100	3.15e−2±9.37e−3	0.79±0.09	*11.1±1.4*	*4374.1±1173.1*	*61.2±18.3*
	200	*2.15*e−2±*8.22*e−*3*	*0.80±0.13*	19.5±1.06	8599.5±830.7	161.5±10.8
	400	3.71e−2±7.60e−3	0.79±0.08	70.3±12.3	6498.8±1005.8	354.0±29.2
	600	4.37e−2±1.01e−2	0.75±0.05	126.1±12.1	8404.8±876.3	498.3±41.8

Comparison of mean responses and standard deviation grouped by type of V1 neurons and dataset. Best performing in each cell for each dataset and error metric is highlighted in italics

Fig. 7 Object representation with Multi-scale network varying the number of V1 neurons: 100, 200, 400 and 600 neurons. The number below each image indicates the reconstruction error (MSE) for that particular image. The black frame highlights the image with the smallest error
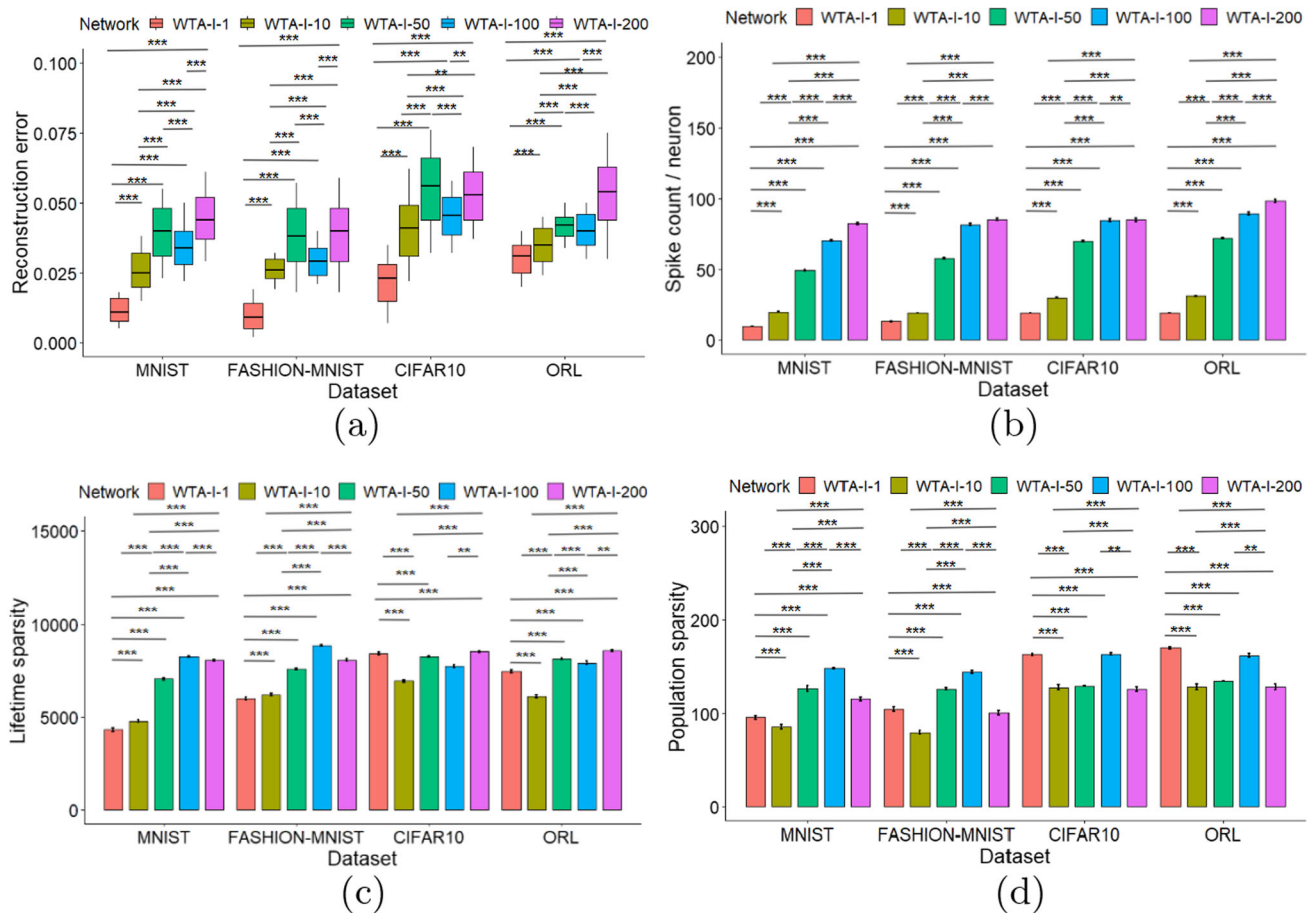
(a)

(b)

(c)

