Patterns

Article

A machine learning framework to optimize optic nerve electrical stimulation for vision restoration

Highlights

- A framework to optimize optic nerve stimulation protocols has been implemented
- A physiologically constrained convolutional neural network models the visual system
- A genetic algorithm evolves optimal stimulations to match cortical activation
- Our protocols elicit the right stimulus classes in static and dynamic scenarios

Authors

Simone Romeni, Davide Zoccolan, Silvestro Micera

Correspondence

silvestro.micera@epfl.ch

In brief

We formulated a computational framework for the optimization of optic nerve stimulation patterns. We have implemented a model of the primate visual system, and an algorithm that allows the evolution of an optic nerve stimulation protocol that induces activation corresponding to natural visual stimuli in a given brain region and, consequently, a specified visual sensation. This could pave the way for novel machine-learning-based optimization of optic nerve stimulation to produce naturalistic sensations in blind patients.

Romeni et al., 2021, Patterns 2, 100286
July 9, 2021 © 2021 The Author(s).
https://doi.org/10.1016/j.patter.2021.100286

Article

A machine learning framework to optimize optic nerve electrical stimulation for vision restoration

Simone Romeni,[1] Davide Zoccolan,[2] and Silvestro Micera[1,3,4,*]

[1]Bertarelli Foundation Chair in Translational NeuroEngineering, Center for Neuroprosthetics and Institute of Bioengineering, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
[2]Visual Neuroscience Lab, International School for Advanced Studies (SISSA), Trieste, Italy
[3]The Biorobotics Institute and Department of Excellence in Robotics and AI, Scuola Superiore Sant'Anna, Pontedera, Italy
[4]Lead contact
*Correspondence: silvestro.micera@epfl.ch

https://doi.org/10.1016/j.patter.2021.100286

THE BIGGER PICTURE Electrical stimulation of the optic nerve can allow the restoration of lost visual functions in an effective and clinically exploitable way. To achieve this goal, it is crucial to develop a suitable approach to target selectively nerve fiber subpopulations that mediate different sensations but share similar locations in the nerve. In the present work, we use a simple computational model of the primate visual system to show that it is possible to optimize the stimulation at the level of the optic nerve to replicate a pattern of activity in a cortical region, producing, at the same time, reliable sensations. This result could produce nerve stimulation patterns that exploit the convergent nature of the visual system to "correct" the representation error introduced at the nerve level. In the long term, this would lead to eliciting naturalistic sensations from non-intuitive protocols that exploit machine learning to overcome the technological limits of nerve interfaces.

1 **2** 3 4 5 **Proof-of-Concept:** Data science output has been formulated, implemented, and tested for one domain/problem

SUMMARY

Optic nerve electrical stimulation is a promising technique to restore vision in blind subjects. Machine learning methods can be used to select effective stimulation protocols, but they require a model of the stimulated system to generate enough training data. Here, we use a convolutional neural network (CNN) as a model of the ventral visual stream. A genetic algorithm drives the activation of the units in a layer of the CNN representing a cortical region toward a desired pattern, by refining the activation imposed at a layer representing the optic nerve. To simulate the pattern of activation elicited by the sites of an electrode array, a simple point-source model was introduced and its optimization process was investigated for static and dynamic scenes. Psychophysical data confirm that our stimulation evolution framework produces results compatible with natural vision. Machine learning approaches could become a very powerful tool to optimize and personalize neuroprosthetic systems.

INTRODUCTION

The feasibility of neuroprosthetic devices aiming at sensory restoration via nerve electrical stimulation has been shown in the past years by both acute and chronic clinical tests for touch,[1–4] proprioception,[5] and vision[6,7] restoration. In particular, optic nerve stimulation can elicit controllable visual responses in subjects who exhibit outer retinal degeneration diseases and cannot be implanted with a retinal device.[6,7] Concurrently, optic nerve implants are less invasive than cortical implants and could allow control of the sensations elicited inside a wide portion of the visual field with a modest number of stimulation sources, in contrast to cortical and retinal implants.[6,8]

However, the effectiveness of a neuroprosthetic device depends on the possibility of identifying the best stimulation parameters, an optimization process living in a huge search space. The current approach has been to restrict the attention to "reasonable" stimulation protocols. In Raspopovic et al.,[1] a simple amplitude modulation was applied to a sensor readout, and the corresponding intensity profile was used as a waveform for the stimulating sites. A more complex alternative, which we will call the "biomimetic approach," aims at replicating as loyally as possible the natural patterns of activation of the targeted nerves.[9,10]

The state-of-the-art biomimetic approaches, nonetheless, limit themselves to extracting some very general, *a priori* features from the simulated afferent activity, using them to modulate stimulation. These methods operate thus an "open loop," as they assume that if an appropriate model of nerve activation is available, the activity in the nerve and thus in all the downstream regions in a sensory system will be replicated. The main limitation of this approach is given by the imperfect selectivity of the current stimulation technology, which allows only the concurrent stimulation of bundles of neighboring fibers, and introduces an unavoidable error in the replication of nerve activation patterns.

One idea to quantify and correct this systematic bias would be to "close the loop," checking the proposed stimulation patterns with their effect on neural activation at the level of the nerve or of more downstream regions. Nonetheless, comparing the artificially elicited neural activation with a natural target, one needs to be able to record peripheral nerve activity with single-fiber resolution, which is currently an open challenge. Because it is instead possible to achieve single-unit recordings in the cortex using a microelectrode array, we could monitor the effects of nerve stimulation and produce nerve-stimulation protocols that activate it toward a natural pattern of activation. This new idea of "closing the loop on the cortex" could also lead to "non-biomimetic," optimal stimulation protocols. It is indeed possible that imperfect selectivity of current neural interfaces may cause the existence of stimulation patterns that are suboptimal in the replication of nerve activation but produce better replication of more downstream (i.e., cortical) activation patterns, because of the high degree of convergence along the sensory stream. Moreover, as higher layers in the cortex encode more complex features, it could be easier to produce stimulation patterns that exploit low-dimensional stimulation protocols to reproduce complex sensations. The natural framework for the required multipolar optimization is provided by machine learning (ML) techniques, whose need for a large amount of training data can be mitigated, in this case, through the definition of an appropriate model of the sensory processing hierarchy and of the stimulation procedure. Exploiting the controllability of such a model, we can also compare our idea of closing the loop on the cortex to the experimentally infeasible "closing the loop on the nerve," which produces the stimulation protocols that replicate best nerve activation. This last proposed strategy can be thought of as the ideal, unattainable, best-case scenario for the biomimetic approach.

The primate visual system involved in object recognition (ventral visual pathway)[11–13] is constituted by a hierarchy of visual processing areas in the cortex that produce more and more complex representations of the external world. Before entering the cortex, visual information is transduced to electrical signals by approximately a hundred million (in humans) photoreceptors[14] and, after being processed by the retinal circuits, enters into the optic nerve through the axons of retinal ganglion cells (approximately a million in humans).[15] Since more than a hundred million (in humans) neurons are present in the primary visual cortex (V1),[16] the optic nerve strongly constrains the information that can be conveyed to the cortex acting as an anatomical bottleneck.

Recently, convolutional neural networks (CNNs)[17–19] have emerged as promising models of the visual system (and in particular the ventral stream),[11] allowing the simulation of the natural single-unit cortical response to ecological visual stimuli. In our study, we exploited a CNN variation proposed by Lindsey and colleagues,[20] which includes a single, very-low-dimensional hidden layer whose units are meant to explicitly represent optic nerve fibers (we will call it the "optic nerve" layer in the following). The downstream layers, instead, are meant to represent different "cortical" layers.[20]

We simulated the technological limitations of current neural interfaces by introducing a model of "imperfect-selectivity" control of activation. The stimulation is applied by a number of current sources—representing the active sites of a generic intraneural electrode—characterized by a location in the optic nerve section and by an intensity value. The optic nerve excitation pattern is obtained assuming that the nerve is a homogeneous isotropic medium. The corresponding activation pattern is obtained by passing the excitation values through a sigmoidal non-linearity compatible with the inputs expected by the optic nerve layer in our CNN model, which simulates the typical sigmoidal frequency-current characteristic of neurons. The described simple framework produces activation patterns naturally clustered around the stimulating sources. To assess the influence of this constraint on our stimulation optimization routines, we also implemented a "perfect-selectivity" control strategy, where we control perfectly the activity of single optic nerve fibers.

We then used an evolutionary heuristic to optimize the stimulation pattern applied to the optic nerve layer of the network. Simple evolutionary heuristics have been already used, for example, to craft input stimuli to a very simple CNN in order to replicate the activation of the last hidden layer of the network.[21] They are robust to modifications of the underlying models because they have been conceived to optimize black-box functions, and they allow adding constraints to the optimization process straightforwardly. In our study, a visual scene is input to the CNN and produces an activation pattern in the target cortical layer. Our ML framework computes the "artificial" activation that should be imposed at the optic nerve layer in order to replicate the target cortical activation pattern. It does so by evolving a population of candidate stimulations along a given number of generations, from an initial random population. The fitness of a candidate stimulation is given by the distance between the cortical activation obtained by applying it to the optic nerve layer and a target cortical activation. At each generation, the best-fitting individual candidates are retained and merged to a set of random perturbations of the best individuals (mutated individuals) and a set of random individuals (immigrants). Retaining the best-fitting individuals allows maintaining the memory of the best

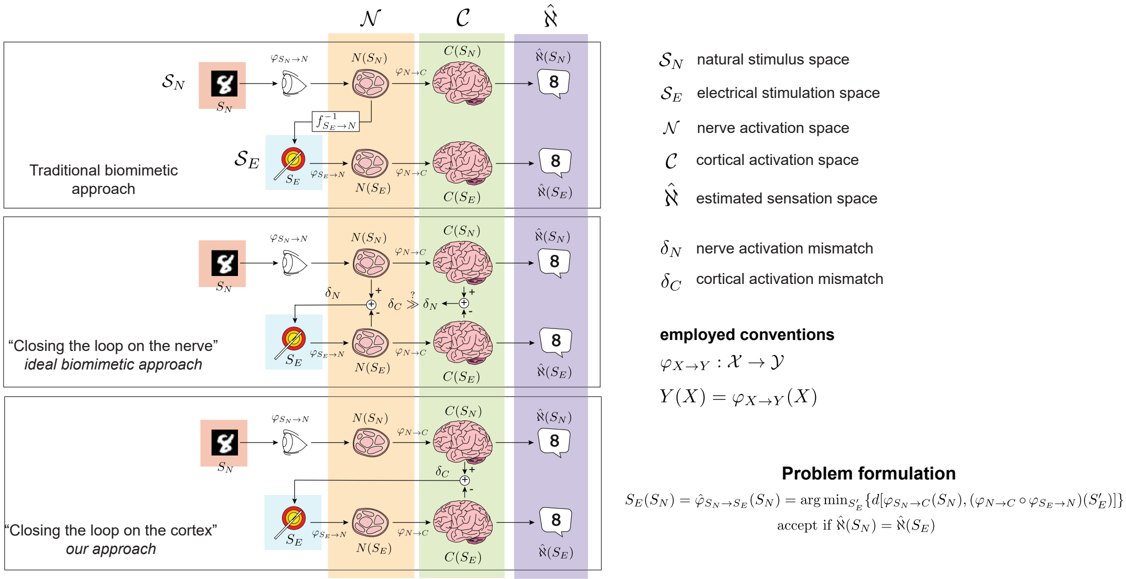

$\mathcal{N}$
$\mathcal{C}$
$\hat{\aleph}$
$\mathcal{S}_N$
S_N
$\varphi_{S_N \to N}$
$N(S_N)$
$\varphi_{N \to C}$
$C(S_N)$
$\hat{\aleph}(S_N)$
8
$f^{-1}_{S_E \to N}$
$\mathcal{S}_E$
S_E
$\varphi_{S_E \to N}$
$N(S_E)$
$C(S_E)$
$\hat{\aleph}(S_E)$
Traditional biomimetic approach
δ_N
$\delta_C \overset{?}{\gg} \delta_N$
"Closing the loop on the nerve"
ideal biomimetic approach
δ_C
"Closing the loop on the cortex"
our approach
$\mathcal{S}_N$ natural stimulus space
$\mathcal{S}_E$ electrical stimulation space
$\mathcal{N}$ nerve activation space
$\mathcal{C}$ cortical activation space
$\hat{\aleph}$ estimated sensation space
δ_N nerve activation mismatch
δ_C cortical activation mismatch
employed conventions
$\varphi_{X \to Y} : \mathcal{X} \to \mathcal{Y}$
$Y(X) = \varphi_{X \to Y}(X)$
Problem formulation
$S_E(S_N) = \hat{\varphi}_{S_N \to S_E}(S_N) = \arg\min_{S'_E}\{d[\varphi_{S_N \to C}(S_N), (\varphi_{N \to C} \circ \varphi_{S_E \to N})(S'_E)]\}$
accept if $\hat{\aleph}(S_N) = \hat{\aleph}(S_E)$

Figure 1. Abstract formulation of the optimization problem

(A) "Traditional biomimetic" approach, in which the electrical stimulation is computed "open loop," using global features of a simulated nerve activation pattern.

(B) "Closing the loop on the nerve" approach, in which the stimulation optimization "loop" is closed on nerve activation (ideal best-case scenario for the biomimetic approach). Such methods could lead to important error amplification because of sensory stream high non-linearity.

(C) Our "closing the loop on the cortex" approach, where the stimulation optimization loop is closed on cortical activation.

stimulation ever produced, and the relative sizes of the mutated and immigrant populations regulate the trade-off between global exploration and local exploitation.

After a given number of generations, we obtain a "best candidate" stimulation protocol, which produces the most similar target activation patter. We assess ex-post the quality of such reconstruction, comparing the classification produced by the network for both the natural and the best candidate stimulation. If they coincide, we have the two stimulations that produce patterns of activation interpreted as being produced by the same class of stimuli; i.e., they are "perceptually" equivalent.

Because the stimulus to be replicated does not change during the evolution time, we call the above optimization settings "static landscape" optimization. In addition to this static landscape optimization problem, we simulated "natural vision" experiments by proposing two "dynamic landscape" settings, where the stimulus to be reconstructed periodically changes. We proposed adequate turnover strategies in our evolutionary heuristic that can deal with such time-varying visual stimuli.

We have shown that it is possible to exploit a model of the visual system to evolve optic nerve stimulation patterns that replicate the chosen region cortical activation for static and dynamic stimuli, with a stimulus-identification accuracy that is acceptable notwithstanding the evident limitations of employing a fixed, low number of stimulating sites.

RESULTS

Our theoretical framework is depicted in Figure 1. Figure 1A shows the traditional biomimetic approach to sensory restoration; Figure 1B depicts the ideal closing the loop on the nerve, which is the best-case scenario for the biomimetic approach; finally, Figure 1C displays our closing the loop on the cortex optimization strategy. The choice of closing the loop on higher regions along the visual system allows us to exploit the higher, more invariant representation of the stimulus provided more complex areas, but has the immediate drawback that a model of the visual system from the stimulation site to the target region must be provided.

Our computational framework is depicted in Figure 2. In Figure 2A, it is shown how the CNN model of the visual system and the evolutionary heuristic are interfaced. In Figures 2B–2D we describe the simplified models for the association of CNN units with optic nerve fibers and the assumptions employed to compute their activation in perfect-selectivity and imperfect-selectivity control paradigms.

We have performed a number of *in silico* experiments to assess and interpret the performance of our framework in static and dynamic settings. In the following paragraph, we will briefly introduce the assessment measures that will be used to report our results throughout the entire article. A detailed account of the computation of these quantities can be found in the experimental procedures.

