Science Advances
AAAS

advances.sciencemag.org/cgi/content/full/6/42/eabd4205/DC1

Supplementary Materials for

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

K. Vinken*, X. Boix, G. Kreiman

*Corresponding author. Email: kasper.vinken@childrens.harvard.edu

Published 14 October 2020, *Sci. Adv.* **6**, eabd4205 (2020)
DOI: 10.1126/sciadv.abd4205

This PDF file includes:

Sections S1 to S7
Figs. S1 to S13

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 ($N = 43,264$).

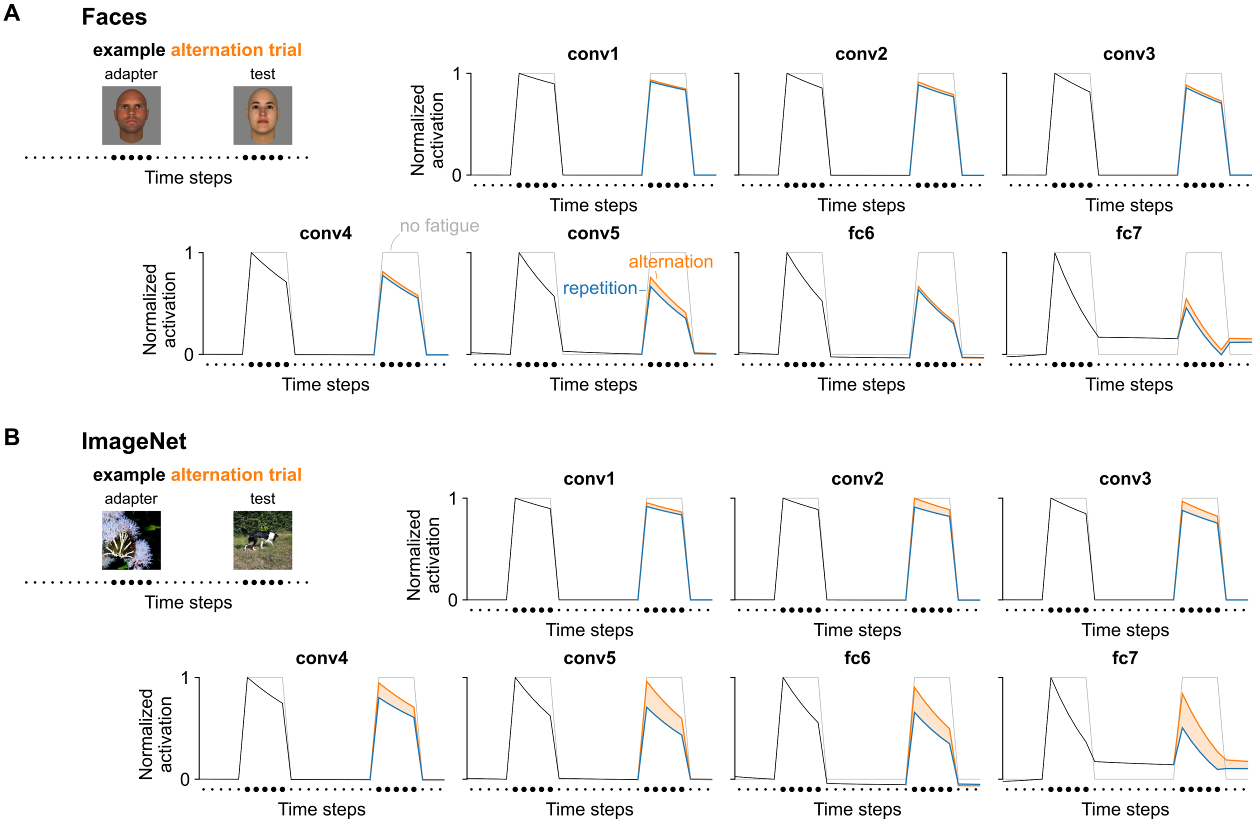

Fig. S1 Stimulus-specific repetition suppression strength varies across model layers and stimulus sets (expanding on Fig. 2). (**A**), Population stimulus-specific repetition suppression in the proposed computational model for a random sub-sample of 500 face pairs (out of 25,000 used in (37), created with FaceGen: `facegen.com`). 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 were added for display purposes; photo credit: Kasper Vinken, Boston Children's Hospital, Harvard Medical School).

