Incorporating intrinsic suppression in deep neural networks captures dynamics of adaptation in neurophysiology and perception

Kasper Vinken[a,b,c,*], **Xavier Boix**[a,b,d], **and Gabriel Kreiman**[a,b]

[a] *Boston Children's Hospital, Harvard Medical School, Boston, MA 02115*

[b] *Center for Brains, Minds and Machines, Cambridge, MA 02139*

[c] *Laboratory for Neuro- and Psychophysiology, Department of Neurosciences, KU Leuven, 3000, Leuven, Belgium*

[d] *Department of Brain and Cognitive Sciences, MIT, Cambridge, MA 02139*

[*] *Correspondence: kasper.vinken@childrens.harvard.edu*

Abstract

Adaptation is a fundamental property of sensory systems that can change subjective experiences in the context of recent information. Adaptation has been postulated to arise from recurrent circuit mechanisms, or as a consequence of neuronally intrinsic suppression. However, it is unclear whether intrinsic suppression by itself can account for effects beyond reduced responses. Here, we test the hypothesis that complex adaptation phenomena can emerge from intrinsic suppression cascading through a feedforward model of visual processing. A deep convolutional neural network with intrinsic suppression captured neural signatures of adaptation including novelty detection, enhancement, and tuning curve shifts, while producing aftereffects consistent with human perception. When adaptation was trained in a task where repeated input impacts recognition performance, an intrinsic mechanism generalized better than a recurrent neural network. Our results demonstrate that feedforward propagation of intrinsic suppression changes the functional state of the network, reproducing key neurophysiological and perceptual properties of adaptation.

Introduction

The way we process and perceive the environment around us is not static, but is continuously modulated by the incoming sensory information itself. This property of sensory systems is called *adaptation* and can dramatically alter our perceptual experience, such as the illusory perception of upwards motion after watching a waterfall for some time (1). In the brain, neural responses adapt to the recent stimulus history in a remarkably similar way across sensory modalities and across species, suggesting that neural adaptation is governed by fundamental and conserved underlying mechanisms (2). The effects of adaptation on both the neural and perceptual levels have been most extensively studied in the visual system, where they appear to play a central role in the integration of temporal context (3–5). Therefore, to understand vision under natural, dynamic conditions, we must consider the neural processes that contribute to visual adaptation, and how these processes generate emergent functional

states in neural networks. Yet, we do not have a comprehensive understanding of what the underlying mechanisms of adaptation are, and how they give rise to changes in perception.

A fundamental question is whether the dynamics of adaptation are implemented by recurrent interactions in the neural network (6), or whether they can arise from established intrinsic biophysical mechanisms operating within each individual neuron (2). An important argument for the role of intrinsic cellular mechanisms in adaptation is that contrast adaptation in cat visual cortex leads to a strong afterhyperpolarization of the membrane potential (7). In other words, the more a neuron fires, the more its excitability is reduced, which is why the phenomenon is sometimes called neuronal fatigue (8). In this scenario, adaptation is caused by intrinsic properties of individual neurons that reduce their responsiveness proportional to their previous activation. Throughout the paper, we use the term *intrinsic suppression* to refer to such neuronally intrinsic mechanisms, which suppress responses based on recent activation.

However, adaptation phenomena in the brain go well beyond firing-rate-based suppression and it is not always clear whether those phenomena can be accounted for by intrinsic neuronal properties. First, the amount of suppression does not just depend on the preceding firing-rate, but can be stimulus-specific, i.e., suppression depends on whether the current stimulus is a repetition or not (9). Second, adaptation can also lead to response enhancement of single neurons (5, 8, 10, 11), sometimes even at the population level (12). Finally, adaptation can cause a shift in the neuron's tuning function for a particular stimulus dimension such as orientation(13, 14), direction (15), or spatial and temporal frequency (16, 17). Tuning shifts include both response suppression and enhancement (13) and have been linked to perceptual aftereffects where adaptation produces a shift in the perception of a stimulus property (15). Complex adaptation phenomena such as tuning shifts have fueled the argument that recurrent network mechanisms should be involved (13, 15, 16, 18). The putative involvement of recurrent signals is supported by computational models, which implemented adaptation by changing recurrent interactions between orientation tuned channels in order to successfully produce peak shifts (18–20).

Adaptation effects cascade through the visual system and can alter the network interactions in unexpected ways (2, 21). Indeed, adaptation-induced shifts in spatial tuning of primary visual cortex (V1) neurons can be explained by a two-layer model where changes in response gain in lateral geniculate nucleus cascade to V1 through a fixed weighting (22). These findings highlight the need for deeper, multi-layer models to capture the effects of adaptation, because previous models which lack the characteristic hierarchical depth and complexity of the visual cortex may not be sufficient to demonstrate the feedforward potential of intrinsic neuronal mechanisms. Moreover, the units in previous models are only designed to encode a particular stimulus dimension, such as orientation, and thus cannot provide a comprehensive framework of visual adaptation. In contrast, deep convolutional neural networks have recently come forward as a powerful new tool to model biological vision (23–26) (see however discussion in (27)). When trained to classify natural images, these models describe the stages of ventral visual stream processing of brief stimulus presentations with unprecedented accuracy (28–33), while capturing essential aspects of object recognition behavior and perceived shape similarity (29, 31, 34). In this study, we exploit another advantage of deep neural networks, i.e., their ability to demonstrate how complex properties can emerge from the introduction of biophysically inspired neural mechanisms. We implemented activation-based intrinsic suppression in a feedforward convolutional neural network (35), and tested the hypothesis that complex adaptation phenomena readily emerge *without dedicated recurrent mechanisms*.

A comprehensive model of visual adaptation should not only capture the neurophysiological dynamics of adaptation, but it should also produce the perceptual consequences. Therefore, we evaluated the proposed computational model implementing intrinsic suppression with critical neurophysiological and psychophysical experiments. We first show that the model captures the fundamental neurophysiological hallmarks of repetition suppression, including stimulus-specific suppression, not only from one image to the next, but also across several image presentations (5). Second, we show that the model readily produces the two fundamental perceptual aftereffects of adaptation, namely a perceptual bias in the estimate

of a stimulus parameter and an enhanced discriminability between parameter levels (3). In contrast with previous models which were constrained to one low-level property such as orientation, we demonstrate these effects using face-gender (36) as a stimulus parameter, to highlight the general applicability of the model. Third, we show that perceptual aftereffects coincided with response enhancements as well as tuning peak shifts, phenomena which are often considered to need the involvement of recurrent network mechanisms (13, 15, 16, 18). Interestingly, response magnitude changes contributed mostly to the perceptual bias, but tuning changes were required to explain enhanced discriminability. Finally, we show that a trained intrinsic neural mechanism is less likely to over-fit and thus provided a less complex solution than a recurrent network mechanism. Overall, while not ruling out any role of recurrent processes in the brain, these results demonstrate that the hallmark neural and perceptual effects of adaptation can be accounted for by activation-based suppression cascading through a complex feedforward sensory system.

Results

We investigate whether complex adaptation phenomena readily emerge from the propagation of activation-based intrinsic suppression, in a feedforward neural network model of ventral stream processing. We used a pre-trained convolutional neural network (35) (**Fig. 1A**) as a bottom-up computational model of vision, and introduced an exponentially decaying intrinsic adaptation state into each unit of each layer of the network, with its parameters set to impose suppression (**Fig. 1B**; **Materials and Methods**). Importantly, the two neural adaptation parameters α and β (**equations (1)** and (**2**)) *were not trained* to fit the neuronal responses or behavioral results, but were the same for each unit and were chosen to lead to a gradual build-up and recovery of the adapted state over several time steps (**Fig. 1B**). Throughout the paper, we use $\alpha = 0.96$ and $\beta = 0.7$, unless indicated otherwise. Due to the intrinsic suppression mechanism, the model units display temporally evolving responses (**Fig. 1C**) and their activations can be directly compared to the neurophysiological dynamics.

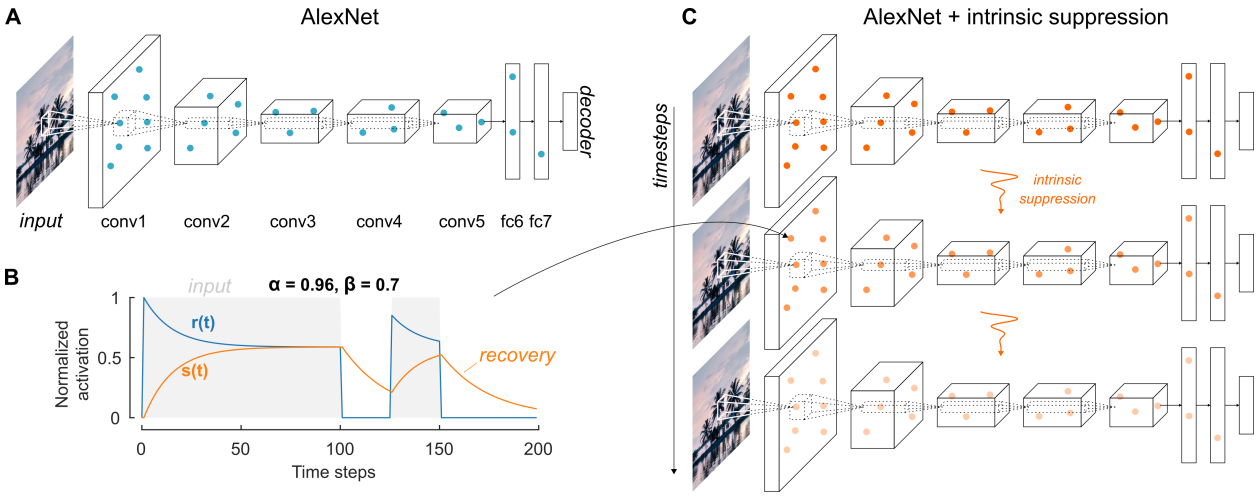

Fig. 1 Neural network architecture and incorporation of activation-based intrinsic suppression. (**A**), Architecture of a static deep convolutional neural network, in this case AlexNet (35). AlexNet contains five convolutional layers (conv1-5) and three fully connected layers (fc6, fc7, and the decoder fc8). The unit activations in each layer, and therefore the output of the network, are a fixed function of the input image. (**B**), Intrinsic suppression was implemented for each unit using an intrinsic adaptation state $s(t)$ (orange), which modulates the response $r(t)$ (blue) and is updated at each time step based on the previous response $r(t-1)$ (**equations (1)** and (**2**)). The parameter values $\alpha = 0.96$ and $\beta = 0.7$ were chosen to impose a response suppression ($\beta > 0$) which gradually builds up over time: for constant input (gray shaded areas), the value of the intrinsic state $s(t)$ gradually increases, leading to a reduction in the response $r(t)$. The intrinsic adaptation state recovers in the absence of input (non-shaded areas). (**C**), Expansion over time of the network in (A), where the activation of each unit is a function of its inputs *and* its activation at the previous time step (**equations (1)** and (**2**)).

A neural network incorporating intrinsic suppression captures temporal dynamics of adaptation at the neurophysiological level

We start with the most prominent characteristic of neural adaptation: repetition suppression, which refers to a reduction in the neuronal responses when a stimulus is repeated. We illustrate this phenomenon using an experiment where face stimuli were presented to a macaque monkey in pairs of two: an adapter followed by a test stimulus (**Fig. 2A** (37)). In repetition trials, the test stimulus was identical to the adapter whereas, in alternation trials, the adapter and test stimuli were different. Neurons recorded in the middle lateral face patch of inferior temporal (IT) cortex showed a decrease in the response during stimulus presentation and after stimulus offset. In addition, the neurons showed a lower response to a

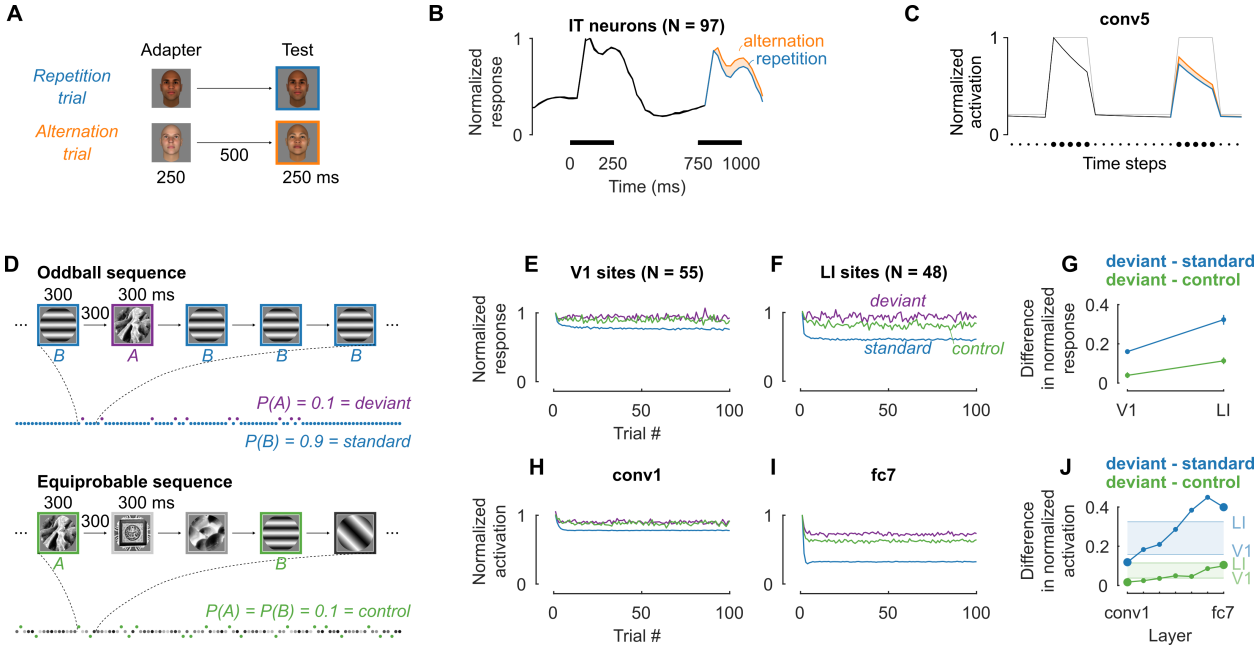

Fig. 2 Activation-based intrinsic suppression in a neural network captures the attenuation in neurophysiological responses during repetition suppression. (**A**), Face stimuli were presented in *repetition* trials (adapter = test) and *alternation* trials (adapter ≠ test). (**B**), Responses in IT cortex ($N = 97$, shown normalized to average peak activity) are suppressed more for a repeated stimulus (blue) than for a new stimulus (orange, data from (37)). Black bars indicate stimulus presentation. (**C**), The same experiment as in (A-B) produced similar repetition suppression in the model with intrinsic suppression (black, blue and orange lines; grey: no adaptation mechanism; average activity after ReLU of all $N = 43,264$ conv5 units). The x-axis units are time steps, mapping to bins of 50 ms in (B). (**D**), Example oddball sequence (top) with a high probability *standard* (blue) and a low probability *deviant* (purple) and example equiprobable sequence (bottom) as control (green). (**E, F**), Average neural responses in rat V1 ($N = 55$, (E)) and LI ($N = 48$, (F)) (12) for the standard (blue), deviant (purple), and control (green) conditions (normalized by the response at trial one). (**G**), Deviant - standard (blue) and deviant - control (green) response differences increase from V1 to LI (error bars: 95% bootstrap CI, assuming no inter-animal difference). (**H-J**), Running the experiment in the model captures response dynamics similar to rat visual cortex. (H) and (I) show the results for conv1 and fc7 (indicated by larger markers in (J)), respectively. Green and blue horizontal lines and shading in (J) indicate the neural data averages of (G).

face stimulus when it was a repetition trial (blue) compared to an alternation trial (orange; **Fig. 2B**).

We evaluated the average time courses of the model unit activations for the same experiment (**Fig. 2C**, mean of all $N = 43,264$ units in layer conv5). The model units showed a decrease in the response during the course of stimulus presentation. Consistent with repetition suppression in biological neurons, the response of model units to the test stimulus was lower for repetition than alternation trials. For this stimulus set, the largest difference between repetition and alternation trials was observed for layer conv5 (see other layers in **Fig. S1A**).

The model units demonstrated several key features of adaptation at two time scales: (i) during presentation of any stimulus, including the first stimulus, there was a decrease in the response with time; (ii) the overall response to the second stimulus was smaller than the overall response to the first stimulus; (iii) the response to the second stimulus was attenuated more when it was a repetition. However, the model did not capture more complex dynamics such as the second peak in neural responses. The model responses showed a smaller difference between repetitions and alternations than biological neurons: the average alternation-repetition difference was 0.07, SD=0.12 (model, 5 test time steps), and 0.11, SD=0.15 (IT neurons, 850-1000 ms window) in the normalized scale of **Fig. 2B,C**.

We hypothesized that the computer generated faces were too similar for the model to display the full range of adaptation effects. Therefore, we ran the same experiment using natural images with more variability. Indeed, natural stimuli resulted in a considerably larger difference between repetition and alternation trials (**Fig. S1B**), suggesting that the selectivity of adaptation at least partially reflects stimulus similarity in the model representations. Consistent with this idea, the stimulus similarity in pre-adaptation activation patterns for different adapter and test images was positively correlated with the amount of suppression for most layers (**Fig. S2**).

An important property of repetition suppression in macaque IT, is stimulus specificity: even for two adapters that equally activate the same neuron, the suppression for an image

repetition is still stronger than for an alternation (9). It is not straightforward to see how a neuronally intrinsic mechanism could account for this phenomenon, because an intrinsic firing-rate-based mechanism is by itself not stimulus selective (5). However, **Fig. S3** demonstrates that when activation-based suppression propagates through the layers of the network, neural adaptation of single units becomes increasingly less dependent on their previous activation, until stimulus-specific suppression is present for the majority of single units in fully connected layers.