For each evolution generation, we will call the best individual stimulation pattern in the current population the candidate stimulation pattern. Applying this candidate stimulation pattern at the optic nerve layer, we can obtain through our CNN model an activation at the target cortical layer and an output array of probabilities for each stimulus class (from the softmax output of the network). From the cortical layer activation, we can obtain a fitness error value, while from the softmax output we can obtain an output loss value and an output class. End-generation output classes will be used to obtain confusion matrices and accuracy values (end-generation accuracy). Finally, we propose the

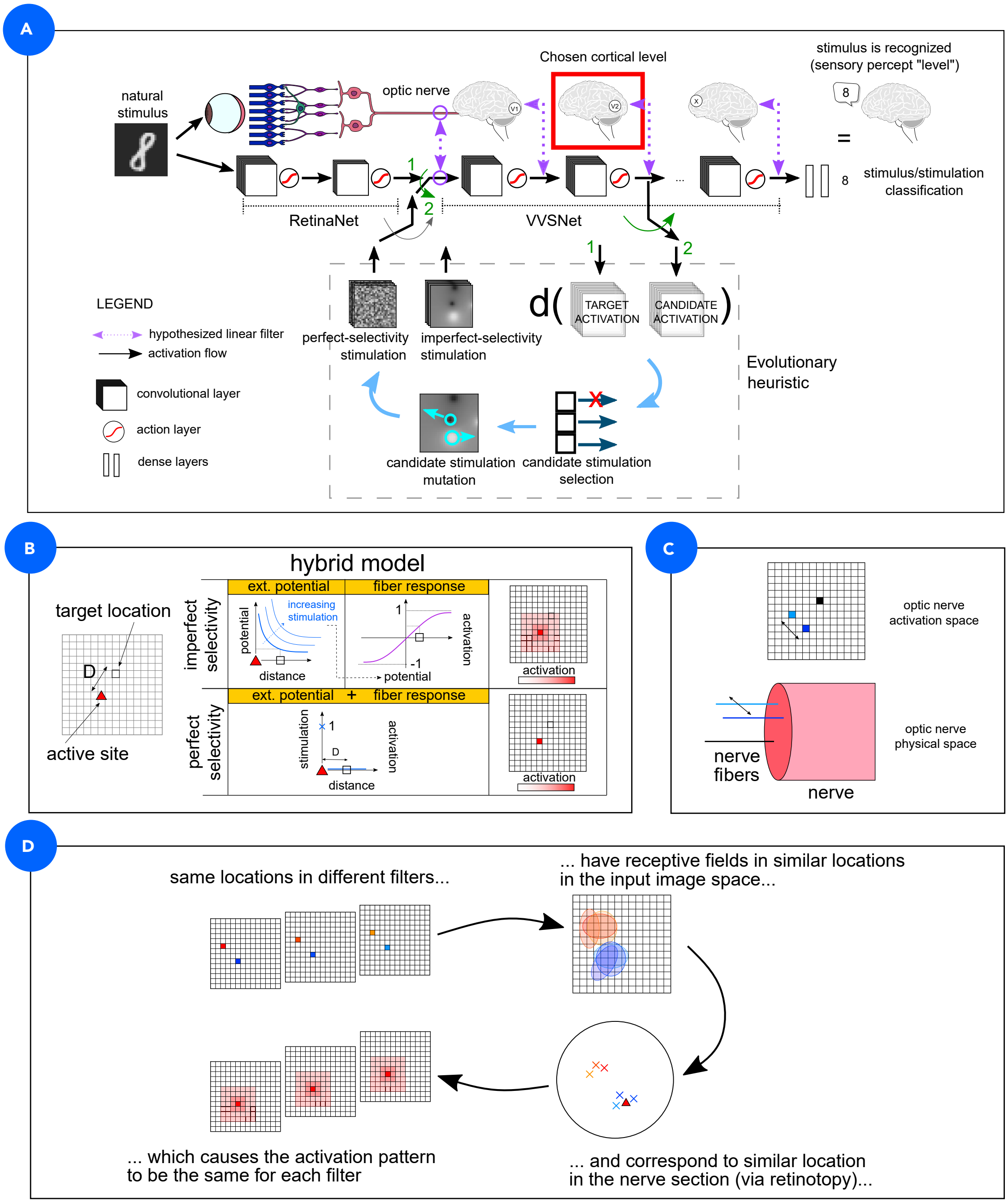

A
natural stimulus
optic nerve
Chosen cortical level
V1
V2
X
stimulus is recognized (sensory percept "level")
8
=
8
stimulus/stimulation classification
1
2
RetinaNet
VVSNet
1
2
LEGEND
hypothesized linear filter
activation flow
convolutional layer
action layer
dense layers
perfect-selectivity stimulation
imperfect-selectivity stimulation
d(TARGET ACTIVATION CANDIDATE ACTIVATION)
Evolutionary heuristic
candidate stimulation mutation
candidate stimulation selection
B
hybrid model
target location
D
active site
imperfect selectivity
perfect selectivity
ext. potential
fiber response
ext. potential + fiber response
potential
increasing stimulation
distance
1
-1
potential
activation
activation
stimulation
1
D
distance
activation
activation
C
optic nerve activation space
optic nerve physical space
nerve fibers
nerve
D
same locations in different filters...
... have receptive fields in similar locations in the input image space...
... which causes the activation pattern to be the same for each filter
... and correspond to similar location in the nerve section (via retinotopy)...

Figure 2. Outline of our computational framework

(A) An evolutionary heuristic is used to evolve activation patterns at the level of the optic nerve such that a given response is attained at a chosen hidden cortical layer. Target and candidate activation patterns are obtained by commuting the green switches in position 1 or 2, respectively. The aim of the whole routine is to produce activation patterns that produce the same classification of a given natural stimulus fed to the network.

(B) Given an active site (red triangle), the corresponding activation map for a perfect-selectivity control is concentrated at the location of the active site. For imperfect-selectivity control, activation is obtained computing a potential according to a 1/distance characteristic and by passing it to a sigmoidal activation function. This produces multiunit, correlated activation maps.

(C) We suppose that the distance between two fibers in the optic nerve is proportional to the distance between the corresponding units in the optic nerve layer (we assume that an appropriate normalization process has been carried out).

(D) We assume that the activation imposed by an active site is the same on all the filters making up the optic nerve. This is because units in the same location in different filters of a CNN are “hard-wired” to have receptive fields in similar locations in the input image space. Assuming retinotopy, this means that the corresponding optic nerve fibers will be located in similar locations in the optic nerve section and will be targeted in the same way by stimulating active sites.

introduction of one-match accuracy, which measures the proportion of runs that produced at least one class-identifying stimulation protocol for the given stimulus along the evolution (see the experimental procedures for further explanations). Unless specified otherwise, generation-wise fitness error and output loss curves will be averaged across the entire proposed stimulus dataset.

Static landscape settings

In the following, we outline the *in silico* experiment carried out to characterize the base behavior of our computational framework for "static stimulus" optimization, and a number of variations to assess the influence of different factors on our routine performance. For every experiment in this section, 25 different stimuli are selected from each stimulus class (10 classes), one evolution run consists of $n_{gen} = 200$ generations, each run is initialized independent of the others, and the stimulated layer is always the optic nerve layer. The chosen values for the evolutionary heuristic parameters can be found in Table 3 (see also the experimental procedures section).

Base scenario

To investigate the base behavior of our framework, the second cortical layer was employed as the target layer, a single-filter optic nerve layer was employed, MNIST stimuli were employed, and 15 sources were used in the optimization. The resulting confusion matrices and generation-wise average fitness error and output loss plots are shown in Figure 3A.

The one-match accuracies were 0.99 and 0.94 for perfect-selectivity and imperfect-selectivity control, respectively. In contrast, the end-generation accuracies were 0.95 and 0.82 for perfect-selectivity and imperfect-selectivity control, respectively. Thus, for imperfect-selectivity control, we were able to determine class-identifying stimulations for 94% of the stimuli. These class-identifying stimulations did not always correspond to the evolution end generation. Indeed, the end-generation stimulation produced class identification for 82% of the stimuli. Runs that produced at least one time along evolution stimulation patterns "evoking" the right class perception produced a class-identifying end-generation stimulation 82%/94% = 87% of the time. This confirms that when a class-identifying stimulation pattern emerges along the evolution process, this can be usually linked to having "identified" the underlying stimulus class.

The interquartile ranges of the fitness errors at the start and end generations are disjoint and the output loss decreases substantially, confirming the goodness of the evolution procedure.

Changing the stimulus dataset

We employed fashion MNIST (FMNIST) stimuli[22] to assess how employing a different, more complex set of visual stimuli would affect our optimization performance. The one-match accuracies were 0.92 and 0.63 for perfect-selectivity and imperfect-selectivity control, respectively. In contrast, the end-generation accuracies were 0.52 and 0.36 for perfect-selectivity and imperfect-selectivity control, respectively. The accuracies stayed above chance but there was a substantial degradation of the performance from MNIST to FMNIST (Figures 3A and 3B), highlighting a limited generalization capacity of our simple network and algorithm to more complex datasets. When FMNIST instances were grouped into four macroclasses (representing shirts, trousers, bags, and shoes, see experimental procedures), the accuracies rose to values compatible with MNIST results. Specifically, the one-match accuracies were 0.96 and 0.94 for perfect-selectivity and imperfect-selectivity control, respectively. In contrast, the end-generation accuracies were 0.74 and 0.75 for perfect-selectivity and imperfect-selectivity control, respectively. Imperfect-selectivity end-generation accuracy increased by more than 100% passing from FMNIST classes to macroclasses.

Changing the target layer

To study the influence of the target-layer choice on the performance of our evolution routines, we performed imperfect-selectivity control, static landscape optimization using different cortical layers as target layers. For each target cortical layer, we performed five independent runs to assess the robustness of our findings with respect to random initialization and mutation. To compare the end-generation accuracies between different cortical layers, we performed the Kruskal-Wallis test obtaining $p = 0.015$. To check which pairs of simulations produced significantly different accuracies, we performed a Conover *post hoc* test with Holm-Bonferroni correction. The obtained corrected p values were $p_{12} = 0.37$, $p_{13} = 0.005$, and $p_{23} = 0.02$, where p_{xy} is the p value obtained comparing cortical layer x with cortical layer y. In addition, we performed one run using as a target the optic nerve layer itself, thus simulating the biomimetic approach referred to in the introduction and the experimental procedures. We notice that the biomimetic approach end-generation accuracy value of 0.82 is outside the range of the accuracy values obtained for cortical layer 3 (0.85, 0.88). Figure 4A shows the resulting accuracies.

We decided to assess whether closing the loop on the nerve layer led to a significant error amplification in cortical layers. We computed the error at the level of cortical layer 3 when the target was the optic nerve and when the target was cortical layer 3 and obtained comparable values. In addition, we computed the error at the other cortical layers when the target was the optic nerve and we found that the error increases as expected monotonically with the distance from the optic nerve layer. Finally, we computed the errors at the optic nerve layer and at the other cortical layers when the target was cortical layer 3, obtaining errors comparable to the case when the target is the optic nerve. The results are shown in Figure 4B.

Changing the optic nerve model

Then, we wanted to assess the influence that changing different features of the optic nerve model has on our results. We investigated the effect of modifying the number of filters in the optic nerve layer and the effect of different extents of retinotopy linking the stimulation pattern to the optic nerve activation.

In Figures 5A–5D, we can compare the optimization routines performed using a three-filter and a single-filter optic nerve layer. In the hypothesis of perfect retinotopy, units in the same location in different filters correspond to approximately the same location in the nerve physical space as they are linked to the same units in the input stimulus space. This leads the units to share the received excitation and to display the same activation in the imperfect-selectivity control settings. This limitation is not shared with the perfect selectivity control scenario, where each fiber is stimulated independently, irrespective of its location in the nerve physical space. Figure 5A schematically depicts this activation sharing and an example of the resulting activation patterns.

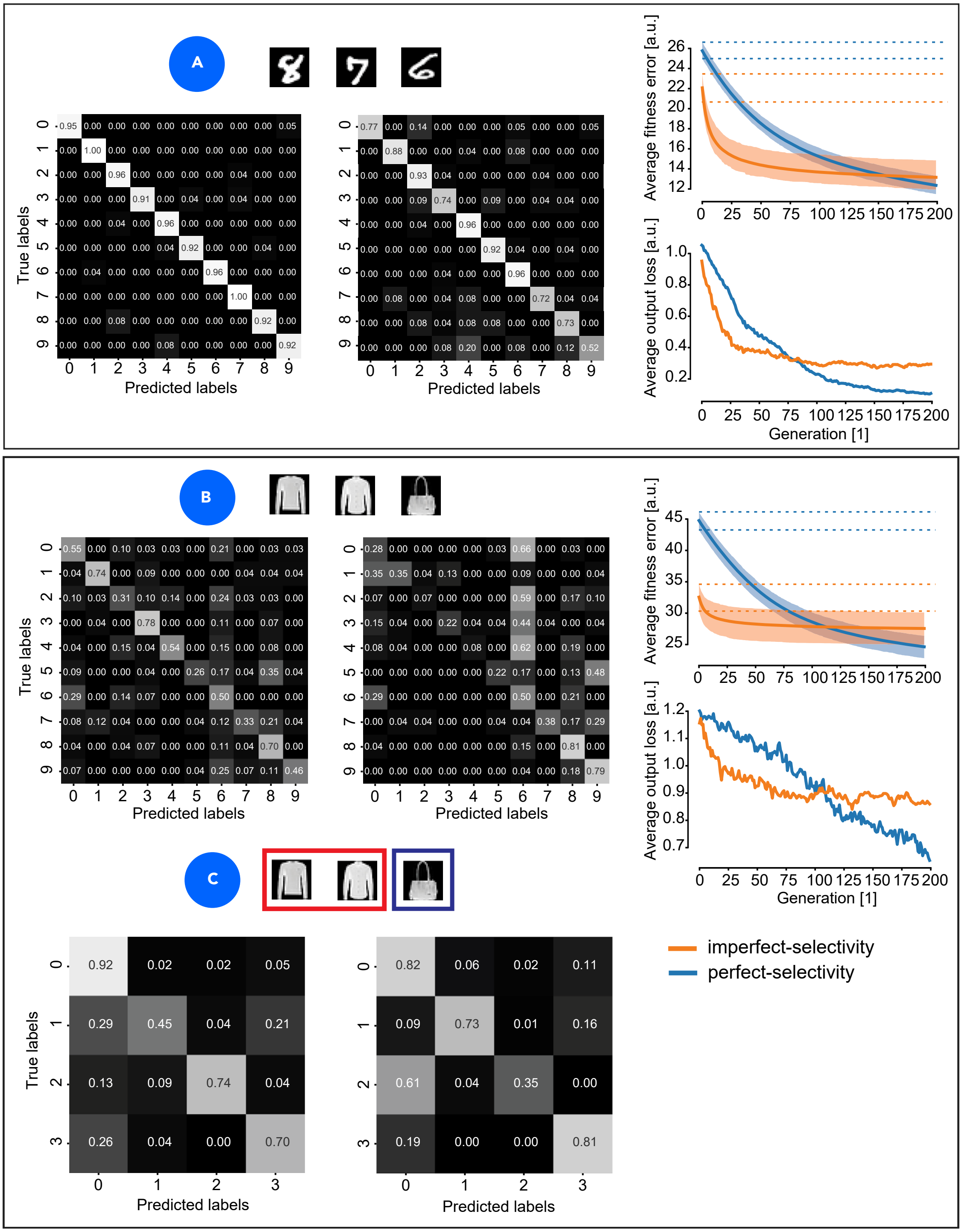


Figure 3. Static landscape optimization, base scenario

(A–C) Confusion matrices, average fitness error, and output loss curves for (A) MNIST and (B) FMNIST datasets. (C) For the macroclass FMNIST classification, only confusion matrices are meaningful. Average fitness error curves are displayed with their interquartile range. The first-generation interquartile range is reported across all generations for comparison.

