Title: Deep neural networks for modeling visual perceptual learning

Abbreviated title: DNNs for VPL

Authors:

Li Wenliang (Corresponding author):
Gatsby Computational Neuroscience Unit, University College London; 25 Howland Street, London W1T 4JG, United Kingdom
+44 (0) 20 3108 8100
kevinli@gatsby.ucl.ac.uk

Aaron R Seitz:
Department of Psychology, University of California–Riverside; 900 University Avenue, Riverside, CA 92521
+1 (951) 827-6422
aseitz@ucr.edu

Number of pages: 47

Number of

figures: 24
tables: 16

Number of words for

Abstract: 249
Introduction: 638
Discussion: 1526

Conflict of Interest: The authors declare no competing financial interests

Acknowledgements: The authors are supported by the Gatsby Charitable Foundation (LW) and National Institute for Health (NEI 1R01EY023582) (AS). We are very grateful to Peter Dayan for extensive discussions on the use of deep neural networks and Merav Ahissar for discussions on the relationship of this work to the Reverse Hierarchy Theory; they also provided helpful comments on earlier versions of the manuscript. We thank Kirsty McNaught and Sanjeevan Ahilan for proofreading the manuscript and suggestions on clarifications. We also thank two anonymous reviewers for their valuable comments.

Journal section: Behavioral/Cognitive

Abstract

Understanding visual perceptual learning (VPL) has become increasingly more challenging as new phenomena are discovered with novel stimuli and training paradigms. While existing models aid our knowledge of critical aspects of VPL, the connections shown by these models between behavioral learning and plasticity across different brain areas are typically superficial. Most models explain VPL as readout from simple perceptual representations to decision areas and are not easily adaptable to explain new findings. Here, we show that a well-known instance of deep neural network (DNN), while not designed specifically for VPL, provides a computational model of VPL with enough complexity to be studied at many levels of analyses. After learning a Gabor orientation discrimination task, the DNN model reproduced key behavioral results, including increasing specificity with higher task precision, and also suggested that learning precise discriminations could asymmetrically transfer to coarse discriminations when the stimulus conditions varied. In line with the behavioral findings, the distribution of plasticity moved towards lower layers when task precision increased, and this distribution was also modulated by tasks with different stimulus types. Furthermore, learning in the network units demonstrated close resemblance to extant electrophysiological recordings in monkey visual areas. Altogether, the DNN fulfilled predictions of existing theories regarding specificity and plasticity, and reproduced findings of tuning changes in neurons of the primate visual areas. Although the comparisons were mostly qualitative, the DNN provides a new method of studying VPL and can serve as a testbed for theories and assist in generating predictions for physiological investigations.

Significance statement

Visual perceptual learning (VPL) has been found to cause changes at multiple stages of the visual hierarchy. We found that training a deep neural network (DNN) on an orientation discrimination task produced similar behavioral and physiological patterns found in human and monkey experiments. Unlike existing VPL models, the DNN was pre-trained on natural images to reach high performance in object recognition but was not designed specifically for VPL, and yet it fulfilled predictions of existing theories regarding specificity and plasticity, and reproduced findings of tuning changes in neurons of the primate visual areas. When used with care, this unbiased and deep-hierarchical model can provide new ways of studying VPL from behavior to physiology.

Introduction

Visual Perceptual Learning (VPL) refers to changes in sensitivity to visual stimuli through training or experience, and has been demonstrated in the discrimination of simple features such as orientation, contrast and dot motion direction as well as more complicated patterns (Fiorentini and Berardi, 1980; Ball and Sekuler, 1982; Karni and Sagi, 1991; Ahissar and Hochstein, 1997; Mastropasqua et al., 2015). A common characteristic of VPL is its lack of transfer to untrained stimulus conditions, such as when rotated by 90° (Fiorentini and Berardi, 1981; Schoups et al., 1995; Crist et al., 1997). Due to their retinotopic mapping and orientation tuning (Hubel and Wiesel, 1968; Blasdel, 1992; Tootell et al., 1998), early visual areas have been hypothesized to contribute to VPL and its specificity (Fahle, 2004). Despite numerous examples supporting this hypothesis (Schoups et al., 2001; Bejjanki et al., 2011; Sagi, 2011; Jehee et al., 2012; Yu et al., 2016), there is substantial evidence that specificity does not require low-level changes (Dosher and Lu, 1998; Ghose et al., 2002) and there is great controversy regarding where learning happens in the visual hierarchy (Wang et al., 2016; Maniglia and Seitz, 2018).

Most models of VPL are artificial neural networks with user-parametrized receptive fields and shallow network structures. Trained using Hebbian-like learning rules (Sotiropoulos et al., 2011; Herzog et al., 2012; Dosher et al., 2013) or optimal decoding methods (Zhaoping et al., 2003), these models can reproduce and predict behavior but rarely account for physiological data. Moreover, a key limitation of these models is that they do not address how the multiple known visual areas may jointly contribute to learning (Hung and Seitz, 2014). Other more conceptual models, such as the Reverse Hierarchy Theory (RHT) (Ahissar and Hochstein, 1997; Ahissar and Hochstein, 2004) and the Dual Plasticity Model (Watanabe and Sasaki, 2015), make predictions regarding what types of learning scenarios may lead to differential plasticity across visual areas, but these descriptive models do not predict specific changes in tuning properties of neurons. Thus, there is a substantial need for a hierarchical model that can simulate learning and simultaneously produce behavioral and neurological outcomes.

A novel approach to modeling VPL can be found in deep neural networks (DNNs) which is readily adapted to learn different tasks. These DNNs have shown impressive correspondences to human behaviors and neural data from early visual areas

and inferior temporal cortex (IT) (Khaligh-Razavi and Kriegeskorte, 2014; Yamins et al., 2014; Guclu and van Gerven, 2015; Cichy et al., 2016; Kheradpisheh et al., 2016; Eickenberg et al., 2017). This hierarchical system opens up new opportunities for VPL research (Kriegeskorte, 2015). As a start, Lee & Saxe (2014) produced theoretical analyses that resembled RHT predictions in a 3-layer linear network. Recently, Cohen and Weinshall (2017) used a shallow network to replicate relative performances of different training conditions for a wide range of behavioral data. To date, the extent to which DNNs can appropriately model physiological data of VPL remains unexplored.

In the present paper, we trained a DNN model modified from AlexNet (Krizhevsky et al., 2012) to perform Gabor orientation and face gender discriminations. The network reflected human behavioral characteristics, such as the dependence of specificity on stimulus precision (Ahissar and Hochstein, 1997; Jeter et al., 2009). Furthermore, the distribution of plasticity moved towards lower layers when task precision increased, and this distribution was also modulated by tasks with different types of stimulus. Most impressively, for orientation discrimination, the network units showed similar changes to neurons in primate visual cortex and helped reconcile divergent physiological findings in the literature (Schoups et al., 2001; Ghose et al., 2002). These results suggest that DNNs can serve as a computational model for studying the relationship between behavioral learning and plasticity across the visual hierarchy during VPL, and how patterns of learning vary as a function of training parameters (Maniglia and Seitz, 2018).

Materials and Methods

Model

An AlexNet-based DNN was used to simulate VPL. We briefly describe the network architecture here and refer readers to the original paper (Krizhevsky et al., 2012) for details. The original AlexNet consists of eight layers of artificial neurons (units) connected through feedforward weights. In the first five layers, each unit is connected locally to a small patch of retinotopically arranged units in the previous layer or the input image. These connections are replicated spatially so that the same set of features is extracted at all locations through weight sharing. The operation that combines local connection with weight sharing is known as convolution. Activations in the last convolutional layer are sent to three fully connected layers with the last layer corresponding to object labels in the original object classification task. Unit activations are normalized in the first two convolutional layers to mimic lateral inhibition.

To construct the DNN model, we took only the first five convolutional layers of AlexNet and discarded the three fully connected layers to reduce model complexity. An additional readout unit is added to fully connect with the units in the last layer, forming a scalar representation of the stimulus in layer 6. We removed the last three layers of AlexNet because they exhibited low representational similarity to early visual areas but high similarity to Inferior Temporal Cortex (IT) and hence may be more relevant to object classification (Khaligh-Razavi and Kriegeskorte, 2014; Guclu and van Gerven, 2015), and we assume that early visual areas play a more critical role for Gabor orientation discrimination. We kept all the five convolutional layers because one of our objectives was to study how learning was distributed over a visual hierarchy with more levels than most VPL models which are usually limited to 2-3 levels. In addition, activations in these five layers have been suggested to correspond to neural activities in V1-4 following a similar ascending order (Guclu and van Gerven, 2015).

The resulting 6-layer network was further modified to model decision making in the two-interval two-alternative forced choice (2I-2AFC) paradigm (Figure 1A). In this paradigm, a reference stimulus is first shown before a second target stimulus, and the network has to judge whether the second stimulus is more clockwise or more counter-clockwise compared with the reference. In our DNN model, each of the reference and target images is processed by the same 6-layer network that yields a scalar readout in layer 6, and the decision is made based on the difference between the representations with some noise. More precisely, two identical streams of the 6-layer network produce scalar representations for the reference and target images, denoted by h_r and h_t, respectively. The network outputs a clockwise decision with probability (or confidence) p by passing the difference $\Delta h = h_t - h_r$ through a logistic function

$$p = \frac{\exp(\Delta h)}{\exp(\Delta h) + 1} \quad (1)$$

This construction assumes perfect memory about the two representations which are computed using the same model architecture and parameters, and each choice is made with some decision noise. An advantage of using this 2I-2AFC architecture is that, when tested under transfer conditions (such as a new reference orientation), the network can still compare

the target with the reference by taking the difference between the representations; whereas if only one stream exists, the model cannot know the new reference orientation. We note that while this training paradigm is suitable for this network and was thus kept consistent throughout the simulations, it is different from those used in the physiological studies (Schoups et al., 2001; Ghose et al., 2002; Yang and Maunsell, 2004; Raiguel et al., 2006) with which we compare our network in the Results section. Learning could also be influenced by details of those experiments beyond what is accounted for by the present simulations.

Task and stimuli

All stimuli in the two experiments below were centered on 8-bit 227×227-pixel images with gray background.

Experiment 1

The network was trained to classify whether the target Gabor stimulus was tilted clockwise or counter-clockwise compared to the reference. We trained the network on all combinations of the following parameters:

- Orientation of reference Gabor: from 0° to 165° at steps of 15°;
- Angle separation between reference and target (0.5°, 1.0°, 2.0°, 5.0°, 10.0°);
- Spatial wavelength (10, 20, 30, 40 pixels).

To simulate the short period of stimulus presentation and sensory noise, we kept the contrast low and added to each image isotropic Gaussian noise with the following parameters:

- Signal contrast (20%, 30%, 40% to 50% of the dynamic range);
- Standard deviation of the Gaussian additive noise (5 to 15 in steps of 5).

In addition, the standard deviation of the Gabor Gaussian window was 50 pixels. Noise was generated at run time independently for each image. An example of a Gabor stimulus pair is shown in Figure 1A.

Experiment 2

The network was trained on face gender discrimination and Gabor orientation discrimination. For the face task, the network was trained to classify whether the target face was more masculine or feminine (closer to the original male or female in warping distance) than the reference face. Stimuli were face images with gender labels from the Photoface Dataset (Zafeiriou et al., 2011). 647 male images and 74 female images with minimal facial hair were manually selected from the dataset, captured in the frontal pose with the blank stare emotion. The bias in subject gender was addressed by subsampling to form balanced training and testing sets. The facemorpher toolbox (https://pypi.python.org/pypi/facemorpher/1.0.1) was used to create a reference that is halfway between a male and a female image.

To manipulate task difficulty, target images were created that varied in dissimilarity to the reference image ranging from 1 (closest to the reference) to 5 (the original male or female image) by adjusting a warping (mixing) parameter. The reference and target were morphed from the same pair of the original faces. The network was trained and tested using 12-fold cross-validation to simulate inter-observer variability. Each fold consists of images morphed from 49 males and 49 females for training (2401 pairs) and 25 males and 25 females for testing (625 pairs) randomly sampled from the full dataset. Examples of face stimuli at the five dissimilarity levels are shown in Figure 1B.

The Gabor stimulus had a wavelength of 20 pixels and the standard deviation of the Gaussian window was 50. The reference angle ranged from 0° to 165° at steps of 15° and the target image deviated from the reference by 0.5, 1.0, 2.0, 5.0 or 10.0°. For both the face and Gabor tasks in this experiment, contrast was set to 50% and noise standard deviation was set to 5.

Training procedure

In both experiments, network weights were initialized such that the readout weights in layer 6 were zeros and weights in the other lower layers were copied from a pre-trained AlexNet (downloaded from http://dl.caffe.berkeleyvision.org/bvlc_alexnet.caffemodel). The learning algorithm was stochastic gradient descent (SGD) whereby the weights were changed to minimize the discrepancy between network output and stimulus label

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \boldsymbol{v}_{t+1} \quad (2)$$

$$\boldsymbol{v}_{t+1} = \mu \boldsymbol{v}_t - \alpha \nabla_{\boldsymbol{\theta}_t} L(\boldsymbol{\theta}_t, \boldsymbol{I}, \boldsymbol{L}) \quad (3)$$

where α (0.0001) and μ (0.9) are learning rate and momentum, respectively, which are held constant through training, $\boldsymbol{\theta}_t$ is the network weights at iteration t, and $\boldsymbol{v}_t$ is the corresponding weight change. $L(\boldsymbol{\theta}, \boldsymbol{I}, \boldsymbol{L})$ is the cross-entropy loss which depends on the network weights, input image batch $\boldsymbol{I}$ of size 20 pairs and the corresponding labels $\boldsymbol{L}$. Gradients were obtained by backpropagation of the loss through layers (Rumelhart et al., 1986).