In addition to the two temporal scales illustrated in **Fig. 2A-C**, adaptation also operates at longer time scales. For example, repetition suppression typically accumulates across multiple trials and can survive intervening stimuli (9). To illustrate this longer time scale, we present multi-unit data from rat visual cortex (12), recorded during an *oddball paradigm* where two stimuli, say A and B, were presented in a random sequence with different probabilities (**Fig. 2D**): a *standard* stimulus was shown with high probability ($P = 0.9$; blue), and a *deviant* stimulus was shown with a low probability ($P = 0.1$; purple). Stimulus (A or B) and condition (standard or deviant) were counterbalanced for each neural recording. The standard stimulus was far more likely to be repeated in the sequence, allowing adaptation to build up and therefore causing a decrease in the response for later trials in the sequence (**Fig. 2E,F**, blue). Adaptation was evident both in primary visual cortex (V1) and in the extrastriate latero-intermediate visual cortex (LI).

We evaluated the model in the oddball paradigm, *without any tuning or parameter changes*. The model qualitatively captured the response difference between standard and deviant stimuli (**Fig. 2H,I**). Comparing **Fig. 2E** versus **F**, the effect of adaptation was stronger in LI compared to V1 (**Fig. 2G**). An increase in adaptation along the visual hierarchy is consistent with the idea of adaptation cascading through the visual system, with additional contributions at multiple stages. Like the neural data, the model showed increasing adaptation effects from one layer to the next (**Fig. 2J**), and this increase only occurred when intrinsic suppression was incorporated in multiple layers (**Fig. S7**).

In the original experiment, the images A and B were also presented in separate equiprob-

able control sequences, where each stimulus was presented with an equally low probability ($P = 0.1$) together with eight additional stimuli (**Fig. 2D**) (12). Equiprobable sequences are typically used to distinguish repetition from surprise effects, because the probability of a repetition in the control condition is the same as for the deviant, yet no image is more likely or unlikely than the others. Thus, if neural responses also signal the unexpectedness of the deviant, then the response to a deviant stimulus should be larger than the control condition, which was observed for recording sites in downstream visual area LI (**Fig. 2F**, purple > green). The model also showed a difference in response between deviant and equiprobable control conditions in higher layers (**Fig. 2I,J**). Because the model only incorporates feedforward dynamics of intrinsic suppression, this response difference cannot be attributed to an explicit encoding of expectation. Instead, the lower response for the control condition results from higher cross-stimulus adaptation from the additional stimuli in the equiprobable sequences. This observation means that intrinsic suppression in a feedforward neural network not only captures response differences due to the repetition frequency of a stimulus itself (deviant versus standard), but also differences related to the occurrence-probability of other stimuli (deviant surrounded by high probability standard versus surrounded by several equiprobable stimuli).

A neural network incorporating intrinsic suppression produces perceptual aftereffects

A comprehensive model of visual adaptation should not only capture the neural properties of repetition suppression, but should also explain perceptual aftereffects of adaptation. Aftereffects occur when recent exposure to an adapter stimulus biases or otherwise alters the perception of a subsequently presented test stimulus. For example, previous exposure to a male face will make another face appear more female to an observer, and vice versa (**Fig. 3A**). In other words, adaptation biases the decision boundary for perceptual face-gender discrimination towards the adapter. A defining property of this type of aftereffect is that no perceptual bias should occur when the adapter corresponds to the original boundary

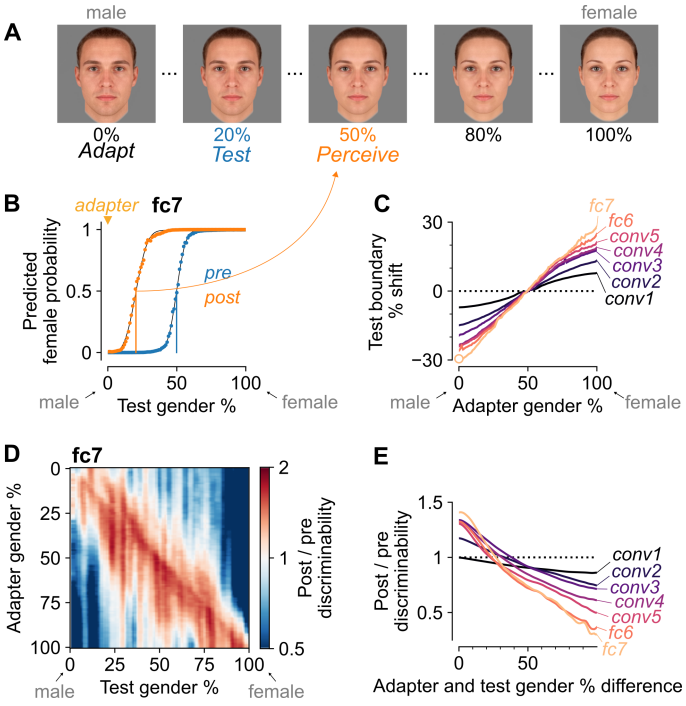

Fig. 3 A neural network incorporating intrinsic suppression produces the perceptual bias and enhanced discriminability of aftereffects. (**A**), Examples of the face-gender morph stimuli used in our simulated experiments. After exposure to a male adapter face, the gender decision boundary shifts towards the adapter and an observer perceives a subsequent test face as more female, and vice versa (36). The example *adapt*, *test*, and *perceive* morph levels were picked based on the estimated boundary shift shown in (B). (**B**), Decision boundaries pre (blue) versus post (orange) exposure to a male (0%) adapter based on the top layer (fc7) of the model with intrinsic suppression. Markers show class probabilities for each test stimulus, full lines indicate the corresponding psychometric functions, and vertical lines denote the classification boundaries. Adaptation to a 0% (male) face leads to a shift in the decision boundary towards male faces, hence perceiving the 20% test stimulus as gender-neutral (50%). (**C**), Decision boundary shifts for the test stimulus as a function of the adapter morph level per layer. The round marker indicates the boundary shift plotted in (B). (**D**), Relative face-gender discriminability (**Materials and Methods**, values > 1 signify increased discriminability and values < 1 decreased discriminability) for fc7 as a function of adapter and test morph level. See color scale on right. The red diagonal indicates that face-gender discriminability is increased for morph levels close to the adapter. (**E**), Average changes in face-gender discriminability per layer as a function of the absolute difference in face-gender morph level between adapter and test stimulus.

stimulus (e.g., a gender-neutral face). Here, we focus on the face-gender dimension, but similar results for the tilt aftereffect (38) with gratings are shown in **Fig. S4**.

To evaluate whether the model can describe perceptual aftereffects, we created a set of face stimuli that morphed from average male to average female, and measured the category decision boundary for each layer of the model pre- and post- adaptation (**Materials and Methods**). Once again, we considered the same model from the previous section without any parameter changes. Exposing the model to an adapter face biased the decision boundary towards the adapter. Before adaptation, the predicted female probabilities for the model fc7 layer showed a typical sigmoidal curve centered around the gender-neutral face stimulus with morph level 50% (**Fig. 3B**, blue). After adapting to a male face with morph level 0%, the decision boundary shifted $\sim$ 30% percentage values towards the gender of the adapter (**Fig. 3B**, orange). **Fig. 3C** shows that for all layers, adaptation to a face stimulus resulted in a boundary shift towards the adapter. Consistent with perceptual aftereffects in human subjects, adapting to the original gender-neutral boundary stimulus with morph level 50% had no effect on the decision boundary (**Fig. 3C**). The perceptual bias did not suddenly emerge in later layers, but slowly built up with increasing layers (**Fig. 3C**, from black to purple to yellow colors), and already occurred within the first layer with intrinsic suppression (**Fig. S8A**).

Although adapting to the boundary stimulus did not shift the decision boundary, it did increase the slope of the psychometric function for fc7 from 0.077 to 0.099 (29%; for layers conv1 - fc6 the slope changes were -3%, 11%, 9%, 12%, 16%, 31%, respectively). An increase in slope signifies a repulsion of more female and more male stimuli away from the adapter. This result is inconsistent with the perceptual renormalization hypothesis, which predicts that adaptation uniformly shifts the norm of the representational space towards the adapter and thus that adapting to the original norm (i.e., the boundary stimulus) should have no effect whatsoever (see Figure 3 of 39). A series of previous experiments has shown that both tilt and face aftereffects involve repulsion rather than renormalization (40), which is consistent with the computational model proposed here.

Besides biasing the perception of a stimulus property, adaptation is also thought to increase sensitivity of the system for small differences from the current prevailing input characteristics, which could serve to maintain good stimulus discriminability (3, 4). In line with this hypothesis, Yang *et al.* (41) found that adapting to a female/male face improved gender discrimination around the face-gender morph level of the adapter. We evaluated whether intrinsic suppression in the model could account for such improved discrimination (**Materials and Methods**). Adaptation in the model indeed enhanced face-gender discriminability at morph levels close to the adapter (red diagonal in **Fig. 3D**), while decreasing discriminability at morph levels different from the adapter (blue). Like the perceptual bias (**Fig. 3C**), and consistent with the results shown in **Fig. 2G,J**, the discriminability effects built up monotonically across successive layers (**Fig. 3E**; see **Fig. S4D-E** for similar results with oriented gratings). Unlike boundary shifts, discriminability enhancements first occurred downstream of the first layer with intrinsic suppression (**Fig. S8B**).

Overall, the proposed model shows that activation-based suppression can account for discriminability improvements close to the adapter without any other specialized mechanisms and without introducing any model changes.

Response enhancement and tuning curve shifts emerge from intrinsic suppression propagating to deeper layers

To better understand the mechanisms underlying perceptual aftereffects, we investigated how adaptation impacts the responses of individual units in the face-gender experiment (see **Fig. S5** for analyses of the tilt aftereffect). **Fig. 4A** shows the pre-adaptation activation of each responsive fc7 unit across the female/male dimension (column 1), and how each unit's activation strength changed as a function of the adapter (columns 2 through 6). The rows in each heatmap are sorted according to the gender selectivity index (SI_g; **Materials and Methods**), ranging from more responsive to male faces ($SI_g < 0$, units shown at the top), to more responsive to female faces ($SI_g > 0$, units shown at the bottom). After adaptation, most units showed an overall suppressed response (blue), regardless of the adapter gender

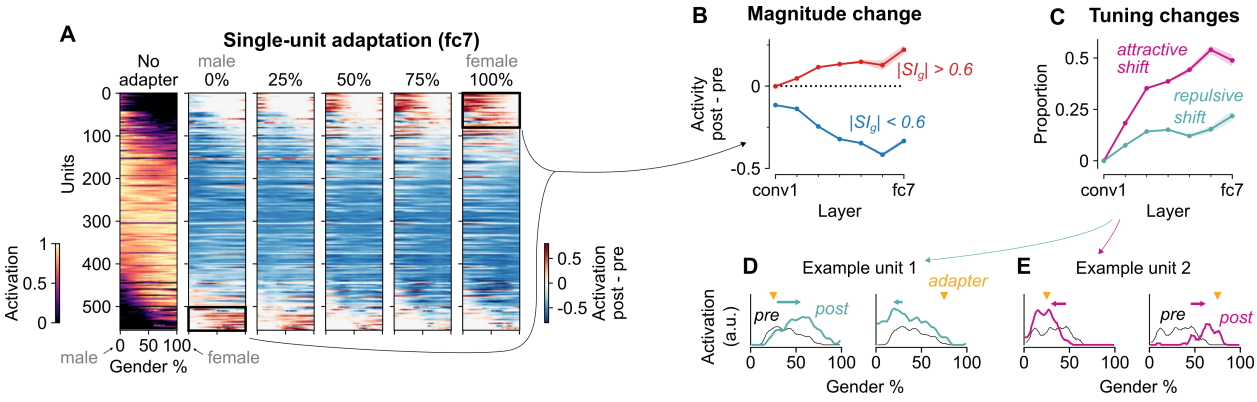

Fig. 4 Response enhancements and tuning shifts emerge in deeper layers of a network incorporating intrinsic suppression. (**A**), Effects of adapting to female/male faces on the activation strength of single units. Left: heatmap showing the activation normalized to the maximum of all 556 responsive fc7 units (rows) for all face-gender morph images (columns). See color scale on bottom left. Rows are sorted according to the gender selectivity index (SI_g; **equation (3)**). The remaining five heatmaps show the difference (post - pre adaptation) in single-unit activations after adapting to five different adapters. See color scale on bottom right. (**B**), Mean response change (activity post - activity pre) across responsive units for each layer (shaded area = 95%CI). For highly gender-selective units (red), the magnitude change (averaged across stimuli) was taken after adapting to a gender stimulus opposite to the unit's preferred gender (0% adapter for $SI_g > 0.6$, 100% adapter for $SI_g < -0.6$; black rectangles in (A)). For less gender-selective units (blue), the magnitude change after both 0% and 100% adapters was used. (**C**), Proportion of adapters causing the preferred morph level to shift towards (attractive, pink) or away (repulsive, green) from the adapter, averaged across units (shaded area = 95% CI). (**D**), An example unit showing a repulsive shift in tuning curves for the 25% (left) and 75% (right) adapters (the y-axes depict activation in arbitrary units; black: pre adaptation tuning curve; green: post adaptation tuning curve; yellow marker: adapter morph level). (**E**), An example unit showing an attractive shift in tuning curves (pink: post adaptation tuning curve; same conventions as (D)).

morph level. However, units with a strong preference for male faces (top rows) showed an enhanced response (red) after neutral to female adapters (columns 3-5), whereas units with a strong preference for female faces (bottom rows) showed the opposite effect (columns 1-3). Thus, highly selective units showed response enhancement after adapting to the opposite of their preferred gender. This response enhancement can be explained by disinhibition (8), where adaptation reduces the inhibitory input for units that prefer morph levels further away from the adapter, much like response enhancements of middle temporal cells for their preferred direction, after adapting to the opposite direction (42).

To quantify and compare this response enhancement for all layers, we considered highly gender-selective units ($|SI_g| > 0.6$) and calculated their response enhancement (averaged across all stimuli) after adapting to the opposite of their preferred gender. **Fig. 4B** shows that the response enhancement for highly selective units (red) emerged in deeper layers (always downstream of the first layer with intrinsic suppression **Fig. S9A**), whereas less selective units mostly showed response suppression (blue) throughout all the layers.

Adaptation can lead to changes in response strength, but it can also cause a shift in the peak of a neuron's tuning curve. For example, in orientation selective neurons, adapting to an oriented grating can produce a shift in the tuning curve's peak either towards the adapter (*attractive shift*(13, 14, 43)), or away from the adapter (*repulsive shift*(13, 18)). Adaptation in the model produced both types of peak shifts in tuning curves (**Fig. 4D**, **E**). For each unit, we calculated the proportion of adapters that produced an attractive shift or a repulsive shift (**Fig. 4C**). Adaptation-induced peak shifts emerged in deeper layers of the network, downstream from the first layer with intrinsic suppression (**Fig. S9B**). Attractive shifts were more common overall, culminating to a proportion of ~ 0.5 in the last layers.

Tuning changes are thought to be necessary for producing perceptual aftereffects. For example, it has been argued that a repulsive perceptual bias, where the decision boundary shifts toward the adapter, requires tuning curves that shift toward the adapter (15, 19). The fact that intrinsic suppression in the model produces mostly attractive shifts (**Fig. 4C**) while also capturing boundary shifts (**Fig. 3C**) seems consistent with this idea. To disentangle

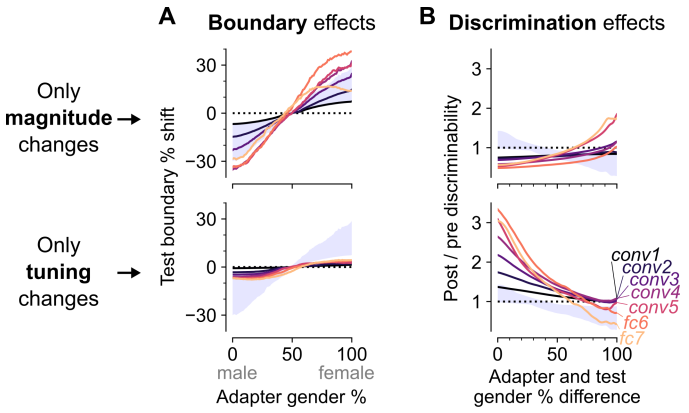

Fig. 5 Response magnitude and tuning changes in the model differentially explain perceptual boundary shifts and discriminability changes. (**A**), Face-gender boundary shifts towards the adapter were produced both by magnitude changes without tuning changes (top) as well as by tuning changes without magnitude changes (bottom). Grey shading indicates the range of original layer effects shown in **Fig. 3C**. (**B**), Face-gender discriminability enhancement for morph-levels close to the adapter was produced by tuning changes without magnitude changes (bottom), but not by magnitude changes without tuning changes (top). Grey shading indicates the range of original layer effects shown in **Fig. 3E**.

the separate contributions of tuning changes and response magnitude changes to the perceptual adaptation effects produced by the model, we manipulated the post-adaptation layer activations to only contain either tuning changes or magnitude changes (**Materials and Methods**; **Fig. 5**). Changes restricted to response magnitude without tuning changes led to even larger boundary shifts than the original model, whereas changes restricted to tuning without any changes in response magnitude led to smaller boundary shifts (**Fig. 5A**). This observation suggests that while the perceptual bias of aftereffects might be the result of a complex interaction between changes in responsivity and tuning, the perceptual bias does not necessarily require attractive shifts as suggested by previous models (15, 19). On the other hand, an increased face-gender discriminability for morph levels close to the adapter did require changes in the tuning response patterns of single-units. Magnitude changes only produced the opposite effect, with increased discriminability for morph levels furthest from the adapter (**Fig. 5B**).

Intrinsic adaptation can be optimized by maximizing recognition performance

Thus far, we have considered a model with an intrinsic adaptation state for each unit, and the adaptation parameters α and β (**equations (1)** and (**2**)) were chosen to impose response suppression. This leaves open the question of whether such adaptation mechanisms can be optimized or learnt in a deep learning framework given a certain task goal. We considered two possible ways in which adaptation could be learned by artificial neural networks: (i) optimize α and β by training a feedforward network with intrinsic adaptation state on a task where adaptation is useful for biological vision; and (ii) train a recurrent network without an intrinsic adaptation state on the same task.