In the case of three-filter optic nerve optimization, the one-match accuracies were 0.99 and 0.82 for perfect-selectivity and imperfect-selectivity control, respectively. The end-generation accuracies were 0.95 and 0.67 for perfect-selectivity and imperfect-selectivity control, respectively (see Figure 5B). It is interesting to compare the evolution processes in the case of single- and three-filter optic nerve (see Figures 5C and 5D). Imperfect-selectivity control of a multiple-layer and of a single-

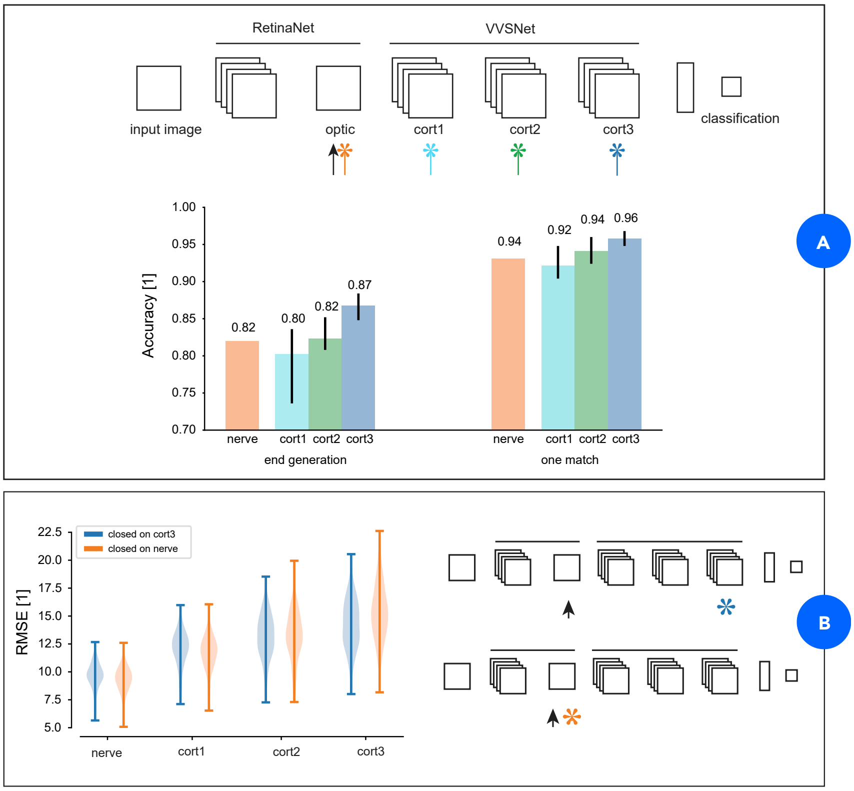


Figure 4. Static landscape optimization, changing the target layer

(A) Accuracy bar chart giving the end-generation (200 generations) and one-match accuracies obtained employing different target layers. Error bars correspond to the min-max range across five re-initializations.

(B) Violin plots of the root-mean-square error (RMSE) at each layer (with respect to its target activation pattern) employing the optic nerve layer or the third cortical layer as target layer. Arrows refer to the layers to which stimulation is applied, while asterisks refer to the target layers and are color coded with the bar and box plots.

layer optic nerve leads to different plateaus in both average fitness error and output loss. In the case of perfect-selectivity control, they settle on the same output loss plateau, resulting in comparable accuracies between controls in multiple- and single-filter optic nerves. As we expected, imperfect-selectivity control of a multiple-layer optic nerve led to higher average fitness error and output loss values and lower classification accuracies than that of a single-layer optic nerve. Interestingly, the temporal dynamics for the convergence of the evolution was the same for the two optic nerve models in the case of imperfect-selectivity control.

We then built a simplified model of retinotopic assignment, in which two parameters control the association between the unit locations in the optic nerve filters (which follow a perfect retinotopy assignment) and the locations of the corresponding units in the optic nerve physical space, thus establishing the distance between each unit and each stimulating source (see experimental procedures and Figure 5E). These two variables control respectively the probability of swapping two pixels in the physical nerve space and the average distance between two swapped pixels. We ran optimizations with increasing mixing (base, $p = 0$, $r = 0$; variations, $p = 0.1$, $r = 4$; $p = 0.25$, $r = 7$; $p = 0.5$, $r = 14$; $p = 1$, $r = 28$) and we additionally tried the cases of a very high probability, very low radius mixing ($p = 1$, $r = 4$) and of a very low probability, very high radius mixing ($p = 0.1$, $r = 28$); see Figure 5F. In Figures 5G and Table 1 we can see the resulting performance measures. As was expected, increasing levels of optic nerve fiber mixing led to a performance decrease. Interestingly, it seems that the parameter affecting the performance drop more is the mixing probability and not the average swapping distance.

Changing the number of employed sources

In all the previous experiments, a maximum number of 15 sources have been employed. This is compatible with the number of active sites normally implanted in nerves, which ranges from 10 to 20. We performed additional optimization runs with 1, 5, and 50 sources, representing, respectively, monopolar stimulation, multipolar stimulation using a number of sites lower than the number of available sites, and multipolar stimulation using a very high number of sites. The last case can help us understand if we need to push technology toward significantly increasing the number of sites in our current electrodes. The results of our analysis are shown in Figure 6. We can see that, as expected, an increasing number of available sources resulted in increased accuracies for all the datasets. Furthermore, restricting to 15 sources leads to a reduction in accuracy of only 10% with respect to the case of 50 sources for the MNIST and macroclass FMNIST problems and of 40% for the FMNIST problem. Surprisingly, for the MNIST and macroclass FMNIST problems, monopolar stimulation performed substantially better than chance. Finally, a multipolar stimulation employing only 5 sources produced relatively low performance deterioration with respect to the case of 15 sources. The fitness curves for both MNIST and FMNIST stimuli show that increasing the number of sources leads to a higher first-generation average fitness error and to a higher improvement in fitness along generations. The trade-off between these quantities is such that ultimately using more sources always leads to a lower end-generation average fitness error. Another interesting thing to observe is that for FMNIST stimuli the crossings of the average fitness error curves between the same numbers of sources happen systematically at earlier generations with respect to MNIST stimuli, which shows how increasing the number of sources generally leads to proportionally better fitting capabilities on more complex stimuli.

Dynamic landscape settings

Once we assessed the stimulation optimization performance for static visual stimuli, we wanted to set the ground for confronting

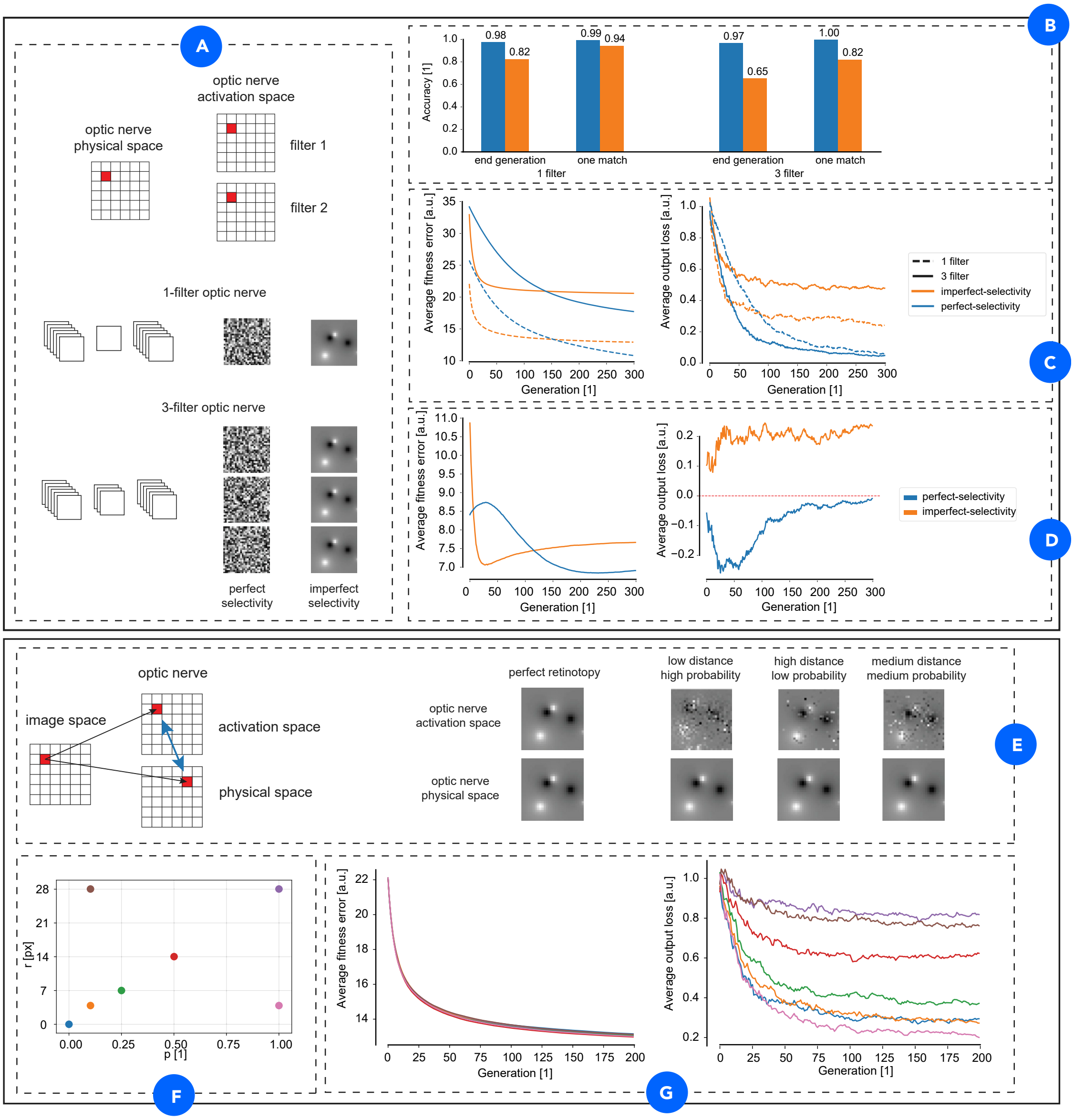


Figure 5. Static landscape optimization, changing the nerve model

(A–D) Changing the number of filters in the nerve layer. (A) Correspondence between optic nerve physical space and activation space and resulting filter activation patterns for perfect- and imperfect-selectivity control; units in the same location in different filters correspond to approximately the same location in the physical space and undergo the same stimulation when selectivity is not perfect. (B) Accuracy bar charts for one- and three-filter optic nerve evolution. (C) Average fitness error and output loss curves for one- and three-filter optic nerve optimization. (D) Plots of the difference between average fitness error and output loss between one- and three-filter optic nerve.

(E–G) Changing retinotopy. (E) Correspondence between optic nerve physical space and activation space and resulting filter activation patterns for different degrees of retinotopy; imperfect retinotopic assignment can be implemented by swapping units in the optic nerve activation space. (F) Retinotopy parameter combinations for which an optimization run has been executed. (G) Average fitness error and output loss curves for the different degrees of retinotopy.

with dynamic vision. Here, we performed two *in silico* experiments, corresponding to gradually and abruptly changing visual stimuli. In the gradual variation scenario, we wanted to determine whether we could adapt stimulation to stimuli rigidly moving in the field of view. In the abrupt variation scenario, we wanted to determine whether it is possible to leverage the capability of static reconstruction to cope with sudden visual stimulus changes. For simplicity, we always used the single-filter optic nerve model and the second cortical layer as the target layer.

Gradual variation

In gradual variation, we chose a stimulus and generated its translated copies with translations corresponding to a maximum of 5 pixels in both directions. We then formed a stimulus family selecting among these derived images the ones that the network correctly classified as belonging to the class of the parent image. For each

Table 1. Evolution heuristic parameters

Variable	Value	Description
$n_{samples}$	250 (25 × 10)	number of stimuli to replicate
n_{best}	50	number of best individuals selected at each generation
n_{imm}	100	number of random immigrant individuals per generation
Γ	0.6	Zipf law parameter (determines number of mutated individuals)
$n_{sources}$	15	number of sources in source-wise control
n_{units}	$28 \times 28 \times n_{filters}$	number of controllable units in point-wise control
p_{xy}	0.5	proportion of mutated source locations
p_{curr}	0.5	proportion of mutated source currents
p_{zero}^{src}	0.05	proportion of source currents set to zero
p_{mut}	0.1	proportion of mutated units
p_{zero}^{pts}	0.01	proportion of units set to zero
xy_{start}	U(0, 28)	source location initialization
$curr_{start}$	U(−3, 3)	source current initialization
Δxy	U(−0.5, 0.5)	source location mutation step
$\Delta curr$	U(−0.25, 0.25)	source current mutation step
act_{start}	U(0, 1)	unit activation initialization
Δact	U(−0.1, 0.1)	unit activation mutation step

stimulus family, we performed an evolution run for a number n_{gen} of generations. Every $n_{genperswitch}$ generations, we chose one of the available stimuli from the given family, and substituted the corresponding target to the current one (see Figure 7A). Because the change of target stimulus modifies the evolution fitness function, the individuals from the past generation are reevaluated and the best are chosen according to the new target stimulus.

We performed a simulation on 250 parent stimuli, with $n_{genperswitch}$ equal to 5 or 25 generations, for a total of n_{gen} = 200. This resulted in, respectively, 40 and 8 stimuli for each parent stimulus. The results can be seen in Figures 7B–7D. The one-match accuracies were 0.99 and 1.00 for 5 and 25 generation variations, respectively. The end-generation accuracies were 0.73 and 0.77 for 5 and 25 generation variations, respectively. The end-generation fitness error intervals are compatible with the one found for static landscape evolution. In Figure 7C, we compare the generation-wise average fitness error curves for the case of $n_{genperswitch}$ equal to 5 or 25 generations. After-switch fitness errors are comparable, while the effect of longer inter-switch time was mainly on the last generation before switch average fitness error. This gained representation advantage, nonetheless, does not have a strong repercussion on the first generation after switch error. In Figure 7D, the generation-wise average output loss curves for the case of $n_{genperswitch}$ equal to 5 or 25 generations and for the static landscape case are compared.

Abrupt variation

In abrupt variation, we selected a sequence of stimuli and switched between them every $n_{genperswitch}$ generations. We proposed three possible turnover variants to manage this dynamic setting. The first possibility (no archive, variation 1) corresponds to the standard turnover performed in the case of static landscape optimization, which was enough to grant good convergence also for a gradually varying landscape. The second possibility (no archive, variation 2) is to reinitialize the population when the stimulus is changed; the whole archived population is substituted to the current population, with the addition of a random immigrant population. The new population is evaluated with respect to the new fitness function, and the best individuals are chosen according to it. This corresponds to passing to the next generation only the archived individuals that capture best the features of the stimulus to be replicated. In addition, no knowledge of the stimulus class is required.

We performed a simulation on 250 samples, with $n_{genperswitch}$ equal to 5. The results can be seen in Figures 8B–8E. The end-generation accuracies were 0.41, 0.44, and 0.71 for no-archive variations 1 and 2 and the archive variation, respectively. Notice that the end-generation fitness error interval for the archive variation is compatible with the one found for static landscape evolution. In Figure 8C, the first generation after switch fitness errors (non-averaged) is shown for 70 following stimuli for the three turnover variants. In Figures 8D and 8E, the average fitness error and output loss curves are shown for a switch duration (5 generations) for the three turnover variants.