(d)
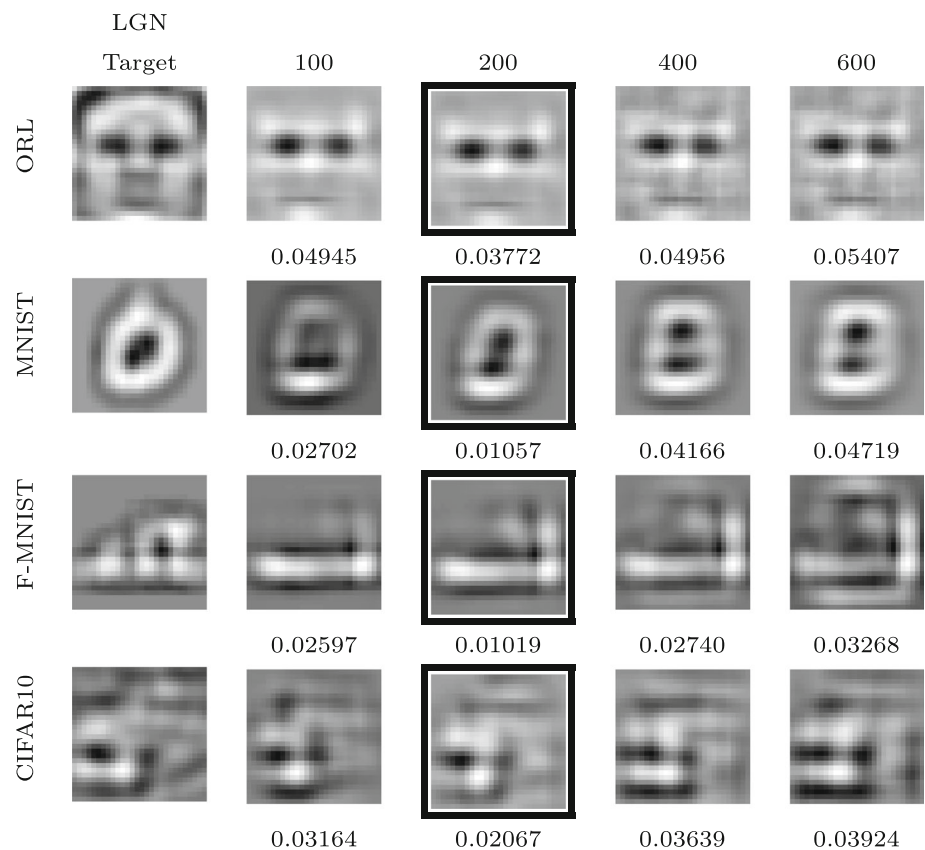
Fig. 8 WTA-I schemes. (**a**) Reconstruction error (MSE) in the test phase as a function of the number of spikes included in the STDP algorithm (WTA-I) for 200 V1 neurons. (**b**) Spike count per neuron: number of spikes fired by an active neuron. (**c**) Lifetime sparsity: active stimuli during the lifetime of a neuron. (**d**) Population sparsity: neurons active at any point in time. Mean responses and standard deviation grouped by the WTA-I schemes. Error bars have been averaged across neurons for lifetime sparsity and averaged across images for population sparsity

datasets, needed the highest number of spikes per neuron for all networks.

Figure 4c shows the number of distinct stimuli the neuron responds to during the lifetime of a neuron. The multi-scale network showed a higher number of active stimuli for all datasets compared to the single networks. Moreover, we found significant differences between the networks, being more significant for medium-/multi-scale and high-/multi-scale. The same trend was found for the population sparsity, where the multi-scale presented more active neurons than the low-, medium- and high-scale networks and significant differences were found between them (see Fig. 4d).

4.2 Object representation using multi-scale network with varying number of V1 neurons

Figure 6a shows the reconstruction error after training for the test set using different numbers of V1 neurons (see Table 2). We found that the reconstruction error went through a min-

imum (at roughly 200 V1 neurons) for all databases, which is consistent with the bias-variance dilemma (Beyeler et al. 2019). It seems that using a larger number of neurons with our multi-scale network leads to overfitting and a less sharp reconstruction, as can be seen in Fig. 7.

In addition, the number of neurons needed to represent an object increased with the number of V1 neurons, nearly tripling the spikes from 200 to 400 neurons and quintupling from 200 to 600 (Fig. 6c). Increasing the V1 population beyond 200 neurons did therefore not lead to any visible benefits in reconstruction error (Fig. 7). We therefore limited our V1 population to 200 neurons for all subsequent simulations and analyses.