To assess whether adaptation could be learnt and to compare the two possible network mechanisms, we needed a task objective with a suitable goal where adaptation could impact visual performance. As mentioned earlier, one of the proposed computational roles of neural adaptation is to increase sensitivity to small changes in the sensory environment (3, 4). A system could increase sensitivity by decreasing the salience of recently seen stimuli or features (5, 21). Thus, we developed a task where the end goal was object classification, but the objects were hidden in a temporally repeated noise pattern. If adaptation serves to reduce the salience of a recent stimulus, adapting to a noise pattern should increase the ability to recognize a subsequently presented target object embedded in the same noise pattern, and a network trained on this task could learn to reduce the salience of previously presented input. To keep the networks relatively lightweight, we chose a classification task with low resolution hand drawn doodles rather than natural images (**Fig. 6A**).

Before training any network, we evaluated human recognition performance in this task. For this experiment, adaptation to the noise pattern at early levels of processing is likely sufficient to enhance the object information of the doodle. We ran a psychophysics experiment where participants were exposed to an adapter image, and then classified a test image (**Fig. 6B**, **Materials and Methods**). Recognizing the doodles in this task is not trivial:

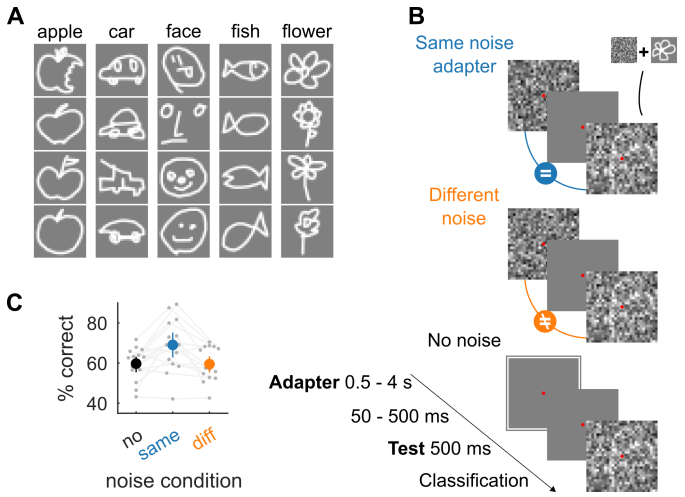

Fig. 6 Adapting to prevailing but interfering input enhances object recognition performance. (**A**), Representative examples for each of the five doodle categories from the total set of 540 selected images. (**B**), Schematic illustration of the conditions used in the doodle experiment. In each trial participants or the model had to classify a hand drawn doodle hidden in noise (test), after adapting to the same (middle), a different (right), or no (left) noise pattern. The trials with different or no noise adapters were control conditions where we expected to see no effect of adaptation. (**C**), Participants showed an increase in categorization performance after adapting to the same noise pattern. Gray circles and lines denote individual participants ($N = 15$). The colored circles show average categorization performance, error bars indicate 95% bootstrap confidence intervals. Chance = 20%.

whereas subjects can readily recognize the doodles in isolation, when they are embedded in noise and in the absence of any adapter, categorization performance was 59.7% ($SD = 8.1\%$) where chance is 20%. As conjectured, adapting to the same noise pattern increased categorization performance by 9.3% (**Fig. 6E**, $p = 0.0043$, Wilcoxon signed rank test, $N = 15$ subjects). This increase in categorization performance was contingent upon the noise pattern presented during the test stimulus being the same as the noise pattern in the adapter: performance in the same-noise condition was 9.6% higher than in the different-noise condition ($p = 0.0015$, Wilcoxon signed rank test, $N = 15$ subjects).

After establishing that adapting to the repeated noise pattern indeed improves the ability to recognize the target objects, we considered whether this behavior could be captured by the model. First, we considered the same model used in previous sections without any tuning. The same pattern of results was captured by the model with α and β fixed to impose activation-based suppression (**Fig. S10**). Next, we asked whether it is feasible to fit intrinsic adaptation parameters α and β in the doodle experiment using recognition performance as the objective. We built a smaller network with an AlexNet-like architecture (**Fig. 7A**, without the recurrent connections shown in blue, which are discussed in the next paragraph; **Materials and Methods**). Each unit (excluding the decoder layer) had an exponentially decaying intrinsic adaptation state as defined by **equations (1)** and (**2**). For simplicity, the trials were presented in three time steps: the adapter, a blank frame, and the test image (**Fig. 7A**). In addition to training the feedforward weights, we simultaneously optimized one α and one β parameter per layer. The value of α determines how fast the intrinsic adaptation state updates, ranging from no update ($\alpha = 1$) to completely renewing at each time step ($\alpha = 0$). The value of β determines whether the intrinsic adaptation state is used for suppression ($\beta > 0$), enhancement ($\beta < 0$) or nothing at all ($\beta = 0$).

After training using 30 random initializations on same-noise trials, the resulting parameters revealed response suppression which was particularly strong for convolutional layers 1 and 2, as indicated by the *positive* high β and low α values (**Fig. 7B**). The average categorization performance on the test set was 97.9% (blue), compared to 74.8% when no

intrinsic adaptation state was included (black; **Fig. 7C**). Thus, when a network with intrinsic adaptation state was trained on an object recognition task with a temporally prevailing but irrelevant input pattern, the resulting adaptation parameters showed activation-based suppression.

A common way to model temporal dynamics in the visual system is by adding recurrent weights to a feedforward network (44–46). Recurrent neural networks can demonstrate phenomena similar to adaptation (47). Recurrent neural networks are the standard architectures used to process input sequences and should be able to perform well in the noisy doodle categorization task. To compare the intrinsic suppression mechanism with a recurrent circuit solution, we considered a network without intrinsic adaptation state and added lateral recurrent connections illustrated in blue in **Fig. 7A** (**Materials and Methods**, **Learning adaptation**). After training on same-noise and different-noise trials, the recurrent architecture achieved the same categorization performance on the test set as the architecture with intrinsic adaptation (**Fig. 7C**). Thus, as expected, the recurrent network performed on par with the network with trained intrinsic adaptation.

Next, we asked whether there are any advantages of implementing adaptation via an intrinsic cellular mechanism versus lateral recurrent network mechanisms. We reasoned that a trained intrinsic suppression mechanism should generalize well across different input features or statistics, whereas the circuit-based solution learned by a recurrent neural network might be less robust. Therefore, we considered situations where the distribution of noise patterns used during training and testing was different. Indeed, the recurrent network failed to generalize well to higher standard deviations of Gaussian noise (**Fig. 7D**), and failed dramatically when tested with uniformly distributed noise (**Fig. 7E**), or Gaussian noise with an offset (**Fig. 7F**). In stark contrast, the intrinsic mechanism generalized well across all of these different input noise changes (**Fig. 7D-F**, magenta). This over-fitting cannot just be explained by a difference in the number of parameters, and also occurs when the number of parameters is equalized between the two networks (**Fig. S11**). Furthermore, depending on the number of parameters, the recurrent network did not necessarily demonstrate the

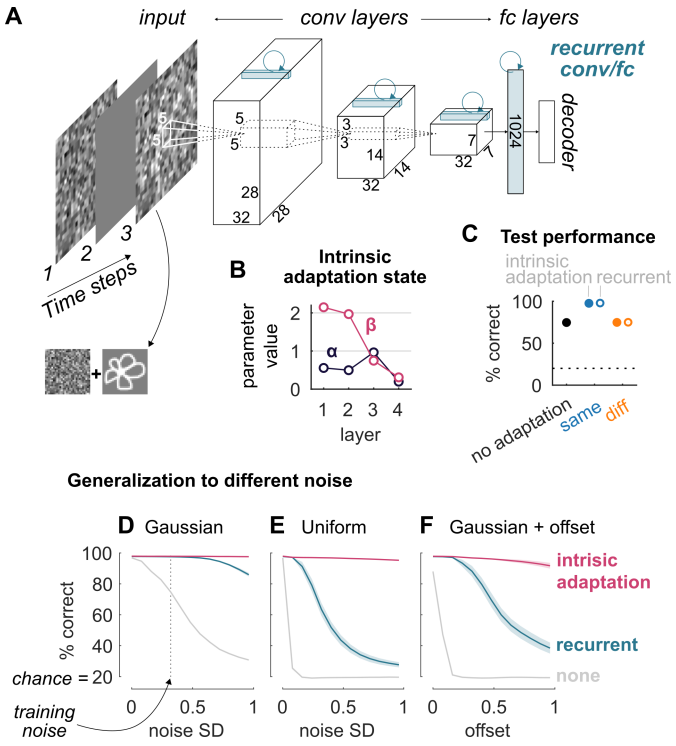

Fig. 7 Intrinsic adaptation can be trained by maximizing recognition performance and is more robust to over-fitting than a recurrent neural network. (**A**), A convolutional neural network with an AlexNet-like feedforward architecture. For the adaptation version, an exponentially decaying hidden state was added to each unit according to **equations (1)** and (**2**) (except for the decoder). For the recurrent version, fully recurrent weights were added for the fully connected layer and convolutional recurrent kernels for the three convolutional layers (see drawings in blue; **Materials and Methods**). (**B**), Average fitted parameters α and β for each layer after training 30 random initializations of the network with intrinsic adaptation state on same noise trials (standard error of the mean bars are smaller than the markers). (**C**), Test categorization performance on trials with the same Gaussian noise distribution as during training. Full markers: average categorization performance after training 30 random initializations on the same noise trials without intrinsic adaptation state (black), after training with intrinsic adaptation state on same noise trials (blue) or on different noise trials (orange). Empty markers: same as full markers but for the recurrent neural network. Standard error of the mean bars are smaller than the markers. Chance = 20%, indicated by horizontal dotted line. (**D-F**), Average generalization performance of the networks with an intrinsic intrinsic adaptation state (pink), recurrent weights (blue), or neither (grey) for same noise trials under noise conditions that differed from training. Performance is plotted as a function of increasing standard deviations (x-axis) of Gaussian noise ((D), the vertical line indicates the SD = 0.32 used during training), of uniform noise (E), or as a function of increasing offset values added to Gaussian noise ((F), SD = 0.32, same as training). Error bounds indicate standard error of the mean.

hallmark property of repetition suppression (**Fig. S12**). In sum, while a recurrent network implementation can learn to solve the same task, the solution is less robust than an intrinsic mechanism to deviations from the particular statistics of the adapter noise used for training the network. These results suggest that intrinsic neuronal mechanisms could provide sensory systems in the brain with a well regularized solution to reduce salience of recent input, which is computationally simple and readily generalizes to novel sensory conditions.

Discussion

We examined whether the paradigmatic neurophysiological and perceptual signatures of adaptation can be explained by a biologically-inspired, activation-based, intrinsic suppression mechanism (7) in a feedforward deep network. The proposed computational model bridges the fundamental levels at which adaptation phenomena have been described: from intrinsic cellular mechanisms, to responses of neurons within a network, to perception. By implementing activation-based suppression (**Fig. 1**), our model exhibited stimulus-specific repetition suppression (4, 5), which recovers over time but also builds up across repeats despite intervening stimuli (48), and increases over stages of processing (12, 49) (**Fig. 2**). Without any fine-tuning of parameters, the same model could explain classical perceptual aftereffects of adaptation (**Fig. 3**), such as the prototypical shift in perceptual bias towards the adapter (36, 38), and enhanced discriminability around the adapter (41, 50), thus suggesting that adaptation modulated the functional state of the network similarly to our visual system. In single units, perceptual aftereffects were associated with changes in overall responsivity (including response enhancements) as well as changes in neural tuning (**Fig. 4** and **5**). In addition, both intrinsic and recurrent circuit adaptation mechanisms can be trained in a task where reducing the salience of repeated but irrelevant input directly impacts recognition performance (**Fig. 6**). However, the recurrent neural network converged on a circuit solution that was less robust to different noise conditions than the proposed model with intrinsic neuronal adaptation (**Fig. 7**). Together, these results show that a neuronally intrinsic

suppression mechanism can robustly account for adaptation effects at the neurophysiological and perceptual levels.

The proposed computational model differs in fundamental ways from previous models of adaptation. Traditionally, adaptation has been modeled using multiple-channel models, where a fixed stimulus dimension such as orientation is encoded by a set of bell-shaped tuning functions (6, 19, 20). The core difference is that here we implemented adaptation in a deep, convolutional neural network model trained on object recognition (35). Even though current convolutional neural networks differ from biological vision in many ways (27), they constitute a reasonable first-order approximation for modeling ventral stream processing, and provide an exciting opportunity for building general and comprehensive models of adaptation. First, in contrast with channel-based models, deep neural networks can operate on any arbitrary image, from simple gratings to complex natural images. Second, the features encoded by the deep neural network model units are not hand-crafted tuning functions restricted to one particular stimulus dimension, but consist of a rich set of increasingly complex features optimized for object recognition, which map reasonably well onto the features encoded by neurons along the primate ventral stream (28–32). A set of bell-shaped tuning curves might be a reasonable approximation of the encoding of oriented gratings in V1, but this scheme might not be appropriate for other visual areas or more complex natural images. Third, the realization that adaptation should be considered in the context of deep networks, where the effects can propagate from one stage of processing to the next (2, 21), calls for complex multi-layer models which can capture the cascading of adaptation. Finally, whereas several models implement adaptation by adjusting recurrent weights between channels (19, 20), we implemented an intrinsic suppression property for each unit and allowed adaptation effects to emerge from the feedforward interactions of differentially adapted units.

The goal was not to fit the model on specific datasets, but to generally capture the phenomenology of adaptation in a model by giving its artificial neurons a biophysically plausible mechanism. The adaptation parameters α and β were not fine-tuned for each simulated experiment, and had the same value for each unit, showing that the ability of the

model to produce adaptation phenomena did nor hinge upon a carefully picked combination of parameters.

By using a feedforward deep neural network as the base for our computational model, we were able to empirically study the role of intrinsic suppression, without any contribution of recurrent interactions. These results should not be interpreted to imply that recurrent computations are irrelevant in adaptation. The results show that complex neural adaptation phenomena readily emerged in deeper layers, arguing that, in principle, they do not need to depend on recurrent mechanisms. Amongst the neural adaptation effects were enhanced responses of single units, as well as shifts in tuning curves, which are often thought to require recurrent network mechanisms (13, 15, 16, 18). Any effect of intrinsic suppression could also be implemented by lateral inhibitory connections in the circuit, leaving open the question of why the brain would prefer one solution over the other. The generalization tests in **Fig. 7** point to an intriguing possibility, which is that intrinsic suppression provides a simpler solution that is more constrained, yet sufficient to implement the goals of adaptation. In contrast, recurrent mechanisms require a complex combination of weights to achieve the same goals and tended to over-fit to the specific training conditions.

There are several functional goals which have been attributed to adaptation. Activation-based suppression could serve to decrease salience of recently seen stimuli or features (5, 21). We successfully exploited this principle to train adaptation in neural networks on a task with temporally repeated but irrelevant noise patterns. Reducing the salience of recently seen features has functional consequences beyond these artificial conditions. By selectively reducing the sensitivity of the system based on previous exposure, adaptation effectively changes the subjective experience of an observer, leading, for example, to a perceptual bias in the face-gender aftereffect. These changes in perception may more broadly reflect mechanisms that serve to maintain perceptual constancy by compensating for variations in the environment (51). The introduction of activation-based, intrinsic suppression to an artificial neural network, made the network subject to the same perceptual biases characterizing perceptual aftereffects in humans (**Fig. 3B,C**), suggesting that intrinsic suppression changed

the model's functional state in a way that is similar to how exposure changes the functional state of our visual system.

Another proposed benefit of reducing sensitivity for recently seen stimuli may be to improve the detection of novel or less frequently occurring stimuli (12, 48). For example, by selectively decreasing responses for more frequent stimuli, adaptation can account for the encoding of object occurrence probability, described in macaque IT (52, 53). Consistent with these observations, intrinsic suppression in the proposed computational model decreased the response strength for a given stimulus proportional to its probability of occurrence (**Fig. 2H-J**). Interestingly, the model also produced stronger responses to a deviant stimulus compared to an equiprobable control condition. Thus, response strength in the model not only captured differences in occurrence probability (standard versus deviant), but also *relative* differences in occurrence probability (control versus deviant): compared to the control condition, the deviant is equally likely in terms of absolute occurrence probability, but it was unexpected merely by virtue of the higher occurrence probability of the standard stimulus.

Adaptation has also been suggested to increase coding efficiency of single neurons by normalizing their responses for the current sensory conditions (4). Neurons have a limited dynamic range with respect to the feature they encode and a limited number of response levels. Adaptation can maximize the information carried by a neuron by re-centering tuning around the prevailing conditions and thus increasing sensitivity and preventing response saturation (51). While AlexNet has ReLU activation functions, which do not suffer from the saturation problem, we did observe an abundance of attractive shifts of tuning curves (**Fig. 4C**). The collective result of these changes in tuning curves was an increased discriminability between stimuli similar to the adapter (**Fig. 4D**), consistent with reports for orientation, motion direction, and face-gender discrimination in humans (41, 50).

Besides direct functional benefits, adaptation may also serve an important role in optimizing the efficiency of the neural population code. Neurons use large amounts of energy to generate action potentials, which constrains neural representations (54). When a particular feature combination is common, the metabolic efficiency of the neural code can be improved

by decorrelating responses of the activated cells and reducing responsiveness. Adaptation has been shown to maintain existing response correlations and equality in time-averaged responses across the population (55), possibly resulting from intrinsic suppression at an earlier cortical stage, which we confirmed by running these experiments in the proposed computational model (**Fig. S13**).

There are several possible extensions to the current model, including the incorporation of multiple time scales and recurrent circuit mechanisms. Adaptation operates over a range of time scales and thus may be best described by a scale-invariant power-law, which could be approximated in the model using a sum of exponential processes (56). More importantly, because we focused on the feedforward propagation of intrinsic suppression, our model did not include any recurrent dynamics. Yet, recurrent connections are abundant in sensory systems, and most likely do contribute to adaptation. There is some evidence suggesting that recurrent mechanisms contribute to adaptation at very short time scales of up to 100 ms (57). During the first 50-100 ms after exposure, adaptation to an oriented grating produces a perceptual boundary shift in the opposite direction of the classical tilt aftereffect (58). Interestingly, this observation was predicted by a recurrent V1 model that only predicted repulsive tuning shifts (6). Repulsive shifts are indeed more common in V1 when each test stimulus is immediately preceded by an adapter (13, 18), whereas adaptation seems to produce mostly attractive shifts at longer gaps (14, 43, 59), consistent with the effects of intrinsic suppression in the proposed model (**Fig. 4**, **Fig. S5**, although repulsive shifts were more common in highly responsive units, **Fig. S6**). These results seem to suggest that recurrent interactions contribute in the first (few) 100 ms, whereas qualitatively different longer adaptation effects might be best accounted for by intrinsic suppression.