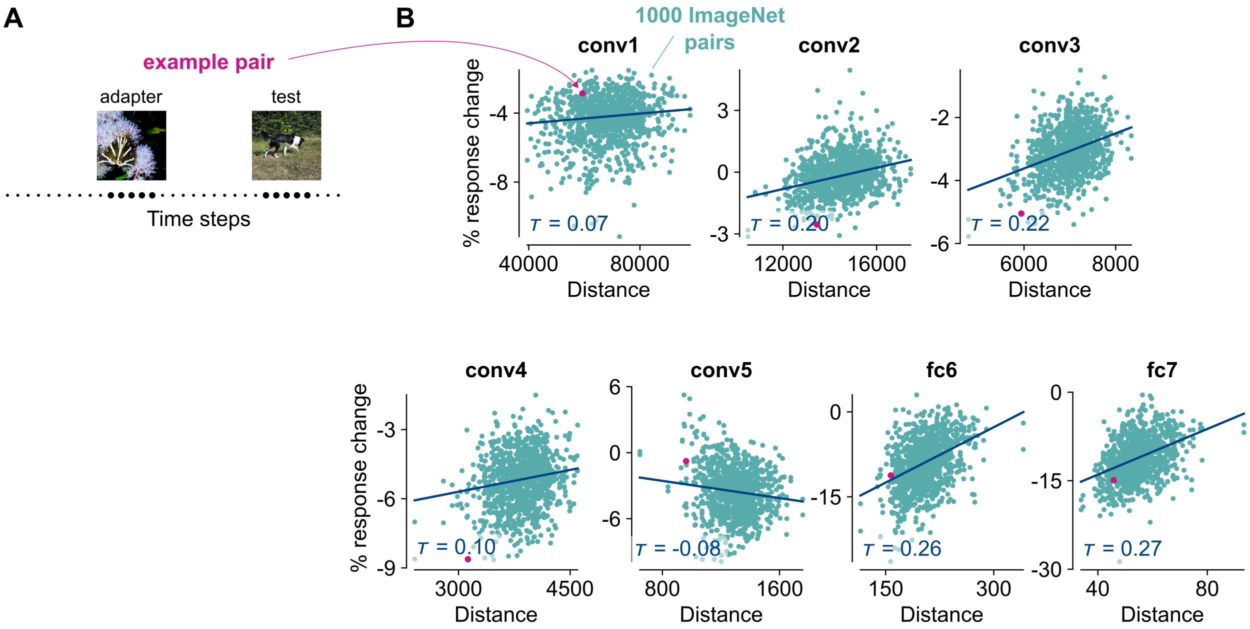

Fig. S2 The amount of activation suppression for a stimulus is related to its similarity with the preceding stimulus (expanding on Fig. 2). (**A**), Illustration of the trial sequence used to investigate the effect of an adapter on the population response suppression for the test image in the model. This experiment was run using a random sample of 1,000 images from the ImageNet test set (61), and the two images from the example trial were added for display purposes. Photo credit: Kasper Vinken (Boston Children's Hospital, Harvard Medical School). 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).

