SCIENTIFIC REPORTS

OPEN

Characterisation of nonlinear receptive fields of visual neurons by convolutional neural network

Received: 4 September 2018
Accepted: 19 February 2019
Published online: 07 March 2019

Jumpei Ukita[1], Takashi Yoshida[1,2] & Kenichi Ohki[1,2,3]

A comprehensive understanding of the stimulus-response properties of individual neurons is necessary to crack the neural code of sensory cortices. However, a barrier to achieving this goal is the difficulty of analysing the nonlinearity of neuronal responses. Here, by incorporating convolutional neural network (CNN) for encoding models of neurons in the visual cortex, we developed a new method of nonlinear response characterisation, especially nonlinear estimation of receptive fields (RFs), without assumptions regarding the type of nonlinearity. Briefly, after training CNN to predict the visual responses to natural images, we synthesised the RF image such that the image would predictively evoke a maximum response. We first demonstrated the proof-of-principle using a dataset of simulated cells with various types of nonlinearity. We could visualise RFs with various types of nonlinearity, such as shift-invariant RFs or rotation-invariant RFs, suggesting that the method may be applicable to neurons with complex nonlinearities in higher visual areas. Next, we applied the method to a dataset of neurons in mouse V1. We could visualise simple-cell-like or complex-cell-like (shift-invariant) RFs and quantify the degree of shift-invariance. These results suggest that CNN encoding model is useful in nonlinear response analyses of visual neurons and potentially of any sensory neurons.

A goal of sensory neuroscience is to comprehensively understand the stimulus-response properties of neuronal populations. In the visual cortex, such properties were first characterised by Hubel and Wiesel, who discovered the orientation and direction selectivity of simple cells in the primary visual cortex (V1) using simple bar stimuli[1]. Later studies revealed that the responses of many visual neurons, including even simple cells[2–5], display nonlinearity, such as shift-invariance in V1 complex cells[6]; size, position, and rotation-invariance in inferotemporal cortex[7–9]; and viewpoint-invariance in a face patch[10]. Nevertheless, nonlinear response analyses of visual neurons have been limited thus far, and existing analysis methods are often designed to address specific types of nonlinearity underlying the neuronal responses. For example, the spike-triggered average[11] assumes linearity; moreover, the second-order Wiener kernel[12] and spike-triggered covariance[13–15] address second-order nonlinearity at most. In this study, we aim to analyse visual neuronal responses using an encoding model that does not assume the type of nonlinearity.

An encoding model that is useful for nonlinear response analyses of visual neurons must capture the nonlinear stimulus-response relationships of neurons. Thus, the model should be able to predict neuronal responses to stimulus images with high performance[16] even if the responses are nonlinear. In addition, the features that the encoding model represents should be visualised at least in part so that we can understand the neural computations underlying the responses. Artificial neural networks are promising candidates that may meet these criteria. Neural networks are mathematically universal approximators in that even one-hidden-layer neural network with many hidden units can approximate any smooth function[17]. In computer vision, neural networks trained with large-scale datasets have yielded state-of-the-art and sometimes human-level performance in digit classification[18], image classification[19], and image generation[20], demonstrating that neural networks, especially convolutional neural networks (CNNs)[21,22], capture the higher-order statistics of natural images through hierarchical information processing. In addition, recent studies in computer vision have provided techniques to extract and visualise the features learned in neural networks[23–26].

[1]Department of Physiology, The University of Tokyo School of Medicine, Bunkyo-ku, Tokyo, Japan. [2]Department of Molecular Physiology, Graduate School of Medical Sciences, Kyushu University, Higashi-ku, Fukuoka, Japan. [3]International Research Center for Neurointelligence (WPI-IRCN), The University of Tokyo, Bunkyo-ku, Tokyo, Japan. Correspondence and requests for materials should be addressed to J.U. (email: jukita@m.u-tokyo.ac.jp) or K.O. (email: kohki@m.u-tokyo.ac.jp)