The results of the noisy-doodle experiment in humans (**Fig. 6**) could be explained by local light adaptation to the adapter noise patterns. It is unclear where in the visual system such local light adaptation would take place. In principle adaptation could be partly or totally at the level of photoreceptors in the retina. However, given that each noise pixel was only 0.3x0.3 visual degrees and given that luminance was distributed independently

across noise-pixels, inherent variability in the gaze of a fixating subject poses a limit on the contribution of photoreceptor adaptation (60). Most likely, the increased performance observed in the behavioral data results from a combination of adaptation at different stages of processing, including the retina. The proposed computational model does not incorporate adaptation at the receptor level (i.e., pixels), but future models could incorporate adaptation both in the input layer as well as later processing layers.

Overall, the current framework connects systems to cellular neuroscience in one comprehensive multi-level model by including an activation-based, intrinsic suppression mechanism in a deep neural network. Response suppression cascading through a feedforward hierarchical network changed the functional state of the network similar to visual adaptation, producing complex downstream neural adaptation effects as well as perceptual aftereffects. These results demonstrate that intrinsic neural mechanisms may contribute substantially to the dynamics of sensory processing and perception in a temporal context.

Materials and Methods

Computational Models

Implementing intrinsic suppression

We used the AlexNet architecture (35) (**Fig. 1A**), with weights pre-trained on the ImageNet dataset (61) as a model for the ventral visual stream. We implemented an exponentially decaying intrinsic adaptation state (62) to simulate neuronally intrinsic suppression. Specifically, in all layers (except the decoder), each unit had an intrinsic adaptation state s_t, which was updated at each time step t based on its previous state s_{t-1} and the previous response r_{t-1} (i.e. activation after the ReLU rectification and linearization operation):

$$s_t = \alpha s_{t-1} + (1 - \alpha) r_{t-1} \tag{1}$$

where α is a constant in [0,1] determining the time scale of the decay (**Fig. 1B**). This intrinsic adaptation state is then subtracted from the unit's current input x_t (given weights W and bias b) before applying the rectifier activation function σ, so that:

$$r_t = \sigma(b + Wx_t - \beta s_t) \tag{2}$$

where β is a constant that scales the amount of suppression. Thus, strictly speaking, **equation (2)** modifies the bias and thus responsivity of the unit, before applying σ, to avoid negative activations. For $\beta > 0$, these model updating rules result in an exponentially decaying response for constant input which recovers in case of no input (**Fig. 1B**), simulating an activation-based suppression mechanism intrinsic to each individual neuron. Note that $\beta < 0$ would lead to response enhancement and $\beta = 0$ would leave the response unchanged. By implementing this mechanism across discrete time steps in AlexNet, we introduced a temporal dimension to the network (**Fig. 1C**). This model was implemented using TensorFlow v1.11 in Python. Throughout the paper, we use $\alpha = 0.96$ and $\beta = 0.7$ unless indicated otherwise (in **Fig. 7**, those parameters are tuned).

Decision boundaries

Perceptual aftereffects are typically measured by computing shifts in the decision boundary along a stimulus dimension. We evaluated boundary shifts in the model using a set of face stimuli that morphed from average male to average female in 100 steps (using Webmorph, https://webmorph.org/), and measured category decision boundaries pre- and post- adaptation using the 101 face-morph images **Fig. 3A-C**. The experiments were simulated by exposing the model to an adapter image for 100 time steps, followed by a gap of uniform grey input for 10 time steps before presenting the test image. The results were qualitatively similar when the number of time steps was changed.

To measure the pre- and post-adaptation decision boundaries for a given layer, we trained a logistic regression classifier to discriminate between male and female faces using the pre-

adaptation activations of responsive units for the full stimulus set. After training, the classifier can output female/male class probability estimates for any given activation pattern. Thus, we used the trained classifier to provide female/male probability estimates for each morph level, based on either the pre- or post- adaptation activation patterns. The decision boundary is then given by the morph level associated with a female/male class probability of $P = 0.5$, which was estimated by fitting a psychometric function on the class probabilities (average R^2 of at least 0.99 per layer).

Face-gender discriminability

To assess model changes in face-gender discriminability in **Fig. 3J**, we calculated the stimulus discriminability at each morph level of the stimulus dimension before and after adaptation. An increased discriminability between morph levels can be conceptualized as an increased perceived change in morph levels with respect to a certain physical change in morph level. Thus, to quantify discriminability, a linear mapping was fit to predict stimulus morph levels from pre-adaptation unit activations using partial least squares regression (using 4 components). We then used this linear mapping to predict morph levels from activation patterns pre- and post-adaptation. If adaptation increases discriminability, then the change in model-estimated morph level y with respect to a physical change in morph level m, should also increase. Thus, to quantify the change in discriminability at morph level m, we calculated the absolute derivative of the predicted post-adaptation morph-level (y_m^{post}), normalized by the absolute derivative of the predicted pre-adaptation morph-level (y_m^{pre}): $|\Delta y_m^{post}|/|\Delta y_m^{pre}|$.

Selectively retaining tuning or magnitude changes

For **Fig. 4B** we manipulated the post-adaptation layer activations to only contain either tuning changes or magnitude changes. To retain only tuning changes, we started with the post-adaptation activation patterns and multiplied the activation of each unit by a constant so that the resulting mean activation matched the pre-adaptation mean value. On the other hand, to retain only magnitude changes, we started with the pre-adaptation activation

patterns and multiplied the activation of each unit by a constant so that the resulting mean activation matched the post-adaptation mean value.

Learning adaptation

In **Fig. 7** we present two models where adaptation is learnt for the noisy doodle classification task: a model with intrinsic adaptation state and a recurrent neural network model. The base feedforward part of the model was based on the AlexNet architecture (35) for the two networks, consisting of three convolutional layers and a fully connected layer followed by a fully connected decoder. The first convolutional layer filters a $28 \times 28 \times 1$ input image with 32 kernels of size $5 \times 5 \times 1$ with a stride of 1 pixel. The second convolutional layer filters the pooled (kernel $= 2 \times 2$, stride $= 2$) output of the first convolutional layer with 32 kernels of size $5 \times 5 \times 32$ (stride $= 1$). The third convolutional layer filters the pooled (kernel $= 2 \times 2$, stride $= 2$) output of the second convolutional layer with 32 kernels of size $3 \times 3 \times 32$ (stride $= 1$). The fully connected layer has 1024 units that process the output of the third convolutional layer with 50% dropout during training.

The recurrent version was extended with lateral recurrent weights. For convolutional layers, lateral recurrence was implemented as 32 kernels of size $1 \times 1 \times 32$ (stride $= 1$), which filtered the non-pooled outputs of the layer at time step $t - 1$ (after ReLu) and were added to the feedforward-filtered inputs of the same layer at time step t (before ReLu). The fully connected layer was recurrent in an all-to-all fashion.

The intrinsic adaptation version was extended with adaptation states, as described in **Materials and Methods**, **Implementing intrinsic suppression**, of which the α and β parameters were now also trained using back-propagation. The β parameters were initialized at 0 (i.e. no adaptation) and the α parameters were initialized using a uniform distribution ranging from 0 to 1.

Both the recurrent and intrinsic adaptation models were trained on the doodle classification task using TensorFlow v1.11 in Python. We used a training set of $500,000$ doodle images ($100,000$ per category; https://github.com/googlecreativelab/quickdraw-dataset),

with a separate set of $1,000$ images to select hyperparameters and evaluate the loss and accuracy during training. We used the Adam optimization algorithm (63) with a learning rate of 0.001, the sparse softmax cross entropy between logits and labels cost function, a batch size of 100, and 50% training dropout in fully connected layers. For the weights, we used Gaussian initialization, with the scale correction proposed by (64). Each model was trained for 5 epochs on the training set, which was sufficient for the loss and accuracy to saturate. Generalization performance was then tested on a third independent set of 5,000 images.

Neurophysiology

We present neurophysiological data from two previously published studies in order to compare them with the neural adaptation effects of the proposed computational model: single cell recordings from inferior temporal (IT; $N = 97$) cortex of one macaque monkey G (37) and multi-unit recordings from primary visual cortex (V1; $N = 55$) and latero-intermediate visual area (LI; $N = 48$) of three rats (12). For methodological details about the recordings and the tasks, we refer to the original papers.

Psychophysics

Before starting the data collection, we preregistered the study design and hypothesis on the Open Science Framework at `https://osf.io/tdb37/` where all the source code and data can be retrieved.

Participants

A total of 17 volunteers (10 female, ages 19-50) participated in our doodle categorization experiments (**Fig. 6**). In accordance with our preregistered data exclusion rule, two male participants were excluded from analyses because we could not record eye tracking data. All subjects gave informed consent and the studies were approved by the Institutional Review

Board at Children's Hospital, Harvard Medical School.

Stimuli

The stimulus set consisted of hand drawn doodles of apples, cars, faces, fish, and flowers from the *Quick, Draw!* dataset (https://github.com/googlecreativelab/quickdraw-dataset). We selected a total of 540 doodles (108 from each of the five categories) that were judged complete and identifiable. We lowered the contrast of each doodle image (28x28 pixels) to either 22 or 29% of the original contrast, before adding a Gaussian noise pattern (SD = 0.165 in normalized pixel values) of the same resolution. The higher contrast level (29%) was chosen as a control so that the doodle was relatively visible in 1/6 of the trials, and was not included in the analyses. The average categorization performance on these high contrast trials was 74% ($SD = 8.3\%$), versus 63% ($SD = 8.9\%$) in the low contrast trials.

Experimental protocol

Participants had to fixate a cross at the center of the screen in order to start a trial. Next, an adapter image was presented (for 0.5, 2, or 4 s), followed by a blank interval (of 50, 250, or 500 ms), a test image (for 500 ms), and finally a response prompt screen. The test images were noisy doodles described in the above paragraph. The adapter image could either be: an empty frame (defined by a white square filled with the background color), the same mosaic noise pattern as the one of the subsequent test image, or a randomly generated different noise pattern (**Fig. 6**). Participants were asked to keep looking at the fixation cross, which remained visible throughout the entire trial, until they were prompted to classify the test image using keyboard keys 1-5. All images were presented at 9 x 9° from a viewing distance of approximately 52 cm on a 19 inch CRT monitor (Sony Multiscan G520, 1024 × 1280 resolution), while we continuously tracked eye movements using a video-based eye tracker (EyeLink 1000, SR Research, Canada). Trials where the root-mean-square deviation of the eye-movements exceeded 1 degree of visual angle during adapter presentation were excluded from further analyses. The experiment was controlled by custom code written in MATLAB

using Psychophysics Toolbox Version 3.0 (65).

Data Analysis

Selectivity index

For the face-gender experiments we calculated a selectivity index based on the average activation of a unit to male (morph level $< 50\%$) and female (morph level $> 50\%$) faces:

$$SI_g = (A_M - A_F)/(A_M + A_F) \tag{3}$$

A value > 0 indicates stronger activation for male faces and a value < 0 stronger activation for female faces.

Acknowledgements

This work was supported by Research Foundation Flanders, Belgium (fellowship of K.V.), by NIH grant R01EY026025 and by the Center for Brains, Minds and Machines, funded by NSF Science and Technology Centers Award CCF-1231216. We thank Thomas P. O'Connell for helpful discussions on this work.

Contributions

K.V. conceived the model and experiment; K.V., X.B., and G.K. designed the model and experiment; K.V. collected the data, implemented the model, and carried out analyses; K.V. and G.K. wrote the manuscript, with contributions from X.B.

Competing interests

The authors declare that they have no competing interests.

Data availability statement

All the psychophysics data and source code are available at https://osf.io/tdb37/

References

[1] R Addams. An account of a peculiar optical phenomenon seen after having looked at a moving body. *London and Edinburgh Philosophical Magazine and Journal of Science*, 5:373–374, 1834.

[2] Clarissa J. Whitmire and Garrett B. Stanley. Rapid Sensory Adaptation Redux: A Circuit Perspective. *Neuron*, 92(2):298–315, 2016. ISSN 08966273. doi: 10.1016/j.neuron.2016.09.046. URL http://linkinghub.elsevier.com/retrieve/pii/S0896627316306511.

[3] Colin W G Clifford, Michael A. Webster, Garrett B. Stanley, Alan A. Stocker, Adam Kohn, Tatyana O. Sharpee, and Odelia Schwartz. Visual adaptation: Neural, psychological and computational aspects. *Vision Research*, 47(25):3125–3131, 2007. ISSN 00426989. doi: 10.1016/j.visres.2007.08.023.

[4] Adam Kohn. Visual Adaptation: Physiology, Mechanisms, and Functional Benefits. *Journal of Neurophysiology*, 10461:3155–3164, 2007. ISSN 0022-3077. doi: 10.1152/jn.00086.2007.

[5] Rufin Vogels. Sources of adaptation of inferior temporal cortical responses. *Cortex*, 80: 185–195, 7 2016. ISSN 00109452. doi: 10.1016/j.cortex.2015.08.024. URL http://dx.doi.org/10.1016/j.cortex.2015.08.024http://linkinghub.elsevier.com/retrieve/pii/S0010945215003342https://linkinghub.elsevier.com/retrieve/pii/S0010945215003342.

[6] Mar Quiroga, Adam P Morris, and Bart Krekelberg. Article Adaptation without Plasticity. *Cell Reports*, 17(1):58–68, 2016. ISSN 2211-1247. doi:

10.1016/j.celrep.2016.08.089. URL http://dx.doi.org/10.1016/j.celrep.2016.08.089.

[7] M V Sanchez-Vives, L G Nowak, and David A McCormick. Membrane mechanisms underlying contrast adaptation in cat area 17 in vivo. *Journal of Neuroscience*, 20(11): 4267–4285, 2000. ISSN 1529-2401. doi: 20/11/4267[pii]. URL http://www.ncbi.nlm.nih.gov/pubmed/10818163.

[8] Bart Krekelberg, Geoffrey M Boynton, and Richard J a van Wezel. Adaptation: from single cells to BOLD signals. *Trends in neurosciences*, 29(5):250–6, 5 2006. ISSN 0166-2236. doi: 10.1016/j.tins.2006.02.008. URL http://www.ncbi.nlm.nih.gov/pubmed/16529826.

[9] Hiromasa Sawamura, Guy A. Orban, and Rufin Vogels. Selectivity of neuronal adaptation does not match response selectivity: a single-cell study of the FMRI adaptation paradigm. *Neuron*, 49(2):307–318, 1 2006. ISSN 0896-6273. doi: 10.1016/j.neuron.2005.11.028. URL http://www.ncbi.nlm.nih.gov/pubmed/16423703.

[10] Dzmitry A. Kaliukhovich and Rufin Vogels. Divisive Normalization Predicts Adaptation-Induced Response Changes in Macaque Inferior Temporal Cortex. *The Journal of Neuroscience*, 36(22):6116–6128, 2016. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.2011-15.2016. URL http://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.2011-15.2016.

[11] Stephanie C Wissig and Adam Kohn. The influence of surround suppression on adaptation effects in primary visual cortex. *Journal of Neurophysiology*, 107(12): 3370–3384, 6 2012. ISSN 1522-1598. doi: 10.1152/jn.00739.2011. URL http://www.ncbi.nlm.nih.gov/pubmed/22423001.

[12] Kasper Vinken, Rufin Vogels, and Hans Op de Beeck. Recent Visual Experience Shapes Visual Processing in Rats through Stimulus-Specific Adaptation and Response

Enhancement. *Current Biology*, 27(6):914–919, 3 2017. ISSN 09609822. doi: 10.1016/j.cub.2017.02.024. URL http://dx.doi.org/10.1016/j.cub.2017.02.024http://linkinghub.elsevier.com/retrieve/pii/S0960982217301616.

[13] Valentin Dragoi, Jitendra Sharma, and Mriganka Sur. Adaptation-induced plasticity of orientation tuning in adult visual cortex. *Neuron*, 28(1):287–298, 2000. ISSN 08966273. doi: 10.1016/S0896-6273(00)00103-3.

[14] Jeyadarshan Jeyabalaratnam, Vishal Bharmauria, Lyes Bachatene, Sarah Cattan, Annie Angers, and Stéphane Molotchnikoff. Adaptation shifts preferred orientation of tuning curve in the mouse visual cortex. *PloS one*, 8(5):e64294, 1 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0064294. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3662720&tool=pmcentrez&rendertype=abstract.

[15] Adam Kohn and J. Anthony Movshon. Adaptation changes the direction tuning of macaque MT neurons. *Nature Neuroscience*, 7(7):764–772, 2004. ISSN 10976256. doi: 10.1038/nn1267.

[16] A. B. Saul and M. S. Cynader. Adaptation in single units in visual cortex: The tuning of aftereffects in the spatial domain. *Visual Neuroscience*, 2(6):593–607, 6 1989. ISSN 0952-5238. doi: 10.1017/S0952523800003527. URL https://www.cambridge.org/core/product/identifier/S0952523800003539/type/journal_articlehttps://www.cambridge.org/core/product/identifier/S0952523800003527/type/journal_article.

[17] A. B. Saul and M. S. Cynader. Adaptation in single units in visual cortex: The tuning of aftereffects in the temporal domain. *Visual Neuroscience*, 2(6):609–620, 6 1989. ISSN 0952-5238. doi: 10.1017/S0952523800003539. URL https://www.cambridge.org/core/product/identifier/S0952523800003539/type/journal_article.

[18] Gidon Felsen, Yao Song Shen, Haishan Yao, Gareth Spor, Chaoyi Li, and Yang Dan. Dynamic modification of cortical orientation tuning mediated by recurrent connections. *Neuron*, 36(5):945–954, 2002. ISSN 08966273. doi: 10.1016/S0896-6273(02)01011-5.

[19] Andrew F. Teich and Ning Qian. Learning and adaptation in a recurrent model of V1 orientation selectivity. *Journal of Neurophysiology*, 89(4):2086–2100, 2003. ISSN 00223077. doi: 10.1152/jn.00970.2002.