Psychophysics experiments

We performed psychophysics experiments in order to provide a sort of benchmark for the performance achieved *in silico* for imperfect-selectivity control for static and dynamic landscape optimization. In the case of static landscape optimization, we wanted to investigate how the lower resolution of the stimuli produced by imperfect-selectivity stimulation alone could cause a drop in classification accuracy, regardless of the power of our optimization algorithm. Thus, we administered to the human subjects degraded (blurred) input images of the two image sets (indicated as ‘‘mnist’’ and ‘‘fmnist’’ in Figure 9) to check the extent to which this degradation of resolution could yield a drop in discrimination performance. If the classification error from healthy subjects confronted with natural images constituted by a low number of intensity blobs (our ‘‘sources’’) is comparable to the algorithmic one, then we can assume that the latter can be explained mostly by taking into account the imperfect-selectivity control constraint and not by some issue in our optimization. The number of sources used to build the blurred images was equal to the one employed in our base evolutionary routine (which is compatible with the current technological limits). See the experimental procedures for a detailed account of stimulus generation. In gradual variation dynamic landscape settings, we asked ourselves whether and how much the evolution of class-identifying stimulation protocols was affected by smoothly moving the visual stimulus to be reconstructed. We tried to translate it in a psychophysical scenario by showing the blurred images described above for a limited time span of 2 s, smoothly changing the stimulus locations every 0.1 s (indicated as ‘‘grad’’ in Figure 9). Our question here was whether the fact that the blurred image was constantly moving would have affected the classification capabilities of healthy subjects. As for abrupt variation settings, we were interested in characterizing the time

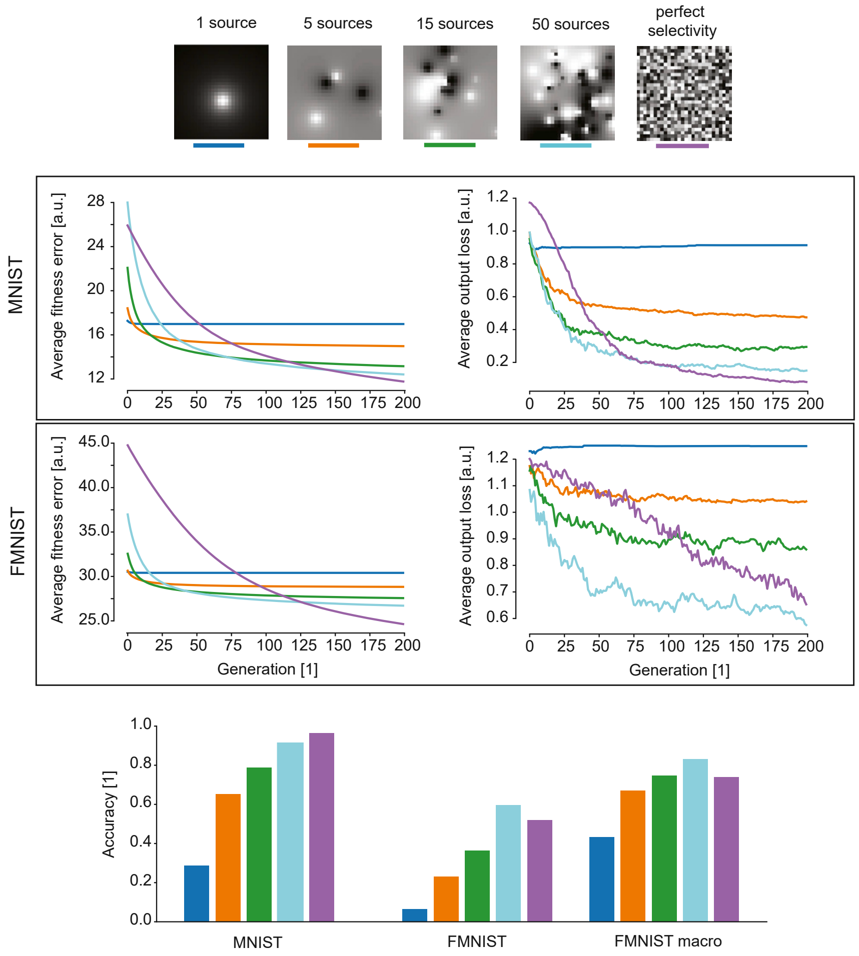


Figure 6. Static landscape optimization, changing the number of employed sources

Average fitness error, output loss curves, and accuracy bar charts for optimization runs employing different numbers of sources and MNIST/FMNIST datasets.

response of our optimization system to rapid discontinuities of the visual stimulus. In our psychophysical setting, we presented subjects with a sequence of static blurred stimuli with one stimulus every second or every 1.5 s (indicated as "abr 10" and "abr 15," respectively, in Figure 9). Finally, we proposed the original MNIST stimuli in an abr 10 setting, to assess if subjects were able to classify non-blurred stimuli in the short given time span. All the experimental paradigms are depicted schematically in Figure 9A.

We tested n = 10 subjects, with median age 23.5 years (minimum 22 years, maximum 61 years). The accuracy of classification for healthy subjects was computed for MNIST stimuli classification and for FMNIST class and macroclass (deduced *a posteriori*) classifications. In Figure 9B, the accuracy results for healthy subjects are displayed and compared with those obtained through our computational framework. Healthy subjects presented lower classification accuracies for the FMNIST dataset with respect to the MNIST dataset, and the classification into FMNIST macroclasses performed substantially better. This pattern is consistent with what we found with our framework and can be justified with the different complexities of the classification problems. Nonetheless, we observed that healthy subjects performed consistently worse than our framework in classifying MNIST stimuli (algorithm performs comparable to the most accurate individual) and consistently better than our framework in classifying FMNIST stimuli (algorithm performs comparable to the least accurate individuals).

We applied repeated-measurement ANOVA to the set of mnist, grad, abr 10, and abr 15 experiment accuracies and found that there were significant differences between the different conditions (p = 0.01). We applied a *post hoc* paired t test with Holm correction and found that abr 10 was significantly different from mnist and grad (p = 0.02 and p = 0.03, respectively).

We characterized inter-subject variability presenting subject-wise accuracy values (Table 2), experiment-wise accuracy boxplots (Figure 9B), and a homogeneity score (Figure 9C). Because every subject was shown a randomization of the same stimuli, we could compute the proportion of homogeneous classifications between subjects (both subjects attributed the same class to the stimulus, irrespective of its correctness). It interesting to notice that average inter-subject homogeneity has a drop between the static MNIST stimulus set (0.69) and both the FMNIST stimulus set (0.61) and the dynamic MNIST stimulus sets (0.56). Because during the static and dynamic MNIST stimulations each subject was presented a different randomization of the same stimuli, we computed the within-subject homogeneity (or the agreement between different classifications of the same stimulus in different conditions by the same subject). We can see that in general the within-subject homogeneity is higher than the inter-subject homogeneity (with a notable outlier), but that on average, almost one-third of the stimulus classifications disagreed.

DISCUSSION

We have presented an ML framework (described in Figures 1 and 2) to optimize optic nerve electrical stimulation for vision restoration. We have outlined a number of *in silico* experiments to evaluate the performance and suitability of our framework for scene reconstruction under static and dynamic (see Figures 7A and 8A) conditions. In the following, we will in turn discuss our results, comment on several modeling choices and their consequences, and, finally, illustrate the limitations and the corresponding future actions to improve our modeling framework.

Discussion of our results

In general, as expected, imperfect-selectivity control is systematically more difficult than perfect-selectivity control when using

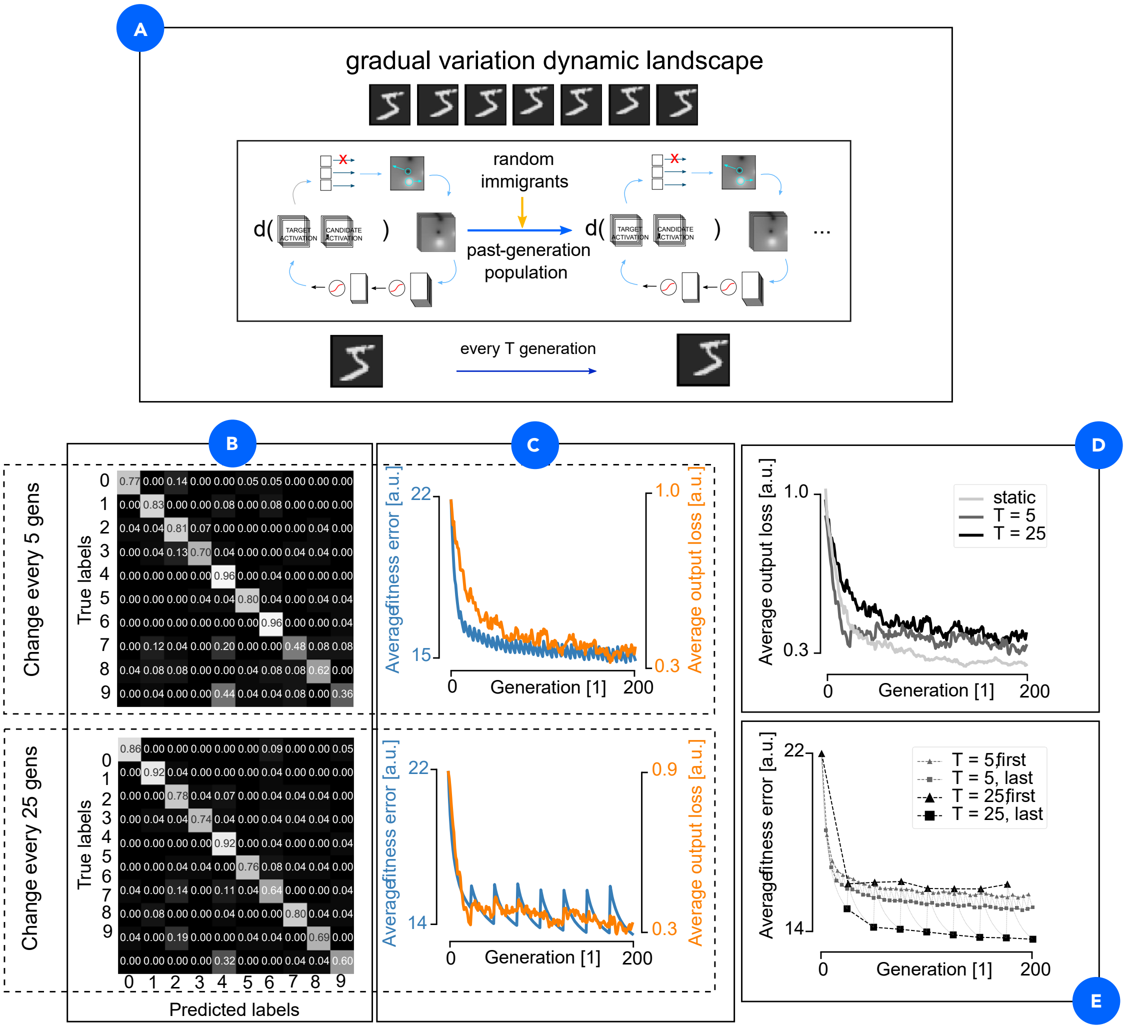


Figure 7. Dynamic landscape optimization, gradual variation

(A) Schematic of the adopted turnover strategy.

(B) Confusion matrices for the two proposed variants.

(C) Performance plots for the two proposed variants.

(D) Comparison of the average output loss curves for the two variants and static landscape optimization.

(E) Comparison of the average fitness error curves for the two variants, where the first and last generations after switch have been highlighted by triangle and square markers, respectively.

a number of sources compatible with the current technology. Still, imperfect-selectivity control evolution routines allow evolving stimulation patterns that can adapt or respond adequately (with an accuracy of stimulus class identification around 0.7) to changing scenes.

Static landscape

In the base scenario, it is possible to obtain good last-generation decoding for perfect-selectivity and imperfect-selectivity control, and almost perfect decoding in a one-match paradigm.

Changing dataset. In contrast to what happened for the network training on MNIST, in this case we obtained good accuracy values but very high validation and test accuracy loss (see Figure 10). This can partially justify our poor results and should convince us that more complex networks should be deployed to grant full generalizability to our approach. The very large improvement in classification accuracy when passing from the full FMNIST classification to macroclass classification, and the fact that perfect-selectivity control was less affected by the merging of different similar classes, could be linked to the intrinsic limitation of imperfect-selectivity stimulation protocols. Indeed, we expect that current electrical stimulation technology exhibits some kind of threshold for replicable stimulus detail. However, the fact that both imperfect-selectivity and perfect-selectivity control routines are visibly affected (even if in different measures) by the transition from MNIST to FMNIST can be explained also through the partial inadequacy of our simple search algorithm. The investigation of this very important scientific question will be one of our goals in the near future.

Changing the target layer. The performance of cortical layer-driven evolution improves the more we move toward the output

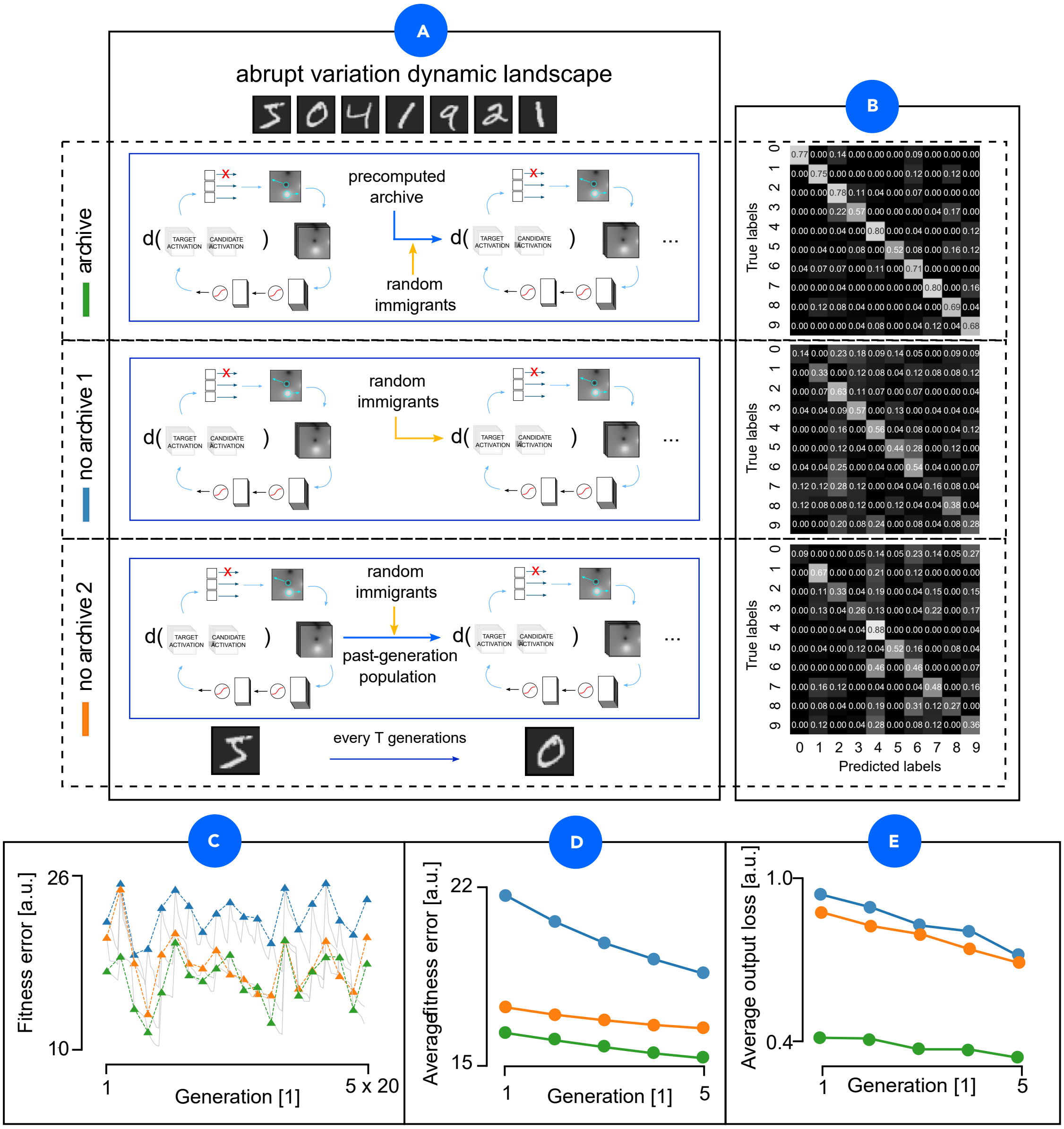


Figure 8. Dynamic landscape optimization, abrupt variation

(A) Schematics of the adopted turnover strategies.