Under this learning rule, zero initialization in the readout weights prevents the other layers from changing in the first iteration, since the weights in those layers cannot affect performance when the readout weights are zeros and hence have zero gradients. This initialization can be interpreted as receiving instruction by subjects because all stimulus representations in the lower layers are fixed while the network briefly learns the task on the highest decision layer. After the first iteration, the readout weights will not be optimal due to small learning rate, and so weights in the lower layers will start to change. Under each stimulus condition, the network was trained for 1000 iterations of 20-image batches so that one iteration is analogous to a small training block for human subjects. Independent noise was generated to each image at each iteration. We outline the limitations of the model in the Discussion section.

Behavioral performance

The network's behavioral performance was estimated as the classification confidence (Equation 1) of the correct label averaged over test trials. For the Gabor task, we tested the network's performance on stimuli generated from the same parameters as in training (trained condition), and also tested on stimuli generated from different parameters (transfer conditions) including rotating the reference by 45 and 90°, halving or doubling the spatial frequencies and changing angle separations between the reference and target (precisions). 200 pairs of Gabor stimuli were used in each test condition. For the face task, performance was tested on 625 unseen validation images. Performance was measured at 20 roughly logarithmically spaced iterations from 1 to 1000.

In Experiment 2, the contribution of DNN layers to performance was estimated using an accuracy drop measure defined as follows. Under each stimulus condition, we recorded the iteration at which the fully plastic network reached 95% accuracy, denoted by t_{95}; we then trained a new network again from pretrained AlexNet weights under the same stimulus condition while freezing successively more layers from layer 1, and used the accuracy drop at t_{95} compared to the all plastic network as the contribution of the frozen layers. For example, suppose that the fully plastic network reached 95% accuracy at 100th iteration (t_{95}=100), and at this iteration, the network trained with layer 1 frozen had 90% accuracy, the network trained with first two layers frozen had 85% accuracy, then the first layer contributed 5% and the first two layers together contributed 10%. This accuracy drop does not indicate the contribution of each layer in isolation, but allows for different interactions within the plastic higher layers when varying the number of frozen lower layers.

Estimating learning in layers and neurons

After training, weights at each layer were treated as a single vector, and learning was measured based on the difference from pre-trained values. Specifically, for a particular layer with N total connections to its lower layer, we denote the original N-dimensional weight vector as $\boldsymbol{w}$ (N and $\boldsymbol{w}$ specified in the pre-trained AlexNet), the change in this vector from training as $\delta\boldsymbol{w}$, and define the layer change as

$$d_{rel1} = \frac{\sum_i^N |\delta w_i|}{\sum_i^N |w_i|} \quad (4)$$

where i indexes each element in the weight vector. Under this measure, scaling the weight vector by a constant gives the same change regardless of dimensionality, reducing the effect of unequal dimensionalities on the magnitude of weight change. For the final readout layer that was initialized as zeros, the denominator in Equation 4 was set to N, effectively measuring the average change per connection in this layer. When comparing weight change across layers, we focus on the first five layers unless otherwise stated. In addition, the following alternative layer change measures were also used

$$d_{rel2} = \frac{\sqrt{\sum_i^N |\delta w_i^2|}}{\sqrt{\sum_i^N |w_i^2|}} \quad (4a)$$

$$d_{m1} = \frac{\sum_i^N |\delta w_i|}{N} \quad (4b)$$

$$d_{m2} = \sqrt{\frac{\sum_i^N |\delta w_i^2|}{N}} \quad (4c)$$

which produced different values, but they did not change the general effects of stimulus conditions on distribution of learning in weights. For the results in the main text, we report weight change in terms of d_{rel1} unless otherwise stated. Due to weight sharing, the layer change defined above is equal to the change in the subset of weights that share the same receptive field. To measure learning of a single unit, we used the same equation but with $\boldsymbol{w}$ being the filter of each unit and N being the size of the filter.

Tuning curves

For each unit in the network, we "recorded" its tuning curve before and after training by measuring its responses to Gabor stimuli presented at 100 orientations evenly spaced over the 180° cycle. The stimuli had noise standard deviation 15, contrast 20% and wavelength 10 pixels, and one period of the sinusoidal component of the Gabor stimulus was contained within the receptive field of a layer-1 unit. The mean and standard deviation at each test orientation were obtained by presenting the network with 50 realizations of noisy stimuli, followed by smoothing with a circular Gaussian kernel. The gradients of tuning curves were computed by filtering the mean response with a circular Laplace filter. Both the Gaussian and Laplace filters had a standard deviation of 1°. Since the receptive fields were shared across locations, we only chose the units with receptive fields at the center of the image; thus, the number of units measured at each layer equals the number of filter types (channels) in that layer. Also, to ensure that units were properly driven by the stimulus, we excluded units that had mean activation over all orientations less than 1.0. The same procedure was repeated under 5 precisions and 12 reference orientations before and after training. No curve fitting was employed. On average, each training condition produced the following number of units after training for analyses: 79.4 out of 96 in layer 1, 91.0 out of 256 in layer 2, 237.3 out of 384 in layer 3, 100.3 out of 384 in layer 4, and 16.0 out of 256 in layer 5. These numbers were roughly the same for the naïve populations. Units were pooled together for the 12 reference orientations.

To compare with electrophysiological data in the literature, we measured the following attributes from tuning curves nonparametrically. The preferred orientation was determined by the orientation at which a unit attained its peak response. Tuning amplitude was taken to be the difference between the highest and lowest responses. Following (Raiguel et al., 2006), the selectivity index, a measure of tuning sharpness, was measured as

$$SI = \frac{\sqrt{[\sum_i s_i \sin(2\alpha_i)]^2 + [\sum_i s_i \cos(2\alpha_i)]^2}}{\sum_i s_i} \quad (5)$$

where s_i is the mean activation of a unit to a Gabor stimulus presented at orientation α_i for the i'th measured orientation (from 1 to 100). The normalized variance (or Fano factor, variance ratio) of the response at a particular orientation was taken as the ratio of response variance to the mean. Following (Yang and Maunsell, 2004), we measured the best discriminability of a unit by taking the minimum, over orientation, of response variance divided by tuning curve gradient squared.

To measure how much information about orientation was contained in a layer per neuron, we computed the average linear Fisher information (FI) (Seriès et al., 2004; Kanitscheider et al., 2015) at a particular orientation as

$$FI(\alpha) = \frac{1}{n} \boldsymbol{f}'(\alpha) \cdot \boldsymbol{\Sigma}(\alpha)^{-1} \cdot \boldsymbol{f}'(\alpha) \quad (6)$$

where $\boldsymbol{f}'(\alpha)$ is a vector of tuning curve gradients at orientation α for n units in that layer (those with receptive fields at the center of the image), and $\boldsymbol{\Sigma}(\alpha)$ is the corresponding response covariance matrix. In addition, independently for each unit, we measured its FI as its tuning curve gradient squared divided by response variance at the measured orientation. For FI calculation, units with activity less than 1.0 at the measured orientation were excluded to avoid very low response variance.

Experimental design and statistical analyses

In Experiment 1, the network was trained on Gabor orientation discrimination under 2880 conditions (12 reference orientation, 4 contrasts, 4 wavelengths, 3 noise levels and 5 angular separations). In Experiment 2, the network was trained

on 360 conditions in each of the Gabor and face tasks (12 reference orientations or training-testing data splits, 5 dissimilarity levels and 0 to 5 frozen layers).

We carried out our analyses on three levels. On the behavioral level, the effects of training and test angle separation on performance was tested using linear regression. On the layer level, the effects of layer number and training angle separation on layer change were tested also using linear regression. To see whether the distribution of learning differed between tasks, we used two-way ANOVA to test whether there was an effect of task on layer change. At the unit level, we tested for significant changes in various tuning curve attributes recorded in the literature, using K-S for distributional changes, two-sample Mann-Whitney U for changes in tuning curve attributes from training, and two-way ANOVA when neurons were grouped according to naïve/trained and their preferred orientations. Finally, to test whether there was a relationship between the network's initial sensitivity to the trained orientation, we used a regression model described with the results. The significance level for all tests was 0.01. Bonferroni corrections were applied for multiple comparisons.

All code was written in Python with Caffe (http://caffe.berkeleyvision.org) for DNN stimulations and statsmodel (Seabold and Perktold, 2010) for statistical analyses. Code and stimulated data are available on request.

Results

Behavior

The network was trained to discriminate whether the target Gabor patch was more clockwise or more counter-clockwise to the reference, repeated in 2880 conditions (12 reference orientation, 4 contrasts, 4 wavelengths, 3 noise levels and 5 precisions or angle separations). The performance trajectories grouped by precision are shown in Figure 2A (top row) for the trained and transfer conditions of rotated reference orientations (clockwise by 45 or 90°) and scaled spatial frequencies (halved or doubled). The accuracy measured at the first iteration (analogous to naïve real subjects) indicates the initial performance when only the readout layer changed. Both initial performance and learning rate under the trained condition were superior for the less precise tasks, consistent with findings from the human literature (Ahissar and Hochstein, 1997; Jeter et al., 2009). Percentage correct increased in a similar way to human data on Vernier discrimination (Herzog and Fahlet, 1997) and motion direction discrimination (Liu and Weinshall, 2000). Convergence of performance for the transfer stimuli required more training (note the logarithmic x-axis in Figure 2A) than for the trained stimuli, which may imply that much more training examples are necessary to achieve mastery on the transfer stimuli, consistent with some studies of tactile perceptual learning (Dempsey-Jones et al., 2016).

Moreover, we characterized the dynamics of transfer by calculating the transfer index as the ratio of transfer accuracy to the corresponding trained accuracy. As shown in Figure 2A (bottom row), this ratio decreased initially but started to rise slowly for all conditions, and the trajectory for orthogonal transfer (ori+90) under the highest precision were almost flat towards the end of training. Similar reduction in transfer index with increasing training sessions has been demonstrated in human experiments (Jeter et al., 2010).

Figure 2B shows the final transfer performance grouped by transfer conditions and training precision. All transfer accuracies were below 1.0, especially for the orientation transfers, indicating learning specificity. A linear regression on transfer accuracy showed a significant positive main effect of log angle separation in all four transfer conditions ($p<0.0001$; see Extended Data Table 2-1 for details), implying better transfer performance for less precise training. This result is consistent with experimental data and theoretical prediction that greater specificity happens in finer precision tasks (Ahissar and Hochstein, 1997; Liu, 1999; Liu and Weinshall, 2000).

However, the test precisions were the same as used during the respective training conditions and thus varied in intrinsic discriminability, which could have determined the observed transfer pattern. We then tested each trained network on all angle separations to see whether we would reproduce human data (Jeter et al., 2009) where a difference in test precision affects transfer more than training precision. Indeed, Figure 2C shows a strong positive main effect of test separation ($p<0.001$ and $R^2>0.35$ for all transfer conditions); however, we also found that *training* separation had a significant effect ($p<0.001$) in all transfer conditions, and the effect size was the smallest in orthogonal orientation transfer (see Extended Data Table 2-1 for details). The more substantial transfer to 45° from trained orientations compared to 90° could be due to a larger overlap between the transfer orientation and the trained orientation representation of the network units. Thus, despite

the observation of diminishing transfer with increasing precision when the training and test precisions are equal (diagonal lines in Figure 2C), the analyses across *all* precision combinations predict that transfer is more pronounced from precise to coarse discrimination than vice-versa, although transfer can be very small at the orthogonal orientation.

Overall, these behavioral findings are consistent with extant behavioral and modeling results of perceptual learning. Furthermore, the DNN model can simulate a wide number of training conditions and makes predictions regarding the relative performances and learning rates of the trained stimuli compared to that of transfer stimuli. However, it is expected that some details of this DNN's behavioral results will necessarily differ from that of experimental data.

Distribution of learning across layers

Learning in the weight space

We next examined the timecourse of learning across the layers as a function of precision, calculated using Equation 4 and shown in Figure 3A. Overall, all layers changed faster at the beginning of training in coarse than in precise angle separations, training precise angle separations produced greater overall changes. Of note, while the highest readout layer (layer 6) changed faster than the other layers (Saxe, 2015), this is likely a consequence of zero-initialization in the readout weights. This suggests that when performance was at chance level, due to naivety to the task, information about the stimulus label cannot be passed on to lower layers while the performance is close to chance. Hence, we focus our discussion on layers 1 to 5, whose weights were copied from a pre-trained AlexNet.

To characterize learning across layers, we studied when and how much each layer changed during training. To quantify when significant learning happened in each layer, we estimated the iteration at which the gradient of a trajectory reached its peak (peak speed iteration, PSI) which are shown in Figure 3B. In layers 1 to 5, we observed significant negative main effects of log angle separation (β=-46.02, t(14396)=-52.93, p≈0.0, R^2=0.40) and layer number (β=-2.07, t(14396)=-13.65, $p=3.6\times10^{-42}$, R^2=0.0031) and positive interaction of the two (β=2.93, t(14396)=11.16, $p=8.0\times10^{-29}$, R^2=0.0050) on PSI, suggesting that layer change started to asymptote later for lower layers and smaller angle separations. For individual precision conditions, a linear regression analysis showed a significant negative effect of layer number on PSI only in the two most precise tasks ($p<0.0001$, see Extended Data Table 3-1 for details). Hence, under high precisions, the order of change across layers is consistent with the Reverse Hierarchy Theory prediction that higher visual areas change before earlier ones (Ahissar and Hochstein, 1997).

The final layer change at the end of training is shown in Figure 3C. For a better visual comparison, we calculated the relative layer change under each stimulus condition by taking the ratio of layer change against the change resulted from training at the coarsest angle separation, keeping other stimulus conditions fixed (Figure 3D). The results suggest that changes in lower layers increased by a larger factor than higher ones except layer 5. A linear regression analysis on the changes in layers 1-5 revealed significant negative main effects of log angle separation (β=-0.0060, t=-87.25, p≈0.0, R^2=0.34) and layer number (8.6×10^{-4}, p≈0.0, R^2=0.092) and positive interaction of the two (0.00010, t=49.23, p≈0.0, R^2=0.082). The interaction of angle separation × layer number on layer change is consistent with the prediction that higher-precision training induces more change lower in the hierarchy (Ahissar and Hochstein, 1997).