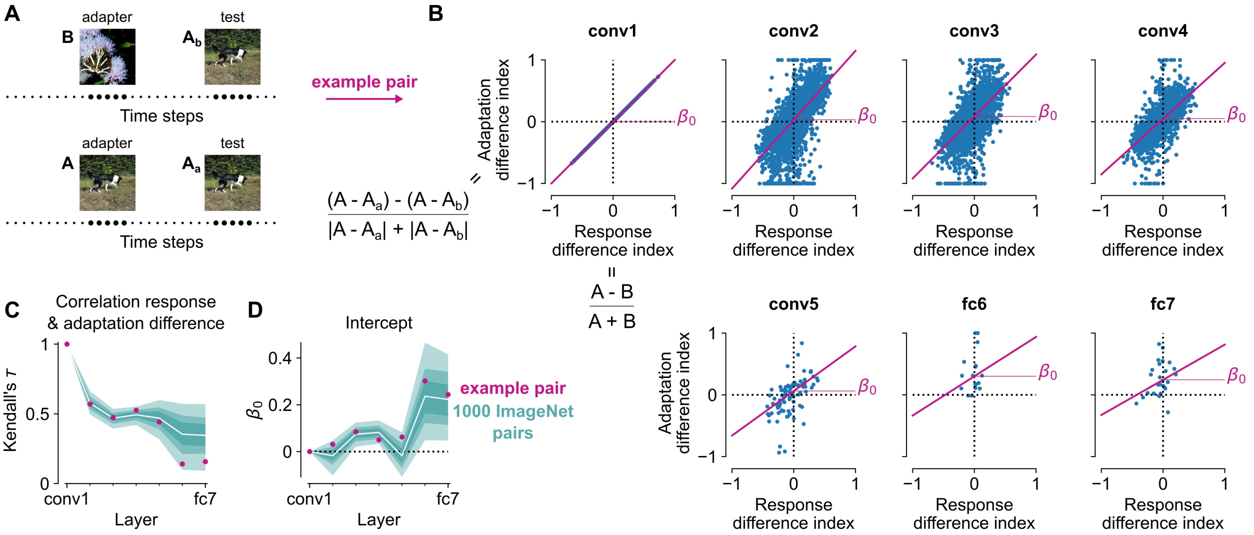

Fig. S3 Stimulus-specific suppression in single units emerges in deeper layers even for two adapter images that equally activated the unit (expanding on Fig. 2). (**A**), Illustration of the trial sequences used to investigate the relation between the activation strength for the adapter and amount of 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). Photo credit: Kasper Vinken (Boston Children's Hospital, Harvard Medical School). (**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 50th, 75th, and 95th (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.

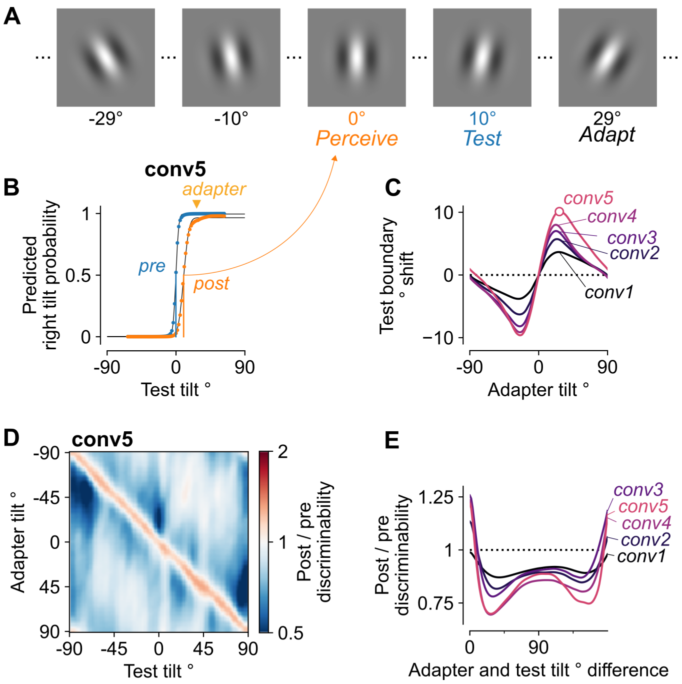

Fig. S4 Perceptual bias and discriminability changes for the tilt aftereffect in the proposed computational model (expanding on Fig. 3). (**A**), Examples of the stimuli used in our simulated experiments: a set of gratings that ranged from -90° (left tilt) to 90° (right tilt) in 100 steps. The example *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.