### 4.3 Object representation using soft WTA-I schemes

We also tested object representation using various soft WTA-I schemes, where we varied the number of V1 neurons allowed to be active for each training image (see Fig. 8). Fig-

Table 3 Global results for WTA-I schemes

Dataset	WTA-I	RE (MSE)	RE (SSIM)	SC	LS	PS
ORL	WTA-I-1	*2.91*e−2±*6.07*e−*3*	*0.76±0.09*	19.5±1.8	7320.0±847.0	169.4±11.7
	WTA-I-10	3.06e−2±6.19e−3	0.76±0.15	*19.2±1.8*	7461.8±745.6	169.5±12.1
	WTA-I-50	3.48e−2±6.48e−3	0.72±0.12	31.7±1.8	*6254.8±716.5*	*124.7±27.3*
	WTA-I-100	3.90e−2±6.10e−3	0.70±0.09	72.2±4.2	8643.1±517.1	134.3±29.9
	WTA-I-200	4.16e−2±4.90e−3	0.70±0.12	90.1±11.6	8107.6±405.9	134.9±3.13
MNIST	WTA-I-1	*1.16*e−2±*3.77*e−*3*	*0.89±0.05*	10.0±0.8	4500.1±782.7	95.0±19.9
	WTA-I-10	1.18e−2±4.33e−3	0.88±0.09	*9.8±0.8*	*4319.9±773.7*	95.7±22.4
	WTA-I-50	2.67e−2±1.18e−2	0.86±0.10	20.1±1.3	4695.9±536.7	*87.5±25.4*
	WTA-I-100	3.41e−2±7.48e−3	0.84±0.08	49.9±5.5	7919.3±558.42	109.2±28.3
	WTA-I-200	3.93e−2±9.61e−3	0.84±0.15	70.1±5.4	7094.6±458.8	128.9±28.1
FASHION-MNIST	WTA-I-1	*9.34*e−3±*5.15*e−*3*	*0.69±0.12*	13.6±1.7	6105.0±907.9	102.5±25.2
	WTA-I-10	9.72e−3±5.31e−3	0.68±0.12	*13.2±1.7*	*5968.1±822.1*	104.6±23.7
	WTA-I-50	2.56e−2±3.82e−3	0.64±0.16	19.0±1.3	6337.5±693.2	*78.1±22.9*
	WTA-I-100	3.14e−2±5.82e−3	0.65±0.14	56.9±4.6	8020.3±455.9	101.2±25.3
	WTA-I-200	3.73e−2±1.10e−2	0.67±0.13	82.1±8.8	7530.2±321.6	131.8±11.1
CIFAR10	WTA-I-1	*2.15*e−2±*8.22*e−*3*	*0.80±0.13*	19.5±1.1	8599.5±830.7	161.5±10.8
	WTA-I-10	2.20e−2±8.13e−3	0.80±0.16	*19.3±1.1*	8401.2±928.3	162.2±10.3
	WTA-I-50	4.09e−2±7.14e−3	0.78±0.13	30.12±1.4	*6884.8±642.5*	*129.3±19.7*
	WTA-I-100	4.51e−2±8.14e−3	0.79±0.10	68.9±4.5	8500.2±625.1	136.2±22.1
	WTA-I-200	5.38e−2±1.24e−2	0.77±0.11	84.5±10.1	8269.7±408.53	130.0±5.37

Comparison of mean responses and standard deviation grouped by type of WTA-I schemes and dataset. Best performance in each cell for each dataset and error metric is highlighted in italics

Fig. 9 Object representation using different WTA-I schemes, where between 1 (harder WTA-I 1) and 200 (softer WTA-I 200) neurons were active for each training sample. The number below each image indicates the reconstruction error for that particular image. Target and prediction images were normalized in [0, 1]. The black frame highlights the image with the smallest error in each row
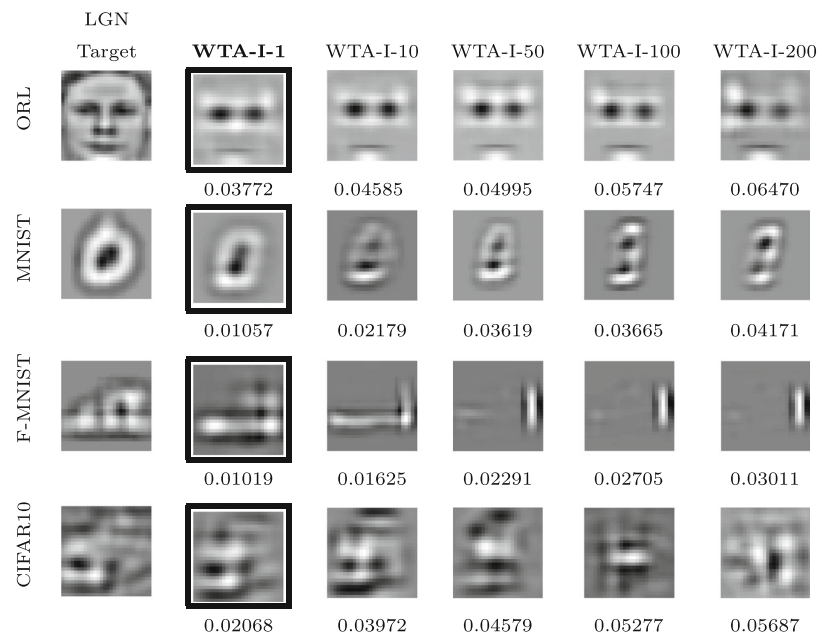
ure 8a shows the reconstruction error on the test set across the range of possible WTA-I schemes, ranging from hard (where only one neuron was active per image) to soft (where all 200 neurons were active).