Change of information about orientation

While we have considered thus far the changes in the weights of the DNN, there is still a question of how the information about orientation changed across layers, and how this may vary as a function of training precision. We address this by showing in Figure 4 the covariance-weighted linear Fisher information (FI, Equation 6) of the trained and naïve unit population at each layer and each test orientation when trained at each angle separation. The tuning curves used to evaluate FI were obtained from the units with receptive fields at the center of the image (see Materials and Methods). A prominent observation is that FI increased most dramatically at the highest layers under the most precise task and diminished towards lower layers and coarser precisions. Lower layers saw significant FI increase only in the most precise tasks, whereas higher layers increased FI in all precisions.

However, the quantitative trend of FI increase was contrary to the layer change where more substantial learning happened in the lower layers. Notably, despite the large change in layer 1 weights (Figure 3C), there was no visible change in FI. The large FI increase in higher layers may not be surprising due to a single hierarchy with readout on the top layer and the accumulation of weight changes from the lower layers.

In addition, we observed patterns of FI over orientations. After training at the finest precision, the top layers exhibited significant and substantial increase of FI around the trained orientation; FI fell off around 20° away from trained orientation but remained noticeable until the orthogonal orientation. This could account for the transfer behavior predicted by the network (Figure 2) where learning transferred more substantially if the Gabor stimulus was rotated by 45°, rather than 90° which was a common transfer condition tested in experiments.

These data show that the increase of information about orientation over network layers changes as a function of training precision. Later, we will discuss how this information in the pre-trained network may affect learning (Figure 10).

Effect of feature complexity on distribution of learning

Are the observed layer changes prescribed solely by the network structure and learning rule regardless of the task? To find out whether these patterns were task specific, we simulated learning of a "higher-level" face gender discrimination task and investigated its effect on the distribution of learning in network weights compared with Gabor orientation discrimination. In the face task, difficulty was manipulated by morphing between male and female face images, and the network was trained to discriminate whether the target was more masculine or feminine compared to the reference. Both tasks were repeated under 360 conditions (12 reference orientations for the Gabor task or 12 training-testing data splits for the face task, 5 dissimilarity levels and 0 to 5 frozen layers). By the end of training, the fully plastic network reached test accuracy above 95% for all stimulus conditions. In this section, we assume that learning in stimulus representation happened in the lower 5 layers and ignore the changes in layer 6.

Due to the hierarchical representation from earlier to later visual areas, one may hypothesize that learning in lower layers of the DNN would play a more important role in performance for the Gabor task relative to the face task. To quantify the contributions of layers, we measured how much accuracy dropped when more lower layers were frozen (keeping weights fixed) at a particular iteration during training (see Methods and Materials). Results are shown in Figure 5A. Performance in the Gabor task dropped considerably when freezing layer 2 onwards; whereas in the face task, learning was significantly impaired only when freezing all the four layers or more. While this freezing technique is unnatural, and it is possible that compensatory changes occurred that did not reflect properties of learning in the fully plastic network, these results support the hypothesis that the higher layers are more informative for more complex stimulus judgements and the earlier layers are so for more precise and simpler ones.

To further test whether the distribution of learning depended on task, we calculated the proportions of layer change in the first five layers by normalizing these changes against their sum (Figure 5B). Since the first layer did not have a large contribution to performance (Figure 5A), and response normalization happened after the first layer, we also compared the layer change proportions after training while freezing the weights in layer 1 at pre-trained values. A two-way ANOVA revealed a significant interaction of layer × task on layer change proportion in both network conditions ($p<0.0001$, see Extended Data Table 5-1 for details), suggesting that task indeed changed the distribution of learning over layers. Post-hoc analysis showed a significant increase in weight change proportions in layers 4 and 5 and a significant decrease in layer 2 (Mann-Whitney U, threshold $p=0.01$, Bonferroni-corrected for 5 or 4 layers). Therefore, more weight change happened in lower layers when learning the "low-level" Gabor task and higher layers acquired more change in the "high-level" face task, consistent with theories of VPL (Ahissar and Hochstein, 1997; Ahissar and Hochstein, 2004; Watanabe and Sasaki, 2015).

One should be careful when interpreting values of layer change defined by Equation 5. For instance, the layer with maximum change varied between layers 1-3 depending on how these changes were calculated, although the general effects of precision and task on layer change were consistent under other weight change measures (Extended Data Figure 5-1). In addition, it may be tempting to infer the relationship between layer contribution and change; however, freezing early layers created different interactions between the higher plastic layers, which makes it difficult to compare layer contribution with the layer change obtained by the fully plastic network.

Thus, by analyzing the weight changes in the network layers, we have shown that the distribution of learning over the network hierarchy moves towards lower layers for more precise discriminations of simple features, and to higher layers for less precise or more complex stimuli, such as faces. To see to what extent this DNN model can reflect changes in the brain during perceptual learning, we compare activations of individual units in the network with activities of real neurons in the brain recorded by electrophysiology in the following section.

Tuning changes of single units

Single units in different layers of the DNN model were examined to see whether the changes in these units were similar to those in monkey neurons after VPL. A key target of this investigation was to address computationally some of the significant findings in the literature that had led to diverging interpretations of plasticity within the visual hierarchy. Previous research on DNNs found that the representational geometries of layers 2 and 3 (but not layer 1) in this network are analogous to those in human V1-3 (Khaligh-Razavi and Kriegeskorte, 2014), so we focused our analyses on layers 2 and higher.

We compared the network units with V1-2 and V4 neurons recorded in four electrophysiological studies; animals in these studies were trained to discriminate orientations of Gabor patches. Schoups *et al.* (2001) discovered an increase in the tuning curve slope at the trained orientation for units tuned to 12-20° away from trained orientation. The same group (Raiguel et al., 2006) later found in V4 a similar change along with other effects of training. These studies used a single-interval 2AFC training paradigm with implicit reference that was never shown. On the other hand, Ghose et al. (2012) used a two-interval 2AFC training paradigm in which an irrelevant feature of the stimulus (spatial frequency) varied between two values through training. Contrary to Schoups et al. (2001), they did not find significant changes in V1-2 regarding orientation tuning (except one case), but the final discrimination thresholds reached by the subjects were higher. This group later revealed several changes in V4 using the same training paradigm (Yang and Maunsell, 2004). We hypothesized that the differences between these studies may be simply explained by differences in precision during training. To test this, we trained the network on a common task, the 2I-2AFC Gabor discrimination paradigm, and tested whether changing training precision, holding everything else constant, was sufficient to reconcile the gross differences observed between these studies.

Overall, it appears that V1-2 only changed when the discrimination threshold was small (0.5-1.2° by Schoups et al., smaller than 3.3-7.3° by Ghose et al.), and V4 changed in both studies though the discrimination thresholds were similar to each other (1.9-3.0° by Raiguel et al. and 1.9-5.4° by Yang and Maunsell). Given the pattern of change in layer Fisher information demonstrated earlier (Figure 4), we hypothesized that

a) the contradictory results in V1-2 (corresponding to lower layers in the network) were due to the mismatch of the final thresholds reached by the subjects in the two V1-2 studies, and
b) the change in V4 (corresponding to higher layers in the network) should persist from fine to coarse precisions in which V1-2 did not change.

To test the first hypothesis using the DNN model, we roughly matched the angle separations trained on the network with the discrimination thresholds reached by the monkeys, 1.0° for high and 5.0° for low precisions, and compared network units in layer 2 (layer 3 in Extended Data) with real V1-2 neurons. We then compared the tuning attributes of units in layer 5 (layer 4 in Extended Data) with those of V4 neurons to test our second hypothesis. These choices were mainly motivated by previous research on comparing this particular network with the visual brain areas (Khaligh-Razavi and Kriegeskorte, 2014; Guclu and van Gerven, 2015; Eickenberg et al., 2017). It should be noted that we used the same spatial frequency for the reference and target (more similar to Schoups et al. (2001) than to Ghose et al. (2002)); thus, while we found that modelling the differences in thresholds is sufficient to account for many differences in physiological findings between the four studies, it is likely that other task and stimulus differences also contributed to the different profiles of learning.

Tuning curves were obtained from the units with receptive fields at the center of the image and are shown in Figure 6A for layers 2 and 5. Many of the layer-2 units showed bell-shaped tuning curves with clear orientation preferences, mirroring the reported similarity between these two layers with human early visual. In addition, there were also intriguing tuning curves that showed more than one tuning peaks, which may not be physiologically plausible. Tuning curves in layer 5 were harder to interpret with some units showing clear orientation tuning and the rest likely tuned to features other than Gabor patches.

Lower layers trained under high precision (Schoups et al., 2001)

We first investigated whether this DNN model could reproduce findings of Schoups et al. (2001) on V1 neurons, who found that monkeys obtained very small thresholds (around 1°) and V1 neurons showed a change in the slope of tuning curve at trained orientation for neurons tuned between 12-20° from trained orientation (Figure 6B). Similar to their procedure, we took the orientation of the maximum response of a unit as its preferred orientation and grouped all units by preferred orientation relative to trained orientation. The slope of the tuning curve at trained orientation was evaluated after normalization by maximum activation (as done for neural data in Schoups et al. (2011)). These results for layer-2 units

(Figure 6C) showed a similar slope increase for units tuned away from trained orientation, overlapping but broader than the range found in V1 neurons, when trained under high precision but not low precision.

Despite a change in tuning slope, Schoups et al. (2001) found that the preferred orientations of neurons were evenly distributed over all orientations before and after training (Figure 6D). This was also the case for the network units which showed no significant difference between trained and naïve distributions of preferred orientation in either of the precisions (Figure 6E, $p>0.9$, K-S, see Extended Data Table 6-2 for details).

Overall, these data from lower-layer units demonstrated an impressive qualitative similarity to data from Schoups et al. (2001) when the network was trained on high precision. We next look at the changes in low precision training.

Lower areas trained under low precision (Ghose et al., 2002)

Contrary to Schoups et al. (2001), Ghose et al. (2001) found very little change in V1-2 neurons after training. The tuning amplitude of V1 and V2 neurons tuned around the trained orientation did not differ significantly from other neurons after training (Figure 7A). However, the monkeys trained by Ghose et al. (2002) achieved relatively poorer discrimination thresholds (around 5°), and when we modelled this as low precision training, we also found no significant effect of preferred orientation on tuning amplitude at layers 2-3 (Figure 7B, $p=0.30$, Mann-Whitney U, see Extended Data Table 7-1 for details). In addition, Ghose et al. (2002) found no significant change in the variance ratio for V1-2 neurons (Figure 7C), which was replicated by our network units at both precisions (Figure 7D, $p>0.6$, Mann-Whitney U, see Extended Data Table 7-2 for details).

Ghose et al. (2002) did observe a decrease in the number of neurons tuned to the trained orientation in V1 but not in V2 (not shown), contrary to Schoups et al. (2001). In our model, no such change was found in either high or low precision (Figure 6D); hence, here the model did not agree with the data.

Thus, under low precision, we did not find significant changes in tuning attributes in lower layers, which was also the case for (Ghose et al., 2002) except for the preferred orientation distribution in V1. Together with the comparisons under high precision in the previous section, our first hypothesis was supported by our simulations. The key in replicating these data is the observation that the precision of training has a profound effect on the distribution of learning across layers. By accounting for the different orientation thresholds found across studies and labs, the DNN model can well address the different observations. In addition, the partial specificity of the network trained under low precisions (Figure 2B) did not require orientation-specific changes in lower layers, in line with previous models and data (Petrov et al., 2005; Sotiropoulos et al., 2011; Dosher et al., 2013).

Higher layers compared to data of Raiguel et al. (2006)

While neurons in primary visual cortex showed plasticity that was largely limited to high-precision conditions, neurons in V4 generally showed more susceptibility to VPL (Yang and Maunsell, 2004; Raiguel et al., 2006). We hypothesized that changes in V4 neurons should happen in both low and high precisions, and tested this hypothesis by comparing the tuning attributes of units in layer 5 (layer 4 in Extended Data) with recordings in those studies.

We first compared the network units with the result of Raiguel et al. (2006) that, similar to V1 (Figure 6B), V4 neurons tuned 22-67° away from trained orientation significantly increased their slopes at trained orientation (Figure 8A). This effect was replicated in the model units (Figure 8B) which, unlike layer 2 (Figure 6C), showed significant increase in tuning slope not only in high but also in low precisions ($p<0.0002$, two-way ANOVA, see Extended Data Table 8-1 for details), although these units were tuned much closer to trained orientation than V4 neurons.

Furthermore, in monkey V4, the distribution of preferred orientation became non-uniform after training (Figure 8C). In our model, this distribution also became significantly different from a uniform distribution as revealed by a K-S test in both conditions ($p<0.003$, K-S, see Extended Data Table 8-2 for details). This is in contrast to layers 2 which only show such a change for high precision (Figure 6E). There was also a substantial increase in the number of neurons tuned very close to trained orientation.

The strength of orientation tuning, measured by the selectivity index defined in Equation 5, was found to increase after training in V4 (Figure 8E), and similar results were found in the model (Figure 8F) where SI increased significantly when

trained under high precision ($p<0.0001$) but not under low precision ($p=0.10$, Mann-Whitney U, see Extended Data Table 8-3 for details). While these results suggest that higher-layer units became sharper after training when the precision was high, the shape of this distribution in layers 4 and 5 did not match real V4 neurons.

Raiguel et al. (2006) also discovered that training reduced response variability at the preferred orientation quantified by the normalized variance (Figure 8G). We found the same in our model units (Figure 8H) at both precisions ($p<0.001$, Mann-Whitney U, see Extended Data Table 8-4 for details), suggesting that noisy units in the lower layers might be rejected by higher ones (Dosher and Lu, 1998).