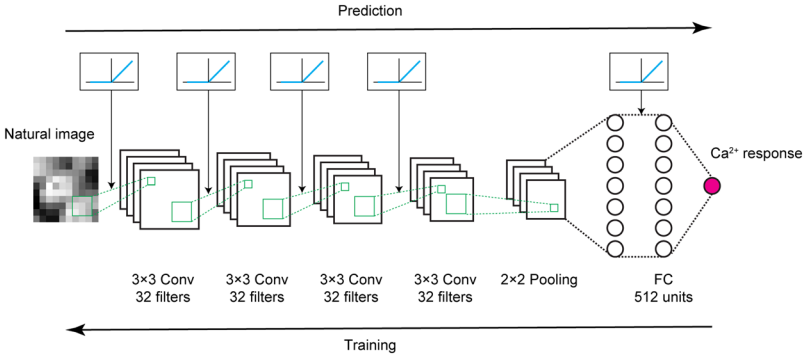

Figure 1. Scheme of CNN encoding model. The Ca^{2+} response to a natural image was predicted by convolutional neural network (CNN) consisting of 4 successive convolutional layers, one pooling layer, one fully connected layer, and the output layer (magenta circle). See Methods for details. Briefly, a convolutional layer calculates a 3×3 convolution of the previous layer followed by a rectified linear (ReLU) transformation. The pooling layer calculates max-pooling of 2×2 regions in the previous layer. The fully connected layer calculates the weighted sum of the previous layer followed by a ReLU transformation. The output layer calculates the weighted sum of the previous layer followed by a sigmoidal transformation. During training, parameters were updated by backpropagation to reduce the mean squared error between the predicted responses and actual responses.

Several previous studies have used artificial neural networks as encoding models of visual neurons. These studies showed that artificial neural networks are highly capable of predicting neuronal responses with respect to low-dimensional stimuli such as bars and textures[27,28] or to complex stimuli such as natural stimuli[29–36]. Furthermore, receptive fields (RFs) were visualised by the principal components of the network weights between the input and hidden layer[29], by linearization[31], and by inversion of the network to evoke at most 80% of maximum responses[32]. However, these indirect RFs are not guaranteed to evoke the highest response of the target neuron.

In this study, we first investigated whether nonlinear RFs could be directly estimated by CNN encoding models (Fig. 1) using a dataset of simulated cells with various types of nonlinearities. We confirmed that CNN yielded the best prediction among several encoding models in predicting visual responses to natural images. Moreover, by synthesising the image such that it would predictively evoke a maximum response ("maximization-of-activation" method), nonlinear RFs could be accurately estimated. Specifically, by repeatedly estimating RFs for each cell, we could visualise various types of nonlinearity underlying the responses without any explicit assumptions, suggesting that this method may be applicable to neurons with complex nonlinearities, such as rotation-invariant neurons in higher visual areas. Next, we applied the same procedures to a dataset of mouse V1 neurons, showing that CNN again yielded the best prediction among several encoding models and that shift-invariant RFs with Gabor-like shapes could be estimated for some cells from the CNNs. Furthermore, by quantifying the degree of shift-invariance of each cell using the estimated RFs, we classified V1 neurons as shift-variant (simple) cells and shift-invariant (complex-like) cells. Finally, these cells were not spatially clustered in cortical space. These results verify that nonlinear RFs of visual neurons can be characterised using CNN encoding models.

Results

Nonlinear RFs could be estimated by CNN encoding models for simulated cells with various types of nonlinearities. We generated a dataset comprising the stimulus natural images (2200 images) and the corresponding responses of simulated cells. To investigate the ability of CNN to handle various types of nonlinearities, we incorporated various basic nonlinearities for the data generation, including rectification, shift-invariance, and in-plane rotation-invariance, which were found in V1 simple cells[2], V1 complex cells[6], and inferotemporal cortex[9], respectively. We generated the responses of simple cells ($N = 30$), complex cells ($N = 70$), and rotation-invariant cells ($N = 10$) using the linear-nonlinear model[2], energy model[37,38], and rotation-invariant model, respectively (Figs 2a,b and 3a; see Methods for details). The responses were generated using one Gabor-shaped filter for a simple cell, two phase-shifted Gabor-shaped filters for a complex cell, and 36 rotated Gabor-shaped filters for a rotation-invariant cell. We also added some noise sampled from a Gaussian distribution such that the trial-to-trial variability of simulated data was similar to that of real data.