We found that the softer the WTA-I scheme, the higher the reconstruction error (see Table 3). The reason for this became evident when we visualized the resulting object representations (Fig. 9). WTA-I schemes where at most 10 neurons were allowed to be active were instrumental in maintaining competition among neurons. In the absence of a strong WTA-I scheme, multiple neurons ended up learning similar visual features, which resulted in poor object reconstruction (right half of Fig. 9). Also, due to this overlap between neurons, the final feature set was quite limited.

We also found that both the active stimuli during the lifetime of a neuron and the active neurons increased with the number of V1 neurons allowed to be active during training (see Fig. 8c, d). Furthermore, the number of spikes needed to represent an object showed the same trend (Fig. 8b).

5 Discussion

In this work, we have proposed an SNN model that uses spike-latency coding and WTA-I to efficiently represent visual stimuli using multi-scale parallel processing. In particular, this paper developed an extension of earlier work (Chauhan et al. 2018, 2021; Sanchez-Garcia et al. 2022) to investigate how the quality of the represented objects changes under different schemes of the primary visual system processing with subsets of neurons tuned to different SF scales.

We found that the multi-scale network outperformed all three single-scale networks across all datasets (Fig. 4), sacrificing sparsity for a lower reconstruction error. However, it is interesting to note that the multi-scale network used the smallest average number of spikes per neuron (Fig. 4b) across all datasets, indicating that it favored a code where many neurons were weakly activated. In all cases, the learned receptive fields (Fig. 3) were in agreement with nonnegative sparse coding (NSC), which is an efficient population coding scheme based on dimensionality reduction and sparsity constraints that promotes sparse and parts-based population codes (Beyeler et al. 2019).

We also studied how the number of V1 neurons in the network affected reconstruction error and sparsity of the learned population code. In agreement with previous work on NSC (Beyeler et al. 2016, 2019), we found that the reconstruction error (on the test set) goes through a minimum as a function of network size (Fig. 6a). This minimum is though to indicate

the optimal model complexity according to the bias-variance dilemma, that is, the point at which the model's generalization error is minimized. Curiously, this "sweet spot" was found to be at roughly 200 V1 neurons for all tested datasets (Fig. 7). On the other hand, sparsity increased monotonically with network size (Fig. 6b–d), which is more in line with the traditional sparse coding literature (Olshausen and Field 1997).

We also implemented various soft WTA-I schemes to investigate how the quality of represented objects changed (Fig. 8). The WTA-I soft schemes consisted of 10, 50, 100, and 200 (i.e., all) neurons firing during a given iteration, while all other neurons were silent. We found that the softer the WTA-I scheme, the larger the reconstruction error (Fig. 8a) and the number of spikes needed to represent an object (Fig. 8b). The reason for this became clear when we visualized the resulting object representations (Fig. 9). In the absence of a strong WTA-I scheme, multiple neurons ended up learning similar visual features, thus resulting in poor object reconstructions (Fig. 9).

Although our network was able to efficiently represent images from various datasets, an important issue that we did not address in this paper is a comparison with other SNNs with other forms of STDP (e.g., with an additive instead of a multiplicative rule) and/or to SNNs trained with other learning scheme (e.g., SNNs trained with the surrogate gradient). In addition, a future extension of the model might focus on deeper architectures with parallel processing with multiple scales and more challenging visual stimuli.