Higher layers compared to data of Yang & Maunsell (2004)

Finally, we address whether the network units also replicated the findings of Yang & Maunsell (2004) where monkeys achieved a threshold comparable with (Raiguel et al., 2006). Overall, in contrast to the V1 study from the same group (Ghose et al., 2002), Yang and Maunsell (2004) found many tuning changes in V4. First, tuning amplitude of V4 neurons increased significantly after training (Figure 9A), and the same was observed in the model under both precisions ($p<0.0001$, Mann-Whitney U, see Extended Data Table 9-1 for details). Second, V4 neurons significantly lower their best discriminability (Figure 9C) after training (lower implies better discriminability), suggesting that finer orientation differences could be detected. Units in the model (Figure 9C) reproduced the same change in both precision levels ($p<0.0005$, Mann-Whitney U, see Extended Data Table 9-2 for details).

Yang & Maunsell (2004) went on to show that these changes were not simply a result of the scaling, but rather the narrowing, of tuning curves (Figure 9E). In layer 5 of the model, the tuning widths of naïve units were already smaller than trained V4 neurons; nonetheless, we found that, under high precision, the mean activation of layer-5 units (Figure 9F) in the non-preferred orientation range (45° away from preferred orientation) was significantly more reduced than that in the preferred orientation range (within 45° of preferred orientation, $p<0.0001$, two-way ANOVA, see Extended Data Table 9-3 for details).

More importantly, the mismatch in the tuning width between network units and real neurons could explain several quantitative discrepancies between model units and V4 neurons seen previously, including the different group of units that increased the tuning slope (Figure 8A and B), the sharp peaks in preferred orientation distributions (Figure 8C and D), and the higher selectivity index distributions than real neurons (Figure 8E and F). The existence of these narrow tuning curves and its consequences may require more accurate physiological measurements to verify.

Therefore, multiple changes found in layer 5 at low precision provide strong evidence supporting our second hypothesis. To conclude the comparisons with physiological studies, we find that the DNN model replicates a number of the single cell results found in extant studies of primate visual cortex. In general, it appears that the network units increased their responses at orientations close to the trained orientation, providing more informative response gradients essential for performance. Changes are more substantial in higher layers through feedforward connections, resulting in sharper tuning curves, larger tuning amplitudes and a significant accumulation of tuning preference close to the trained orientation. Also, noisy neurons in lower layers may be rejected by higher ones after training, reducing response variability (Dosher and Lu, 1998). Nonetheless, it is important to note the quantitative differences in data between the DNN model and primates as noted in the comparisons above.

Linking initial sensitivities to weight changes

A key question to understanding learning in the DNN is the extent to which learning depends on the initial conditions of the network. Here, we focus on the first five layers and explore whether there is relationship between initial sensitivity to the trained orientation and weight changes.

To test whether a larger *layer* sensitivity to the trained stimulus may give rise to more learning in the weights of this layer, we correlated the untrained layer FI with layer change measured by Equation 4. Although the network's initial state was the same for all simulations, the reference orientations varied in 12 values to which the network were differentially sensitive. We used the following regression model to test the contributions of various factors

$$d_{rel1} = \beta_0 + \beta_{FI}FI + \sum_{l=1}^{5}[\beta_l + \beta_{l\times FI}FI] + \sum_{s=1}^{5}[\beta_s + \beta_{s\times FI}FI] \quad (7)$$

where β_{FI} is the linear coefficient for FI, β_l and β_s are the main effects of layer and angle separation, respectively, each with 5 categorical levels, and $\beta_{l\times FI}$ and $\beta_{s\times FI}$ are the interactions of layer × FI and angle separation × FI, respectively. ANOVA on this model showed significant main effects of β_l and β_s ($p<0.0001$) which accounted for 21% and 46% of variance, but the three effects involving FI were insignificant ($p>0.1$, see Extended Data Table 10-1 for details). This means that the weights in a layer did not change more when it was more sensitive to the trained reference orientation.

To test whether a larger *unit* sensitivity to the trained stimulus may give rise to more learning in the weights of the unit, we used the untrained fisher information of the units (tuning gradient squared divided by variance at the trained orientation) as a proxy for sensitivity, and correlated this with their weight changes. We show in Figure 10B the relationship between these two quantities for each training condition after rescaling. The correlations were generally small except at layer 5, and despite the positive correlation (consistent to that observed in Raiguel et al. (2006)), there was a tendency for less sensitive units to also change substantially. Neurons in layer 3 showed the lowest correlation compared with those in other layers even though this layer had the highest pre-train FI. The same regression analyses above revealed that all effects were significant ($p<0.0001$, see Extended Data Table 10-2 for details), but layer and angle separation explained 2.6% and 12.4% of the variance in weight change, whereas the effects involving FI (β_{FI}, $\beta_{l\times FI}$ and $\beta_{s\times FI}$) together explained 1.1%. In addition, we also saw that the distribution of FI became more spread-out after training (Figure 10C), particularly for higher layers, suggesting that training did not improve FI for all units equally. Under the finest precision, there was a positive correlation between the initial untrained FI and change in FI in the five layers (0.1-0.5, $p<0.0001$).

Therefore, although the network's initial sensitivity to the trained orientation might influence the magnitude of learning, its effect size was less considerable compared to layer and training precision. On the layer level, a higher initial sensitive lowered the amount of learning; and on the unit level, the effect of initial sensitivity on learning was mixed.

Discussion

We find that the DNN model studied here is a highly suitable model to investigate visual perceptual learning. On the behavioral level, when the network was trained on Gabor orientation discrimination, the network's initial performance, learning rate and degree of transfer to other reference orientations or spatial frequencies depended on the precision of the stimuli in a similar manner as found in behavioral and theoretical accounts of perceptual learning. We found slower learning with less transfer under finer training precision as predicted by Reverse Hierarchy Theory (Ahissar and Hochstein, 1997; Ahissar and Hochstein, 2004); however, the model also suggests that test precision had a major influence on the transfer performance and was able to account for greater transfer from precise to coarse stimuli (Jeter et al., 2009). This model makes the novel prediction that high-precision training transfers more broadly to untrained and coarse orientation discriminations than low-precision training. On the layer level, increasing the task precision resulted in slower but more substantial changes in lower layers. In addition, learning to discriminate more complex features (e.g. face gender discrimination) resulted in relatively greater changes in higher layers of the network. On the unit level, only high-precision tasks significantly changed tuning curve attributes at lower layers, whereas units in the higher layers showed more robust changes across precisions. Various changes found in the network units mirrored many electrophysiological findings of neurons in monkey visual cortex. Overall, this DNN model, while not originally designed for VPL, arrived at impressively convergent solutions to behavioral, layer-level and unit-level effects of VPL found in the extant literature of theories and experiments.

The present findings help reconcile disparate observations in the literature regarding the plasticity in early visual cortex. While Schoups et al. (2001) found changes in orientation tuning curves of neurons in primary visual cortex, these results were not replicated by Ghose et al. (2002). The DNN model provides a parsimonious explanation for these results, accounting for the discrepancy as related to the different discrimination thresholds reached by the subjects between those experiments. In Schoups et al. (2001), the subjects reached lower thresholds than those in Ghose et al. (2002), which was sufficient to move plasticity down to the lower layers of the DNN model. Furthermore, through a number of observations on higher-layer units that were similar to V4 neurons, the model verified our hypothesis that these neurons are more susceptible to changes after VPL even under low precisions.

Compared to a shallower model that would have no problem learning the tasks, a deeper network has the advantage of demonstrating the distribution of learning over finer levels of hierarchy. Also, it serves as an example where recurrent or feedback processes may not be necessary to capture the distribution and order of learning over layers, as lower layers changed before higher layers in the absence of inhomogeneous learning rate or attentional mechanisms. Moreover, despite its biological implausibility, weight sharing in the convolutional layers did not result in substantial unwanted transfer over reference orientation or spatial frequency in this study; location specificity was also demonstrated in a shallow convolutional network (Cohen and Weinshall, 2017), although breaking this weight sharing may still be necessary to better interpret learning.

Despite the striking resemblance between the DNN model output and primate data, it is worth noting that a number of choices we made regarding training paradigm, noise injection, learning rule and learning rate may have significant consequences to the results reported here. First, the network was trained on fixed differences which differed from using staircases as done in many VPL studies; the 2I-2AFC procedure, which avoided explicit definition of label in transfer, may produce different learning outcomes compared to the 2AFC without a reference used by Schoups et al. (2001) or the varied spatial frequencies used by Ghose et al. (2002). Second, there was no source of internal noise in the middle layers to generate behavioral variability (Acerbi et al., 2014), and the readout weights were zero-initialized to minimize learning variability in contrast to previous network models that used random initialization to simulate multiple subjects (Sotiropoulos et al., 2011; Talluri et al., 2015). Third, our learning rule SGD does not compare favorably with Hebbian-like learning methods (Sotiropoulos et al., 2011; Dosher et al., 2013; Talluri et al., 2015) that are more biologically plausible, although more biologically plausible versions have been proposed (Lillicrap et al., 2016; Scellier and Bengio, 2017). Other studies suggested that small differences in training paradigms, including precision as shown in the present data, can have a significant impact on learning and plasticity (Hung and Seitz, 2014), and it will be important target for future studies to research the contributions of the numerous other differences between these studies.

Another issue regards the distinction between representation learning in the lower layers and task or decision learning in the readout layer. This DNN can perform very well even if only the final layer is plastic, in which case both forms of learning are mixed into the same layer. In our simulations, the use of a small learning rate was necessary to ensure learning stability on precise tasks, but this also had the effect of distributing more change to the lower layers. Other schemes can be used to control learning between layers, such as pre-training on an easier task or readout weight regularization. Direct connections from lower layers to the readout layer are also possible given a reasonable weight initialization. Future research will be necessary to examine the consequence of such alternative schemes on the predictions regarding distribution of learning.

A long-standing topic in research of neural coding regards the efficiency of sparse representations in visual cortex (Barlow, 1961; Olshausen and Field, 1997). This raises a question of whether perceptual learning "raises all boats" or "makes the rich get richer" and mostly the best tuned neurons change the most. Some support for the latter possibility is found in physiological studies where neural responses changed primarily in the most informative neurons. The present DNN model appears insufficient to well-address sparsity in learning. While we found (Figure 10) small positive correlations between initial Fisher information and weights changes from training, consistent with the notion of "the rich get richer", there were also substantial changes in many of the insensitive units across layers. This may be due to the fact that the network was only trained on one task and was not consistently performing the many visual tasks involving natural stimuli that humans and animals must perform on a daily basis. Hence, to better address learning sparsity, a network may need to be trained simultaneously on a number of tasks.

Furthermore, one must be cautious in inferring homologies between layers of DNNs and areas in the brain. The similarity between the network and visual areas depends on the layer parameters (such as number of filters, receptive field size, etc.) in a subtle manner (Pinto et al., 2009; Yamins and DiCarlo, 2016). It is also unknown how much our analyses on the changes in the weights (instead of unit activity) can inform us about the synaptic changes caused by VPL. Comparing results across different DNNs may help us understand which results are robust against details of model architecture.

The simulations shown in the present manuscript just touched the surface of the vast VPL literature. While beyond the scope of the present study, future modeling targets can be considered in pursuit of many of perceptual learning phenomena. For example, DNNs may be used to replicate other psychophysical phenomena, including disruption (Seitz et al., 2005), roving (Zhang et al., 2008; Tartaglia et al., 2009; Hussain et al., 2012), double training (Xiao et al., 2008; Zhang et al., 2010) and

the effects of attention (Ahissar and Hochstein, 1993; Byers and Serences, 2012; Bays et al., 2015; Donovan et al., 2015) and adaptation (Harris et al., 2012). Moreover, small variations in training procedures can lead to dramatic changes in learning outcome (Hung and Seitz, 2014); therefore, it is important for future simulations to understand how such details may impact learning in DNNs.

In addition, DNNs may provide a straightforward way to model the 'when' and 'where' aspects of VPL that would be otherwise difficult to test experimentally on subjects. As discussed in previous reviews (Seitz, 2017; Watanabe and Sasaki, 2015), it is likely that VPL involves more areas than the two or three-layer models of pattern-matching representation and nonlinear readout that typify the field (Dosher and Lu, 2017). The distribution and timecourse of plasticity could be further examined in other tasks, using other DNNs or layer change measures.

In conclusion, we find that deep neural networks provide an appropriate framework to study VPL. An advantage of DNNs is that they can be flexibly adapted to different tasks, stimulus types and training paradigms. Also, layer- and unit-specific changes resulting from learning can be examined and related to fMRI and electrophysiological data. While some caution is needed in interpreting the relationship between these models and biological systems, the striking similarities with many studies suggests that DNNs may provide solutions to learning and representation problems faced by biological systems and as such may be useful in generating testable predictions to constrain and guide perceptual learning research within living systems.

Reference

Acerbi L, Vijayakumar S, Wolpert DM (2014) On the Origins of Suboptimality in Human Probabilistic Inference. *PLoS Computational Biology* 10:e1003661.

Ahissar M, Hochstein S (1993) Attentional control of early perceptual learning. *Proceedings of the National Academy of Sciences* 90:5718–5722.

Ahissar M, Hochstein S (1997) Task difficulty and the specificity of perceptual learning. *Nature* 387:401–406.

Ahissar M, Hochstein S (2004) The reverse hierarchy theory of visual perceptual learning. *Trends in Cognitive Sciences* 8:457–464.

Ball K, Sekuler R (1982) A specific and enduring improvement in visual motion discrimination. *Science* 218:697–8.

Barlow HB (1961) Possible Principles Underlying the Transformations of Sensory Messages In *Sensory Communication*, pp. 216–234. The MIT Press.