We first used a dataset of simulated simple cells and complex cells and trained the CNN for each cell to predict responses with respect to the natural images (Fig. 1). For comparison, we also constructed the following types of encoding models: an L1-regularised linear regression model (Lasso), L2-regularised linear regression model (Ridge), support vector regression model (SVR) with a radius basis function kernel[39], and hierarchical structural model (HSM)[31]. The prediction similarity, defined as the Pearson correlation coefficient between the predicted responses and actual responses in a 5-fold cross-validation manner, of CNN was high and better than that of other models for both simple cells and complex cells (Fig. 2c), ensuring that the stimulus-response relationships of these cells were successfully captured by CNN.

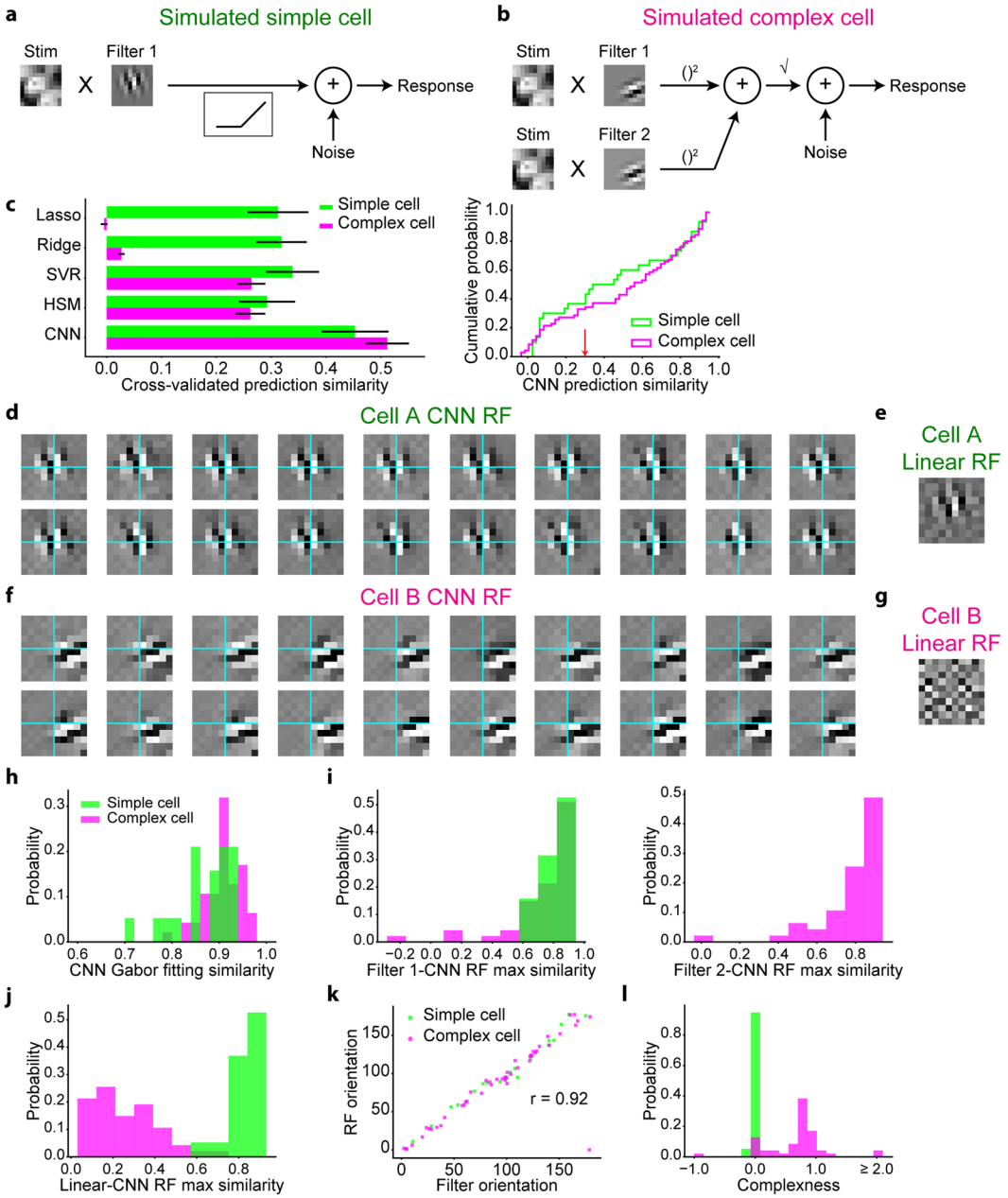

Figure 2. Nonlinear RFs could be estimated by CNN encoding models for simulated simple cells and complex cells. **(a,b)** Scheme of response generation for simulated simple cells (**a**) and simulated complex cells (**b**) (See Methods for details). The Gabor-shaped filters of simulated simple cell A and complex cell B are displayed. **(c)** Left: comparison of the response predictions among the following encoding models: the L1-regularised linear regression model (Lasso), L2-regularised linear regression model (Ridge), support vector regression model (SVR), hierarchical