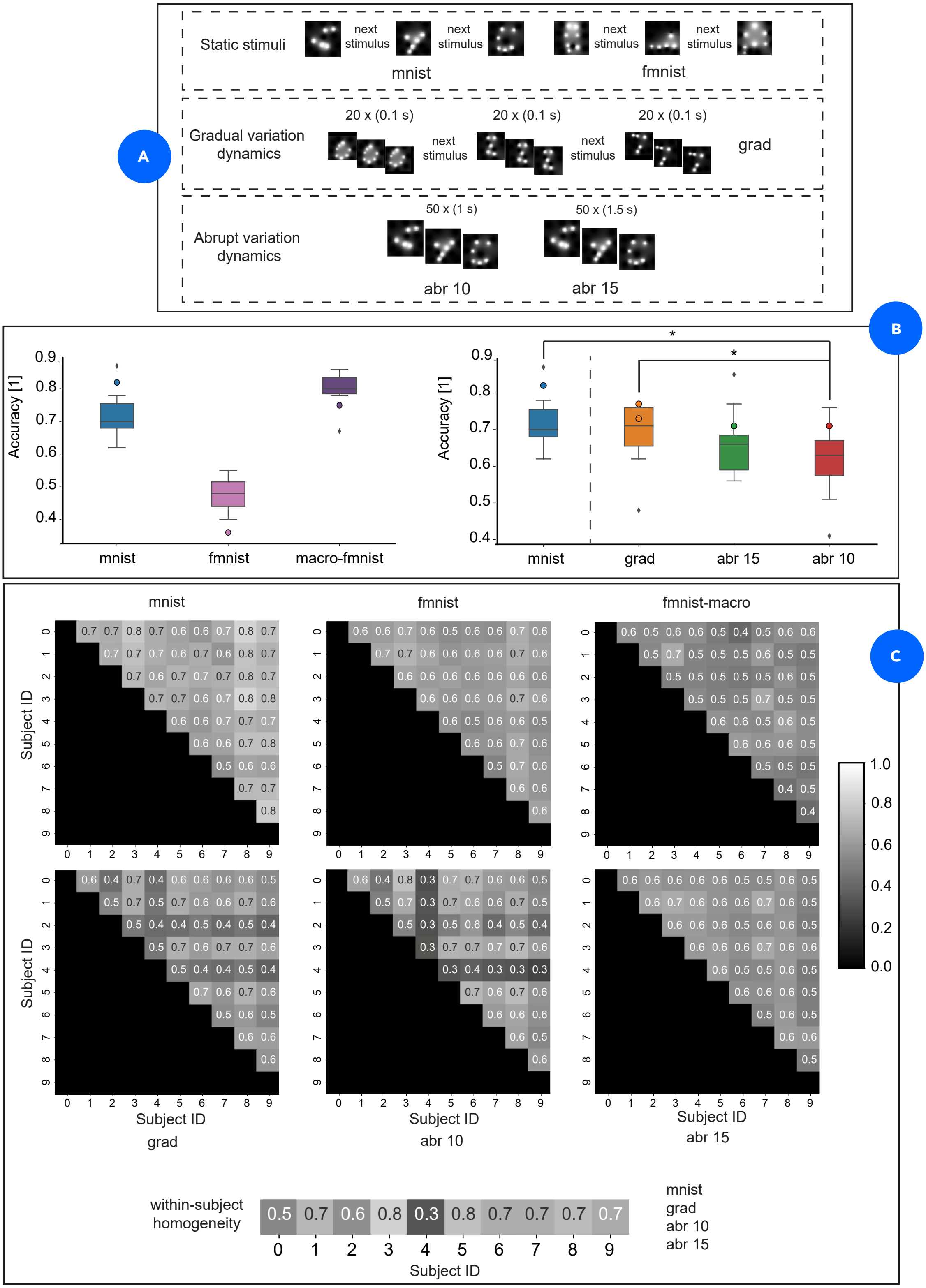
(B) Confusion matrices for the three proposed variants.

(C) Fitness error curves for the three proposed variants across a 70 stimulus evolution.

(D) Average fitness error curves for the three proposed variants.

(E) Average output loss curves for the three proposed variants.

layer. This is reasonable, as deeper layers should produce a hierarchy of representations that are more and more informative and invariant with respect to the classification problem. Interestingly, the performance of "biomimetic" stimulation is located between the performances of the first and the last cortical layer-driven simulations. We conclude that replicating nerve activation patterns seems to be an intrinsically easier task than replicating cortical activation patterns, but that this intrinsic advantage can be compensated for using layers that are nearer to the classification output, whose activation is thus more representative of the target class. This could be the case because of the direct influence of the feedback input on its output and because of the lower dimensionality of the nerve activation space. This lower dimensionality is here simulated by allowing for fewer filters in the optic nerve layer. The natural counterpart of having a lower number of filters is possessing a lower number of independent receptive field types. This is observed naturally, as cortical cell receptive fields are naturally more complex and more diverse than those of retinal ganglion cells.[20,23] Nonetheless, what we can gain from the invariance of higher representations can



mnist

	0	1	2	3	4	5	6	7	8	9
0		0.7	0.7	0.8	0.7	0.6	0.6	0.7	0.8	0.7
1			0.7	0.7	0.7	0.6	0.7	0.6	0.8	0.7
2				0.7	0.6	0.7	0.7	0.7	0.8	0.7
3					0.7	0.7	0.6	0.7	0.8	0.8
4						0.6	0.6	0.7	0.7	0.7
5							0.6	0.6	0.7	0.8
6								0.5	0.6	0.6
7									0.7	0.7
8										0.8
9										

fmnist

	0	1	2	3	4	5	6	7	8	9
0		0.6	0.6	0.7	0.6	0.5	0.6	0.6	0.7	0.6
1			0.7	0.7	0.6	0.6	0.6	0.6	0.7	0.6
2				0.6	0.6	0.6	0.6	0.6	0.6	0.6
3					0.6	0.6	0.6	0.6	0.7	0.6
4						0.6	0.5	0.6	0.6	0.5
5							0.6	0.6	0.7	0.6
6								0.5	0.7	0.6
7									0.6	0.6
8										0.6
9										

fmnist-macro

	0	1	2	3	4	5	6	7	8	9
0		0.6	0.5	0.6	0.6	0.5	0.4	0.5	0.6	0.6
1			0.5	0.7	0.5	0.5	0.5	0.6	0.5	0.5
2				0.5	0.5	0.5	0.5	0.5	0.6	0.5
3					0.5	0.5	0.5	0.7	0.5	0.5
4						0.6	0.6	0.5	0.6	0.5
5							0.6	0.6	0.6	0.5
6								0.5	0.5	0.5
7									0.4	0.5
8										0.4
9										

grad

	0	1	2	3	4	5	6	7	8	9
0		0.6	0.4	0.7	0.4	0.6	0.6	0.6	0.6	0.5
1			0.5	0.7	0.5	0.7	0.6	0.6	0.7	0.6
2				0.5	0.4	0.4	0.5	0.4	0.5	0.4
3					0.5	0.7	0.6	0.7	0.7	0.6
4						0.5	0.4	0.4	0.5	0.4
5							0.7	0.6	0.7	0.6
6								0.5	0.6	0.5
7									0.6	0.6
8										0.6
9										

abr 10

	0	1	2	3	4	5	6	7	8	9
0		0.6	0.4	0.8	0.3	0.7	0.7	0.6	0.6	0.5
1			0.5	0.7	0.3	0.7	0.6	0.6	0.7	0.5
2				0.5	0.3	0.5	0.6	0.4	0.5	0.4
3					0.3	0.7	0.7	0.7	0.7	0.6
4						0.3	0.4	0.3	0.3	0.3
5							0.7	0.6	0.7	0.6
6								0.6	0.6	0.6
7									0.6	0.5
8										0.6
9										

abr 15

	0	1	2	3	4	5	6	7	8	9
0		0.6	0.6	0.6	0.6	0.6	0.5	0.5	0.6	0.5
1			0.6	0.7	0.6	0.6	0.6	0.7	0.6	0.5
2				0.6	0.6	0.6	0.5	0.6	0.6	0.5
3					0.6	0.6	0.6	0.7	0.6	0.6
4						0.6	0.5	0.5	0.6	0.5
5							0.6	0.6	0.6	0.5
6								0.5	0.6	0.5
7									0.6	0.6
8										0.5
9										

within-subject homogeneity (mnist, grad, abr 10, abr 15)

Subject ID	0	1	2	3	4	5	6	7	8	9
within-subject homogeneity	0.5	0.7	0.6	0.8	0.3	0.8	0.7	0.7	0.7	0.7

Figure 9. Psychophysics experiments

(A) Schematics of the experiments performed, with sample stimuli.

(B and C) (B) Boxplots for static and dynamic landscape experiments (10 subjects). Diamonds denote outliers; asterisks indicate p values < 0.05; the circle markers indicate the (end-generation) accuracy values obtained with the presented ML routines. (C) Inter-subject and within-subject homogeneity matrices.

Table 2. Accuracy results for optimization under different retinotopy assumptions

p	R	Accuracy [1]
0	0	0.82
0.1	4	0.84
0.25	7	0.75
0.5	14	0.54
1	28	0.44
1	4	0.44
0.1	28	0.87

ultimately overcome the advantage of optic nerve activation space lower dimensionality. Of course, before being able to conclude anything about true nerve electrical stimulation, we need to properly validate this model and procedure, but it is still an encouraging result. We did not observe error amplification in the downstream layers when using as a target the optic nerve layer. This may suggest that indeed replicating with very high loyalty the activation at the level of the optic nerve will produce cortical activations that are comparable to closing the loop directly on the cortex. On one hand, this paves the way for a quantitative justification of the traditional biomimetic approach. In fact, it suggests that the unavoidable error due to the imperfect selectivity of imperfect-selectivity stimulation produces errors that are not substantially amplified along the sensory stream. On the other hand, it suggests that closing the loop on the cortex may lead to results comparable to a biomimetic approach. This could provide an alternative, technologically feasible approach to the biomimetic approach, whose "closed-loop implementation" is currently limited by the limited resolution of nerve recordings. In principle, it may seem obvious that closing the loop "more downstream" could lead to a performance improvement. Indeed, here we are suggesting that notwithstanding the compactness of nerve representation and the higher complexity and harder interpretability of cortical activation patterns, it may be indeed feasible to close the loop on the cortex, obtaining a performance comparable to an optimal biomimetic implementation. Finally, we remark that the fact that closing the loop on the last cortical layer leads to very low nerve activation errors provides evidence against the claim that we may generate adversarial examples using our optimization routines. Indeed, in that case, we would expect that we could find optic nerve activations very different from the natural one producing very similar cortical activations, which is excluded by the error similar to the one obtained closing directly on the nerve.

Changing the optic nerve model. In imperfect-selectivity multiple-filter optic nerve control, we are assuming that the same stimulation is applied to the same location in different filters. This relies on the fact that different retinal ganglion cell functional populations are arranged in independent mosaics covering the retina and thus, via retinotopy, the optic nerve section. We hypothesize that all units in a given location inside each filter are connected to the same region of the input image space and are thus located in nearby positions in the optic nerve cross section. Thus, they share stimulation intensity because of their spatial proximity. Notice that this constraint is not present in perfect-selectivity control. There, every unit can be activated independently, without considering the filter it belongs to. For this reason, the difference between perfect-selectivity control of stimulation of single- and multiple-filter optic nerves is only quantitative, in that the search space dimension has been increased, but its controllability is unaltered. In contrast, the difference between imperfect-selectivity control of single- and multiple-filter optic nerves is also qualitative because of the "coupling" between units belonging to different filters. This is reflected in the fact that for perfect-selectivity control, the output loss curves reach the same plateau. Interestingly, the same phenomenon cannot be highlighted in the fitness error curves.

Changing the extent of retinotopic organization in the optic nerve (namely, the spatial reorganization between unit assignment between the retina and the optic nerve) showed that our results are actually robust and perfect retinotopy is not a necessary requirement for our algorithm to provide class-identifying stimulation protocols. Indeed, when 25% of the fibers were displaced by an average distance of 25% of the nerve "diameter," the class identification accuracy stayed above 0.7. This result is actually extremely relevant as the precise extent of retinotopic organization in the more proximal sections of the optic nerve is still a matter of debate. Here we have shown that our possible ignorance on such anatomical constraints does not jeopardize the results shown in the present work.

Changing number of sources

We see that, as was expected, decreasing the number of available sources leads to lower accuracy values. Anyway, we observe that this decrease is strongly sublinear, even with a low number of sources. This could be a by-product of the simplicity of the chosen datasets. In fact, we see that, for example, we have an increase of approximately 2 times in passing from one to five sources for MNIST stimuli, but of 3.5 times for FMNIST stimuli. This highlights another reasonable result: the more complex a set of stimuli, the more sources we need to restore at a given level or to improve to a given extent the corresponding sensations. We can observe that monopolar stimulation (only one source) reached end-generation accuracies substantially above chance for MNIST and FMNIST-macro, but it is compatible with chance-level classification for the full FMNIST. A likely explanation for this result may be that the CNN is able to discriminate between simple stimuli using a single well-placed relevant feature. Indeed, the fitness values do not decrease substantially along the generations, which indicates a very limited training extent. Another very interesting result is that when 50 sources were used for FMNIST stimuli optimization, we could attain better end-generation accuracies than when employing perfect-selectivity control. While this may seem unreasonable, because the set of imperfect-selectivity stimulations is a subset of the set of perfect-selectivity stimulations, it is also true that perfect-selectivity control requires optimizing with respect to many more parameters and is thus more prone to local minima. In addition, complex naturalistic images are naturally spatially correlated, which could be leveraged by imperfect-selectivity stimulation, while perfect-selectivity control has to build these structures of correlation from scratch. This also explains why this differential improvement manifests on the FMNIST dataset, which contains more complex and realistic visual stimuli.

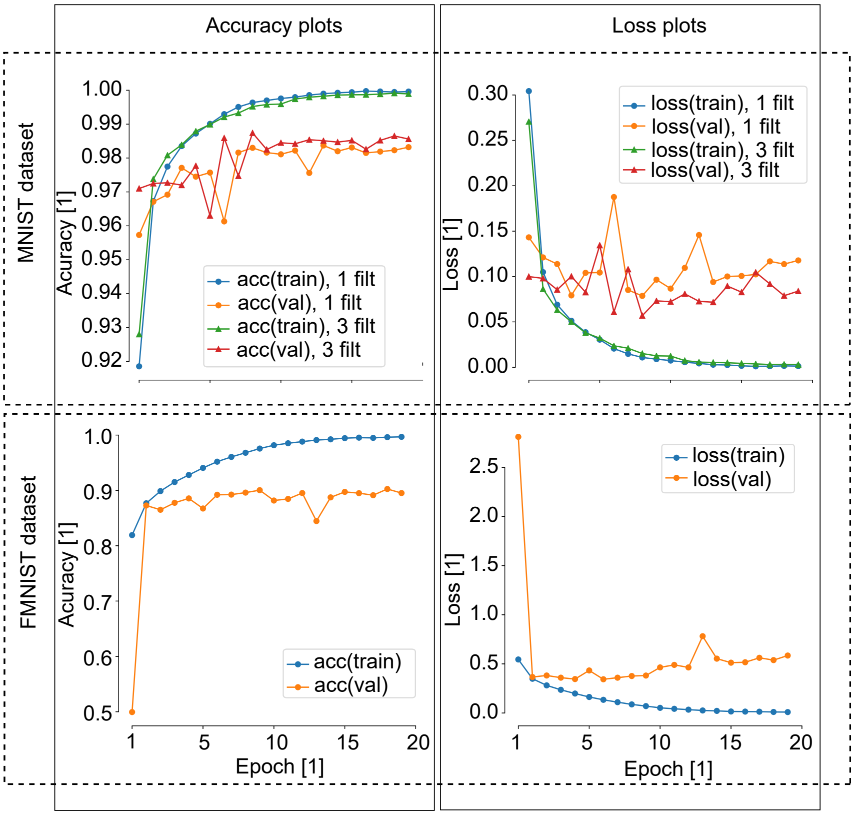


Figure 10. Training of the CNN models

Accuracy and loss curves for the training process of the CNN models on MNIST and FMNIST data, in the cases of single- and multiple-filter optic nerve layers.

Dynamic landscape

Dynamic landscape gradual variation. We showed that imperfect-selectivity evolution can adapt readily to rigid movement of the stimulus exploiting the current evolved population from different translated copies of the stimulus. The average fitness error (Figure 7C) exhibits strong seasonality synchronous with the stimulus changes. If we look at the error values corresponding to the first and last generation for each stimulus replica, we can make the following considerations. Looking at first-generation errors (Figure 7E, triangle markers), we can see that the evolutionary heuristic rapidly builds a population that is able to replicate any variation of the base stimulus with substantially lower error than a random population could. It is interesting to notice that when periodicity is very low, "learning" is smooth and both first- and last-generation errors exhibit a characteristic "power-law-like" behavior, which is observed also in static landscape optimization. High-periodicity landscape variation seems to exhibit a more stable first-generation error. It could emerge from the fact that when the stimulus changes, the population has more time to converge toward a specific sample replica and can gain less from "inter-sample learning."

Average output loss curves (Figure 7D), in contrast, exhibit less marked seasonality. This may be explained by the higher degree of stimulus invariance displayed by higher order layers and whose decreasing behavior may indicate an inter-sample class identification process. The comparison between low- and high-period dynamic landscape and static landscape optimization shows that static optimization is, expectedly, more effective in class identification. Interestingly, high-period dynamics allow a jump to a very loyal representation after the first stimulus replica, and loss decrease is slower after the first stimulus is observed, possibly for some "premature convergence" effect.