Bays BC, Visscher KM, Dantec CCL, Seitz AR (2015) Alpha-band EEG activity in perceptual learning. *Journal of Vision* 15:1–12.

Bejjanki VR, Beck JM, Lu ZL, Pouget A (2011) Perceptual learning as improved probabilistic inference in early sensory areas. *Nature neuroscience* 14:642–648.

Blasdel GG (1992) Orientation selectivity, preference, and continuity in monkey striate cortex. *The Journal of Neuroscience* 12:3139–3161.

Byers A, Serences JT (2012) Exploring the relationship between perceptual learning and top-down attentional control. *Vision Research* 74:30–39.

Cichy RM, Khosla A, Pantazis D, Torralba A, Oliva A (2016) Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports* 6:27755.

Cohen G, Weinshall D (2017) Hidden Layers in Perceptual Learning In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4554–62, Honolulu, HI.

Crist RE, Kapadia MK, Westheimer G, Gilbert CD (1997) Perceptual learning of spatial localization: specificity for orientation, position, and context. *Journal of Neurophysiology* 78:2889–2894.

Dempsey-Jones H, Harrar V, Oliver J, Johansen-Berg H, Spence C, Makin TR (2016) Transfer of tactile perceptual learning to untrained neighboring fingers reflects natural use relationships. *Journal of Neurophysiology* 115:1088–1097.

Donovan I, Szpiro S, Carrasco M (2015) Exogenous attention facilitates location transfer of perceptual learning. *Journal of Vision* 15:11.

Dosher BA, Lu ZL (1998) Perceptual learning reflects external noise filtering and internal noise reduction through channel reweighting. *Proceedings of the National Academy of Sciences* 95:13988–13993.

Dosher BA, Jeter P, Liu J, Lu ZL (2013) An integrated reweighting theory of perceptual learning. *Proceedings of the National Academy of Sciences* 110:13678–13683.

Eickenberg M, Gramfort A, Varoquaux G, Thirion B (2017) Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage* 152:184–194.

Fahle M (2004) Perceptual learning: A case for early selection. *Journal of Vision* 4:4.

Fiorentini A, Berardi N (1980) Perceptual learning specific for orientation and spatial frequency. *Nature* 287:43–44.

Fiorentini A, Berardi N (1981) Learning in grating waveform discrimination: Specificity for orientation and spatial frequency. *Vision Research* 21:1149–1158.

669 Ghose GM, Yang T, Maunsell JHR (2002) Physiological Correlates of Perceptual Learning in Monkey V1 and V2. *J*
670 *Neurophysiol* 87:1867–1888.

671 Guclu U, van Gerven MAJ (2015) Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations
672 across the Ventral Stream. *Journal of Neuroscience* 35:10005–10014.

673 Harris H, Gliksberg M, Sagi D (2012) Generalized Perceptual Learning in the Absence of Sensory Adaptation. *Current*
674 *Biology* 22:1813–1817.

675 Herzog MH, Aberg KC, Frémaux N, Gerstner W, Sprekeler H (2012) Perceptual learning, roving and the unsupervised bias.
676 *Vision Research* 61:95–99.

677 Herzog MH, Fahlet M (1997) The Role of Feedback in Learning a Vernier Discrimination Task. *Vision Research* 37:2133–
678 2141.

679 Hubel DH, Wiesel TN (1968) Receptive fields and functional architecture of monkey striate cortex. *The Journal of*
680 *Physiology* 195:215–243.

681 Hung SC, Seitz AR (2014) Prolonged Training at Threshold Promotes Robust Retinotopic Specificity in Perceptual
682 Learning. *Journal of Neuroscience* 34:8423–8431.

683 Hussain Z, Bennett PJ, Sekuler AB (2012) Versatile perceptual learning of textures after variable exposures. *Vision*
684 *Research* 61:89–94.

685 Janssen D, Schuett H, Wichmann F (2016) Some observations on the psychophysics of Deep Neural Networks. *Journal of*
686 *Vision* 16:963–963.

687 Jehee JFM, Ling S, Swisher JD, van Bergen RS, Tong F (2012) Perceptual Learning Selectively Refines Orientation
688 Representations in Early Visual Cortex. *Journal of Neuroscience* 32:16747–16753.

689 Jeter PE, Dosher BA, Liu SH, Lu ZL (2010) Specificity of perceptual learning increases with increased training. *Vision*
690 *Research* 50:1928–1940.

691 Jeter PE, Dosher BA, Petrov A, Lu ZL (2009) Task precision at transfer determines specificity of perceptual learning.
692 *Journal of Vision* 9:1.

693 Kanitscheider I, Coen-Cagli R, Kohn A, Pouget A (2015) Measuring Fisher Information Accurately in Correlated Neural
694 Populations. *PLoS Computational Biology* 11:e1004218.

695 Karni A, Sagi D (1991) Where practice makes perfect in texture discrimination: evidence for primary visual cortex plasticity.
696 *Proceedings of the National Academy of Sciences* 88:4966–70.

697 Khaligh-Razavi SM, Kriegeskorte N (2014) Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical
698 Representation. *PLoS Computational Biology* 10:e1003915.

699 Kheradpisheh SR, Ghodrati M, Ganjtabesh M, Masquelier T (2016) Deep Networks Can Resemble Human Feed-forward
700 Vision in Invariant Object Recognition. *Scientific Reports* 6:32672.

701 Kriegeskorte N (2015) Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information
702 Processing. *Annual Review of Vision Science* 1:417–446.

703 Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks In *Advances*
704 *in Neural Information Processing Systems 25*, pp. 1097–1105.

705 Lillicrap TP, Cownden D, Tweed DB, Akerman CJ (2016) Random synaptic feedback weights support error
706 backpropagation for deep learning. *Nature Communications* 7:13276.

707 Liu Z (1999) Perceptual learning in motion discrimination that generalizes across motion directions. *Proceedings of the*
708 *National Academy of Sciences* 96:14085–14087.

Liu Z, Weinshall D (2000) Mechanisms of generalization in perceptual learning. *Vision Research* 40:97–109.

Maniglia M, Seitz AR (2018) Towards a whole brain model of Perceptual Learning. *Current Opinion in Behavioral Sciences* 20:47–55.

Mastropasqua T, Galliussi J, Pascucci D, Turatto M (2015) Location transfer of perceptual learning: Passive stimulation and double training. *Vision Research* 108:93–102.

Olshausen BA, Field DJ (1997) Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research* 37:3311–3325.

Petrov AA, Dosher BA, Lu ZL (2005) The Dynamics of Perceptual Learning: An Incremental Reweighting Model. *Psychological Review* 112:715–743.

Pinto N, Doukhan D, DiCarlo JJ, Cox DD (2009) A high-throughput screening approach to discovering good forms of biologically inspired visual representation. *PLoS Computational Biology* 5.

Raiguel S, Vogels R, Mysore SG, Orban Ga (2006) Learning to see the difference specifically alters the most informative V4 neurons. *The Journal of neuroscience* 26:6589–6602.

Robert G, Janssen D, Schütt H, Bethge M, Wichmann F (2017) Of Human Observers and Deep Neural Networks: A Detailed Psychophysical Comparison In *Vision Sciences Society*.

Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. *Nature* 323:533–536.

Sagi D (2011) Perceptual learning in Vision Research. *Vision Research* 51:1552–1566.

Saxe A (2015) The Effect of Pooling in a Deep Learning Model of Perceptual Learning In *Cosyne Abstracts*, Salt Lake City USA.

Scellier B, Bengio Y (2017) Equilibrium Propagation: Bridging the Gap between Energy-Based Models and Backpropagation. *Frontiers in Computational Neuroscience* 11.

Schoups AA, Vogels R, Orban GA (1995) Human perceptual learning in identifying the oblique orientation: retinotopy, orientation specificity and monocularity. *The Journal of Physiology* 483:797–810.

Schoups A, Vogels R, Qian N, Orban G (2001) Practising orientation identification improves orientation coding in V1 neurons. *Nature* 412:549–553.

Seabold S, Perktold J (2010) Statsmodels: econometric and statistical modeling with Python. *Proceedings of the 9th Python in Science Conference* pp. 57–61.

Seitz AR, Yamagishi N, Werner B, Goda N, Kawato M, Watanabe T (2005) Task-specific disruption of perceptual learning. *Proceedings of the National Academy of Sciences* 102:14895–14900.

Seriès P, Latham PE, Pouget A (2004) Tuning curve sharpening for orientation selectivity: coding efficiency and the impact of correlations. *Nature Neuroscience* 7:1129–1135.

Sotiropoulos G, Seitz AR, Seriès P (2011) Perceptual learning in visual hyperacuity: A reweighting model. *Vision Research* 51:585–599.

Talluri BC, Hung SC, Seitz AR, Seriès P (2015) Confidence-based integrated reweighting model of task-difficulty explains location-based specificity in perceptual learning. *Journal of vision* 15:17.

Tartaglia EM, Aberg KC, Herzog MH (2009) Perceptual learning and roving: Stimulus types and overlapping neural populations. *Vision Research* 49:1420–1427.

Tootell RB, Hadjikhani NK, Vanduffel W, Liu AK, Mendola JD, Sereno MI, Dale AM (1998) Functional analysis of primary visual cortex (V1) in humans. *Proceedings of the National Academy of Sciences* 95:811–7.

Wang R, Wang J, Zhang JY, Xie XY, Yang YX, Luo SH, Yu C, Li W (2016) Perceptual Learning at a Conceptual Level. *Journal of Neuroscience* 36:2238–2246.

Watanabe T, Sasaki Y (2015) Perceptual Learning: Toward a Comprehensive Theory. *Annual Review of Psychology* 66:197–221.

Xiao LQ, Zhang JY, Wang R, Klein SA, Levi DM, Yu C (2008) Complete Transfer of Perceptual Learning across Retinal Locations Enabled by Double Training. *Current Biology* 18:1922–1926.

Yamins DLK, DiCarlo JJ (2016) Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience* 19:356–365.

Yamins DLK, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences* 111:8619–8624.

Yang T, Maunsell JH (2004) The effect of perceptual learning on neuronal responses in monkey visual area V4. *J.Neurosci.* 24:1617–1626.

Yu Q, Zhang P, Qiu J, Fang F (2016) Perceptual Learning of Contrast Detection in the Human Lateral Geniculate Nucleus. *Current Biology* 26:3176–3182.

Zafeiriou S, Hansen M, Atkinson G, Argyriou V, Petrou M, Smith M, Smith L (2011) The Photoface database In *CVPR 2011 Workshops*, pp. 132–139. IEEE.

Zhang JY, Kuai SG, Xiao LQ, Klein SA, Levi DM, Yu C (2008) Stimulus Coding Rules for Perceptual Learning. *PLoS Biology* 6:e197.

Zhang T, Xiao LQ, Klein SA, Levi DM, Yu C (2010) Decoupling location specificity from perceptual learning of orientation discrimination. *Vision Research* 50:368–374.

Zhaoping L, Herzog M, Dayan P (2003) Nonlinear ideal observation and recurrent preprocessing in perceptual learning. *Network: Computation in Neural Systems* 14:233–247.

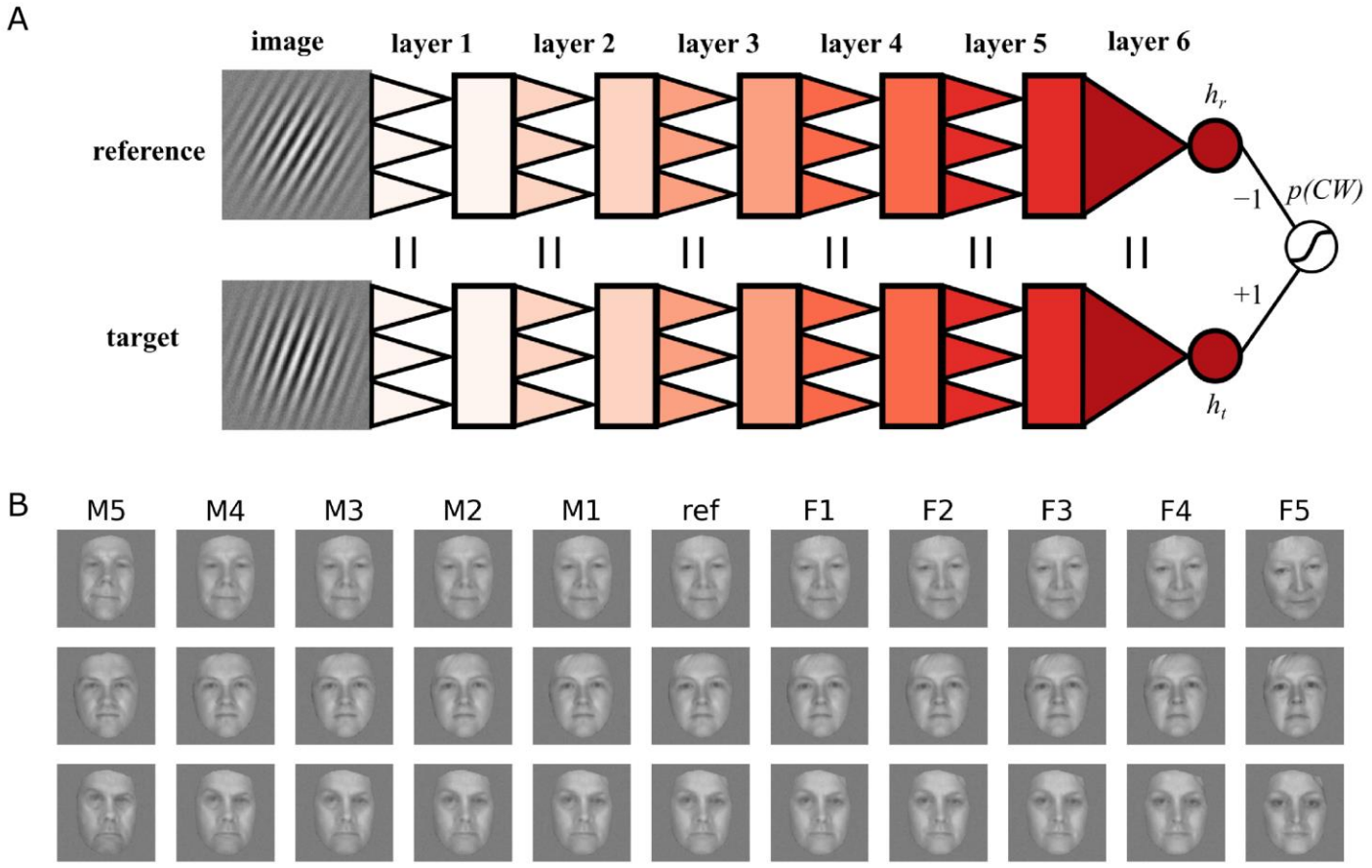

Figure 1. Model structure and stimulus examples. **A**, Model architecture and a pair of Gabor stimuli. The network consists of two identical processing streams producing scalar representations, one for the reference h_r and the other for the target h_t, and the difference of the two is used to obtain a probability of the target being more clockwise *p(CW)* through the sigmoid function. Darker colors indicate higher layers. Layers 1 to 5 consists of multiple units arranged in retinotopic order (rectangles) and convolutional weights (triple triangles, not indicative of the actual filter sizes or counts) to their previous layers or image, and layer 6 has a single unit (dark orange circles) fully connected to layer-5 units (single triangles). Weights at each layer are shared between the two streams so that the representations of the two images are generated by the same parameters. Feedback is provided at the last sigmoidal unit. The Gabor examples have the following parameters: reference orientation: 30°, target orientation: 20°, contrast: 50%, wavelength: 20 pixels, noise standard deviation: 5. **B**, Face examples morphed from three males and three females. The reference (ref) is paired with either a more masculine (M) or more feminine(F) target image, both morphed from the same two originals (M5 and F5) with the reference being the halfway morph. The number following the label indicates dissimilarity with the reference.