structural model (HSM), and CNN. Data are presented as the mean ± s.e.m. (N = 30 simulated simple cells and N = 70 simulated complex cells). Right: cumulative distribution of CNN prediction similarity. Simulated cells with a CNN prediction similarity ≤0.3 (indicated as the red arrow) were removed from the following receptive field (RF) analysis. **(d,f)** Results of iterative CNN RF estimations for simulated simple cell A (**d**) and complex cell B (**f**). Only 20 of the 100 generated RF images are shown in these panels. Grids are depicted in cyan. Although the simulated simple cell A had RFs in nearly identical positions, the simulate complex cell B had RFs in shifted positions. **(e,g)** Linearly estimated RFs (linear RFs) of simulated simple cell A (**e**) and complex cell B (**g**), using a regularised pseudoinverse method. (**h**) Gabor-fitting similarity of CNN RFs, defined as the Pearson correlation coefficient between the CNN RF and fitted Gabor kernel. (**i**) Maximum similarity between each generator filter and 100 CNN RFs. (**j**) Maximum similarity between linear RFs and CNN RFs. Similarity was defined as the normalised pixelwise dot product between the linear RF and CNN RF. (**k**) Relationship of the Gabor orientations between generator filters and CNN RFs. (**l**) Distribution of complexness. Only cells with a CNN prediction similarity >0.3 were analysed in (**h–l**) (N = 19 simple cells and N = 47 complex cells).