Dynamic landscape, abrupt variation. We showed that an archive of "converged" stimulation patterns can be used as a start-point population. The performance employing an archive is better than the performance of an evolution from a random population or from the last-generation population (Figure 8C). This was expected, as it suggests the following fact: consider two sets of stimuli, S_1 and S_2, containing images sampled from the same classes. For a stimulus s_2 in S_2 we can find a stimulus s_1 in S_1 so that: (1) the cortical activation pattern corresponding to s_2 is similar to the one corresponding to s_1 and (2) s_2 and s_1 share the same class.

In Figure 8C, we can visualize the erratic behavior of first-generation (after switch) fitness error, where no ongoing learning process can be spotted. Interestingly, variant 2 of the no-archive setting (current population retained) performs substantially better than variant 1 (current population entirely replaced by random individuals). This is further explored in Figures 8D and 8E, where it can be seen that no-archive variant 2 is nearest to archive in terms of average fitness error but is nearest to no-archive variant 1 in terms of average output loss. This suggests that no-archive variant 2 populations rapidly converge toward and remain in the subspace of possible activations due to a meaningful (cipher) stimulus, while no-archive variant 1 proposes random, potentially "meaningless" stimulations. This explains why the two archive-less variants lead to comparable loss values. For the network, trained only for cipher classification, a different class cipher produces a loss comparable to any other "meaningless" input sample, as it produces in any case a wrong cipher classification.

Psychophysics experiments

In static landscape experiments, the general changes in accuracy among the different classification problems for both human subjects and our algorithm can be justified by the complexity of the classification task (10 class MNIST classification is easier than 10 class FMNIST classification; 4 class FMNIST classification is easier than 10 class FMNIST classification). Nonetheless, we find it encouraging that the performance of our framework is in general comparable to the human one. This was not obvious, as it could have happened that human subjects performed substantially better than our routines on all the tasks or at least on the "easier" ones. In contrast, it seems that in addition to the fact that those accuracy ranges are compatible, human subjects actually displayed a worse average performance in the “easy” MNIST classification task.

One fundamental point motivating these static landscape psychophysics experiments was to “decouple” the influence on our optimization results coming from the limited interface selectivity and the limited capabilities of the optimization routine. Because healthy human subjects did not perform substantially better than our framework, we may advance the hypothesis that to increase stimulus-encoding performance we need to develop more selected neural interfaces, which could at least partly mitigate the intrinsic limitations in imperfect-selectivity control.

In gradual variation dynamic landscape psychophysical experiments, we wanted to investigate whether the robustness of our ML routine with respect to smoothly moving stimuli is a common feature with natural vision. Indeed, we found in healthy subjects that there was no significant performance drop because the stimuli to be classified were moving during the time of presentation. In abrupt variation dynamic landscape psychophysical experiments, we wanted to pave the way for some quantitative comparison of the time dynamics of animal vision and of our evolutionary heuristic. Indeed, the translatability of our heuristic to more ecologically valid scenarios depends on the speed of stimulus identification allowed by the relatively naive algorithm that we have proposed in the present work. We found that when stimuli were presented for a period of 1 s, the performance of healthy subjects was significantly lower than in the static case (where subjects controlled presentation time). When stimulus presentation had a period of 1.5 s instead, the performance change was not significant. This kind of result gives us some indication about the reasonable time dynamics for stimulus class identification given the imperfect-selectivity constraint on stimulus reconstruction, and it could help in the future in setting reasonable evolution durations for online object recognition experiments attaining performances similar to those of healthy subjects. We remark that such a long time is needed for healthy subjects because stimuli were heavily blurred. Indeed, when the true MNIST ciphers were presented at a rate of 1 s, the subjects scored almost perfect accuracies.

Finally, we notice how such psychophysics experiments display large inter-subject and even within-subject variabilities, which complicates any parallel between natural vision and our framework, in that the performance differences between subjects vary widely and show no notable regularity.

Justification of our modeling choices

General significance of dynamic landscape settings

Gradual variation simulates the slight variations occurring when looking at a homogeneous dynamic scene.[24] Abrupt variation simulates sudden changes in a visual scene.[25,26] We hypothesize that physiological vision can be seen as a flow of gradual variations interrupted by point-like abrupt variations. We have shown that it is possible to respond readily to both types of scene variation expected in natural vision, if an archive of precomputed stimulation patterns is available. In the future, we can imagine using a similarity-based classifier (in the space of natural stimuli) to classify each scene variation into (1) negligible variations, (2) gradual variations, or (3) abrupt variations, thus deciding whether to maintain the current population (settings 1 and 2) or to exploit the archive (setting 3).

General significance of imperfect-selectivity control

Our introduction to the imperfect-selectivity control paradigm has to be intended as a first step toward simulating realistic stimulation devices, where it is impossible to stimulate individual fibers independently and stimulation instead must target bundles of fibers. At the current state of the work, we are not yet simulating realistic electrode geometries, as active sites are allowed to distribute themselves inside the nerve section without any geometrical constraint. This is nonetheless compatible with penetrating intraneural electrodes. The model used to convert active-site stimulation intensity to the activation of optic nerve fibers is extremely simple but retains some fundamental physiological mechanisms.

Proneness of our method to “adversarial examples”

Adversarial examples were introduced in Szegedy et al.,[27] where it was observed that “deep neural networks learn input-output mappings that are fairly discontinuous to a significant extent. We can cause the network to misclassify an image by applying a certain hardly perceptible perturbation, which is found by maximizing the network’s prediction error.” Following this definition, we can see that what we have proposed is indeed not related to adversarial examples or “attacks.” First, in the present work, we were interested in replicating the activation pattern in an intermediate layer (representing a target cortical region), and we have never constrained the network to output the right prediction. Second, our main interest was in performing this task while employing the superposition of a set of highly correlated activation maps applied at an upstream layer (ideally representing the optic nerve): the highly correlated nature of the introduced stimulation strongly limits the ability of constructing very complex “hardly perceptible perturbation[s].”

Nonetheless, something similar to adversarial attacks could be happening, but we find it more likely that it may be penalizing our results instead of playing in our favor. We generally assume that the more an intermediate activation pattern resembles a target one leading to a given classification outcome, the higher the probability that such activation pattern will also lead to the same classification. Indeed, the proneness of deep neural networks to adversarial examples means that it is possible to find extremely similar activation patterns leading to different classification outcomes. In contrast, it is generally recognized that adversarial attacks are a weakness that is not shared with biological vision. Indeed, it has even been proven recently that constraining an upstream layer of a CNN to resemble the primate visual cortex leads to improved resistance of the network to adversarial attacks.[28] Thus, it could be possible that here some of the best-fitting activation patterns are actually adversarial examples and lead to the wrong classification in our CNN, whereas they would be class identifying when elicited in animal experiments.

Finally, another concern would be that we could be attacking the cortical layer by crafting *ad hoc* optic nerve activation. We answer this possible claim in two ways. The first is that we think that adversarial attacks loosen their meaning when the output layer is of a higher dimension than the input one, as they are fundamentally based on input convergence. The second is that the possibility of crafting such attacks could suggest that the high dimensionality increase between the spaces of activation in the optic nerve and in the cortex occurs so that it is possible to craft very-low-dimensional non-intuitive (non-biomimetic) stimulation protocols that guide cortical activation (and thus the restored sensation) in non-trivial ways. This, indeed, could also establish the optic nerve as an optimal site of stimulation for vision restoration, as it is low dimensional and also provides the ability to finely control the very-high-dimensional downstream regions to an acceptable extent.

Comparison between our approach and the traditional biomimetic approach

The results obtained with our very simple model suggest that our framework in which stimulation is optimized using cortical activations has the potential to rival a closed-loop optimization in which the stimulation is optimal in replicating nerve activation. As we outlined in the introduction, such optimization strategy can be imagined as the unattainable best-case scenario for traditional biomimetic approaches, which exploit only some features of whole nerve activity. Even supposing that this result will hold for more complex network models and after an *in vivo* validation phase, our approach still presents the relevant drawback of requiring either a cortical recording implant or some method to deduce cortical activation at the desired resolution through less invasive technologies.

Comparison with alternative approaches

To the best of our knowledge, the present work is the first attempt at systematic analysis of the possibility of exploiting a CNN model of the visual system to optimize optic nerve stimulation in a dynamic, imperfect-selectivity setting. There are some noteworthy points of contact between our work and that of Tafazoli et al.,[21] where an evolutionary heuristic was employed on a CNN to hint at the feasibility of online evolution of cortical stimulation patterns in mice. There are nonetheless many notable differences in our aims and methods that suggest to us that a direct, explicit comparison of the two works would not be appropriate or informative. For example, a very different network architecture was employed (presence of a bottleneck, different number of filters at each layer, much larger stimuli in our case). Moreover, a much different evolutionary heuristic is employed in the present work, which does not need to evolve patterns online. Finally, we are mainly interested in the evolution of imperfect-selectivity stimulation at a fixed intermediate layer (perfect-selectivity stimuli were evolved at different locations in the network).

Use of a CNN model instead of a physiologically accurate model

Physiologically accurate models are often employed for similar stimulation-control problems.[29,30] Such models have a number of advantages over the use of the employed CNN models; for example, they allow greater interpretability, in that each model component has a clear, *a priori* connection with a physiological entity. Moreover, depending on their formulation, more elegant and effective control-theoretic methods can be employed to optimize their stimulation. Nonetheless, we think that the current knowledge of the visual system is incompatible with the development of such a model. For example, the individuation and the characterization of the fundamental retinal ganglion cell classes converging into the optic nerve are still a matter of active research[31] (compare them, for example, with the full characterization available for the tactile afferents converging into the median nerve).[32] In contrast, CNNs have proven to be superior to detailed models (given our current knowledge of them) in the ability to explain neural selectivity in a variety of higher cortical regions during ecological tasks (image recognition).[17,18,33,34] In addition, they provide a framework that is relatively easy to expand with the aim of adding representative capabilities, as was recently done by Lindsey et al.[20] for the optic nerve and Dapello et al.[28] for V1.

Study limitations and future developments

Even if promising, our study has some important limitations, which are listed below and will be addressed in the near future.

Hybrid model improvement

The transformation from an imperfect-selectivity stimulation protocol and the corresponding optic nerve activation should come from a less abstract model. The natural choice here would be to insert a hybrid model[35] to compute the response of the optic nerve fibers to electrical stimulation via detailed finite element modeling and neural computation simulation. The setting of an accurate hybrid model for the situation under study would require gaining some further knowledge on the morphology of the optic nerve for the targeted subject population and the definition of viable neuroprosthetic devices.

Generalization to more complex networks and datasets

The advantage of a more complex network architecture is 2-fold. On one side, networks that are more complex attain lower loss values, needed because they indicate stable and informative extracted features, on top of which the whole evolution process builds. On the other side, we are assuming a goal-driven setting,[18] which requires that the network attains animal-like performance on the task on which it is trained. In any case, to be able to translate the optimization results to animal experiments and human patients, it will be necessary to establish a correspondence between neural activity recordings and the activations at the different layers of our CNN model. This will most likely require much more complex network models. Naturally, the passage to such networks will most likely cause the ruggedness of the fitness landscape to increase, thus increasing the difficulty of traversing it via simple isotropic mutation.

"Truly time-varying" stimuli

For the time being, we focused on static scene classification. We tried to add the time dimension by performing different kinds of dynamic scene classification. However, animal vision is not simply high-bandwidth image classification; thus, the next step in this direction is to employ a sensory stream model that naturally encompasses time.[24]

Need for in vivo validation

In the present work, we provided a proof of concept using a very simple CNN model, which we assumed to be a good model of the visual system. Such assumption was motivated by the state of the art, but should be carefully validated for future, more complex network architectures. In particular, it will be relevant to establish the link between the network optic nerve layer, whose position is constrained by the corresponding architectural bottleneck, and the physiological activation space of the animal/patient optic nerve. Finally, the correspondence between cortical activation patterns and evoked percepts needs to be characterized *in vivo* and in comparison with its *in silico* counterpart.

Use of our model and significance of the work

We moved from the hypothesis that goal-driven CNNs can be single-unit predictive models of the primate visual system and provided a proof of concept that if the natural cortical activation corresponding to the perception of a given natural visual stimulus were known, it would be possible to produce via optimization perceptually equivalent optic nerve stimulation protocols for flows of static images. In an experimental setting, this framework could be used thus to propose optic nerve stimulation patterns to evoke an *a priori* established number of visual stimuli. We will briefly outline three possible routes for the future use of our framework.

The first route is to consider CNNs as a simple test bed to perform exploratory feasibility analyses. For example, here we have shown that imperfect-selectivity stimulation of the optic nerve leads to a decrease in stimulation performance that is still acceptable. Because CNNs are in general good models of the visual system, this motivates us to continue in the way of optical nerve stimulation. The stimulation optimization framework that we have presented is a black-box optimization and thus the patient's visual system can be substituted to the CNN used here. The major problem of this approach is the very high number of candidate stimulations to be explored. Nonetheless, here we need only the cortical response to the stimulation (and not, for example, a verbal feedback from the patient), and we could thus think of administering to an anesthetized/sedated patient a very high number of optic nerve stimulations, exploiting protocols similar to rapid-sequence visual presentation (RSVP). A brief review of applications of this technique can be found in Potter et al.[36] Of course, the present analysis will need to be repeated using much more complex CNN models before the translation could work, and this could jeopardize the effectiveness of our very simple evolutionary heuristic, requiring some form of "smarter" optimization.

Even though RSVP protocols theoretically allow testing a very high number of candidate stimulations, it could be argued that evolving stimulation protocols from scratch still requires the evaluation of too many of them. We can envisage two main alternatives to exploit the modeling power of CNNs for reducing their number.

Thus, the second research route would be to work in the direction of the "interpretation" of the optic nerve stimulation patterns obtained by performing optimization on the CNN model. It could be possible to extract a set of constraints and smart initializations for the *in vivo* optimization using RSVP, in a manner similar to what we investigated for the management of abrupt dynamic changes in the target stimulus. Indeed, the generality of our evolutionary heuristics allows the straightforward integration of additional constraints.

Finally, we could use directly the converged stimulation protocols obtained through the CNN model. Because here we would rely heavily on the hypothesis of single-unit predictivity of CNNs, we will need to produce filters that allow obtaining single-unit predictions from a patient's recordings. One possibility is to use RSVP to establish a relationship between the optic nerve-to-cortex transfer function in the CNN and in the patient's nervous system. At that point, the patient-specific CNN can be employed to evolve a set of proposed optimal optic nerve stimulation protocols, which provide initialization points for a second, very fast, round of local exploitation performed using the patient's cortical responses directly (again, using RSVP on the anesthetized patient).

Required recording selectivity in the cortex

In the present work, we have assumed that we have perfect knowledge of all the cortical units in the region corresponding to the target layer of our CNN. It is indeed possible to arrive with reasonable confidence at the extraction of single-unit tracks for relatively high numbers of neurons in the cortex using microelectrode arrays and spike sorting (whereas in nerves this is still a very complex open challenge). Nonetheless, the use of recording microelectrode arrays as a support for optic nerve stimulation in human patients would heavily increase the invasiveness of the neural interface, raising the obvious alternative of performing recording and stimulation in the cortex using a single device. For this reason, once our method has been validated and further explored and optimized with invasive recordings in animals, we will move to estimating cortical activity through non-invasive methods such as electroencephalography (EEG) and adequate electrical inverse models. To obtain an acceptable resolution in the cortex from the very low spatial selectivity of EEG recordings, the characterization of the cortical neurophysiological correlates of natural visual stimulation could be carried out using more invasive setups (e.g., electrocorticography) already implanted for other uses in able-sighted patients.

Conclusions

We have presented an ML framework that allows us to optimize optic nerve stimulation protocols by exploiting a model of nerve electrical stimulation (hybrid model) and a CNN model of a sensory system. This framework allows the development of optimal non-biomimetic (cortically driven) stimulation protocols. We have shown that in a simplified setting we could evolve class-identifying stimulations from different stimulus datasets of varying complexity. Our study paves the way for the development of stimulation optimization routines based on ML approaches.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

The lead contact for this work is Silvestro Micera at silvestro.micera@epfl.ch.

Materials availability

This study did not generate new unique reagents.

Data and code availability

Data and code can be accessed at the following link: https://github.com/s-romeni/CNN-GA-OpticStim.