[20] Zachary M. Westrick, David J. Heeger, and Michael S. Landy. Pattern adaptation and normalization reweighting. *Journal of Neuroscience*, 36(38):9805–9816, 2016. ISSN 15292401. doi: 10.1523/JNEUROSCI.1067-16.2016.

[21] Samuel G. Solomon and Adam Kohn. Moving Sensory Adaptation beyond Suppressive Effects in Single Neurons. *Current Biology*, 24(20):R1012–R1022, 10 2014. ISSN 09609822. doi: 10.1016/j.cub.2014.09.001. URL http://linkinghub.elsevier.com/retrieve/pii/S0960982214011166.

[22] Neel T. Dhruv and Matteo Carandini. Cascaded Effects of Spatial Adaptation in the Early Visual System. *Neuron*, 81(3):529–535, 2014. ISSN 08966273. doi: 10.1016/j.neuron.2013.11.025. URL http://dx.doi.org/10.1016/j.neuron.2013.11.025.

[23] Daniel L K Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, 2016. ISSN 1097-6256. doi: 10.1038/nn.4244. URL http://www.nature.com/doifinder/10.1038/nn.4244.

[24] Blake A. Richards, Timothy P. Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, Rui Ponte Costa, Archy de Berker, Surya Ganguli, Colleen J. Gillon, Danijar Hafner, Adam Kepecs, Nikolaus Kriegeskorte, Peter Latham, Grace W. Lindsay, Kenneth D. Miller, Richard Naud,

Christopher C. Pack, Panayiota Poirazi, Pieter Roelfsema, João Sacramento, Andrew Saxe, Benjamin Scellier, Anna C. Schapiro, Walter Senn, Greg Wayne, Daniel Yamins, Friedemann Zenke, Joel Zylberberg, Denis Therien, and Konrad P. Kording. A deep learning framework for neuroscience. *Nature Neuroscience*, 22(11):1761–1770, 2019. ISSN 15461726. doi: 10.1038/s41593-019-0520-2. URL http://dx.doi.org/10.1038/s41593-019-0520-2.

[25] Tim C. Kietzmann, Patrick McClure, and Nikolaus Kriegeskorte. Deep Neural Networks in Computational Neuroscience. In *Oxford Research Encyclopedia of Neuroscience*, pages 1–28. Oxford University Press, 1 2019. ISBN 9780190264086. doi: 10.1093/acrefore/9780190264086.013.46. URL https://oxfordre.com/neuroscience/view/10.1093/acrefore/9780190264086.001.0001/acrefore-9780190264086-e-46.

[26] Radoslaw M. Cichy and Daniel Kaiser. Deep Neural Networks as Scientific Models. *Trends in Cognitive Sciences*, 23(4):305–317, 2019. ISSN 1879307X. doi: 10.1016/j.tics.2019.01.009.

[27] Thomas Serre. Deep Learning: The Good, the Bad, and the Ugly. *Annual Review of Vision Science*, 5(1):399–426, 9 2019. ISSN 2374-4642. doi: 10.1146/annurev-vision-091718-014951. URL https://www.annualreviews.org/doi/10.1146/annurev-vision-091718-014951.

[28] Charles F. Cadieu, Ha Hong, Daniel L K Yamins, Nicolas Pinto, Diego Ardila, Ethan A. Solomon, Najib J. Majaj, and James J. DiCarlo. Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLoS Computational Biology*, 10(12), 2014. ISSN 15537358. doi: 10.1371/journal.pcbi.1003963.

[29] D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher

visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014. ISSN 0027-8424. doi: 10.1073/pnas.1403112111. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1403112111.

[30] Ioannis Kalfas, Satwant Kumar, and Rufin Vogels. Shape Selectivity of Middle Superior Temporal Sulcus Body Patch Neurons. *Eneuro*, 4(June):0113–17, 2017. ISSN 2373-2822. doi: 10.1523/ENEURO.0113-17.2017. URL http://eneuro.sfn.org/lookup/doi/10.1523/ENEURO.0113-17.2017.

[31] Ioannis Kalfas, Kasper Vinken, and Rufin Vogels. Representations of regular and irregular shapes by deep Convolutional Neural Networks, monkey inferotemporal neurons and human judgments. *PLOS Computational Biology*, 14(10):e1006557, 2018. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1006557. URL http://dx.plos.org/10.1371/journal.pcbi.1006557.

[32] Dean A. Pospisil, Anitha Pasupathy, and Wyeth Bair. 'Artiphysiology' reveals V4-like shape tuning in a deep network trained for image classification. *eLife*, 7:1–31, 2018. ISSN 2050084X. doi: 10.7554/eLife.38242.

[33] Seyed Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Computational Biology*, 10(11), 2014. ISSN 15537358. doi: 10.1371/journal.pcbi.1003915.

[34] Jonas Kubilius, Stefania Bracci, and Hans P. Op de Beeck. Deep Neural Networks as a Computational Model for Human Shape Sensitivity. *PLoS Computational Biology*, 12 (4):1–26, 2016. ISSN 15537358. doi: 10.1371/journal.pcbi.1004896.

[35] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Advances In Neural Information Processing Systems*, pages 1–9, 2012. ISSN 10495258. doi: http://dx.doi.org/10.1016/j.protcy.2014.09.007.

[36] Michael A. Webster, Daniel Kaping, Yoko Mizokami, and Paul Duhamel. Adaptation to natural facial categories. *Nature*, 428(6982):557–561, 4 2004. ISSN 0028-0836. doi: 10.1038/nature02420. URL http://www.nature.com/doifinder/10.1038/nature02420.

[37] Kasper Vinken, Hans P. Op de Beeck, and Rufin Vogels. Face Repetition Probability Does Not Affect Repetition Suppression in Macaque Inferotemporal Cortex. *The Journal of Neuroscience*, 38(34):7492–7504, 8 2018. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.0462-18.2018. URL http://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.0462-18.2018.

[38] J. J. Gibson and Minnie Radner. Adaptation, after-effect and contrast in the perception of tilted lines. I. Quantitative studies. *Journal of Experimental Psychology*, 20(5):453–467, 1937. ISSN 0022-1015. doi: 10.1037/h0059826. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/h0059826.

[39] Michael A. Webster and Donald I.A. Macleod. Visual adaptation and face perception. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1571): 1702–1725, 2011. ISSN 09628436. doi: 10.1098/rstb.2010.0360.

[40] Katherine R. Storrs and Derek H. Arnold. Face aftereffects involve local repulsion, not renormalization. *Journal of Vision*, 15(8):1–18, 2015. ISSN 15347362. doi: 10.1167/15.8.1.

[41] Hua Yang, Jianhong Shen, Juan Chen, and Fang Fang. Face adaptation improves gender discrimination. *Vision Research*, 51(1):105–110, 2011. ISSN 00426989. doi: 10.1016/j.visres.2010.10.006. URL http://dx.doi.org/10.1016/j.visres.2010.10.006.

[42] Steven E. Petersen, James F. Baker, and John M. Allman. Direction-specific adaptation in area MT of the owl monkey. *Brain Research*, 346(1):146–150, 1985. ISSN 00068993. doi: 10.1016/0006-8993(85)91105-9.

[43] N. Ghisovan, A. Nemri, S. Shumikhina, and S. Molotchnikoff. Long adaptation reveals mostly attractive shifts of orientation tuning in cat primary visual cortex. *Neuroscience*, 164(3):1274–1283, 2009. ISSN 03064522. doi: 10.1016/j.neuroscience.2009.09.003. URL http://dx.doi.org/10.1016/j.neuroscience.2009.09.003.

[44] Hanlin Tang, Martin Schrimpf, William Lotter, Charlotte Moerman, Ana Paredes, Josue Ortega Caro, Walter Hardesty, David Cox, and Gabriel Kreiman. Recurrent computations for visual pattern completion. *Proceedings of the National Academy of Sciences*, page 201719397, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1719397115. URL http://www.pnas.org/lookup/doi/10.1073/pnas.1719397115.

[45] Kohitij Kar, Jonas Kubilius, Kailyn Schmidt, Elias B Issa, and James J. DiCarlo. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*, page 354753, 4 2019. ISSN 1097-6256. doi: 10.1038/s41593-019-0392-5. URL http://dx.doi.org/10.1038/s41593-019-0392-5http://www.nature.com/articles/s41593-019-0392-5https://www.biorxiv.org/content/10.1101/354753v1.

[46] Tim C Kietzmann, Courtney J Spoerer, Lynn K A Sörensen, Radoslaw M Cichy, Olaf Hauk, and Nikolaus Kriegeskorte. Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, 116(43):21854–21863, 10 2019. ISSN 0027-8424. doi: 10.1073/pnas.1905544116. URL http://www.pnas.org/lookup/doi/10.1073/pnas.1905544116.

[47] William Lotter, Gabriel Kreiman, and David Cox. A neural network trained for prediction mimics diverse features of biological neurons and perception. *Nature Machine Intelligence*, 2(4):210–219, 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-0170-9. URL http://dx.doi.org/10.1038/s42256-020-0170-9.

[48] Nachum Ulanovsky, Liora Las, and Israel Nelken. Processing of low-probability sounds by cortical neurons. *Nature Neuroscience*, 6(4):391–398, 4 2003. ISSN 1097-6256. doi: 10.1038/nn1032. URL http://www.ncbi.nlm.nih.gov/pubmed/12652303http://www.nature.com/articles/nn1032.

[49] Dzmitry A. Kaliukhovich and Hans Op de Beeck. Hierarchical stimulus processing in rodent primary and lateral visual cortex as assessed through neuronal selectivity and repetition suppression. *Journal of Neurophysiology*, 120(3):926–941, 2018. ISSN 0022-3077. doi: 10.1152/jn.00673.2017.

[50] Colin W.G. Clifford. Perceptual adaptation: Motion parallels orientation. *Trends in Cognitive Sciences*, 6(3):136–143, 2002. ISSN 13646613. doi: 10.1016/S1364-6613(00)01856-8.

[51] Michael A. Webster, John S. Werner, and David J. Field. Adaptation and the Phenomenology of Perception. In *Fitting the Mind to the WorldAdaptation and After-Effects in High-Level Vision*, chapter 10, pages 241–278. Oxford University Press, 5 2005. ISBN 9780199251841. doi: 10.1093/acprof:oso/9780198529699.003.0010. URL http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780198529699.001.0001/acprof-9780198529699-chapter-10.

[52] Andrew H. Bell, Christopher Summerfield, Elyse L. Morin, Nicholas J. Malecek, and Leslie G. Ungerleider. Encoding of Stimulus Probability in Macaque Inferior Temporal Cortex. *Current Biology*, 26(17):2280–2290, 9 2016. ISSN 09609822. doi: 10.1016/j.cub.2016.07.007. URL http://dx.doi.org/10.1016/j.cub.2016.07.007http://linkinghub.elsevier.com/retrieve/pii/S0960982216307588.

[53] Kasper Vinken and Rufin Vogels. Adaptation can explain evidence for encoding of probabilistic information in macaque inferior temporal cortex. *Current Biology*, 27 (22):R1210–R1212, 11 2017. ISSN 09609822. doi: 10.1016/j.cub.2017.09.018. URL

http://linkinghub.elsevier.com/retrieve/pii/S096098221731182Xhttps://linkinghub.elsevier.com/retrieve/pii/S096098221731182X.

[54] Peter Lennie. The Cost of Cortical Computation. *Current Biology*, 13(6):493–497, 3 2003. ISSN 09609822. doi: 10.1016/S0960-9822(03)00135-0. URL https://linkinghub.elsevier.com/retrieve/pii/S0960982203001350.

[55] Andrea Benucci, Aman B Saleem, and Matteo Carandini. Adaptation maintains population homeostasis in primary visual cortex. *Nat Neurosci*, 16(6):724–729, 2013. ISSN 1546-1726. doi: 10.1038/nn.3382. URL http://dx.doi.org/10.1038/nn.3382.

[56] Patrick J Drew and L F Abbott. Models and Properties of Power-Law Adaptation in Neural Systems. *Journal of Neurophysiology*, 96(2):826–833, 8 2006. ISSN 0022-3077. doi: 10.1152/jn.00134.2006. URL http://www.physiology.org/doi/10.1152/jn.00134.2006.

[57] Michael Wehr and Anthony M. Zador. Synaptic mechanisms of forward suppression in rat auditory cortex. *Neuron*, 47(3):437–445, 2005. ISSN 08966273. doi: 10.1016/j.neuron.2005.06.009.

[58] Maria del Mar Quiroga, Adam P. Morris, and Bart Krekelberg. Short-Term Attractive Tilt Aftereffects Predicted by a Recurrent Network Model of Primary Visual Cortex. *Frontiers in Systems Neuroscience*, 13(November):1–14, 2019. ISSN 16625137. doi: 10.3389/fnsys.2019.00067.

[59] Vishal Bharmauria, Lyes Bachatene, and Stéphane Molotchnikoff. The speed of neuronal adaptation: A perspective through the visual cortex. *European Journal of Neuroscience*, 49(10):1215–1219, 2019. ISSN 14609568. doi: 10.1111/ejn.14393.

[60] Michele Rucci and Martina Poletti. Control and Functions of Fixational Eye Movements. *Annual review of vision science*, 1:499–518, 2015. ISSN 23744642. doi: 10.1146/annurev-vision-082114-035742.

[61] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. ISSN 15731405. doi: 10.1007/s11263-015-0816-y.

[62] Guillaume Bellec, Darjan Salaj, Anand Subramoney, Robert Legenstein, and Wolfgang Maass. Long short-term memory and learning-to-learn in networks of spiking neurons. *Conference on Neural Information Processing Systems*, 3 2018. doi: arXiv:1803.09574. URL http://arxiv.org/abs/1803.09574.

[63] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, pages 1–15, 12 2014. URL http://arxiv.org/abs/1412.6980.

[64] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the 13th International Conference On Artificial Intelligence and Statistics*, 9:249–256, 2010.

[65] David H. Brainard. The Psychophysics Toolbox. *Spatial Vision*, 10(4):433–436, 1997. ISSN 01691015. doi: 10.1163/156856897X00357.

Supplementary Materials

1 In-depth investigation of repetition suppression in the proposed computational model (Fig. S1 to S3)

In this section we take a more in-depth look at adaptation/repetition suppression in the proposed computational model. **Fig. S1** shows stimulus specific repetition suppression, which is discussed in **Fig. 2** of the main text, for each layer of the network and for both computer generated face stimuli as well as natural images from the ImageNet dataset (61). In **Fig. S2** we investigate the relation between the amount of suppression and the similarity of the stimulus representations in each layer of the network. Finally, in **Fig. S3** we demonstrate the existence of stimulus-specific adaptation in single units, which cannot be explained exclusively by the activation strength of that unit for the preceding stimulus

In **Fig. S1** we show that the difference between repetition trials and alternation trials in the proposed computational model was larger for stimuli from ImageNet compared to the face stimuli used in the neural recordings. This observation is consistent with the idea that the face stimuli were too similar for the model to display the full range of adaptation effects, which were larger in neural recordings (see **Fig. 2**). However, these neural responses were recorded in a patch of cortex where almost all neurons show significantly stronger responses to a set of face images compared to object images in a localizer experiment (37). To test whether this bias towards face selective units could explain the stronger stimulus specific effect in the neural data, we passed the same localizer images through the proposed computational model and selected only those units that showed on average a substantially larger response to the face images (R_{face}) compared to the object images (R_{object}). Face selectivity was quantified using a face selectivity index: $FSI = (R_{face} - R_{object})/(R_{face} + R_{object})$. Overall, face selective units ($FSI > 0$) did not show a larger stimulus specific effect for face stimuli compared to the other units. For example, even with highly selective conv5 units ($FSI > 0.9$; $N = 2,777$), the average alternation-repetition difference for the test stimulus was 0.06, SD=0.10 (normalized response values), compared to 0.07, SD=0.12 when all units were considered ($N = 43,264$).

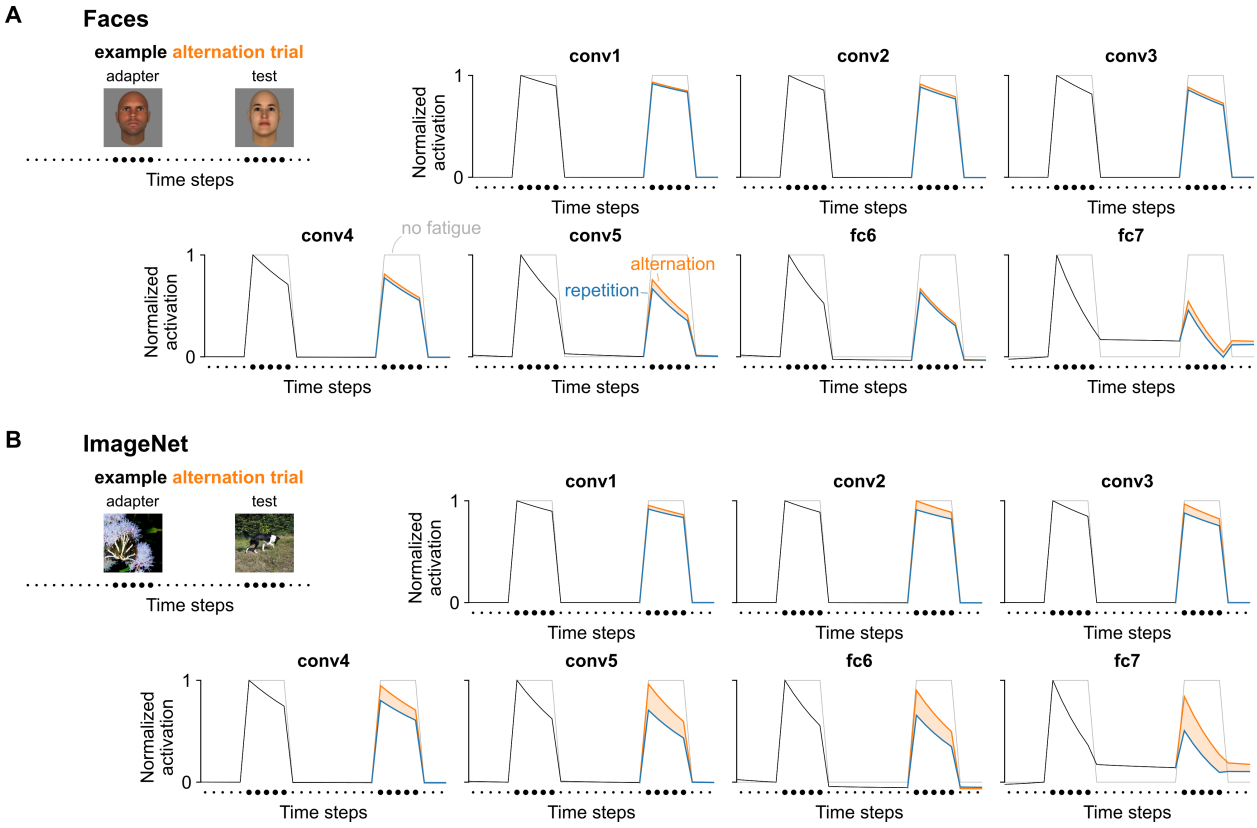

Fig. S1 Stimulus-specific repetition suppression strength varies across model layers and stimulus sets (expanding on Fig. 2). (**A**), Population stimulus-specific repetition suppression in the proposed computational model for a random sub-sample of 500 face pairs (out of 25,000 used in (37)). Adapter and test images were presented for five time steps each (large dots in example alternation trial), preceded by ten time steps of blank (uniform grey) input (small dots). For each trial the network started in an unadapted state. Black: average activity after ReLU across all units and all stimuli in each layer before the presentation of the second stimulus. Blue (repetition): average activity during and after a repeated presentation of the first stimulus. Orange (alternation): average activity during and after the presentation of a different second stimulus. Grey: average activity for AlexNet with no adaptation. (**B**), Same as (A), but for stimulus pairs using a random sample of 1,000 images from the ImageNet test set (61). The ImageNet images are more distinct and therefore reveal stronger stimulus-specific adaptation effects (the two images from the example trial are owned by the first author and used for display purposes only).