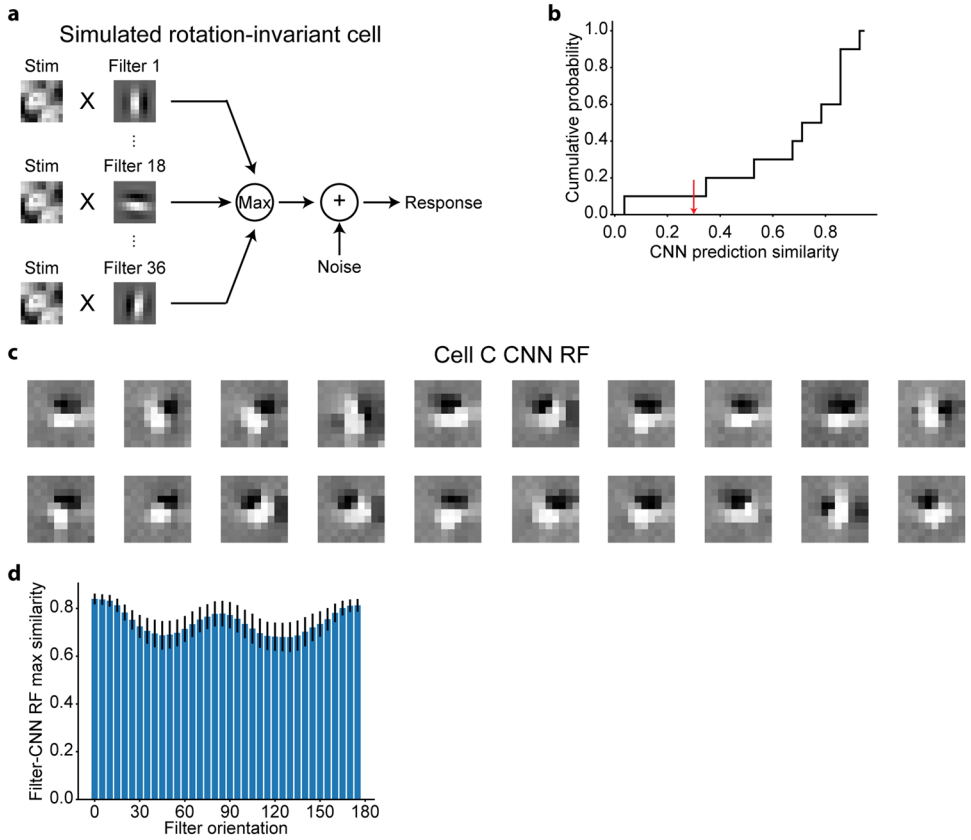

Figure 3. Nonlinear RFs could be estimated by CNN encoding models for simulated rotation-invariant cells. **(a)** Scheme of response generation for simulated rotation-invariant cells. The response to a stimulus was defined as the maximum of the output of 36 subunits followed by an additive Gaussian noise. Each subunit, which had a Gabor-shaped filter with different orientations, calculated the dot product between the stimulus image and the filter (See Methods for details). The filters of simulated cell C are displayed in this panel. **(b)** Cumulative distribution of CNN prediction similarity (N = 10 cells). Simulated cells with a CNN prediction similarity ≤0.3 (indicated as the red arrow) were removed from the following RF analysis. **(c)** Results of iterative CNN RF estimations for simulated cell C. Only 20 of the 1,000 generated RF images are shown in this panel. RF images had Gabor-like shapes but their orientations were different in different iterations. **(d)** Maximum similarity between each generator filter and 1,000 CNN RFs. Only cells with a CNN prediction similarity >0.3 were analysed (N = 9 cells).

Next, we visualised the RF of each cell using the maximization-of-activation approach (see Methods)[23,24] where the RF was regarded as the image that evoked the highest activation of the output layer of the trained CNN. We performed this RF estimation 100 times independently for each cell, utilising the empirical fact that an independent iteration of RF estimation processes creates different RF images by finding different maxima[23]. Figure 2d,f show 20 out of the 100 RF images estimated by the trained CNN (CNN RF images) for a representative simple cell and complex cell, respectively. The predicted responses with respect to these RF images were all >99% of the maximum response in the actual data of each cell, ensuring that the activations of the CNN output layers were indeed maximised. All visualised RF images had clearly segregated ON and OFF subregions, and the structure was close to the Gabor-shaped filters used in the response generations (Fig. 2d vs. Fig. 2a and Fig. 2f vs. Fig. 2b). Furthermore, when RF images were compared within a cell, RF images of cell A had ON and OFF subregions in nearly identical positions, while some RF images of cell B were shifted in relation to one another. These observations are consistent with the assumption that cell A is a simple cell and cell B is a complex cell.