6 Conclusion

In conclusion, we have shown that a network of spiking neurons tuned to different SFs can represent objects with as little as 15 spikes per neuron using spike-latency coding and WTA-I. WTA-I schemes were essential for enforcing competition among neurons, which led to sparser object representations and lower reconstruction errors. Studying how object recognition may be implemented using biologically plausible learning rules in SNNs may not only further our understanding of the brain, but also lead to new efficient artificial vision systems.
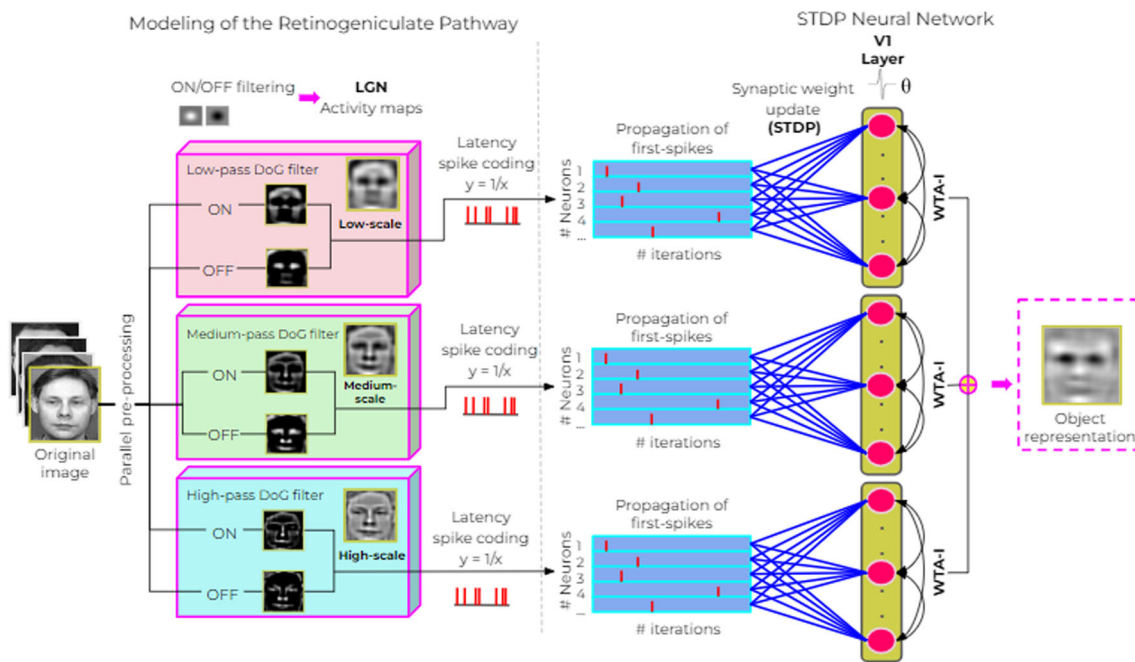
Fig. 10 Lateral-scale network. Images from the ORL dataset (Samaria and Harter 1994) were convolved with ON and OFF center-surround kernels to simulate responses in the LGN. We used three LGN subnetworks processed based on a particular SF: low-scale, medium-scale and high-scale (see Fig. 2). The three LGN responses were converted to spike latencies and fed to a SNN each, resulting in three lateral SNN with plastic synapses implementing STDP and WTA-I. The reconstructed images resulted of the three lateral networks were added at the end for the object reconstruction

7 Comparison between multi-scale and lateral-scale network architectures

We propose another network architecture called 'lateral-scale' that also uses parallel processing of multiple scales (see Fig. 10). In this case, the LGN preprocessing is the same as in the multi-scale network architecture, but now the three LGN responses were converted to spike latencies and fed to a SNN each, resulting in three lateral SNN with plastic synapses implementing STDP and WTA-I. The reconstructed images resulted of the three lateral sub-networks were added at the end of the training for the object representation.

As shown in Fig. 11a, the lateral-scale network results in a lower but very similar reconstruction error than the proposed multi-scale network. This may be because the lateral-scale scheme recognizes a few more details corresponding to fine details in the image (see Fig. 12). Lateral-scale was not significantly better than multi-scale if we refer to the representation of objects (see Fig. 12 but used significantly more spikes (Fig. 12b). The number of spikes required for reconstruction increases by approximately double spikes/neuron in some datasets. One drawback in lateral-scale network is that we are training three lateral sub-networks, that means three times more trainable weights.
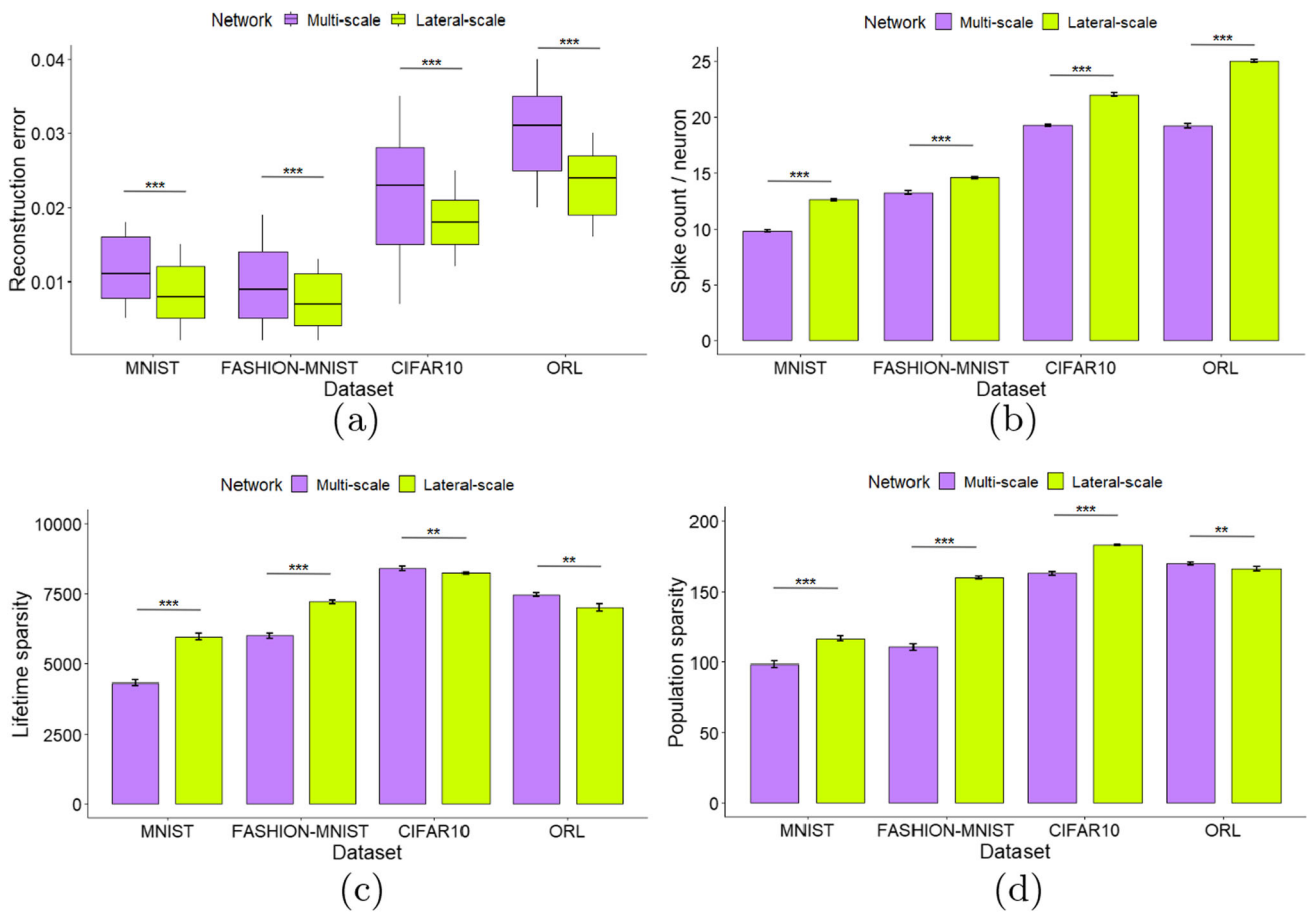
See Table 4.