Problem statement

Let us denote the natural stimulus (i.e., image) space with $\mathcal{S}_{\mathcal{N}}$, the electrical stimulus space with $\mathcal{S}_{\mathcal{E}}$, the cortical activation pattern space with $\mathcal{C}$, and the nerve activation pattern space with $\mathcal{N}$. We are interested in finding a map $\phi_{S_N \to S_E} : \mathcal{S}_{\mathcal{N}} \to \mathcal{S}_{\mathcal{E}}$ that associates an electrical stimulation S_E to a natural stimulus S_N, that is, $S_E(S_N) = \phi_{S_N \to S_E}(S_N)$, so that the application of S_E produces the same sensation of S_N (if we indicate the sensation evoked by a stimulus S—natural or artificial—by $\aleph(S)$, we can write that the final aim of our routine is $\aleph(S_E) = \aleph(S_N)$). Because sensations cannot be "quantified" directly, we will reformulate the problem in terms of replicating the nerve or cortical activation pattern produced by a natural stimulus. At the end of this section, we will introduce a possible quantification of $\aleph(S)$ specific to our ML framework. Such measure will be used in the following to deduce *a posteriori* if the optimization routines have produced sensations compatible with the natural ones.

Table 3. Subject-wise accuracies for psychophysics experiments

	Subj 1	Subj 2	Subj 3	Subj 4	Subj 5	Subj 6	Subj 7	Subj 8	Subj 9	Subj 10
MNIST	0.77	0.69	0.63	0.87	0.71	0.70	0.62	0.68	0.78	0.74
FMNIST	0.53	0.55	0.49	0.55	0.47	0.42	0.40	0.50	0.46	0.42
MFMNIST	0.86	0.78	0.80	0.86	0.83	0.84	0.79	0.67	0.80	0.83
GRAD	0.76	0.71	0.76	0.76	0.76	0.75	0.62	0.69	0.69	0.62
ABR 1.0	0.68	0.57	0.58	0.76	0.66	0.70	0.63	0.65	0.60	0.51
ABR 1.5	0.77	0.66	0.56	0.85	0.67	0.70	0.59	0.66	0.61	0.56
TRUE 1.0	0.99	0.97	0.96	0.95	0.98	0.98	0.99	0.99	0.98	0.97

A biomimetic approach assumes that we can solve the problem by cascading the two maps $\phi^{-1}_{S_E \to N}$ and $\phi_{S_N \to N}$ to obtain:

$$\phi_{S_N \to S_E} = \phi^{-1}_{S_E \to N} \circ \phi_{S_N \to N}. \qquad \text{(Equation 1)}$$

In other words, we determine the nerve activation pattern corresponding to a given natural stimulus, and then we look for the electrical stimulation that replicates it.

The map $\phi_{S_E \to N}$ cannot be inverted, and what is actually done in practice is the solving of the optimization problem:

$$S_E(S_N) = \underset{S'_E}{\operatorname{argmin}} \left\{ d\left[\phi_{S_E \to N}(S'_E),\ \phi_{S_N \to N}(S_N)\right] \right\}, \qquad \text{(Equation 2)}$$

where $d(\bullet, \bullet)$ is a dissimilarity function and the optimal error in the nerve activation pattern is given by:

$$\delta_N = \min_{S'_E} \left\{ d\left[\phi_{S_E \to N}(S'_E),\ \phi_{S_N \to N}(S_N)\right] \right\}. \qquad \text{(Equation 3)}$$

The error δ_N will be in general amplified non-linearly for each processing step from the nerve to the formation of a sensory percept. Because of this process of non-linear error propagation, it is possible that non-biomimetic stimulation protocols will produce sensations more similar to the natural ones than biomimetic protocols.

Since we want to decrease the action of non-linear processing steps, we choose to maintain stimulation at the level of the nerve (where it is simple to cover the whole view field),[6] but we close the loop on a higher region, corresponding to a predefined cortical area C. The problem from Equation (2) becomes thus:

$$S_E(S_N) = \underset{S'_E}{\operatorname{argmin}} \left\{ d\left[\phi_{S_E \to C}(S'_E),\ \phi_{S_N \to C}(S_N)\right] \right\}. \qquad \text{(Equation 4)}$$

We can write:

$$\phi_{S_E \to C} = \phi_{N \to C} \circ \phi_{S_E \to N}, \qquad \text{(Equation 5)}$$

where $\phi_{N \to C}$ is a model of the processing applied by the nervous system to map nerve activation patterns to the corresponding activation patterns at the chosen cortical level, and $\phi_{S_E \to N}$ is a model that converts an electrical stimulation protocol into the consequent nerve activation pattern.

Both the $\phi_{N \to C}$ from Equation (5) and the $\phi_{S_N \to C}$ from Equation (4) will be modeled via the same goal-driven CNN model, which simulates the visual system (and, more specifically, the ventral visual stream). We propose two different control settings to define the map $\phi_{S_E \to N}$: perfect-selectivity and imperfect-selectivity control. In perfect-selectivity control, we imagine being able to control independently the activation of each fiber in the nerve. In imperfect-selectivity control, we employ an elementary homogeneous, isotropic model of the nerve to produce extracellular potential in the nerve section.[37] We then pass it through a sigmoid function to obtain an activation value. This produces multiple-unit activation patterns, similar to what is observed in the standard practice of neuromodulation (see "Candidate solution (individual) definition" and "Nerve activation in imperfect-selectivity control"). The *argmin* search will be carried through an evolutionary heuristic (see "Genetic algorithm").

Finally, to assess the extent to which we are able to fulfill our final aim to find a map $\phi_{S_N \to S_E}$ so that $\varkappa(S_E) = \varkappa(S_N)$, we will employ the classification output of the CNN model as an estimator for the sensation evoked by a natural or artificial stimulus. Our aim will thus be $\widehat{\varkappa}_N(S_N) = \widehat{\varkappa}_E(S_E)$, where $\widehat{\varkappa}_N(S_N)$ corresponds to the classification output provided by the network when given S_N as an input stimulus, and $\widehat{\varkappa}_E(S_E)$ corresponds to the classification output yielded by the network when the activation $N = \phi_{S_E \to N}(S_E)$ is imposed as a nerve activation pattern.

For brevity, in the following we will refer to an S_N as a "stimulus" and to an S_E as a "stimulation pattern."

Convolutional neural network

The CNN used takes inspiration from the one presented in Lindsey et al.[20] It takes as input an "MNIST digits" look-alike image (28 × 28 pixels, gray scale) and then it is organized on a number of 2D convolutional layers. The network can be imagined as divided into two subnetworks, the RetinaNet and the VVSNet, which represent, respectively, the processing from the natural scene to the output of the retinal ganglion cells and the processing occurring along the ventral visual stream (we will not explicitly model the lateral geniculate nucleus here). The RetinaNet consists of two convolutional layers with numbers of filters equal to 10 and N_{BN}, respectively. The number N_{BN} is set lower than 10, to model the physiological bottleneck represented by the fact that the optic nerve has a reduced number of fibers with respect to the number of neurons in the previous retinal layers. The VVSNet consists of a number D_{VVS} of convolutional layers with 10 filters each. Each convolutional layer has kernel size of 3, stride = 1, and symmetric padding granting that no filter dimension reduction occurs between the layers. The network ends with two dense layers with 1,024 and 10 units. A non-linear activation follows every convolutional layer. All the activations but the very last are *tanh* activations; the last is a softmax activation. Training employs categorical cross-entropy and RMSprop optimizer. The network has been trained for 20 epochs with batch size of 64. The network definition and training were performed with the Python (https://www.python.org/) (3.7.6) module Keras (https://keras.io/) (2.3.1) with Tensorflow (https://www.tensorflow.org/) (2.1.0) backend.

Because of the network being explicitly built with the aim of replicating specific entities in the visual stream, we will refer in the following for brevity to the output layer of the RetinaNet as the "optic nerve layer" and to the output of any VVSNet convolutional + activation macrolayer as a "cortical layer."

Genetic algorithm

The values of all the genetic algorithm parameters described in the following section are reported in Table 3.

Candidate solution (individual) definition

In what follows, we will refer to the activation/stimulation patterns applied to the optic nerve layer as "individuals." Two different kinds of individuals are defined: the individuals corresponding to unit-wise stimulation and the ones corresponding to imperfect-selectivity stimulation. A perfect-selectivity individual consists of the activation values imposed on all the units of optic nerve layer, which are fed as inputs to the VVSNet. It thus consists of a number 28 × 28 × N_{BN} of values. An imperfect-selectivity individual consists of the values of x and y locations in the nerve cross section and of the value of the stimulation current for each employed source. It thus consists of a number $3n_{sources}$ of values. See Figure 2B for a schematic. All values are intended to be real numbers. At each generation of the genetic algorithm, a population of individuals is analyzed. A population is an unstructured collection of individuals.

Nerve activation in imperfect-selectivity control

The activation elicited via imperfect-selectivity control is computed via a hybrid model.[35] Each unit in a filter of the optic nerve layer is associated with a fiber in the true nerve section. We compute in turn the extracellular potential generated by the active sites that stimulates the fiber and the fiber response in terms of an abstract "firing rate," normalized so that it can be used directly as the activation value for the given unit. We employ a simple model of point sources in a homogeneous, isotropic, infinite medium.[37] In space, we would have:

$$V_m = \sum_{n=1}^{n_{\mathrm{sources}}} \frac{I_n}{4\pi\sigma \cdot d_{m,n}}, \quad FR_m \propto \mathrm{sigmoid}(V_m), \qquad \text{(Equation 6)}$$

where V_m is the electric potential in location $\boldsymbol{r}_m$=(x_m, y_m), I_n is the current injected by the n-th source (active site), $d_{m,n}$ is the Euclidean distance between $\boldsymbol{r}_m$ and the location $\boldsymbol{r}_n$=(x_n, y_n) of the n-th source, σ is the conductivity of the nerve, and FR_m is the firing rate of a fiber in location $\boldsymbol{r}_m$ in the nerve section.

Here we choose to work in normalized units so that $4\pi\sigma = 1$. We imagine that, because of retinotopy,

$$d[\boldsymbol{r}_m, \boldsymbol{r}_n] \propto d[(i, j), \boldsymbol{r}_n], \qquad \text{(Equation 7)}$$

where (i, j) is the location of the filter of the unit corresponding to the fiber that would be in location $\boldsymbol{r}_m$ in the nerve section. Filter numbers do not enter into establishing the location of a unit, coherent with the Hubel and Wiesel hypothesis of separation of "what" and "where."[38] The firing rate of the true fiber corresponds to the activation imposed on the corresponding unit, and, coherently, we set the sigmoid function in Equation (6) to a *tanh* function, so that the ranges of activations produced by the original network and our model match. We thus obtain the formula:

$$\mathrm{act}_{\mathrm{stim}}^{ijk} = \tanh\left(\sum_{n=1}^{n_{\mathrm{sources}}} \frac{I_n}{\sqrt{(x_n - i)^2 + (y_n - j)^2}}\right). \qquad \text{(Equation 8)}$$

Here, k indicates the filter to which the unit belongs. See Figure 2D for a schematic.

Retinotopy

In the past section, we implicitly hypothesized that there is complete overlap between the stimulus image space (where the activation space coincides with the physical space, or location (i, j) in the input tensor corresponds to location (i, j) in the input image/field of view), the optic nerve activation space, and the optic nerve image space. Thus, we have imagined that an optic nerve fiber in location (i, j) in each filter of the optic nerve corresponds to the location (i, j) in the physical space of the optic nerve, and that the center of the receptive field for the said unit will be in location (i, j) in the image space. Because of imperfect retinotopy, by the way, we expect that some rearrangement happens between the image space and the optic nerve space, so that location (i, j) in the image space will correspond to location (m, n) = (f(i), g(j)) in the optic nerve physical space.

Because the perfect retinotopy assumption is indeed "hard-wired" in CNNs, we decide to swap some locations at the level of the optic nerve, so that the activation and physical spaces of the optic nerve no longer coincide. In this way, image location (i, j) is still the center of the receptive field of optic nerve activation location (i, j), but corresponds to a location in the optic nerve physical space (f(i), g(j)). Because the CNN accesses only activation spaces, it will assume that optic nerve activation location (x, y) will correspond to a location (f^{-1}(x), g^{-1}(y)) in the image space. Because here f() and g() are unit location swaps, it follows that f^{-1}() and g^{-1}() are also unit location swaps, which shows that we are indeed implementing some kind of imperfect retinotopic association. We control the amount of retinotopic "confusion" with two parameters, p and r, which correspond, respectively, to the probability that two units are swapped between the image space and the optic nerve space, and to the average distance between the swapped units.

Fitness function definition

The fitness function associates a fitness value with each individual in a population. Individuals are selected on the basis of their fitness values. The fitness of an individual is defined here as the root-mean-square deviation of the activation elicited in the chosen cortical layer by the stimulation corresponding to the given individual from a target activation corresponding to the stimulus whose sensation has to be elicited via the stimulation:

$$\mathrm{fitness}(\mathrm{individual}) = -\mathrm{RMSE}\big(\mathrm{act}_{\mathrm{stim}}(\mathrm{individual}),\ \mathrm{act}_{\mathrm{targ}}(\mathrm{stimulus})\big). \qquad \text{(Equation 9)}$$

The activation elicited via unit-wise control can be simply obtained through the CNN.

Mutation strategy

Mutation consists of two types of individual perturbations: additive noise and zeroing. Each unit or source of each individual to be mutated is modified with a given probability by adding a disturbance according to a value sampled from a given uniform distribution. In addition, each unit activation or source injected current can be set to zero with a given probability, which can serve as a "long-range" mutation and to refresh values that could have prematurely converged in uninteresting regions of the fitness landscape.

Turnover strategy

Every non-initial generation population is obtained from the preceding one in the following way. A number n_{best} of best individuals is selected from the past-generation population. A number n_{mutated} of mutated individuals are generated via mutation from these best individuals. Finally, a number $n_{\mathrm{immigrants}}$ of immigrant random individuals is generated. The next-generation population is the collection of the

$$n_{\mathrm{individuals}} = n_{\mathrm{best}} + n_{\mathrm{mutated}} + n_{\mathrm{immigrants}} \qquad \text{(Equation 10)}$$

obtained individuals. Because we are applying an elitism strategy by conserving the past best individuals, the best individual fitness will be a monotonic function of the generation number.

Here, we have applied only unary (single-parent) mutations, and we did not define n-ary mutation strategies (the n-parent generalization of crossing over). Parents are constituted by the best individuals of the past generation. Because the learning problem is very difficult, we decided to assign a different number of offspring individuals to each parent individual, according to their fitness rank. Specifically, we employed a Zipf law (power law) distribution so that the least-represented parents provide exactly one mutated candidate, and the better-fitting parents provide mutated candidates according to the given law. Thus, the number n_{mutated} (and, correspondingly, the number $n_{\mathrm{individuals}}$) depends on the power law constant.

Static and dynamic landscape settings

Static landscape optimization consists in evolving for a number n_{gen} of generations the optimal stimulation pattern corresponding to a single static stimulus. The fitness function does not change across the different generations and consequently it defines a static "optimization landscape."

We propose two different types of fitness landscape time variation: (1) gradual variation and (2) abrupt variation. In gradual variation, the same input stimulus is employed, to which a low-magnitude (maximum of 5 pixels in 8 pixel connectivity) translation has been applied (with the appropriate padding at the external boundaries). In contrast, in abrupt variation, different input stimuli are employed (with the possibility of a transition happening between two different stimuli belonging to the same class).

Every variation event is instantaneous and its effects start during the evolution generation immediately following its occurrence. In the current implementation, the system does not have any *a priori* knowledge about these variations and does not build any internal model, neither for the time location of the changes nor for their nature or intensity.

We hypothesize that gradual variation can be adequately managed by standard archive-less evolution, while abrupt variation will benefit for the implementation of an archive. We summarize the methods proposed to deal with dynamic landscape settings in Figures 7A and 8A.

Gradual variation

In gradual variation, we choose a stimulus and generate its translated copies with translations corresponding to a maximum of 5 pixels in both x and y directions. We then select among these derived images the ones that the network correctly classifies as belonging to the class of the parent image. We perform, for each parent stimulus, one trial with a number n_{gen} of generations. Every $n_{\mathrm{genperswitch}}$ generations, one of the available stimuli is chosen randomly and substituted to the current target. When the target stimulus is substituted, the best individuals from the past generation are chosen according to the new target stimulus. This has the obvious consequence of disrupting the monotonicity of the best individual fitness function through the generations.