When the iteratively visualised RF images were comprehensively compared, we found that the RF images can be divided into several clusters (Supplementary Fig S1a left and Supplementary Fig S1d left). The properties of these clusters can be explained mostly by the parallel shifts, since the peak cross-correlations of these RF images were overall high (Supplementary Fig S1b,S1e). This is also evident when the representative RF image of each cluster was visualised, which is shifted with respect to each other (Supplementary Fig S1a right and Supplementary Fig S1d right). Then we quantified the distance with which each pair of RF images within a cell were shifted, where we defined the shifted images as the cross-correlation >0.7, showing that while the RF images of the representative simple cell A were shifted mostly parallelly to the Gabor orientation (Supplementary Fig S1c), the RF images of the representative complex cell B were shifted both parallelly and orthogonally to the Gabor orientation (Supplementary Fig S1f). When populations of cells were analysed, among all RF pairs within each cell, 86.5% ± 15.9% (mean ± SD) of the pairs were shifted. Furthermore, the maximum shift distance orthogonal to

the Gabor orientation was distinct between the simulated simple cells and complex cells (Supplementary Fig S1g). Taken together, these results indicate that the differences of these visualised RFs can be explained mostly by the shifts, and the shift axis is distinct between the simple cells and complex cells.

For complex cells, we expect that RF estimation using linear methods would fail to generate an image with clearly segregated ON and OFF subregions, whereas nonlinear RF estimation would not[14]. Thus, the similarity between a linearly estimated RF image (linear RF) and a nonlinearly estimated RF image is expected to be low for complex cells. We performed linear RF estimations following a previous study[40]. Although the linear RF image and CNN RF image were similar for cell A (Fig. 2e), the linear RF image for cell B was ambiguous, lacked clear subregions, and was in sharp contrast to the CNN RF image (Fig. 2g). These results are again consistent with the assumption that cell A is a simple cell and cell B is a complex cell.

In some studies, researchers have used white noise images, instead of natural images, as the stimuli to map RFs of visual neurons[41,42]. Here we also estimated RFs of these simulated cells using a dataset that consists of the white noise images sampled from normal distributions, and the corresponding responses that were simulated using the same filters as above (Supplementary Fig S2). The estimated RF images of these simulated cells were almost indistinguishable from the RF images estimated using natural images. Therefore, in the following analyses, we analysed the responses with respect to the natural images.

Next, we comprehensively analysed the RFs of populations of simulated simple cells and complex cells. Cells with a CNN prediction similarity $\leq$0.3 (the threshold was determined arbitrarily) were omitted from the analyses (Fig. 2c). First, the maximum similarity between a linear RF image and CNN RF image, measured as the normalised pixelwise dot product between a linear RF image and 100 CNN RF images, was distinctly different between simple cells and complex cells (Fig. 2j), reflecting different degrees of nonlinearity. Second, the similarity of Gabor-kernel fitting of the CNN RF image, measured as the pixelwise Pearson correlation coefficient between a CNN RF image and the fitted Gabor kernel, was high among all analysed cells (Fig. 2h), confirming that the estimated RFs had a shape similar to a Gabor kernel. Third, the maximum similarity between each filter used in the response generation and 100 CNN RF images were high for both simple cells and complex cells (Fig. 2i). Fourth, the orientations of the CNN RF images, estimated by fitting them to Gabor kernels, were nearly identical to the orientations of the filters of the response generators (circular correlation coefficient[43] = 0.92; Fig. 2k). These results suggest that the RFs estimated by the CNN encoding models had similar structure to the ground truth and that the shift-invariant property of complex cells was successfully visualised from iterative RF estimations.

We also performed similar analyses using a dataset of simulated rotation-invariant cells. When trained to predict the responses with respect to the natural images, CNNs again yielded good prediction (Fig. 3b). Next, we estimated RFs using the maximization-of-activation approach independently 1,000 times for each cell. The predicted responses with respect to these RF images were all >99% of the maximum response in the actual data of each cell, ensuring that the activations of CNN output layers were indeed maximised. As shown in Fig. 3c, the visualised RF images of cell C had Gabor shapes close to the filters used in the response generation (Fig. 3a). In addition, some RF images were rotated in relation to one another, consistent with the rotation-invariant response property of this cell. Finally, we quantitively compared the RFs (1,000 RF images for each cell) and the filters of the response generator (36 filters for each cell). For each filter, the maximum similarity with 1,000 CNN RF images was high (Fig. 3d), suggesting that the estimated RFs had various orientations and similar structure to the ground truth. Thus, using the proposed RF estimation approach, RFs were successfully estimated by the CNN encoding models, and various types of nonlinearity could be visualised from multiple RFs without assumptions, although the hyperparameters and layer structures of CNNs were unchanged across cells.