Fig. 11 (**a**) Reconstruction error (MSE) of test set using multi-scale and lateral-scale networks. (**b**) Number of spikes per neuron needed for the object representation using multi-scale and Lateral-scale networks. (**c**) Lifetime sparsity: active stimuli during the lifetime of a neuron. (**d**) Pop-

ulation sparsity: neurons active at any point in time. $*** = p < .001$; $** = p < .01$; $* = p < .05$; $ns = p > .05$. All t tests paired samples, two-tailed

Fig. 12 Object representation for multi-scale and lateral-scale network architectures using 200 V1 neurons. Two examples of object representation (image A and image B) for multi-scale and lateral-scale architectures and for the four databases. The lateral-scale scheme recognizes some finer details in the image compared to multi-scale, where the image details are coarser. The number below each image indicates the reconstruction error (MSE) for that particular image. The black frame highlights the image with the smallest error
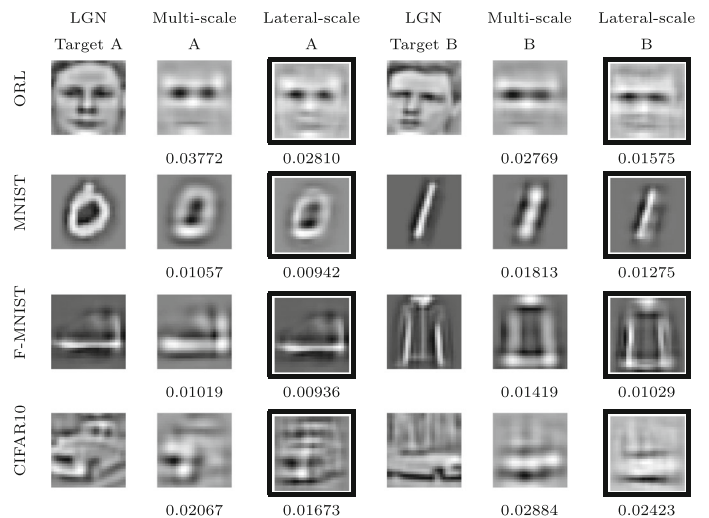
Table 4 Global results for multi- and lateral-scale

Dataset	Network	RE (MSE)	RE (SSIM)	SC	LS	PS
ORL	Multi-scale	2.91e−2±6.07e−3	0.76±0.23	*19.5±1.78*	7320±847.0	169.4±11.7
	Lateral-scale	*2.28e−2±4.84e−3*	*0.80±0.19*	25.1±1.5	*7233.0±1241.5*	*168.3±17.4*
MNIST	Multi-scale	*1.16e−2±3.77e−3*	*0.89±0.05*	*10.0±0.79*	*4500.1±782.7*	*95.0±19.9*
	Lateral-scale	8.27e−3±4.04e−3	0.66±0.18	12.5±1.4	5867.9±444.1	115.6±4.9
FASHION-MNIST	Multi-scale	9.34e−3±5.15e−3	0.59±0.20	*13.6±1.76*	*6105.0±907.9*	*102.5±25.2*
	Lateral-scale	*6.93e−3±3.53e−3*	*0.71±0.14*	14.4±1.2	7160.4±745.1	161.2±11.5
CIFAR10	Multi-scale	2.15e−2±8.22e−3	0.80±0.13	*19.5±1.06*	8599.5±830.7	*161.5±10.8*
	Lateral-scale	*1.87e−2±4.23e−3*	*0.87±0.02*	21.9±1.2	*8258.7±1130.5*	182.4±19.2

Comparison of mean responses and standard deviation grouped by type of Multi and Lateral-scale and dataset

Author Contributions TC and BRC conceived and designed the original study, which was subsequently extended by MSG and MB. TC wrote all the code and MSG ran all the simulations. MSG and MB analyzed and interpreted the results. MSG drafted the manuscript. All authors reviewed and approved the final version of the manuscript.