772

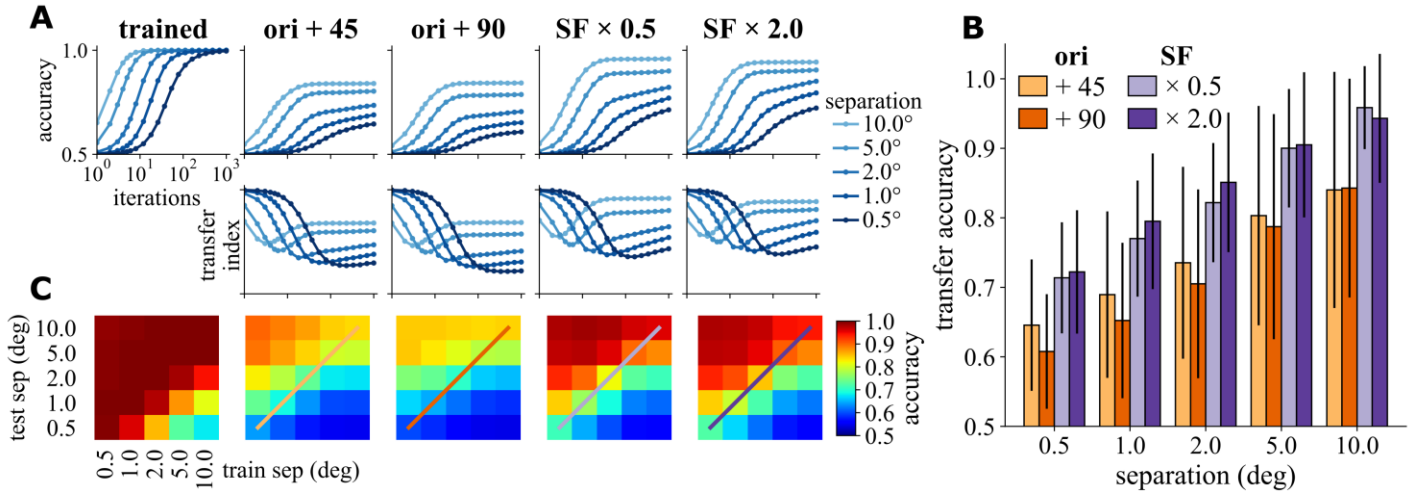

Figure 2. Performance of the model when trained under various angle separations and tested at the trained and transfer conditions: reference orientation rotated clockwise by 45 (ori + 45) or 90° (ori + 90) and spatial frequency halved (SF × 0.5) or doubled (SF × 2.0). See Extended Data Table 2-1 for statistical details. **A**, Accuracy (top row) and transfer index (bottom row) trajectories against training iterations. Darker blue indicates finer precision. Error envelope represents 1 sem and is hardly visible. See Extended Data Figure 2-1 for accuracies plotted as mean±std during the first 50 iterations. **B**, Final performance under the four transfer conditions. There was a significant positive main effect of log angle separation in each of the four transfer conditions ($p<0.0001$, $R^2>0.2$ in all conditions), indicating greater transfer for coarser precisions. Error bar indicates 1 std. **C**, Final mean accuracies when the network was trained and tested on all combinations of training and test precisions. The diagonal lines in the four transfer conditions indicate equal training and test precision for which the accuracies are also shown in panel B. For each transfer condition, there was a strong positive main effect of log test separation ($p<0.0001$, $R^2>0.35$ in all conditions) shown as increasing color gradient from bottom to top, and the log training separation also had a weaker but significant negative effect ($p<0.0001$ in all conditions, $R^2>0.05$ in all conditions except for angle+90 where $R^2=0.018$) shown as decreasing color gradient from left to right. Higher training precisions enhanced performance at transfer to low precisions, shown as higher accuracy on top-left quadrants compared to lower-right quadrants.

773

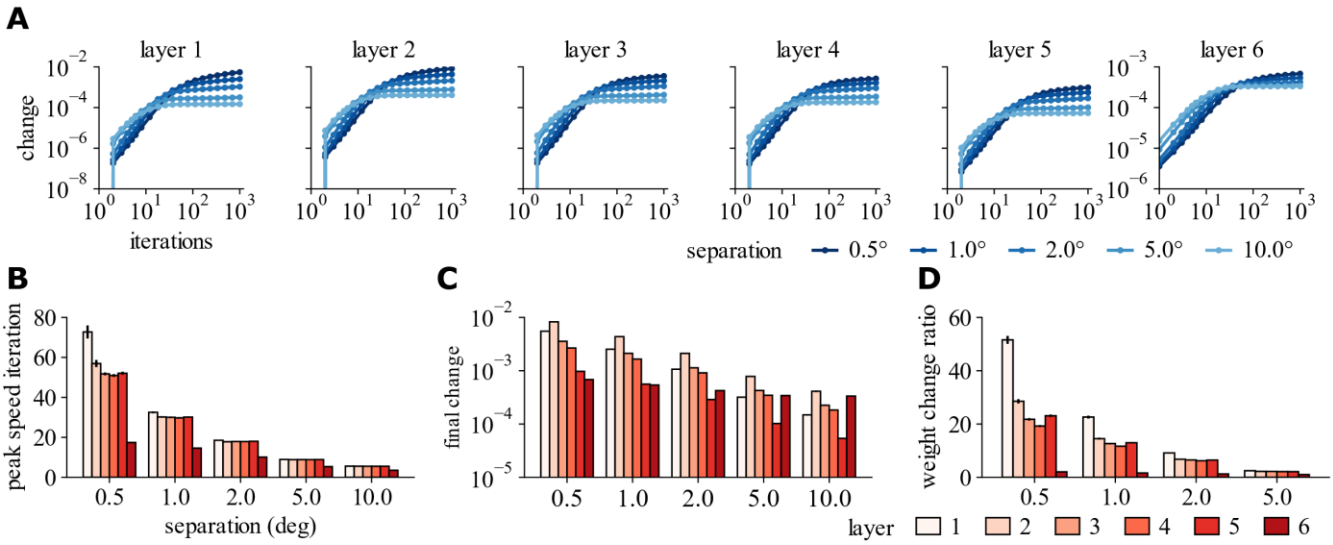

Figure 3. Layer change under different training precisions. **A**, Layer change (Equation 4) trajectories during learning. Lighter colors indicate larger angle separations. One-sem envelops are hardly visible. **B**, Iteration at which the rate of change peaked (peak speed iteration, PSI). Excluding layer 6, there were significant negative main effects of log angle separation (β=-37.24, $t(14397)$=-100.02, $p\approx0.0$, R^2=0.41) and layer number (β=-1.07, $t(14397)$=-8.73, $p=2.9\times10^{-18}$, R^2=0.0031) on PSI, suggesting that layer change started to asymptote earlier in higher layers and finer precisions. For individual precisions, layer number had a significant effect only for the two smallest angle separations ($p<0.0001$, see Table 3-1 for details). **C**, Final layer change. Ignoring layer 6, for which the change was measured differently, a linear regression analysis on the logarithm of layer change yielded significant negative main effects of log angle separation (β=-1.0, $t(14397)$=-208.4, $p\approx0.0$, R^2=0.66) and layer number (β=-0.15, $t(14397)$=-91.2, $p\approx0.0$, R^2=0.13), implying greater layer change in lower layers and finer precisions. **D**, Ratio of final layer change relative to the change under the easiest condition (10.0°). Changes in lower layers increased by a larger factor than higher layers when precision was high. **B-D**, Error bar represents 1 sem.

774

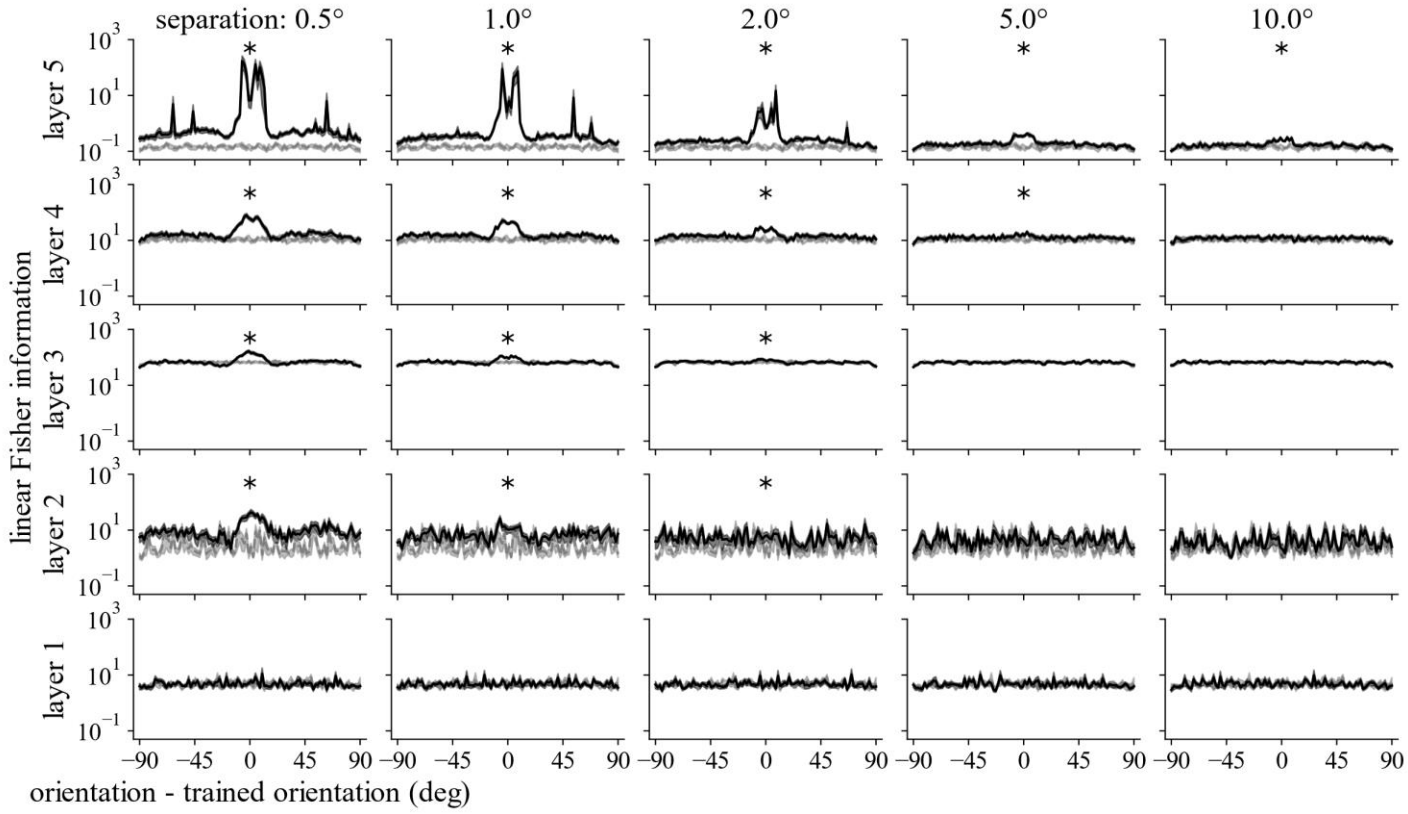

Figure 4. Linear Fisher information (FI) defined in Equation 6 of the trained (black) and naïve (gray) populations at each layer and each test orientation and when trained at each precision. Only units with receptive fields at the center of the image and with minimum activation of 1.0 are included. (*) indicates significant increase in mean FI within 10° of the trained orientation (threshold p=0.01, Mann-Whitney U, Bonferroni-corrected for 5 layers × 5 angle separations). The oscillations were caused by pooling over 12 reference orientations and the inhomogeneity in orientation sensitivity of AlexNet. The FI values at layer 5 do not reflect real discrimination thresholds because the readout was noisy.