CNN yielded the best prediction of the visual response of V1 neurons. Next, we used a dataset comprising the stimulus natural images (200−2200 images) and corresponding real neuronal responses (N = 2465 neurons, 4 planes), which were recorded using two-photon Ca^{2+} imaging from mouse V1 neurons. To investigate whether CNN was able to capture the stimulus-response relationships of V1 neurons, we trained the CNN for each neuron to predict the neuronal responses to the natural images (Fig. 1). The prediction similarity was again measured by the Pearson correlation coefficient between the predicted responses and actual responses of the held-out test data in a 5-fold cross-validation manner (N = 2455 neurons that were not used for the hyperparameter optimisations; see Methods). Comparison of the prediction similarities among several types of encoding models revealed that CNN outperformed other models (Fig. 4a), and the prediction of the CNNs was good (Fig. 4b,c). These results show that the stimulus-response relationships of V1 neurons were successfully captured by CNN, demonstrating the efficacy of using CNN for further RF analyses of V1 neurons.

Estimation of nonlinear RFs of V1 neurons from CNN encoding models. Next, we visualised the RF of each neuron by the maximization-of-activation approach (see Methods)[23,24]. Neurons with a CNN prediction similarity $\leq$0.3 were omitted from this analysis (Fig. 4b). The resultant RF images for two representative neurons are shown in Fig. 5b. Both RF images have clearly segregated ON and OFF subregions and were well fitted with two-dimensional Gabor kernels (Fig. 5c), consistent with known characteristics of simple cells and complex cells in V1[14,44]. The similarity of Gabor-kernel fitting, measured as the pixelwise Pearson correlation coefficient between the RF image and fitted Gabor kernel, was high among all analysed neurons (median r = 0.77; Fig. 5e), suggesting that the RF images generated from the trained CNNs (CNN RF images) successfully captured the Gabor-like structure of RFs observed in V1. We also performed linear RF estimations following a previous study[40]. Although the linear RF image and CNN RF image were similar for neuron D, the linear RF image for neuron E was ambiguous, lacked clear subregions, and was in sharp contrast to the CNN RF image (Fig. 5a,b), suggesting that neuron D would be linear and neuron E would be nonlinear. Supporting this idea, further analysis (see below) revealed that neuron D was a shift-variant (simple) cell, and neuron E was a shift-invariant (complex-like) cell. The similarity between a linear RF image and a CNN RF image, measured as