References

Ales JM, Appelbaum LG, Cottereau BR et al (2013) The time course of shape discrimination in the human brain. Neuroimage 67:77–88

Beyeler M, Dutt ND, Krichmar JL (2013) Categorization and decision-making in a neurobiologically plausible spiking network using a STDP-like learning rule. Neural Netw 48:109–24

Beyeler M, Dutt N, Krichmar JL (2016) 3D visual response properties of MSTd emerge from an efficient, sparse population code. J Neurosci 36(32):8399–8415

Beyeler M, Rounds E, Carlson K et al (2019) Neural correlates of sparse coding and dimensionality reduction. PLoS Comput Biol 15(6):e1006908

Bi GQ, Poo MM (1998) Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. J Neurosci 18(24):10,464-10,472

Bing Z, Baumann I, Jiang Z et al (2019) Supervised learning in snn via reward-modulated spike-timing-dependent plasticity for a target reaching vehicle. Front Neurorobot 13:18

Brzosko Z, Mierau SB, Paulsen O (2019) Neuromodulation of spike-timing-dependent plasticity: past, present, and future. Neuron 103(4):563–581

Campbell, Fergus W. The transmission of spatial information through the visual system. From Theoretical Physics to Biology. Karger Publishers, 1973. 374–384

Caporale N, Dan Y et al (2008) Spike timing-dependent plasticity: a hebbian learning rule. Annu Rev Neurosci 31(1):25–46

Chang L, Tsao DY (2017) The code for facial identity in the primate brain. Cell 169(6):1013–1028

Chauhan T, Masquelier T, Montlibert A et al (2018) Emergence of binocular disparity selectivity through Hebbian learning. J Neurosci 38(44):9563–9578

Chauhan T, Masquelier T, Cottereau BR (2021) Sub-optimality of the early visual system explained through biologically plausible plasticity. Front Neurosci 15:727448

Cichy RM, Pantazis D, Oliva A (2016) Similarity-based fusion of meg and fmri reveals spatio-temporal dynamics in human cortex during visual object recognition. Cereb Cortex 26(8):3563–3579

De Valois RL, Albrecht DG, Thorell LG (1982) Spatial frequency selectivity of cells in macaque visual cortex. Vis Res 22(5):545–559

De Valois RL, Albrecht DG, Thorell LG (1982) Spatial frequency selectivity of cells in macaque visual cortex. Vis Res 22(5):545–559

Delorme A, Thorpe SJ (2001) Face identification using one spike per neuron: resistance to image degradations. Neural Netw 14(6–7):795–803

Derrington A, Lennie P (1982) The influence of temporal frequency and adaptation level on receptive field organization of retinal ganglion cells in cat. J Physiol 333(1):343–366

Derrington A, Lennie P, Wright M (1979) The mechanism of peripherally evoked responses in retinal ganglion cells. J Physiol 289(1):299–310

DiCarlo J, Zoccolan D, Rust N (2012) How does the brain solve visual object recognition? Neuron 73(3):415–434

Diehl PU, Cook M (2015) Unsupervised learning of digit recognition using spike-timing-dependent plasticity. Front Comput Neurosci. https://doi.org/10.3389/fncom.2015.00099

Enroth-Cugell C, Robson JG (1966) The contrast sensitivity of retinal ganglion cells of the cat. J Physiol 187(3):517–552

Falez P, Tirilly P, Bilasco IM, et al (2019) Multi-layered spiking neural network with target timestamp threshold adaptation and stdp. In: 2019 international joint conference on neural networks (IJCNN). IEEE, pp 1–8

Feldman DE (2012) The spike-timing dependence of plasticity. Neuron 75(4):556–571

Field DJ (1987) Relations between the statistics of natural images and the response properties of cortical cells. Josa A 4(12):2379–2394

Fu Q, Dong H (2021) An ensemble unsupervised spiking neural network for objective recognition. Neurocomputing 419:47–58

Gerstner W, Kistler WM (2002) Spiking neuron models: single neurons, populations, plasticity. Cambridge University Press, Cambridge

Ginsburg AP (1986) Spatial filtering and visual form perception. Handbook of Perception and Human Performance, Vol 2 Cognitive Processes and Performance

Goel A, Tung C, Lu YH, et al (2020) A survey of methods for low-power deep learning and computer vision. In: 2020 IEEE 6th world forum on internet of things (WF-IoT). IEEE, pp 1–6

Gütig R, Aharonov R, Rotter S et al (2003) Learning input correlations through nonlinear temporally asymmetric hebbian plasticity. J Neurosci 23(9):3697–3714. https://doi.org/10.1523/JNEUROSCI.23-09-03697.2003