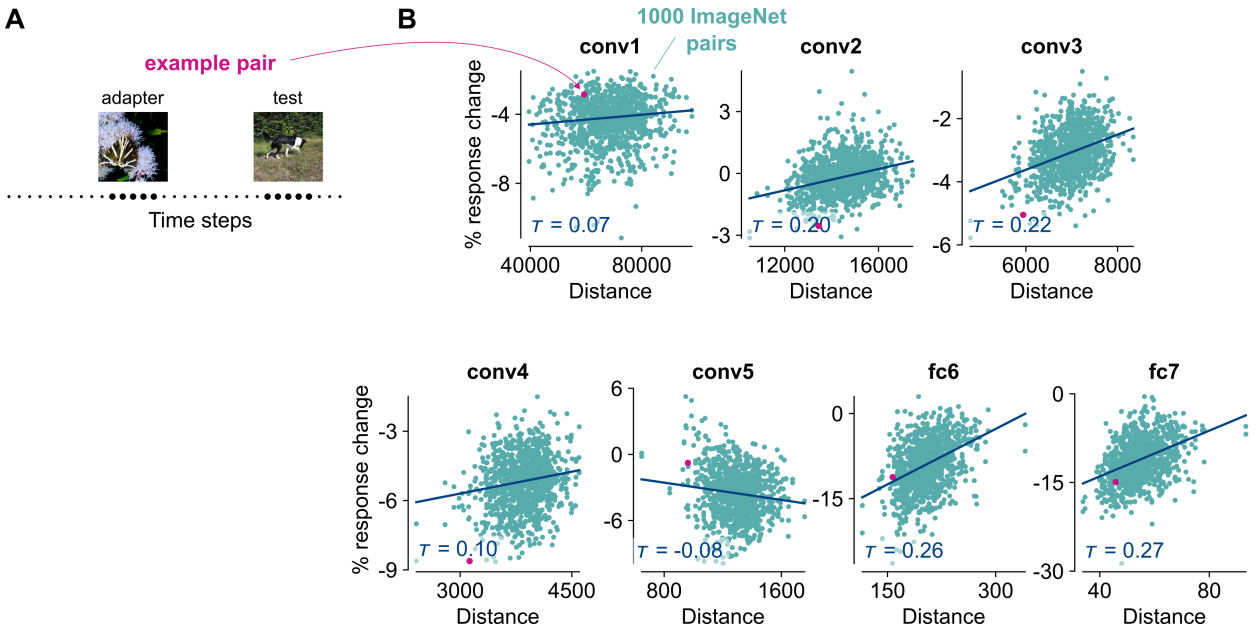

Fig. S2 The amount of activation suppression for a stimulus is related to its similarity with the preceding stimulus (expanding on Fig. 2). (**A**), Illustration of the trial sequence used to investigate the effect of an adapter on the population response suppression for the test image in the model. This experiment was run using a random sample of 1,000 images from the ImageNet test set (61), but the two images from the example trial are owned by the first author and used for display purposes only. Adapter and test images were presented in succession for five time steps each (large dots), and each preceded by ten time steps of blank (uniform grey) input (small dots). For each trial the network started in an unadapted state. (**B**), Scatter plots per layer showing for each stimulus pair the Euclidean distance between the activation patterns for the two images (both calculated without preceding stimulus) and the amount of suppression for the test image (percentage response change averaged across all units of a layer). Negative percentage response change values indicate a response reduction when the test image is preceded by the adapter. Green dots: pairs of ImageNet images; pink dot: example pair from (A). Regression lines show the fit resulting from a robust Theil-Sen estimator, and the inserted τ values are Kendall's correlation coefficient. A positive slope/correlation indicates that the suppression is stronger for image pairs that elicit more similar activation patterns. The correlation is slightly positive for all layers, except conv5 (for unknown reasons).

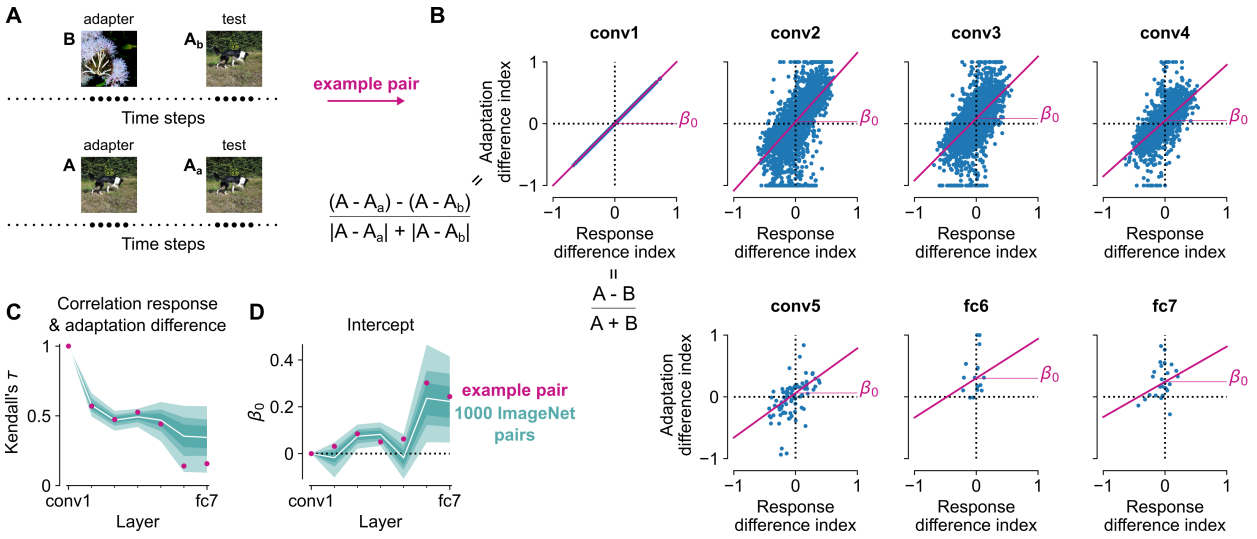

Fig. S3 Stimulus-specific suppression in single units emerges in deeper layers even for two adapter images that equally activated the unit (expanding on Fig. 2). (**A**), Illustration of the trial sequences used to investigate the relation between the activation strength for the adapter and amount of adaptation for a subsequently presented test image. In order to do this, each test image A was randomly paired with a different adapter image B. As in previous physiology investigations (9), the effect of adapting to a different image (BA trial) was compared directly with the effect adapting to the same image (AA trial). (**B**), Scatter plots per layer showing the relation between the adapter response difference index and the adaptation difference index for the image pair in (A). Each dot is a unit that responds significantly to both adapters (activation $> 20\%$ of the unit's maximum activation across the random sample of 1,000 ImageNet images of **Fig. S2**). Regression lines show the fit resulting from a robust Theil-Sen estimator, and the horizontal line labeled β_0 indicates the intercept. In conv1, the difference in adaptation resulting from adapter A versus B is proportional to the response difference between adaptors A and B. From conv2 onward, a richer repertoire of effects emerges: even for units that are activated more by adapter B than A (negative values on the x-axis), adaptation can be stronger for adapter A (positive values on the y-axis). In fact, the positive β_0 intercept in deeper layers (in particular fc6 and fc7) indicates that on average, units that are equally activated by adapters A and B, still show a stronger suppression for a stimulus repetition (AA trial), replicating experimental results for macaque IT neurons (9). (**C**), Correlations (Kendall's τ) between the response difference index and adaptation difference index, averaged (white line) across 1,000 unique pairs of the ImageNet images of **Fig. S2**. Green shaded error bounds indicate the 50^{th}, 75^{th}, and 95^{th} (from dark to lighter green) percentile intervals. Pink markers indicate the values for the example image pair in (A). The reduced correlation in deeper layers means that adaptation strength is increasingly less related to the activation strength of the adapter. (**D**), Intercepts resulting from regressing (Theil-Sen) the adaptation difference index onto the response difference index, averaged across the same image pairs as (C) (white line). Same conventions as (C). A positive intercept, means stronger suppression for a repetition than for an alternation, even for units that were equally activated by the two adapters.

2 Aftereffects with oriented gratings in the proposed computational model (Fig. S4 and S5)

A classic example of an adaptation aftereffect is the tilt aftereffect, which occurs when adapting to an oriented bar or grating causes an observer to perceive a subsequently presented stimulus to be slightly more tilted in the direction opposite to the orientation of the adapter (38). To evaluate whether the model also shows the tilt aftereffect, we created a set of gratings that ranged from left to right (-90° to 90°) in 100 steps (**Fig. S4A**), and measured the boundary shifts analogous to those along the face-gender dimension in **Fig. 3**. For a right tilted adapter (29°), the decision boundary in conv5, that is the orientation at which the predicted right tilt probability was 0.5, shifted 10° towards the tilt of the adapter (**Fig. S4B**). We only present results for the convolutional layers, as the fully connected layers were invariant to the property of left or right tilt (e.g. the representation for a -10° grating was very similar to that for a 10° grating). This mirror-symmetry is likely the result of a form of data augmentation, where horizontal reflections of the training set were used during training (35). As predicted, adaptation to a vertically oriented grating (i.e. the original boundary stimulus) had no effect on the decision boundary.

As for the face-gender stimulus set, we measured orientation discriminability at each test orientation as a function of the adapter orientation. We found that adaptation in the model enhanced orientation discriminability for orientations similar to the adapter (**Fig. S4D**; red diagonal; **Fig. S4E**).

We repeated the analyses on response magnitude and tuning changes for the tilt aftereffect shown in **Fig. 4** and **5** for the tilt aftereffect. The results are presented in **Fig. S5** and are consistent with the results for the face-gender stimulus set.

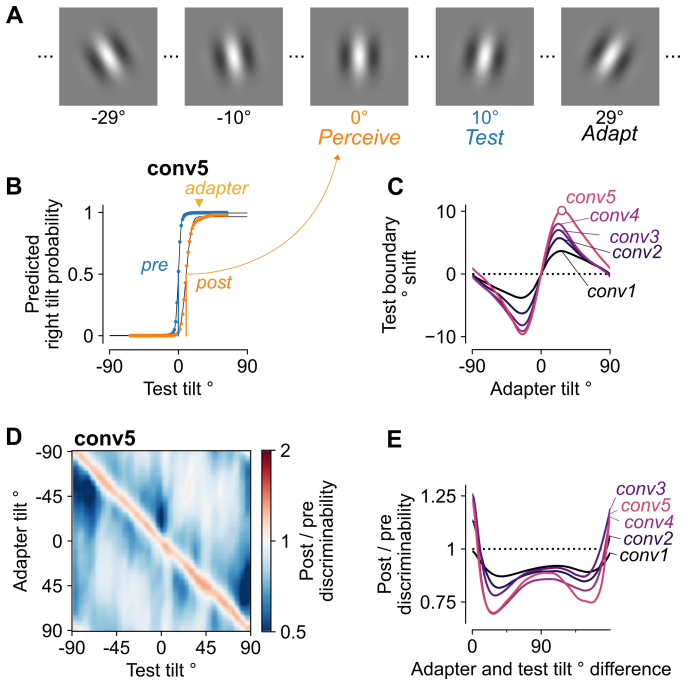

Fig. S4 Perceptual bias and discriminability changes for the tilt aftereffect in the proposed computational model (expanding on Fig. 3). (**A**), Examples of the stimuli used in our simulated experiments: a set of gratings that ranged from -90° (left tilt) to 90° (right tilt) in 100 steps. The example *adapt*, *test*, and *perceive* orientations were picked based on the estimated boundary shift shown in (B). (**B**), Decision boundaries pre (blue) versus post (orange) exposure to a 29° right tilted adapter based on the top convolutional layer (conv5) of the model with intrinsic suppression. Only angles between -63° and 63° were used to fit the psychometric functions to avoid issues with the circularity of the orientation dimension. Markers show class probabilities for each test stimulus, full lines indicate the corresponding psychometric functions, and vertical lines the classification boundaries. Adaptation to a 29° adapter leads to a shift in the decision boundary towards positive (right tilted) orientations, hence perceiving the 10° test stimulus as vertical (0°). (**C**), Decision boundary shifts for the test stimulus as a function of the adapter tilt per layer. The round marker indicates the boundary shift plotted in (B). (**D**), Relative orientation discriminability ($|\Delta y_m^{post}|/|\Delta y_m^{pre}|$) for conv5 as a function of adapter and test tilt. See color scale on right. The red areas indicate where orientation discriminability is increased. (**E**), Average changes in tilt discriminability per layer as a function of the absolute difference in orientation between adapter and test stimulus.

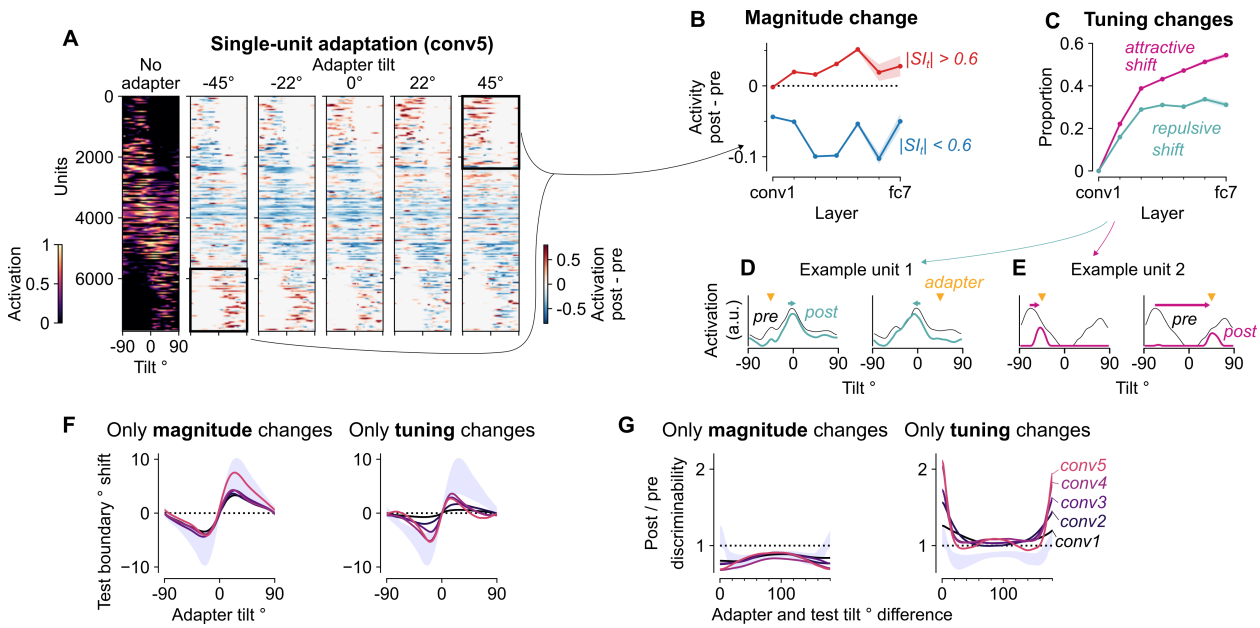

Fig. S5 Response magnitude and tuning changes for the tilt aftereffect in the proposed computational model (expanding on Fig. 4 and 5). (**A**), Effects of adapting to oriented gratings on the activation strength of single units. Left: heatmap showing the activation of all responsive conv5 units (rows) for all oriented gratings (from -90° to 90°; columns). Rows are sorted according to a left versus right tilt selectivity index (SI_t), calculated analogously to the gender selectivity index (**equation (3)**). The remaining five heatmaps show the difference (post - pre adaptation) in single-unit activations after adapting to five different adapters. (**B**), Mean response change (activity post - activity pre) across responsive units for each layer (shaded area = 95%CI). For highly left versus right tilt-selective units (red), the magnitude change (averaged across stimuli) was taken after adapting to a stimulus tilted opposite to the unit's preferred tilt (-45° adapter for $SI_t > 0.6$, 45° adapter for $SI_t < -0.6$; black rectangles in (A)). For less tilt-selective units (blue), the magnitude change after both -45° and 45° adapters was used. (**C**), Proportion of adapters causing the preferred morph level to shift towards (attractive, pink) or away (repulsive, green) from the adapter, averaged across units (shaded area = 95% binomial CI). (**D**), An example unit showing a repulsive shift in tuning curves for the -45° (left) and 45° (right) adapters (the y-axes depict activation in arbitrary units; black: pre adaptation tuning curve; green: post adaptation tuning curve; yellow marker: adapter morph level). (**E**), An example unit showing an attractive shift in tuning curves (pink: post adaptation tuning curve; same conventions as (D)). (**F**), Tilt boundary shifts towards the adapter were produced both by magnitude changes without tuning changes (left) as well as by tuning changes without magnitude changes (right). Grey shading indicates the range of original layer effects shown in **Fig. S4C**. (**G**), Tilt discriminability enhancement for orientations close to the adapter was produced by tuning changes without magnitude changes (right), but not by magnitude changes without tuning changes (left). Grey shading indicates the range of original layer effects shown in **Fig. S4E**.