775

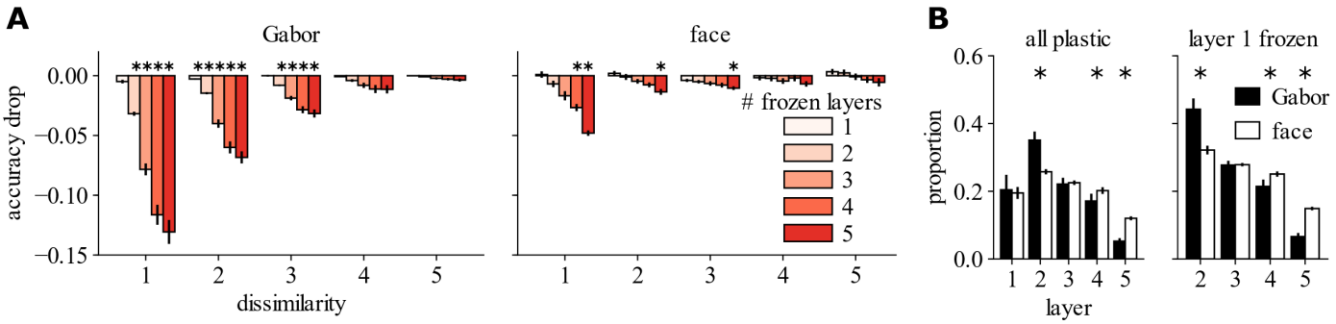

Figure 5. Effect of different tasks (Gabor orientation and face gender discriminations) on performance and layer change. **A**, Accuracy drop as successive low layers were frozen at the iteration where the fully plastic network reached 95% accuracy for the two tasks. (*) indicates significant drop from zero (threshold p=0.01, 1-sample t-test against zero, Bonferroni-corrected for 5 frozen layers × 5 dissimilarities). Performance was impaired when freezing layer 2 onwards in the Gabor task and when freezing layer 4 onwards in the face task. The largest incremental performance drop happened in layer 3 for the Gabor task and layer 5 for the face task. **B**, Distribution of learning over layers when the network was trained on the two tasks if the network was fully plastic (left) or if layer 1 was frozen (right). There was a significant interaction of layer × task on layer change proportion (p<0.0001, 2-way ANOVA, see Extended Data Table 5-1 for details). See Extended Data Figure 5-1 for demonstration of robustness to other measures of layer change. (*) indicates significant difference in the layer change proportion between the two tasks (threshold p=0.01, Mann-Whitney U, Bonferroni-corrected for 5 or 4 layers). Error bar represents 1 sem.

776

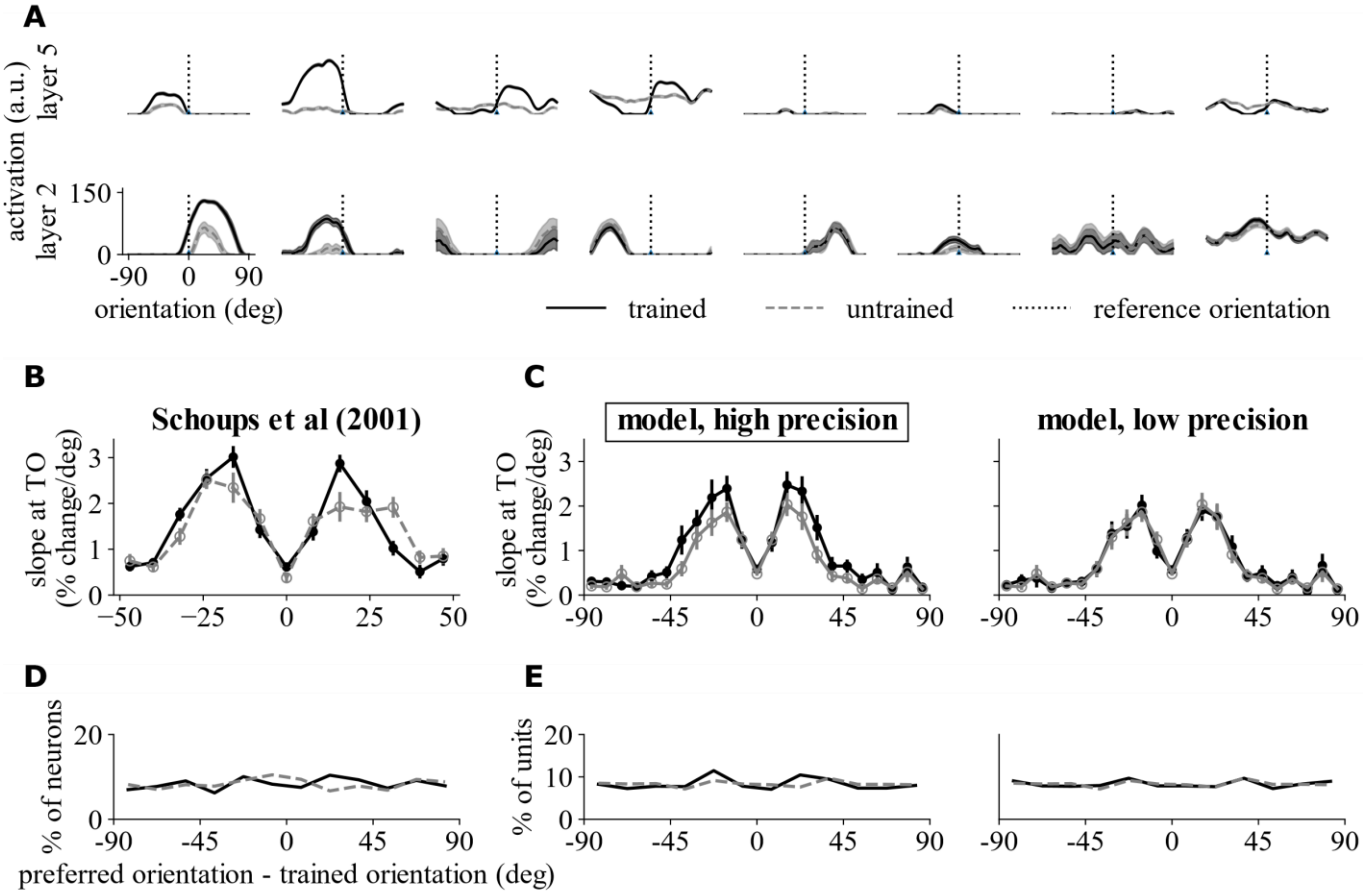

Figure 6. **A**, Tuning curve examples of network units before (gray dashed) and after (black solid) training in layers 2 and 5. **B-E**, Comparison between V1 neurons trained under high precision in Schoups et al. (2001) with model units in layer 2 trained under high (1.0°, matching with experiment) and low (5.0°) precisions. **B**, Slope at trained orientation for trained and naïve V1 neurons grouped according to preferred orientation. **C**, Same as B but from model units. Units tuned around 20° increased their slope magnitude at trained orientation only under high precision. See Extended Data Table 6-1 for statistical test details; see Extended Data Figure 6-1 for layer 3 and other precisions. **D**, Distribution of preferred orientation in V1 was roughly uniform before and after training. **E**, Same as D but from model units. There was no significant difference in the distribution of preferred orientation between the trained and naïve populations under either of the two precisions ($p>0.9$, K-S, see Extended Data Table 6-2 for details; see Extended Data Figure 6-2 for layer 3 and other precisions). **B and D,** Reproduced from Schoups et al. (2001)

777

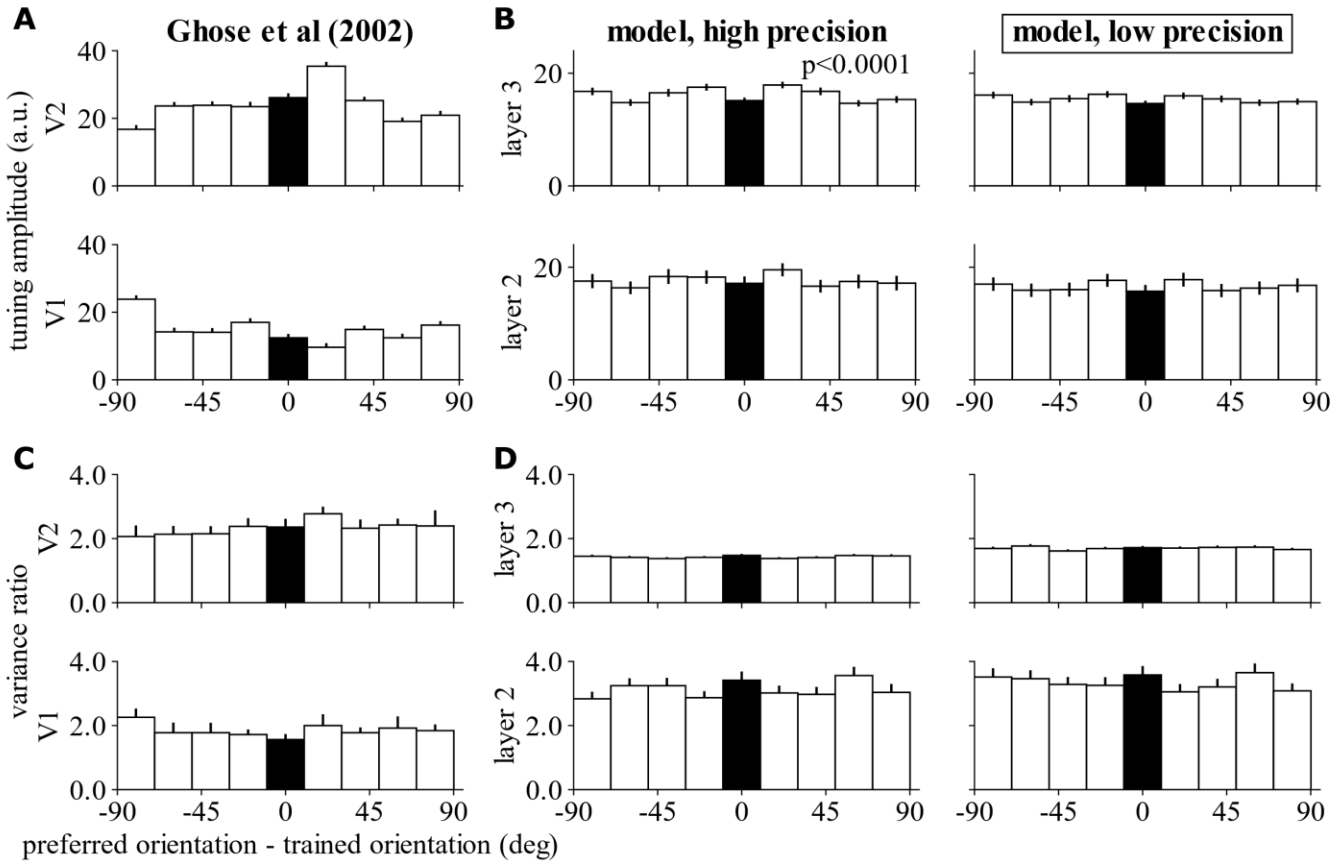

Figure 7. Comparison between V1-2 neurons trained under low precision in Ghose et al. (2002) with model units in layers 2 and 3 trained under high (1.0°) and low (5.0°, matching with experiment) precisions. Black bar indicates the orientation bin that contains the trained orientation. **A,** No significant effect of preferred orientation was found on tuning amplitude. **B**, Same as A but from model units. Preferred orientation had a significant effect only in layer 3 when trained under high precision ($p<0.0001$, one-way ANOVA, see Extended Data Table 7-1 for details; see Extended Data Figure 7-1 for layer 3 and other precisions). **C**, No significant effect of preferred orientation was found on variance ratio. **D**, Same as C but from model units. No significant effect of preferred orientation was found on variance ratio ($p>0.6$, one-way ANOVA, see Extended Data Table 7-2 for details; see Extended Data Figure 7-2 for layer 3 and other precisions) for either layer and precision. **A** and **C**, Reproduced from Ghose et al. (2002)