Gütig R, Aharonov R, Rotter S et al (2003) Learning input correlations through nonlinear temporally asymmetric hebbian plasticity. J Neurosci 23(9):3697–3714

Hao Y, Huang X, Dong M et al (2020) A biologically plausible supervised learning method for spiking neural networks using the symmetric stdp rule. Neural Netw 121:387–395

He K, Zhang X, Ren S, et al (2015) Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision, pp 1026–1034

Henriksson L, Nurminen L, Hyvärinen A et al (2008) Spatial frequency tuning in human retinotopic visual areas. J Vis 8(10):5–5

Hughes HC, Nozawa G, Kitterle F (1996) Global precedence, spatial frequency channels, and the statistics of natural images. J Cognit Neurosci 8(3):197–230

Jiang P, Ergu D, Liu F et al (2022) A review of yolo algorithm developments. Procedia Comput Sci 199:1066–1073

Kauffmann L, Ramanoël S, Peyrin C (2014) The neural bases of spatial frequency processing during scene perception. Front Integr Neurosci 8:37

Kheradpisheh SR, Ganjtabesh M, Thorpe SJ et al (2018) Stdp-based spiking deep convolutional neural networks for object recognition. Neural Netw 99:56–67

Krizhevsky A, Hinton G (2009) Learning multiple layers of features from tiny images. Tech. Rep. 0, University of Toronto, Toronto, Ontario

LeCun Y (1998) The MNIST database of handwritten digits. http://yann.lecun.com/exdb/mnist/

Liu D, Yue S (2016) Visual pattern recognition using unsupervised spike timing dependent plasticity learning. In: 2016 international joint conference on neural networks (IJCNN). IEEE, pp 285–292

Liu Q, Pan G, Ruan H et al (2020) Unsupervised aer object recognition based on multiscale spatio-temporal features and spiking neurons. IEEE Trans Neural Netw Learn Syst 31(12):5300–5311

Maass W (2000) On the computational power of winner-take-all. Neural Comput 12(11):2519–2535

Majaj NJ, Hong H, Solomon EA et al (2015) Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. J Neurosci 35(39):13,402-13,418

Masquelier T, Thorpe S (2007) Unsupervised learning of visual features through spike timing dependent plasticity. PLoS Comput Biol 3(2):e31

Mozafari M, Ganjtabesh M, Nowzari-Dalini A et al (2019) Bio-inspired digit recognition using reward-modulated spike-timing-dependent plasticity in deep convolutional networks. Pattern Recognit 94:87–95

Nassi JJ, Callaway EM (2009) Parallel processing strategies of the primate visual system. Nat Rev Neurosci 10(5):360–372

Olshausen BA, Field DJ (1997) Sparse coding with an overcomplete basis set: a strategy employed by V1? Vis Res 37(23):3311–3325. https://doi.org/10.1016/S0042-6989(97)00169-7

Samaria FS, Harter AC (1994) Parameterisation of a stochastic model for human face identification. In: Proceedings of 1994 IEEE workshop on applications of computer vision. IEEE, pp 138–142

Sanchez-Garcia M, Chauhan T, Cottereau BR, et al (2022) Efficient multi-scale representation of visual objects using a biologically plausible spike-latency code and winner-take-all inhibition. arXiv:2212.00081

Shapley R, Lennie P et al (1985) Spatial frequency analysis in the visual system. Annu Rev Neurosci 8(1):547–581

Solomon SG, White AJ, Martin PR (2002) Extraclassical receptive field properties of parvocellular, magnocellular, and koniocellular cells in the primate lateral geniculate nucleus. J Neurosci 22(1):338–349

Stivaktakis R, Tsagkatakis G, Tsakalides P (2019) Deep learning for multilabel land cover scene categorization using data augmentation. IEEE Geosci Remote Sens Lett 16(7):1031–1035

Stuijt J, Sifalakis M, Yousefzadeh A et al (2021) μbrain: an event-driven and fully synthesizable architecture for spiking neural networks. Front Neurosci 15:538

Sun Y, Liang D, Wang X, et al (2015) Deepid3: face recognition with very deep neural networks. arXiv:1502.00873

Tolhurst DJ, Tadmor Y, Chao T (1992) Amplitude spectra of natural images. Ophthalmic Physiol Opt 12(2):229–232

Vigneron A, Martinet J (2020) A critical survey of stdp in spiking neural networks for pattern recognition. In: 2020 international joint conference on neural networks (IJCNN). IEEE, pp 1–9

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

Xiao H, Rasul K, Vollgraf R (2017) Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. arXiv:1708.07747

Yu Q, Tang H, Tan KC et al (2013) Rapid feedforward computation by temporal encoding and learning with spiking neurons. IEEE Trans Neural Netw Learn Syst 24(10):1539–1552

Zhou Q, Li X (2022) A bio-inspired hierarchical spiking neural network with reward-modulated stdp learning rule for aer object recognition. IEEE Sens J 22(16):16,323-16,338