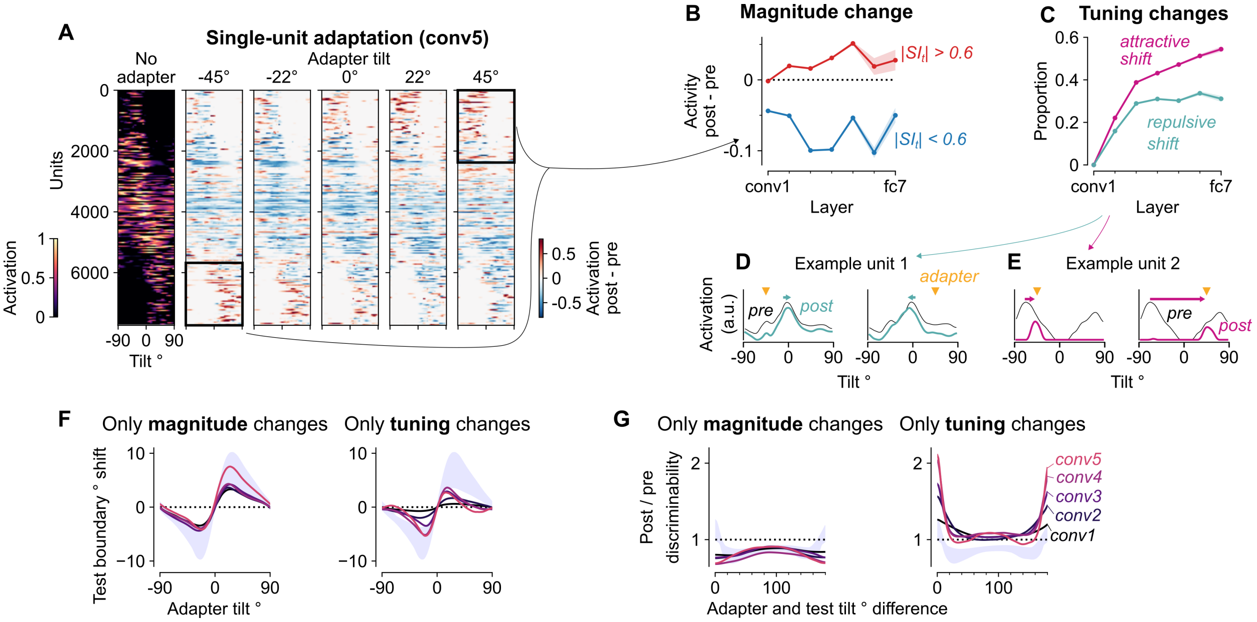

Fig. S5 Response magnitude and tuning changes for the tilt aftereffect in the proposed computational model (expanding on Fig. 4 and 5). (**A**), Effects of adapting to oriented gratings on the activation strength of single units. Left: heatmap showing the activation of all responsive conv5 units (rows) 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)

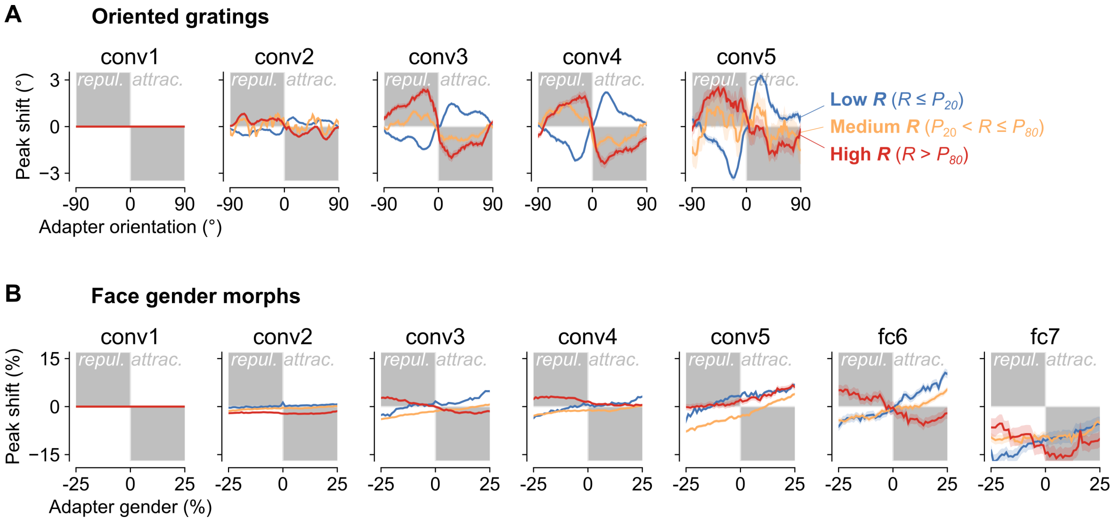

Fig. S6 Peak shift direction separate for most and least responsive units (expanding on Fig. 4). (**A**), Average peak shift of orientation tuning curves as a function of the adapter orientation, relative to the preferred orientation, which is centered at 0. Units were split based on the 20th (P_{20}) and 80th (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 20th and 80th 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)