778

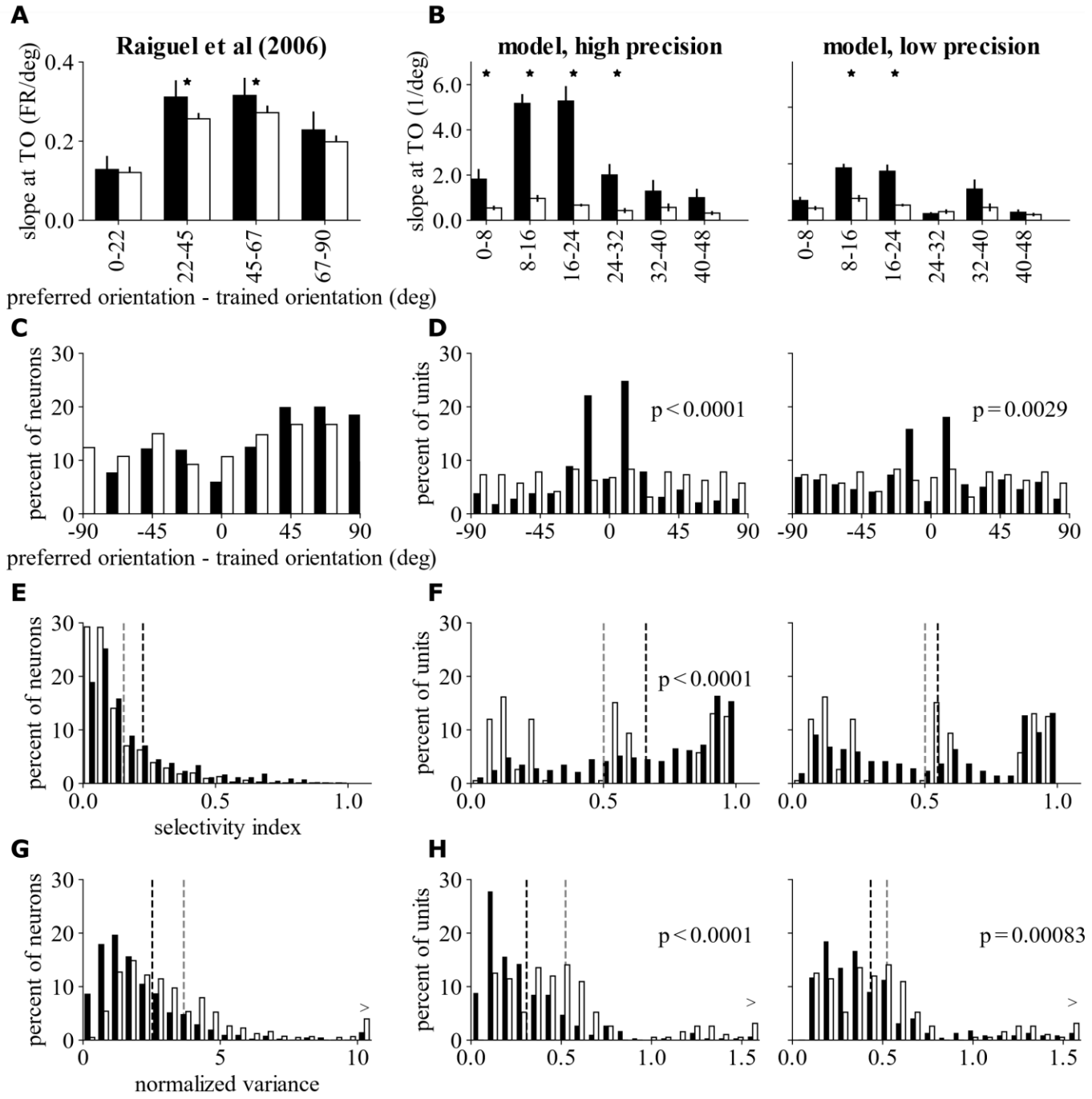

Figure 8. Comparison between V4 neurons in Raiguel et al. (2006) with model units in layer 5 trained under high (1.0°) and low (5.0°) precisions. Black and white bars indicate trained and naïve populations, respectively. **A**, Neurons tuned 22-67° away from trained orientation increased their slopes at trained orientation. (*) indicates significant increase in slope before and after training. **B**, Same as A but from model units. There was a significant interaction of training × preferred orientation under either of the precisions ($p<0.0002$, two-way ANOVA, see Extended Data Table 8-1 for details; see Extended Data Figure 8-1 for layer 4 and other precisions). (*) indicates significant increase in slope before and after training (threshold $p=0.01$, Mann-Whitney U, Bonferroni-corrected for 6 orientation bins). Only neurons tuned within 48° of trained orientation are shown; other neurons did not change significantly after training. **C**, Distribution of preferred orientation shifted away from uniform after training. **D**, Same as C but from model units. The units under both precisions altered their preferred orientation distribution which became significantly different from a uniform distribution ($p<0.003$, K-S, see Extended Data Table 8-2 for details; see Extended Data Figure 8-2 for layer 4 and other precisions). **E-H**, Black and gray dashed lines indicate trained and naïve distribution means, respectively. **E,** Selectivity

index of V4 neurons increased significantly after training. **F**, Same as E but from model units. Training produced a significant increase in selectivity index under high precision ($p<0.0001$, Mann-Whitney U, see Extended Data Table 8-3 for details; see Extended Data Figure 8-3 for layer 4 and other precisions). **G**, Training significantly reduced normalized variance of V4 neurons. **H**, Same as G but from model units. Normalized variance was significantly reduced after training under both precisions ($p<0.001$, Mann-Whitney U, see Extended Data Table 8-4 for details; see Extended Data Figure 8-4 for layer 4 and other precisions). **A**, **C**, **E** and **G,** Reproduced from Raiguel et al. (2006)

779

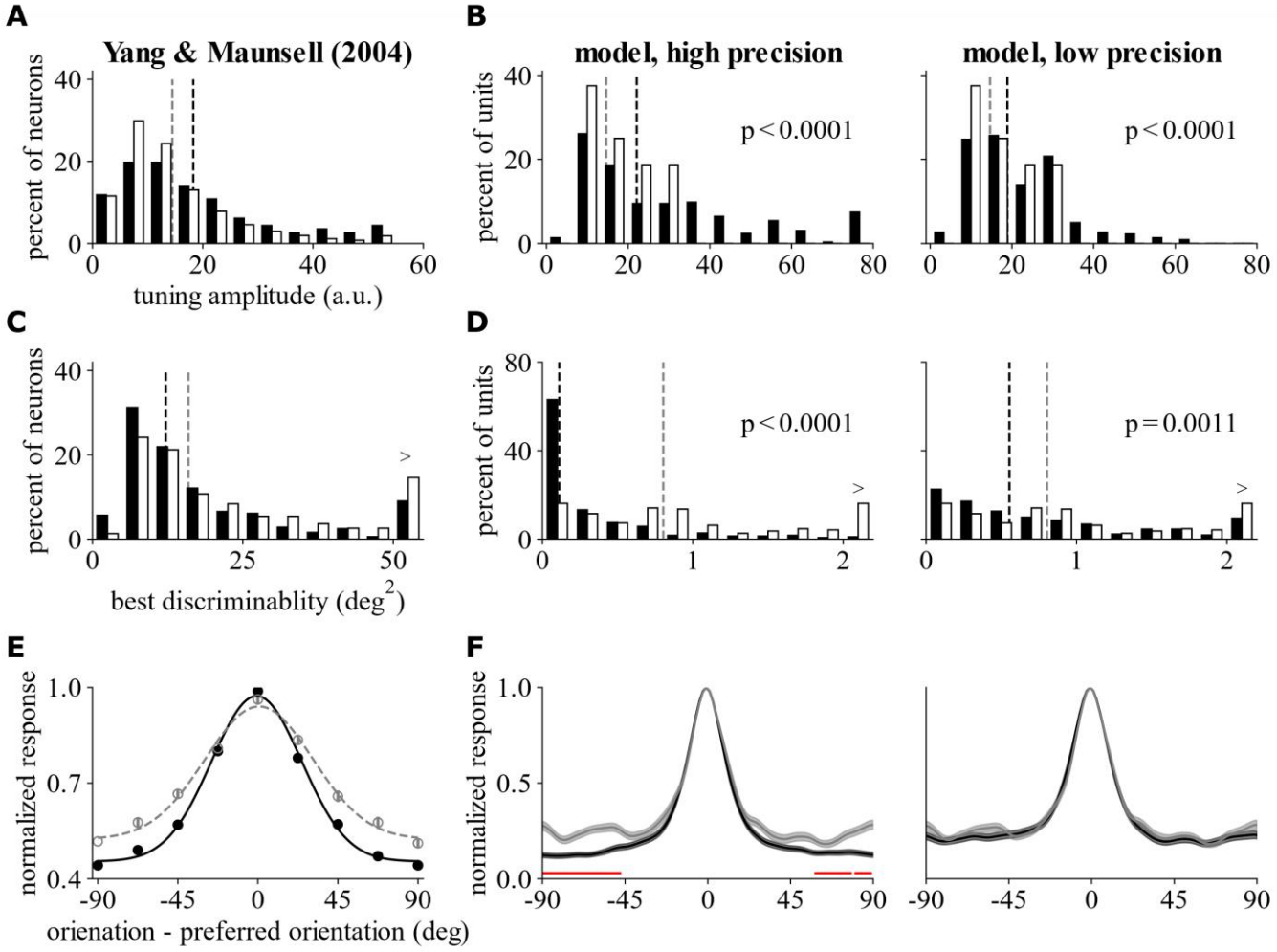

Figure 9. Comparison between V4 neurons in Yang & Maunsell (2004) with model units in layer 5 trained under high (1.0°) and low (5.0°) precisions. **A-D**, Black and white bars indicate trained and naïve populations, respectively; black and gray dashed lines indicate trained and naïve distribution medians, respectively. **A**, Training significantly increased the tuning amplitude of V4 neurons. **B**, Same as A but from model units. Training significantly increased response amplitude for both precisions ($p<0.0001$, Mann-Whitney U, see Extended Data Table 9-1 for details, see Extended Data Figure 9-1 for layer 4 and other precisions). **C**, Training produced a significant reduction in best discriminability (lower indicates better discriminability) for V4 neurons. **D**, Same as B but from model units. Training significantly reduced the best discriminability for both precisions ($p<0.0005$, Mann-Whitney U, see Extended Data Table 9-2 for details; see Extended Data Figure 9-2 for layer 4 and other precisions). **E**, Training resulted in narrower normalize tuning curves (by maximum response). **F**, Same as E but from model units. Activation was significantly lower after training for the non-preferred orientation range (>45° away of trained orientation) than preferred orientation range in high precision ($p<0.0001$, two-way ANOVA, see Extended Data Table 9-3 for details; see Extended Data Figure 9-3 for layer 4 and other precisions). Red lines indicate orientations with significant reduction in activation (threshold $p=0.01$, Mann-Whitney U, Bonferroni-corrected for 100 test orientations). Curves are mean with one-sem envelope. **A**, **C**, **E** and **G,** Reproduced from Yang & Maunsell (2004).

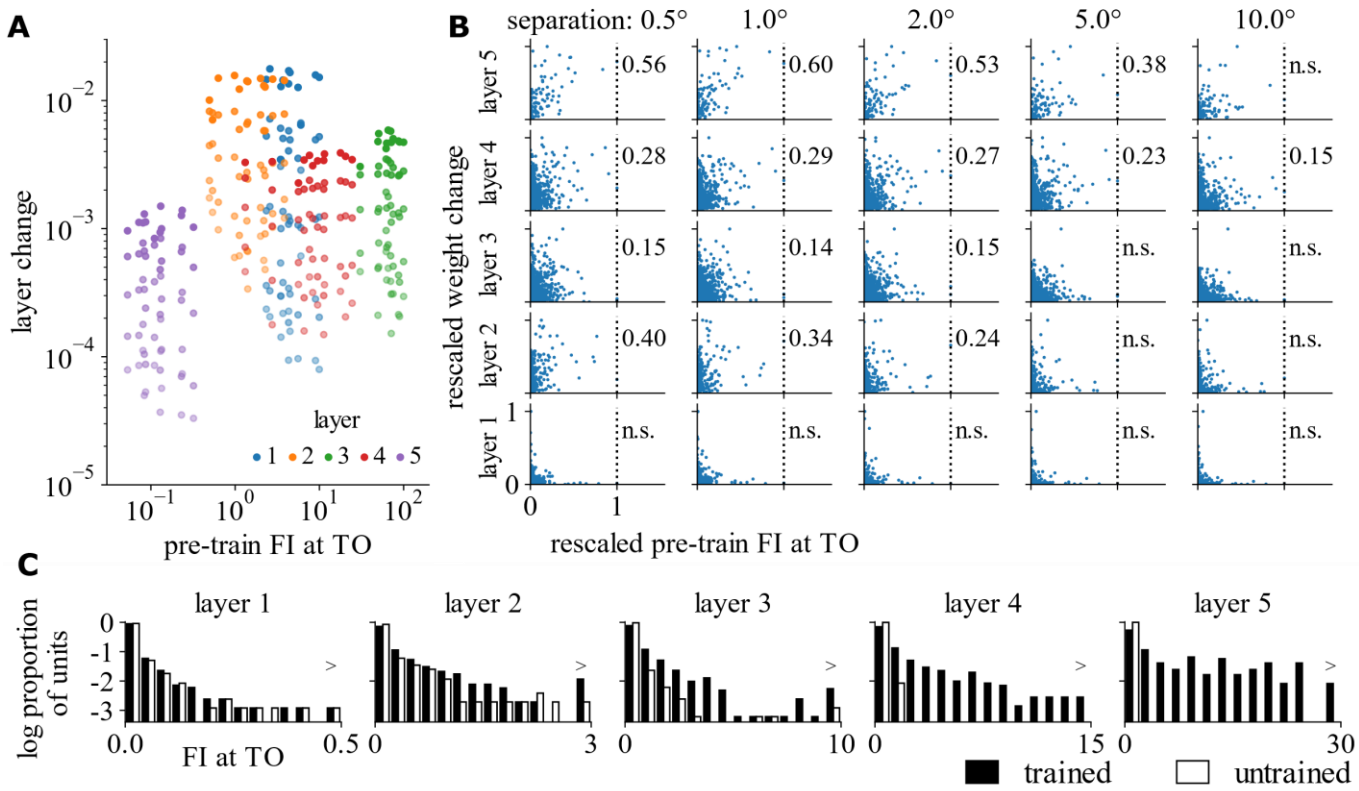

Figure 10. **A-B**, Effect of the network's initial sensitivity to trained orientation (TO) on the magnitude of learning. **A**, Relationship between layer weight change and layer-wise pre-train linear fisher information at TO (FI). Color indicates different layers, and darker color indicates higher training precision. Using a regression analysis on the layer change under the main effects of layer, precision, FI and the interactions of layer × FI and precision × FI, the effects of layer and precision were significant ($p<0.0001$) while the main effect of FI and its two-way interactions with layer and precision were not ($p>0.1$, see Extended Data Table 10-1 for details; see Extended Data Figure 10-1 for results under different measures of weight change). **B**, Relationship between the change in the weights and pre-train FI at TO for network units, both of which are rescaled to between 0 and 1 after dividing by the respective maximum change for each angle separation and each layer. Significant Pearson's correlation is shown for each layer and angle separation (Bonferroni-corrected for 5 layer × 5 angle separations). Despite the general positive correlation, there was a tendency for units with lower FI to change more when training precision increased. **C**, Distribution of network unit fisher information at TO (FI) before (white) and after (black) training. In layers 1 to 2, most units had very low FI before and after training, whereas the distributions for units in layers 3 and 4 are more spread-out and training increased FI for many neurons.

781