3 Adaptation produces mostly repulsive shifts in highly responsive units (Fig. S6)

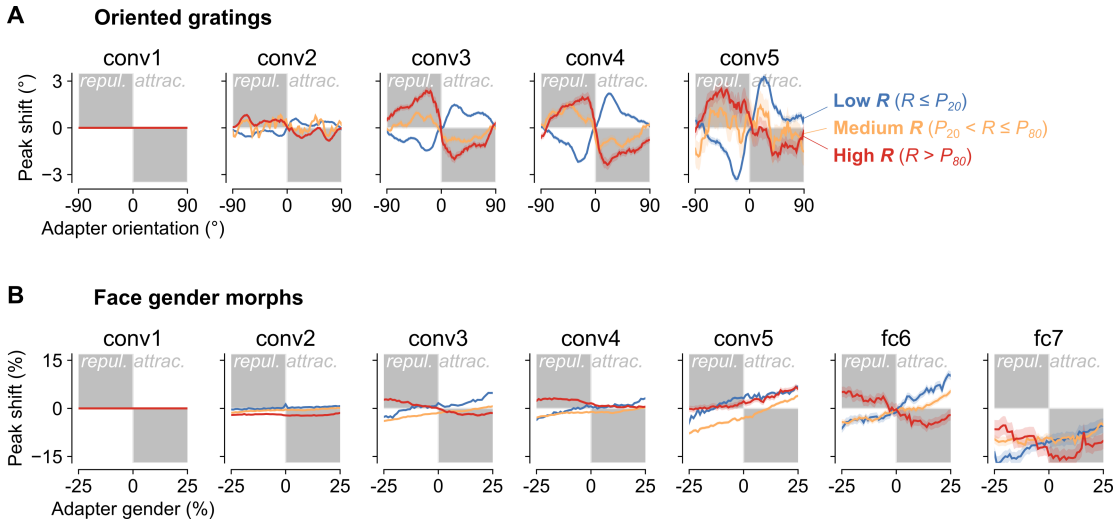

Fig. S6 Peak shift direction separate for most and least responsive units (expanding on Fig. 4). (**A**), Average peak shift of orientation tuning curves as a function of the adapter orientation, relative to the preferred orientation, which is centered at 0. Units were split based on the 20^{th} (P_{20}) and 80^{th} (P_{80}) percentiles of their median pre-adaptation responsivity (R), calculated across all orientations. Highly responsive units (red) undergo on average repulsive peak shifts, whereas the lowest responsive units (blue) undergo on average attractive shifts. (**B**), Average peak shift of face-gender tuning curves as a function of the adapter morph level (gender percentage), relative to the preferred morph level, which is centered at 0. Only units with preferred morph-level between 25% and 75% were considered, in order to be able to have an adapter at equal distances left and right of the peak. Units were split based on the according to the same criterion as (A).

Overall, attractive shifts were more common in the proposed computational model (**Fig. 4**, **Fig. S5**), whereas several studies report mainly repulsive shifts (13, 18). A plausible explanation is that repulsive shifts are caused by recurrent interactions at short timescales of a few 100 ms, whereas adaptation causes more attractive shifts at a longer timescale (**Discussion**). Another possible explanation is that neurons with clear and strong response profile, which are more likely to get isolated and recorded from, are also more likely to show a repulsive shift. Consistent with this idea, we noticed that adaptation produced mostly repulsive shifts for units with higher average activations, particularly for oriented gratings. We demonstrate this by splitting the units per layer into three groups based on the 20^{th} and 80^{th} percentiles

of their median activation, calculated across all morph levels for the face-gender aftereffect, and across all orientations for the tilt aftereffect (Fig. S6).

4 Adaptation in single layers (Fig. S7 to S9)

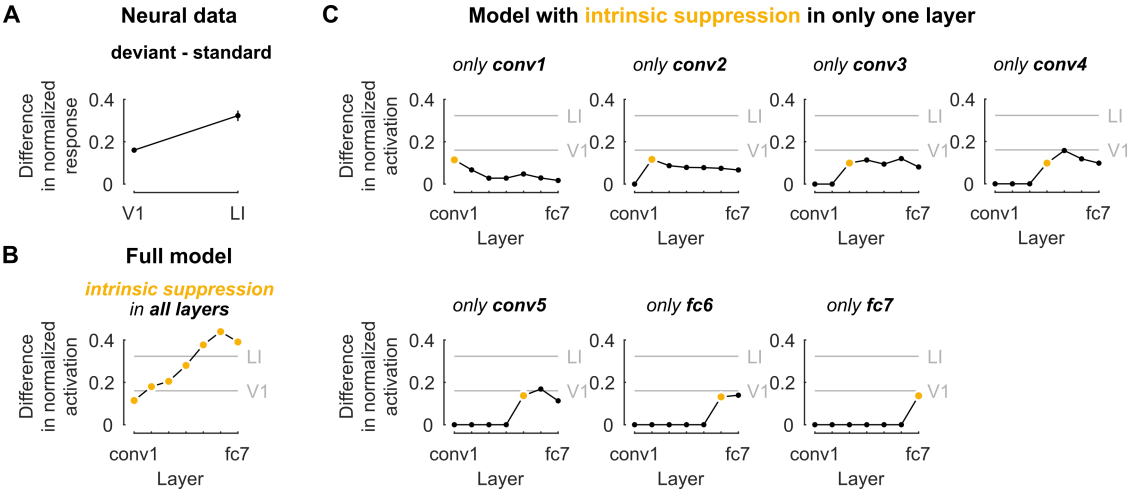

Fig. S7 An increased sensitivity to stimulus presentation frequency in downstream areas requires intrinsic suppression at multiple stages (expanding on Fig. 2). (**A**), Difference (average with 95% bootstrap CI) in response between the low (deviant) and high probability (standard) stimulus in the oddball experiment explained in Fig. 2. The response difference increases from V1 to downstream area LI. (**B**), Difference in average activation for the low and high probability stimulus in a simulated oddball sequence (Fig. 2D), for the full model which has intrinsic suppression implemented in each layer. The response difference builds up across network layers. Grey horizontal lines indicate the neural data averages of (A). (**C**), Same as (B), except that the model has intrinsic suppression only implemented in one layer (yellow markers). The response difference between low and high probability stimuli no longer builds up across multiple layers.

Several adaptation effects in the model increase across consecutive layers or emerge only in deeper layers. This could be because each layer increases adaptation by providing additional activation-based suppression on top of the adapted outputs from the previous layer, but it is also possible that adapted outputs from early layers propagating through the network are sufficient. Here we address this question by recreating several critical figures, using modified models with intrinsic suppression implemented in only one layer at a time (always using the same parameters values $\alpha = 0.96$ and $\beta = 0.7$ that were used for the full model).

In Fig. 2, we showed that repetition suppression in the proposed computational model

accumulated across layers, replicating the increased sensitivity to stimulus frequency in the putative homologue of the rat ventral stream (12). The modified neural networks with intrinsic suppression in only one layer do not show any build-up of repetition suppression across layers (**Fig. S7C**), demonstrating that activation-based suppression implemented at multiple stages of processing is indeed necessary to capture the neural data (**Fig. S7A**).

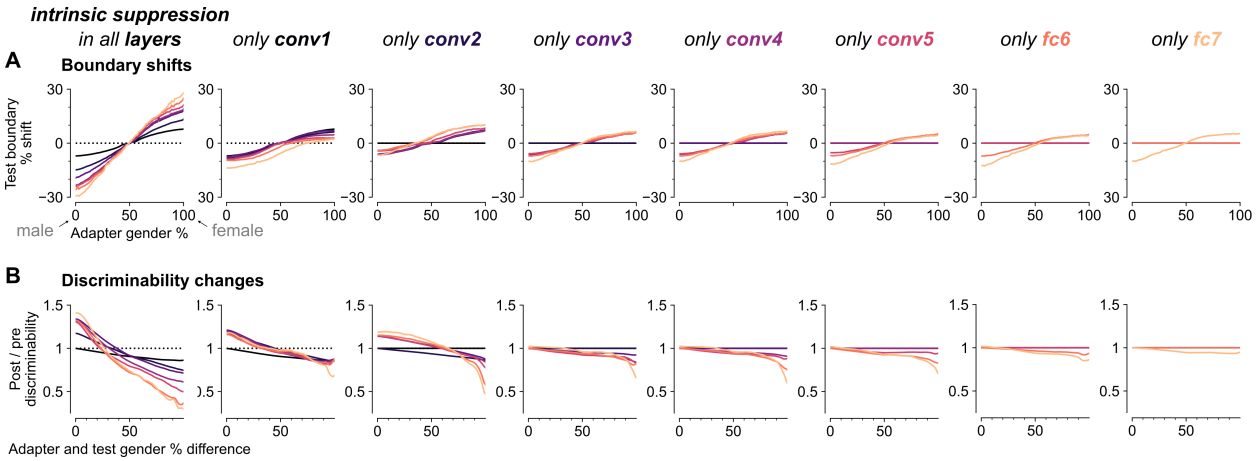

Fig. S8 Intrinsic suppression causes a perceptual bias within the same layer, but only causes discriminability enhancements in downstream layers (expanding on Fig. 4). (**A**), Adapting to a female/male face shifted the face-gender decision boundary towards the adapter morph level (**Fig. 3C**). Left: boundary shifts for a network with intrinsic suppression in all layers. Rest: boundary shifts for networks with intrinsic suppression in only one layer (indicated by the column title). The first layer to show a boundary shift is always the first layer with intrinsic suppression. (**B**), Adapting to a female/male face enhanced face-gender discriminability around the adapter morph level (**Fig. 3E**). Left: discriminability changes for a network with intrinsic suppression in all layers. Rest: discriminability changes for networks with intrinsic suppression in only one layer (indicated by the column title). The first layer to show enhanced discriminability is always downstream of the first layer with intrinsic suppression.

Similar to the accumulation of repetition suppression across layers, the magnitude of perceptual aftereffects (i.e., perceptual bias and discriminability changes) also increased across layers **Fig. 3C** and **E**. **Fig. S8** shows that, consistent with the increase in neural adaptation effects in **Fig. S7**, the increase in magnitude of aftereffects also requires intrinsic suppression in multiple layers. The same analysis also shows that a perceptual bias (i.e., boundary shift) as well as a reduced discriminability (for morph levels further from the adapter) always already occurs in the first layer with intrinsic suppression (**Fig. S8A,B**). In contrast,

the enhanced discriminability effect for face-gender morph levels close to the adapter occurs first in the layer after the one with intrinsic suppression (**Fig. S8B**), suggesting that this aftereffect relies on the downstream propagation of suppressed outputs. Note also that the discriminability effects are smaller when the layer with intrinsic suppression is more downstream.

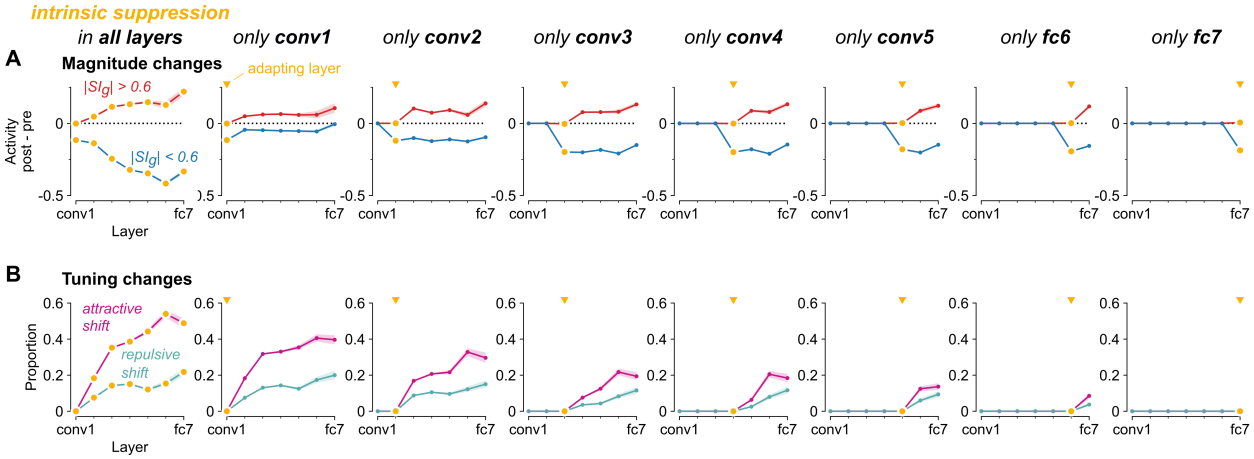

Fig. S9 Intrinsic suppression causes response reductions within the same layer, whereas response enhancements and tuning peak shifts only emerge in downstream layers (expanding on Fig. 4). (**A**), Mean response change after adapting (shaded area: 95% CI). Left: highly gender-selective units ($|SI|_g > 0.6$, red) show response enhancement after adapting to a gender stimulus opposite to their preferred gender; less selective units ($|SI|_g < 0.6$, blue) show response suppression. Left: magnitude changes for a network with intrinsic suppression in all layers (see also **Fig. 4B**). Rest: magnitude changes for networks with intrinsic suppression in only one layer (yellow markers). The first layer to show suppression is always the first layer with intrinsic suppression, but enhancement only emerges downstream. (**B**), Proportion of adapters causing the preferred morph level to shift towards (attractive, pink) or away (repulsive, green) from the adapter, averaged across units (shaded area: 95% CI). Left: peak shifts for a network with intrinsic suppression in all layers (see also **Fig. 4C**). Rest: peak shifts for networks with intrinsic suppression in only one layer (yellow markers). The first layer to show peak shifts is always downstream of the first layer with intrinsic suppression.

The perceptual aftereffects in the model coincided with complex adaptation effects in deeper layers, including response enhancement and tuning curve peak shifts (**Fig. 4**). As expected, in the networks with intrinsic suppression in only one layer, response suppression occurred already within the layer with intrinsic suppression, with little change in subsequent layers (**Fig. S9A**, blue). This is generally consistent with **Fig. S7**. In contrast, complex

adaptation effects (i.e., response enhancements and tuning curve peak shifts) only occurred in layers downstream from the layer with intrinsic suppression (**Fig. S9A**, red; **B**).

5 Intrinsic suppression in the proposed computational model captures the experimental data of Fig. 6 (Fig. S10)

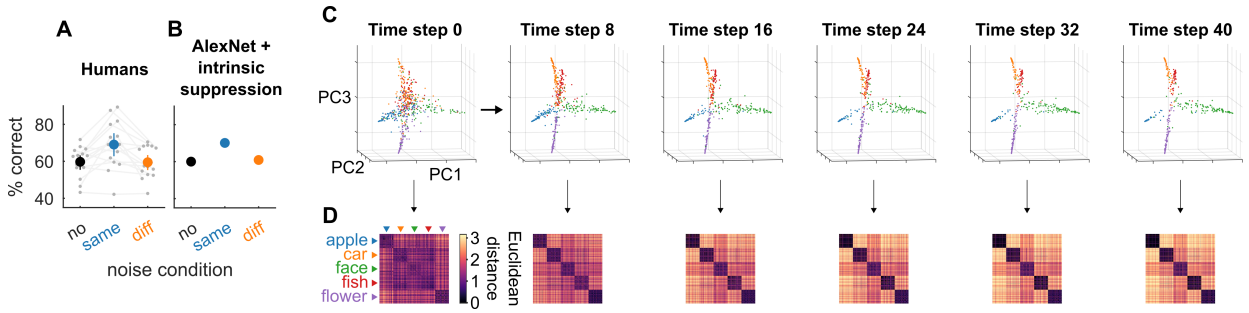

Fig. S10 Adapting to prevailing but interfering input enhances object recognition performance in the proposed computational model (expanding on Fig. 6). (**A**), Participants showed an increase in categorization performance after adapting to the same noise pattern (this is a repeat of **Fig. 6C**). Gray circles and lines denote individual participants ($N = 15$). The colored circles show average categorization performance, error bars indicate 95% bootstrap confidence intervals. Chance = 20%. (**B**), The proposed computational model could capture the effect in (A) with adaptation parameters α and β chosen to impose suppression. To match the performance increase in humans, the suppression scaling constant was lowered to $\beta = 0.1$ (for all other figures it was set to $\beta = 0.7$). (**C**), Adapting the model for 40 time steps to the same-noise condition moved the fc8 representations of the noisy doodles into more separable clusters matching the five doodle categories. The 3 axes correspond to the first 3 principal components of the fc8 layer representation of all the test images. Each dot represents a separate noisy doodle image, the color corresponds to the category (as shown by the text in (D)). (**D**), Dissimilarity matrices for all pairs of images. Entry (i,j) shows the Euclidean distance between image i and image j based on the fc8 features before (time step 0) or after (time step 40) continuous exposure to same-noise. The distance is represented by the color of each point in the matrix (see scale on right). Images are sorted based on their categories. Adaptation leads to an increase in between category distances and a decrease in within category distances as shown by the pairwise distance matrices.

The model with α and β fixed to impose suppression captures same pattern of results as the psychophysics experiment in **Fig. 1**. To simulate the experiment, we fine-tuned the pre-trained fully connected layers of AlexNet to classify high contrast (i.e., 40% as opposed to 22% in the experiment) doodles on a noisy background. We used a set of $50,000$ doodle

images (10,000 per category) that were different from the ones used in the experiment and fine-tuned the fully connected layers of AlexNet (without intrinsic suppression) for 5 epochs (i.e. 5 full cycles through the training images), with every epoch using a different noise background for each image. We used the Adam optimization algorithm (63) with a learning rate of 0.001, the sparse softmax cross entropy between logits and labels cost function, a batch size of 100, and no dropout.