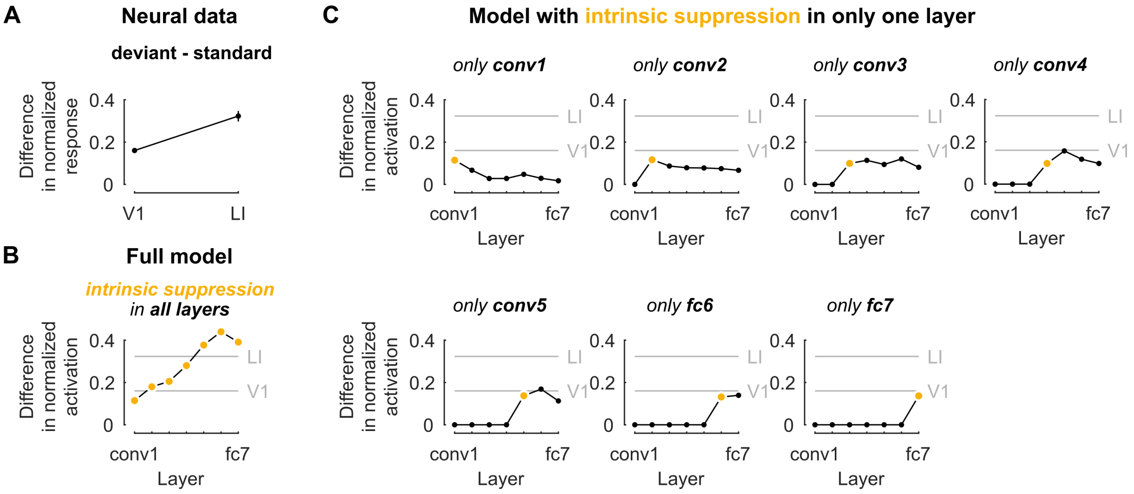

Fig. S7 An increased sensitivity to stimulus presentation frequency in downstream areas requires intrinsic suppression at multiple stages (expanding on Fig. 2). (**A**), Difference (average with 95% bootstrap CI) in response between the low (deviant) and high probability (standard) stimulus in the oddball 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**).

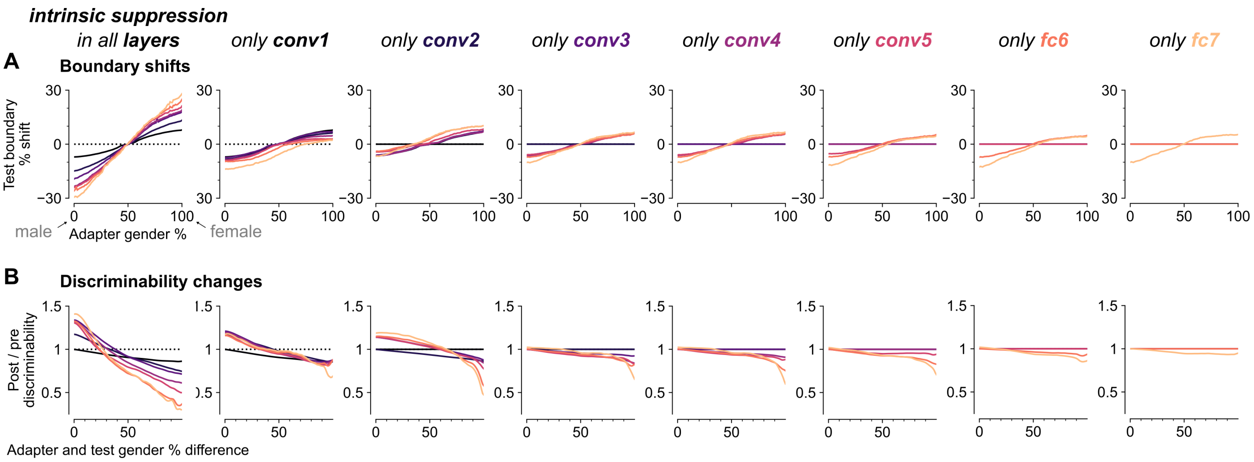

Fig. S8 Intrinsic suppression causes a perceptual bias within the same layer, but only causes discriminability enhancements in downstream layers (expanding on Fig. 4). (**A**), Adapting to a female/male face shifted the face-gender decision boundary towards the adapter morph level (**Fig. 3C**). Left: 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.