Abrupt variation

In abrupt variation, we change the stimulus every $n_{genperswitch}$ generations. We perform two variants of abrupt variation evolution: (1) exploiting an archive of converged stimulation patterns and (2) without any prior knowledge of the stimuli.

In variant (1), we needed to build an archive of stimulation patterns. We employed the MNIST test set stimuli that would not be employed in the following experiment and we archived a stimulation pattern if it produced a right classification at generation 200 (all parameters as in static landscape optimization). We stopped when we had the stimulation pattern corresponding to $n_{stimpercat}$ stimuli per category (corresponding to a total archive dimension of $10 \cdot n_{stimpercat}$ individuals). When the stimulus was changed, the whole archived individual population was substituted to the current population, with the addition of a random immigrant population.

In variant (2), when the stimulus changes we can either (2a) maintain the current population or (2b) replace the current population with a random population. An immigrant random population is added so that in variants (1), (2a), and (2b) the number of individuals for each generation is the same, enabling us to compare the performance of the three variants. The numbers of selected best individuals and of immigrant random individuals for each generation except the initial one are the same in all other simulations.

FMNIST

FMNIST[22] is a drop-in replacement dataset for MNIST. It depicts fashion items divided into 10 categories, namely, (0) T-shirt, (1) trousers, (2) pullover, (3) dress, (4) coat, (5) sandal, (6) shirt, (7) sneaker, (8) bag, and (9) ankle boot. Given the higher complexity of the sample images with respect to standard digit MNIST, in the following, we consider also the classification problem corresponding to the macrocategories (0M) shirt (which are obtained from original classes (0), (2), (3), (4), and (6)), (1M) trousers (from class (1)), (2M) shoe (from classes (5), (7), and (9)), and (3M) bag (from class (8)). The evolution of stimulation patterns for FMNIST stimuli was carried out employing the same network architecture and the same genetic algorithm parameters as for MNIST.

Accuracy estimation

We are posed with the problem of assessing the quality of our "reconstruction" capabilities. In what follows, we will study two different types of classification accuracy: (1) the "end-generation" accuracy, computed after a very high number of generations, and (2) the "one-match" accuracy, which corresponds to the fraction of samples for which a class-identifying stimulation protocol has been used after a given number of generations (in a given time span). These two accuracy definitions correspond to two different experimental settings. In the first setting, we want to reconstruct an unknown class stimulus; the end-generation accuracy can be thought as an estimate of the probability that the "regime" reconstruction (the proposed stimulus reconstruction after many generations) will allow class identification. In the second setting, we want to find a class-identifying stimulation protocol for a given stimulus; the one-match accuracy gives us the probability of finding such a stimulation protocol if we are willing to wait a given number of evolution generations (which is linked to computational time).

Psychophysics experiments

Ten subjects (two females) were enrolled in the study. All participants gave their written informed consent to participate, and the study was approved by the Commission Cantonale d'Éthique de la Recherche Genève. All volunteers had normal or corrected-to-normal vision. All experiments were conducted in accordance with relevant guidelines and regulations.

Experiments

Experiments consisted of a training session and three classification sessions. At the beginning, the subjects were shown a subset of 100 images from MNIST and FMNIST datasets, with the corresponding classification. During the three classification sessions, static and dynamic stimuli were shown to the subjects, who were asked to provide a classification for each presented stimulus. Static stimulus experiments consisted in presenting 100 blurred images from MNIST and 100 blurred images from FMNIST, presented in two series of 50 images. The gradual and abrupt variation dynamic stimulus experiments employed the same 100 blurred images from MNIST. In static stimulus experiments, subjects controlled the flow of stimulation and could use as much time as they needed for the classification. In gradual variation dynamic stimulus experiments, subjects controlled the flow of stimulation, but each stimulus appeared on the screen for 2 s, moving smoothly at 10 frames per second, and then disappeared; subjects could use as much time as they needed for the classification. Abrupt variation dynamic stimulus experiments consisted in the rapid presentation of 50 blurred stimuli from MNIST, without the possibility for the subject to control the flow of presentation. We performed runs with 1 and 1.5 s spacing between stimuli. To check that in the case of 1 s spacing the drop in performance was motivated by the complexity of the proposed stimuli and not by some intrinsic limitation linked to the very fast pace of stimulation, we administered a run of 1 s spaced abrupt variation experiments with true MNIST digits. Stimulus sequences and experiments (static, gradual, abrupt) were randomized between subjects to avoid any (unsupervised) learning effect.

In the case of static stimulation, constraining the maximum duration of observation could have artificially lowered the classification rate, depending on the allowed observation duration. On the other hand, we could not find any reasonable method for the selection of an "appropriate" observation duration, given that the ML accuracies were obtained at convergence, without constraining significantly the time of evolution (see evolution performance plots, where it is evident that at generation 200 evolution has virtually "stopped"). Indeed, in inquiring the "maximum" performance that a healthy subject can attain given the imperfect-selectivity reconstruction constraint, we wanted to set a maximal accuracy that we could reasonably expect from our evolutionary heuristic.

Stimulus generation

Each reconstructed stimulus is generated from an MNIST or FMNIST image. Stimuli are obtained by passing to a sigmoid the superposition of 15 point sources in a manner analogous to what we did in imperfect-selectivity control of the optic nerve activation. The locations and intensities of the employed sources for each stimulus were obtained so that they minimized the Euclidean distance between the source-reconstructed images and the original ones. The employed minimizer was the default Python *scipy.optimize.minimize* minimizer (L-BFGS-B method). Each optimization was repeated for 25 different initialization conditions and the best-candidate solution was retained.

ACKNOWLEDGMENTS

We warmly thank Daniel Caproni for his contribution to the preliminary discussions about this work and Giulio Matteucci for pointing out the model of physiologically constrained CNN by Lindsey and colleagues. This work was partly funded by the Bertarelli Foundation and by the Swiss National Science Foundation (National Competence Center Research in Robotics and CHRONOS SNF project).

AUTHOR CONTRIBUTIONS

Conceptualization, S.M., D.Z., and S.R.; methodology, S.M., D.Z., and S.R.; software, S.R.; formal analysis, S.R.; investigation, S.R.; writing – original draft, S.R.; writing – review & editing, S.M., D.Z., and S.R.; visualization, S.R.; supervision, S.M. and D.Z.; funding acquisition, S.M.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: February 7, 2021
Revised: March 5, 2021
Accepted: May 17, 2021
Published: June 16, 2021

REFERENCES

1. Raspopovic, S., Capogrosso, M., Petrini, F.M., Bonizzato, M., Rigosa, J., Di Pino, G., Carpaneto, J., Controzzi, M., Boretius, T., Fernandez, E., et al. (2014). Restoring natural sensory feedback in real-time bidirectional hand prostheses. Sci. Transl. Med. *6*, 222ra19. https://doi.org/10.1126/scitranslmed.3006820.
2. Tan, D.W., Schiefer, M.A., Keith, M.W., Robert, J., Tyler, J., and Tyler, D.J. (2014). A neural interface provides long-term stable natural touch

perception. Sci. Transl. Med. *6*, 1–25. https://doi.org/10.1126/scitranslmed.3008669.

3. Petrini, F.M., Valle, G., Strauss, I., Granata, G., Di Iorio, R., D'Anna, E., Cvancara, P., Mueller, M., Carpaneto, J., Clemente, F., et al. (2019). Six-month assessment of a hand prosthesis with intraneural tactile feedback. Ann. Neurol. *85*, 137–154. https://doi.org/10.1002/ana.25384.
4. Petrini, F.M., Bumbasirevic, M., Valle, G., Ilic, V., Mijovic, P., Cvancara, P., Barberi, F., Katic, N., Bortolotti, D., Andreu, D., et al. (2019). Sensory feedback restoration in leg amputees improves walking speed, metabolic cost and phantom pain. Nat. Med. *25*, 1356–1363. https://doi.org/10.1038/s41591-019-0567-3.
5. D'Anna, E., Valle, G., Mazzoni, A., Strauss, I., Iberite, F., Patton, J., Petrini, F.M., Raspopovic, S., Granata, G., Di Iorio, R., et al. (2019). A closed-loop hand prosthesis with simultaneous intraneural tactile and position feedback. Sci. Robot. *4*. https://doi.org/10.1126/scirobotics.aau8892.
6. Veraart, C., Raftopoulos, C., Mortimer, J.T., Delbeke, J., Pins, D., Michaux, G., Vanlierde, A., Parrini, S., and Wanet-Defalque, M.C. (1998). Visual sensations produced by optic nerve stimulation using an implanted self-sizing spiral cuff electrode. Brain Res. *813*, 181–186. https://doi.org/10.1016/S0006-8993(98)00977-9.
7. Brelén, M.E., Duret, F., Gérard, B., Delbeke, J., and Veraart, C. (2005). Creating a meaningful visual perception in blind volunteers by optic nerve stimulation. J. Neural Eng. *2*, 21–28. https://doi.org/10.1088/1741-2560/2/1/004.
8. Chen, X., Wang, F., Fernandez, E., and Roelfsema, P.R. (2020). Shape perception via a high-channel-count neuroprosthesis in monkey visual cortex. Science *370*, 1191–1196.
9. Saal, H.P., and Bensmaia, S.J. (2015). Biomimetic approaches to bionic touch through a peripheral nerve interface. Neuropsychologia(Elsevier) *79*, 344–353. https://doi.org/10.1016/j.neuropsychologia.2015.06.010.
10. Valle, G., Mazzoni, A., Iberite, F., D'Anna, E., Strauss, I., Granata, G., Controzzi, M., Clemente, F., Rognini, G., Cipriani, C., et al. (2018). Biomimetic intraneural sensory feedback enhances sensation naturalness, tactile sensitivity, and manual dexterity in a bidirectional prosthesis. Neuron *100*, 37–45.e7. https://doi.org/10.1016/j.neuron.2018.08.033.
11. DiCarlo, J.J., Zoccolan, D., and Rust, N.C. (2012). How does the brain solve visual object recognition? Neuron *73*, 415–434. https://doi.org/10.1016/j.neuron.2012.01.010.How.
12. Logothetis, N.K. (1996). Visual object recognition. Annu. Rev. Neurosci. *19*, 577–621. https://doi.org/10.1146/annurev.neuro.19.1.577.
13. Nassi, J.J., and Callaway, E.M. (2009). Parallel processing strategies of the primate visual system. Nat. Rev. Neurosci. *10*, 360–372. https://doi.org/10.1038/nrn2619.
14. Curcio, C.A., Sloan, K.R., Kalina, R.E., and Hendrickson, A.E. (1990). Human photoreceptor topography. J. Comp. Neurol. 497–523.
15. Curcio, C.A., and Allen, K.A. (1990). Topography of ganglion cells in human retina. J. Comp. Neurol. *300*, 5–25. https://doi.org/10.1002/cne.903000103.
16. Leuba, G., and Kraftsik, R. (1994). Anatomy and Embryolo and total number of neurons of the human primary visual cortex. Anat. Embryol. (Berl) *190*, 351–366.
17. Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. Annu. Rev. Vis. Sci. *1*, 417–446. https://doi.org/10.1146/annurev-vision-082114-035447.
18. Yamins, D.L.K., and DiCarlo, J.J. (2016). Using goal-driven deep learning models to understand sensory cortex. Nat. Neurosci. *19*, 356–365. https://doi.org/10.1038/nn.4244.
19. Richards, B.A., Lillicrap, T.P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R.P., de Berker, A., Ganguli, S., et al. (2019). A deep learning framework for neuroscience. Nat. Neurosci. *22*, 1761–1770. https://doi.org/10.1038/s41593-019-0520-2.
20. Lindsey, J., Ocko, S.A., Ganguli, S., and Deny, S. (2019) A unified theory of early visual representations from retina to cortex through anatomically constrained deep cnNs. 7th Int. Conf. Learn. Represent. ICLR 2019 1–17.
21. Tafazoli, S., MacDowell, C., Che, Z., Letai, K.C., Steinhardt, C.R., and Buschman, T. (2020). Learning to control the brain through adaptive closed-loop patterned stimulation. J. Neural Eng. https://doi.org/10.1088/1741-2552/abb860.
22. Xiao, H., Rasul, K., and Vollgraf, R. (2017). Fashion-mniST: a novel image dataset for benchmarking machine learning algorithms. arXiv, 1–6.
23. Hirsch, J.A., and Martinez, L.M. (2009). Visual cortical and subcortical receptive fields. In Encyclopedia of Neuroscience, M.D. Binder, N. Hirokawa, and U. Windhorst, eds. (Springer Berlin Heidelberg), pp. 4307–4310. https://doi.org/10.1007/978-3-540-29678-2_6348.
24. Piasini, E., Soltuzu, L., Caramellino, R., Balasubramanian, V., and Zoccolan, D. (2019). Intrinsic dynamics enhance temporal stability of stimulus representation along a visual cortical hierarchy. bioRxiv, 1–42. https://doi.org/10.1101/822130.
25. Vinje, W.E., and Gallant, J.L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. Sci. (80- *287*, 1273–1276. https://doi.org/10.1126/science.287.5456.1273.
26. Sheinberg, D.L., and Logothetis, N.K. (2001). Noticing familiar objects in real world scenes: the role of temporal cortical neurons in natural vision. J. Neurosci. *21*, 1340–1350. https://doi.org/10.1523/jneurosci.21-04-01340.2001.
27. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2014) Intriguing properties of neural networks. 2nd Int. Conf. Learn. Represent. ICLR 2014 - Conf. Track Proc. 1–10.
28. Dapello, J., Marques, T., Schrimpf, M., Geiger, F., Cox, D.D., and DiCarlo, J.J. (2020). Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations. bioRxiv, 1–26. https://doi.org/10.1101/2020.06.16.154542.
29. Choi, J.S., Brockmeier, A.J., McNiel, D.B., Von Kraus, L.M., Príncipe, J.C., and Francis, J.T. (2016). Eliciting naturalistic cortical responses with a sensory prosthesis via optimized microstimulation. J. Neural Eng. *13*. https://doi.org/10.1088/1741-2560/13/5/056007.
30. Dura-Bernal, S., Li, K., Neymotin, S.A., Francis, J.T., Principe, J.C., and Lytton, W.W. (2016). Restoring behavior via inverse neurocontroller in a lesioned cortical spiking model driving a virtual arm. Front. Neurosci. *10*, 1–17. https://doi.org/10.3389/fnins.2016.00028.
31. Brackbill, N., Rhoades, C., Kling, A., Shah, N.P., Sher, A., Litke, A.M., and Chichilnisky, E.J. (2020). Reconstruction of natural images from responses of primate retinal ganglion cells. Elife *9*, 1–65. https://doi.org/10.7554/eLife.58516.
32. Saal, H.P., Delhaye, B.P., Rayhaun, B.C., and Bensmaia, S.J. (2017). Simulating tactile signals from the whole hand with millisecond precision. Proc. Natl. Acad. Sci. U. S. A. *114*, E5693–E5702. https://doi.org/10.1073/pnas.1704856114.
33. Yamins, D.L.K., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., and DiCarlo, J.J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc. Natl. Acad. Sci. U S A *111*, 8619–8624. https://doi.org/10.1073/pnas.1403112111.
34. Cadena, S.A., Denfield, G.H., Walker, E.Y., Gatys, L.A., Tolias, A.S., Bethge, M., and Ecker, A.S. (2019). Deep convolutional models improve predictions of macaque V1 responses to natural images. Plos Comput. Biol. *15*, 1–27. https://doi.org/10.1371/journal.pcbi.1006897.
35. Romeni, S., Valle, G., Mazzoni, A., and Micera, S. (2020). Tutorial: a computational framework for the design and optimization of peripheral neural interfaces. Nat. Protoc. *15*, 3129–3153. https://doi.org/10.1038/s41596-020-0377-6.
36. Potter, M.C., Wyble, B., Pandav, R., and Olejarczyk, J. (2010). Picture detection in rapid serial visual presentation: features or identity? J. Exp. Psychol. Hum. Percept. Perform. *36*, 1486–1494. https://doi.org/10.1037/a0018730.
37. McNeal, D.R. (1976). Analysis of a model for excitation of myelinated nerve. IEEE Trans. Biomed. Eng. BME- *23*, 329–337. https://doi.org/10.1109/TBME.1976.324593.
38. Hubel, D.H. (1995). Eye, Brain, and Vision (Scientific American Library/Scientific American Books).