The model demonstrated the same effects as the human participants, showing increased performance for the same-noise condition compared to the no adapter condition or different-noise condition (**Fig. S10B**). Thus, adapting to a prevailing noise pattern improved the ability to recognize test images and this effect could be accounted for by activation-based, intrinsic suppression in a feedforward neural network. To visualize the effect of adaptation for the same-noise condition on the representation of noisy doodles, we plotted each noisy doodle image in a space determined by the first 3 principal components of the fc8 outputs. Before adaptation (at time step 0), the colored dots representing the doodle images were not well separated, because the noise obscures the relevant features of the doodles (**Fig. S10C**, left). After exposing the network to the same-noise adapter for 40 time steps, adaptation decreased the salience of interfering noise features and the representations of the doodle images migrated into distinctly separable clusters (**Fig. S10C**, right). We quantified this separation in feature space by computing dissimilarity matrices for all possible pairs of images (**Fig. S10D**). Adaptation led to increased differentiation of the between-category comparisons (off diagonal squares) and increased similarity between images within each category (diagonal squares) from the initial conditions (left) to the final time step (right).

6 Equalizing the number of parameters for the trained intrinsic suppression and recurrent networks (Fig. S11 and S12)

In **Fig. 7** we showed that a network with intrinsic adaptation state could generalize well to different adapter noise conditions, whereas a recurrent network failed to do so. Here, we investigate whether this difference in generalization performance can be explained by the

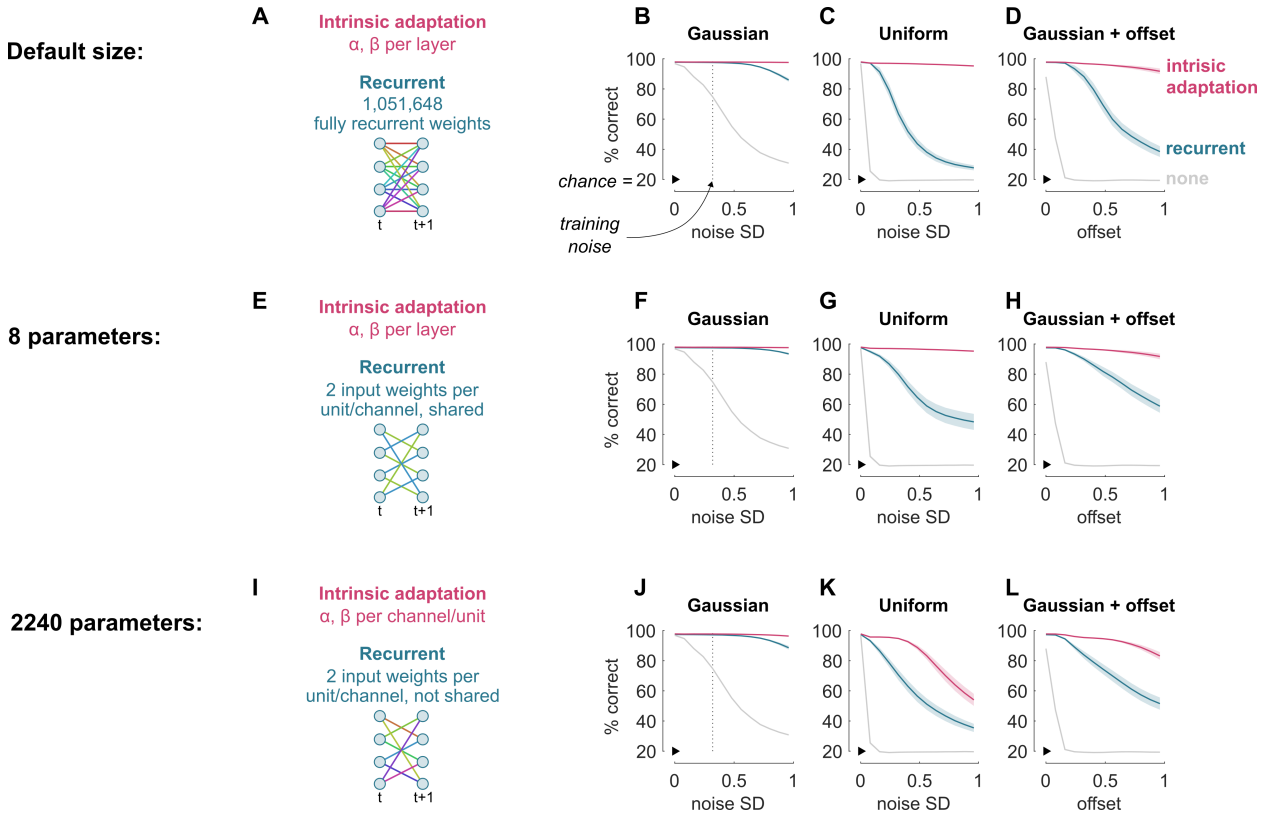

Fig. S11 A trained network with intrinsic adaptation is more robust than a recurrent neural network with the same number of parameters (expanding on Fig. 7). (**A-D**), Results for the default network sizes shown in **Fig. 7**. The network with intrinsic adaptation has 8 adaptation parameters: one α and one β per layer. The recurrent network has 1,051,648 recurrent parameters: within each layer, each channel (convolutional layers) or unit (fully connected layer) projects to all channels/units at the next time step, with no weight sharing. (**E-H**), Results for networks with 8 adaptation/recurrent parameters. The network with intrinsic adaptation is the same as in (A-D). The recurrent network is reduced in size: within each layer, each channel/unit projects to two channels/units at the next time step, and those two weights are shared across channels/units within a layer. (**I-L**), Results for networks with 2240 adaptation/recurrent parameters. The network with intrinsic adaptation is increased in size and has one α and one β per channel (convolutional layers) or unit (fully connected layer). The recurrent network is reduced in size: within each layer, each channel/unit projects to two channels/units at the next time step, with no weight sharing. (B-D), Average generalization performance of the networks (pink: with intrinsic adaptation; green: recurrent) under noise conditions that differed from training (the vertical line in (B) indicates the Gaussian noise with SD = 0.32 that was used during training). Chance level is at 20%, indicated by the black marker. Shaded bounds indicate standard error of the mean (for 30 random initializations per network). Same conventions for (F-H) and (J-L).

difference in the number of parameters used to implement intrinsic adaptation ($N = 8$, i.e., one α and one β per layer) versus lateral recurrence ($N = 1,051,648$ recurrent weights), by: (i) reducing the number of recurrent weights to $N = 8$, (ii) increasing the number of intrinsic adaptation parameters *and* reducing the number of recurrent weights to $N = 2240$.

In the default size recurrent network, each channel (convolutional layers) and each unit (fully connected layer) received lateral input from all within-layer channels/units at the previous time step. To reduce these recurrent weights to 8, we designed an architecture with only 2 recurrent weights per layer: each channel/unit only received lateral input from 2 other channels/units, and the input weights were shared across channels/units within a layer (**Fig. S11E**). Despite the drastic reduction in recurrent weight parameters, the network could generalize well when the adapter noise matched the training noise (**Fig. S11F**, dashed line), but failed to generalize to different adapter noise conditions (**Fig. S11F-H**).

Next, we increased the number of parameters for the intrinsic adaptation network by using a different α and β for each channel (convolutional layers) or each unit (fully connected layer), resulting in a total of 2240 adaptation parameters. For comparison, we created a recurrent network with the same number of parameters: each channel/unit received lateral input from 2 other channels/units, with no sharing of input weights across channels/units (**Fig. S11I**). The intrinsic adaptation network with 2240 parameters showed impaired generalization to uniform noise, yet still performed better than the same-size recurrent network in all noise conditions (**Fig. S11J-L)**). These results suggest that the intrinsic adaptation mechanism provided a less complex solution that generalizes better regardless of the number of parameters.

Finally, we assessed for each of these trained networks whether they also demonstrated repetition suppression for doodle images (without noise), a hallmark property of neural adaptation. We compared the amount of response suppression for a repeated doodle (repetition) with the amount of suppression for a doodle preceded by a different doodle (alternation). In all networks, the response for a doodle repetition was lower than the response for a doodle alternation (**Fig. S12**). However, in contrast with neural repetition suppression, the third

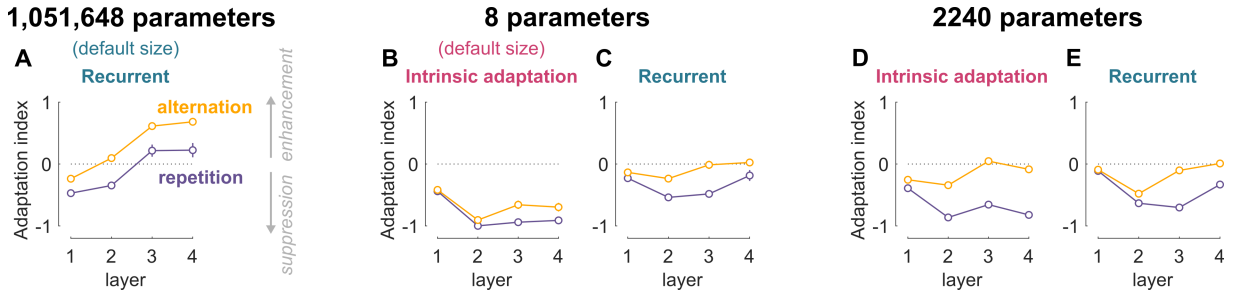

Fig. S12 Adaptation learnt by the recurrent network did not necessarily lead to repetition suppression (expanding on Fig. 7). Average adaptation index per layer for stimulus repetitions and alternations for the trained networks of **Fig. S11** (error bars are standard error of the mean for 30 random initializations per network). Repetition (purple): the same doodle (no noise) was presented on time step 1 and time step 3, with blank input at time step 2. Alternation (yellow): a different doodle was presented on time step 1 and 3. Y-axis: adaptation index, based on the average activation for the second (S2) versus first (S1) stimulus presentation: $(S2 - S1)/(S2 + S1)$. A negative value indicates suppression for the second stimulus presentation, whereas a positive value indicates enhancement. To replicate repetition suppression in the brain, the adaptation index for stimulus repetitions should be negative on average.

and fourth layers of the default size recurrent network showed response *enhancement* for the second stimulus, regardless of whether it was a repetition or alternation (**Fig. S11A**), suggesting that this recurrent network solution differs in a critical way from neural adaptation in the brain.

7 Adaptation maintains population homeostasis (Fig. S13)

Benucci et al.(55) showed that adaptation in cat V1 enforces a tendency toward equality in time-averaged responses and independence in neural activity across the population. The authors showed that, to achieve this, adaptation followed a simple multiplicative rule which depends on stimulus attributes as well as neuronal preference, possibly resulting from intrinsic suppression at an earlier cortical stage. To evaluate whether intrinsic suppression in a feedforward network could indeed capture their results, we simulated the main experiment of (55) in the proposed computational model using the tilted gratings from **Fig. S4** and the face stimuli from **Fig. 3**.

Briefly, the gratings(/faces) were presented in random sequences of 220 presentations,

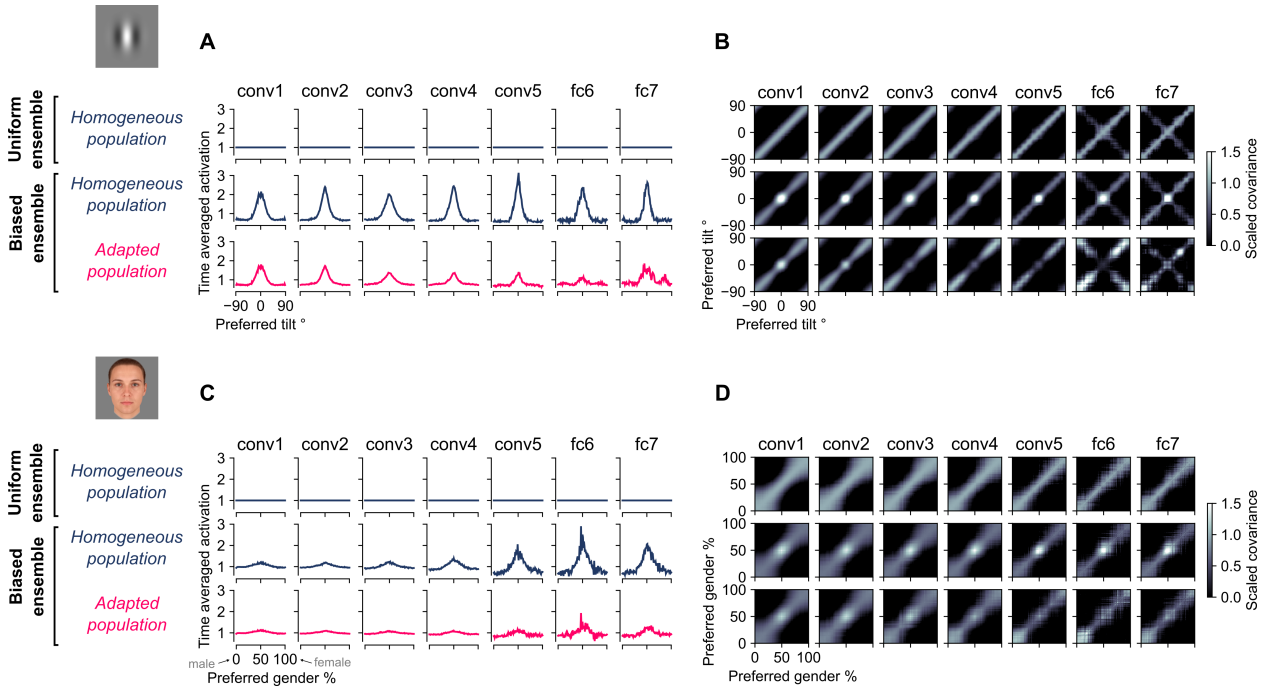

Fig. S13 Intrinsic suppression reduces time-averaged responses and decorrelates responses for biased stimulus ensembles. (**A**), Time-averaged responses for each orientation bin (i.e., units with the same preferred orientation), normalized by the time average of each orientation bin for the homogeneous population in the uniform ensemble. First row: time-averaged responses for the homogeneous population in the uniform stimulus ensemble (each bin is normalized to 1). Middle row: time-averaged responses for the homogeneous population in the biased stimulus ensemble. When the network is adapted to the uniform ensemble (homogeneous population), time-averaged responses in the biased ensemble show a strong peak around the more frequent orientation. Bottom row: time-averaged responses for the adapted population in the biased stimulus ensemble. The higher response for the biased stimulus is much attenuated when the network is allowed to adapt to the biased ensemble. (**B**), Top row: covariance matrices for the homogeneous population in the uniform stimulus ensemble. Diagonals are scaled to 1. Middle row: covariance matrices for the homogeneous population in the biased stimulus ensemble (using the same scaling factors as the top row panels). Population responses are highly correlated for orientation bins with a preferred orientation similar to the more frequent orientation (central peak in covariance matrices). Bottom row: covariance matrices for the adapted population in the biased stimulus ensemble (using the same scaling factors as the top row panels). Adaptation decorrelates population responses for orientation bins with a preferred orientation similar to the biased stimulus (central peak in covariance matrices is much reduced compared to the middle row). (**C,D**), Same as (A,B), but for the face stimuli from **Fig. 3**.

with one stimulus presentation per time step. In uniform ensembles, the probability of each orientation(/face gender morph-level) was equal. In biased ensembles, the vertical oriented grating(/gender neutral face) was presented at a higher probability of $P = 0.5$. The network started in an unadapted state at the beginning of each sequence and adapted throughout the sequence, resulting in a *homogeneous population* for uniform ensembles and an *adapted population* for biased ensembles (terminology was chosen to match (55)). As a control condition, we also simulated the experiment with biased sequences, but with the network adapted to a uniform ensemble (i.e., homogeneous population). Before simulating the experimental conditions, we ran a separate uniform ensemble to determine the preferred orientation(/morph-level) of each unit and divided the population of each layer into bins pooling units with the same preferred orientation(/morph-level).

Like the experimental results in cat V1 (55), adaptation in deeper layers of the proposed computational model reduced time-averaged responses (**Fig. S13A,C**) and decorrelated population responses (**Fig. S13B,D**) for frequent stimuli in biased ensembles, in accordance with the claim that adaptation enforces "a tendency toward equality and independence in neural activity across the population" (55).

Word Counts

```
File: main.tex
Encoding: utf8
Sum count: 9893
Words in text: 8235
Words in headers: 96
Words outside text (captions, etc.): 1425
Number of headers: 23
Number of floats/tables/figures: 7
Number of math inlines: 134
Number of math displayed: 3
Subcounts:
  text+headers+captions (#headers/#floats/#inlines/#displayed)
  0+15+0 (1/0/0/0) _top_
  151+1+0 (1/0/0/0) Abstract
  963+1+0 (1/0/0/0) Section: Introduction
  0+1+0 (1/0/0/0) Section: Results
  158+15+178 (1/1/12/0) fig1
  1052+9+238 (1/1/10/0) fig2
  685+14+249 (1/1/1/0) fig3
  445+0+248 (0/1/11/0) fig4
  506+9+99 (1/1/4/0) fig5
  461+0+128 (0/1/28/0) fig6
  390+0+285 (0/1/3/0) fig7
  1800+1+0 (1/0/2/0) Section: Discussion
  1624+30+0 (14/0/63/3) Section: Materials and Methods


File: output.bbl
Encoding: utf8
Sum count: 0
Words in text: 0
Words in headers: 0
Words outside text (captions, etc.): 0
Number of headers: 0
Number of floats/tables/figures: 0
Number of math inlines: 0
Number of math displayed: 0

Total
Sum count: 9893
Words in text: 8235
Words in headers: 96
Words outside text (captions, etc.): 1425
Number of headers: 23
Number of floats/tables/figures: 7
Number of math inlines: 134
Number of math displayed: 3
Files: 2
Subcounts:
  text+headers+captions (#headers/#floats/#inlines/#displayed)
  8235+96+1425 (23/7/134/3) File(s) total: main.tex
  0+0+0 (0/0/0/0) File(s) total: output.bbl
```

```
(errors:5)
```