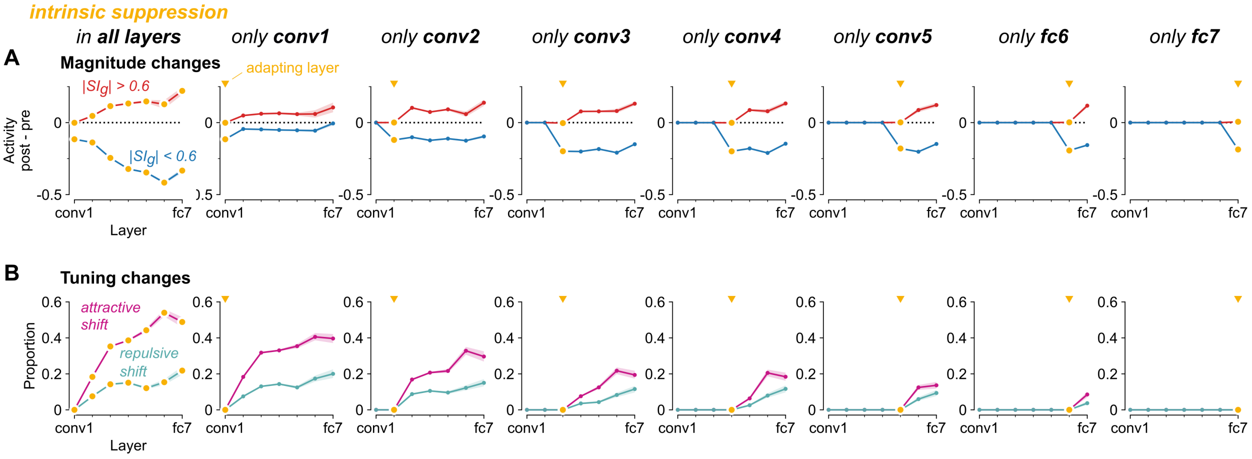

Fig. S9 Intrinsic suppression causes response reductions within the same layer, whereas response enhancements and tuning peak shifts only emerge in downstream layers (expanding on Fig. 4). (**A**), Mean response change after adapting (shaded area: 95% CI). Left: highly gender-selective units ($|SI|_g > 0.6$, 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)

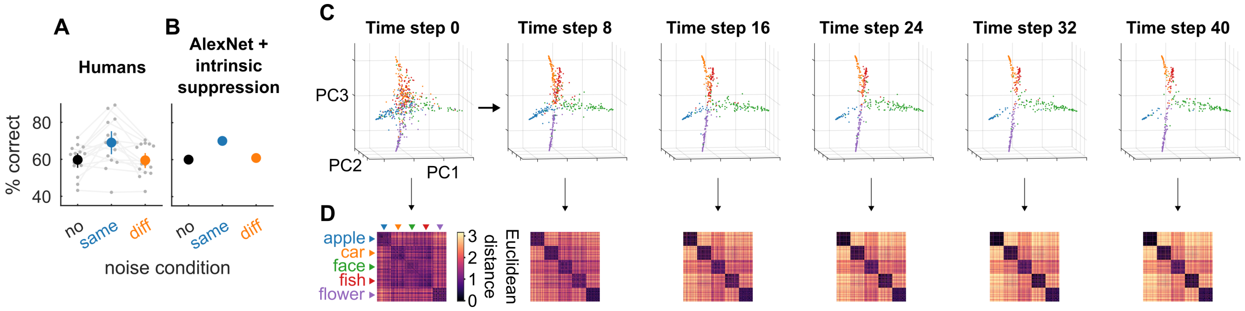

Fig. S10 Adapting to prevailing but interfering input enhances object recognition performance in the proposed computational model (expanding on Fig. 6). (**A**), Participants showed an increase in categorization performance after adapting to the same noise pattern (this is a repeat of **Fig. 6C**). Gray 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

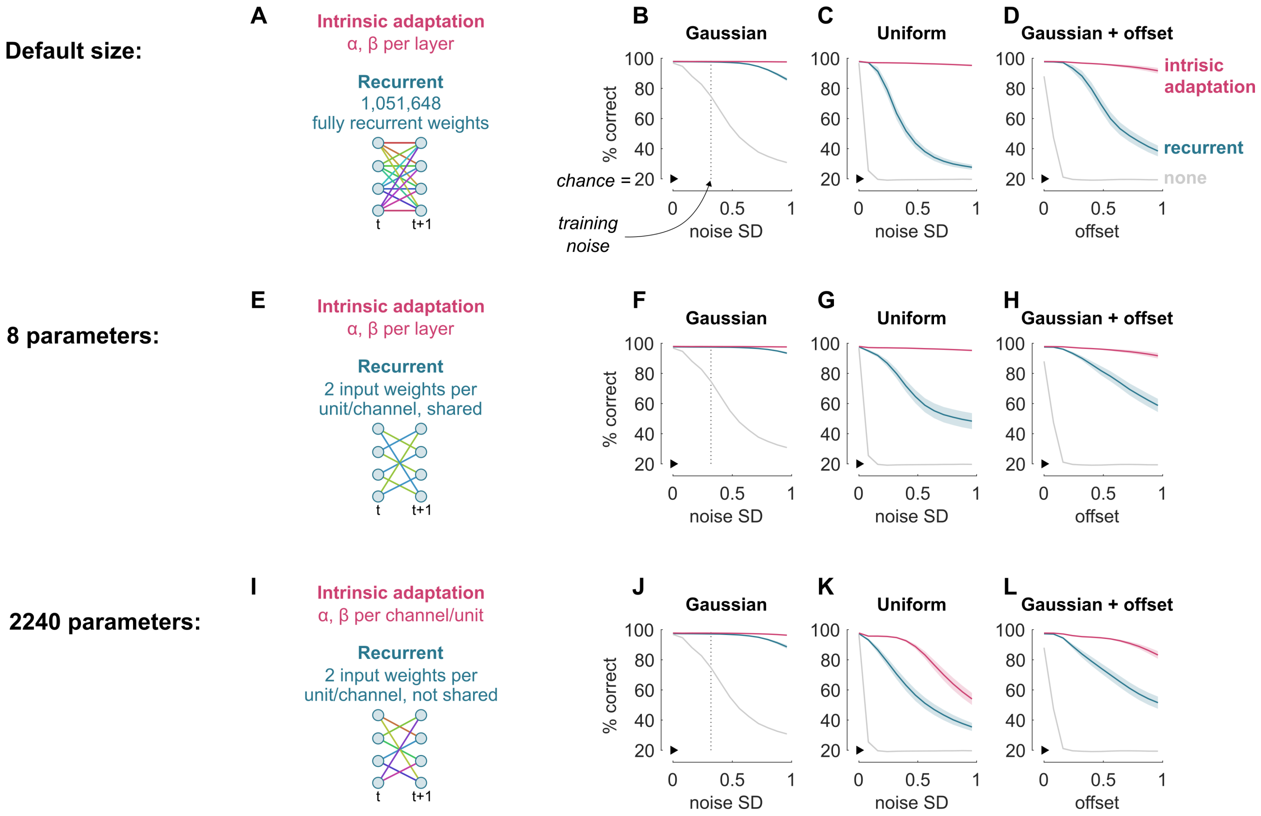

Fig. S11 A trained network with intrinsic adaptation is more robust than a recurrent neural network with the same number of parameters (expanding on Fig. 7). (**A-D**), Results for the default network sizes shown in **Fig. 7**. The network with intrinsic adaptation has 8 adaptation parameters: one α 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

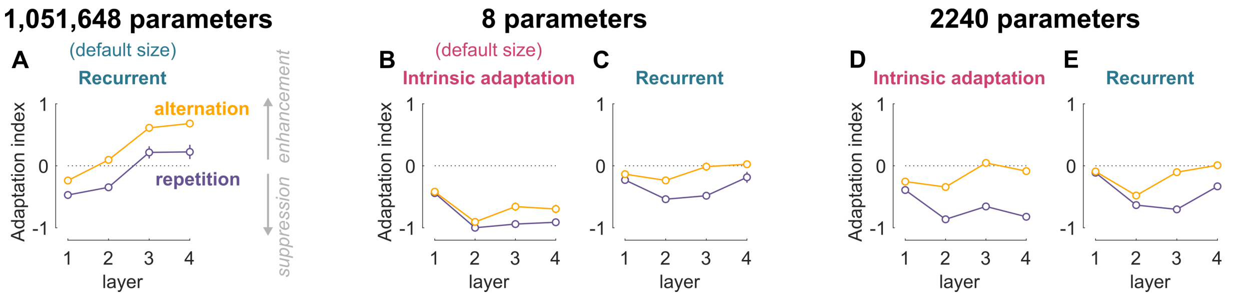

Fig. S12 Adaptation learnt by the recurrent network did not necessarily lead to repetition suppression (expanding on Fig. 7). Average adaptation index per layer for stimulus repetitions and alternations for the trained networks of **Fig. S11** (error bars are standard error of the mean for 30 random 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,

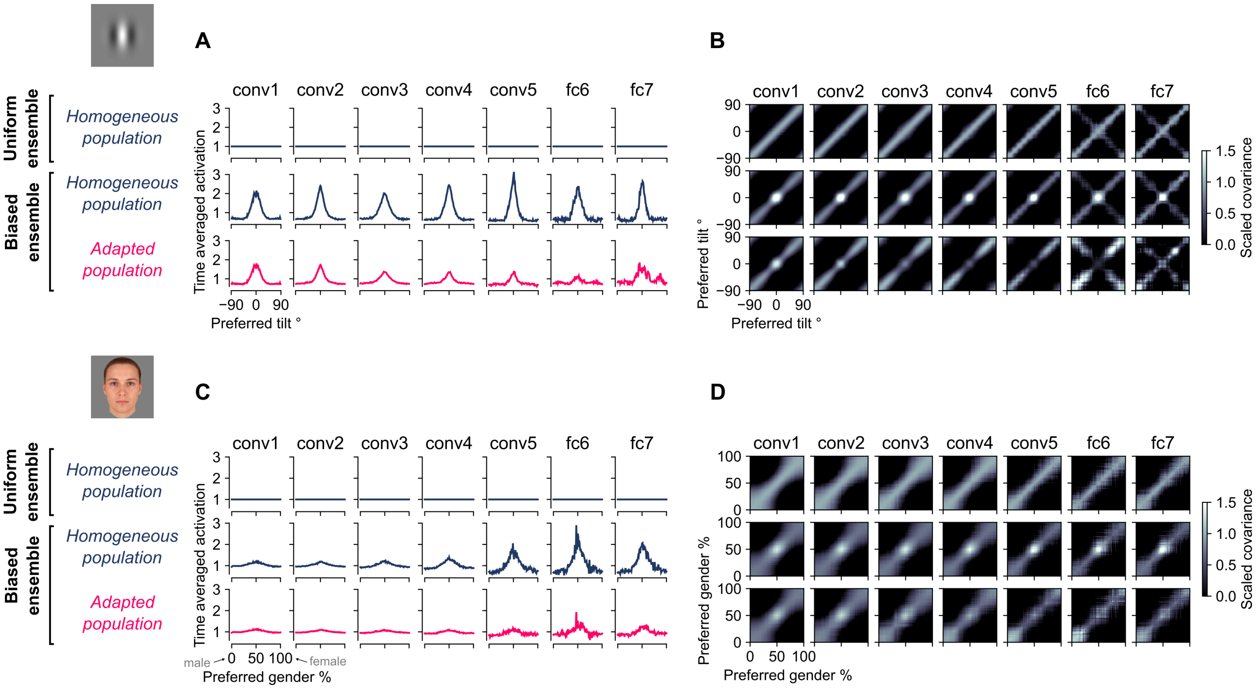

Fig. S13 Intrinsic suppression reduces time-averaged responses and decorrelates responses for biased stimulus ensembles. (**A**), Time-averaged responses for each orientation bin (i.e., units with the same preferred orientation), normalized by the time average of each orientation bin for the homogeneous 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 (created with: `webmorph.org`) 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).