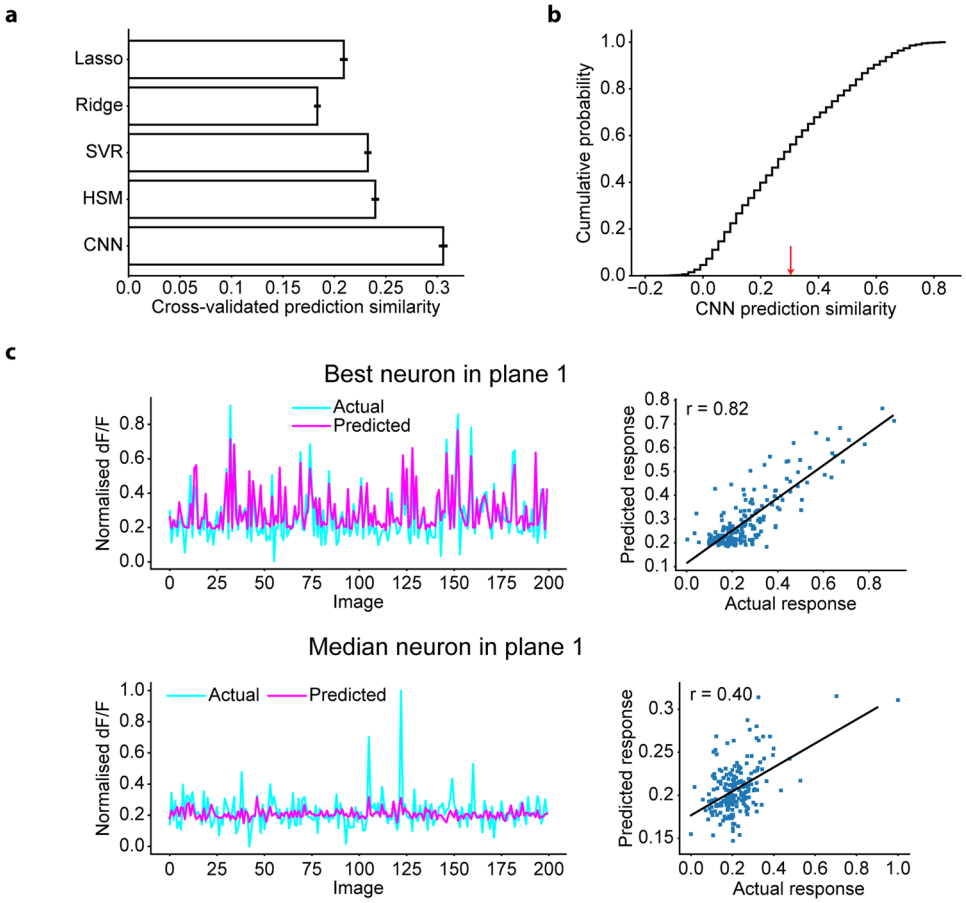

Figure 4. Prediction of the CNN for V1 neurons. **(a)** Comparison of the response predictions among various encoding models: the L1-regularised linear regression model (Lasso), L2-regularised linear regression model (Ridge), SVR, HSM, and CNN. Data are presented as the mean ± s.e.m. (N = 2455 neurons). **(b)** Cumulative distribution of CNN prediction similarity. Neurons with a CNN prediction similarity ≤0.3 (indicated as the red arrow) were removed from the following RF analysis. **(c)** Distributions of actual responses and predicted responses of the neuron with the best prediction similarity in a plane (top) and the neuron with the median prediction similarity in a plane (bottom). Each dot in the right panel indicates data for each stimulus image. Solid lines in the right panels are the linear least-squares fit lines. Only data for 200 images are shown.

the normalised pixelwise dot product between these two images, varied among all analysed neurons (Fig. 5d), reflecting the distributed nonlinearity of V1 neurons. These distributions were minimally affected by the choice of the cell-selection threshold (Supplementary Fig. S3a,b).

One might wonder whether the distinction between the linear RF images and CNN RF images results from the failure of the linear RF estimation method, which was originally based on the electrophysiological data, to capture the nonlinearity in spike-to-calcium signal transformation. Therefore, we also estimated linear RFs using responses deconvolved from the calcium signal traces using the constrained nonnegative matrix factorisation method (Supplementary Fig S4a)[45]. The evoked event was defined as the z-scored difference of the mean activities during the stimulation period and the baseline period. We first confirmed that the evoked events obtained from the deconvolved data were similar to those directly obtained from the calcium signal trace (dF/F data), both for the representative neuron F (Supplementary Fig S4b) and for the populations of neurons on an imaging plane (Supplementary Fig S4d). We then confirmed that the prediction similarity of the L1-regularised linear regression model (Lasso) using the deconvolved data and the dF/F data were also similar, both for the representative neuron F (0.40 when the deconvolved data were used, and 0.41 when the dF/F data were used) and for the populations of neurons (Supplementary Fig S4e). Finally, we compared the linear RF estimated from the calcium dF/F data and that estimated from the deconvolved data, showing that these RFs were nearly identical, both for the representative neuron F (Supplementary Fig S4c) and for the populations of neurons (Supplementary Fig S4f). These results support that little difference was observed on the linear prediction and linear RF between the dF/F data and their deconvolved data.

Estimated RFs of some V1 neurons were shift-invariant. We then performed 100 independent CNN RF estimations for each V1 neuron to characterise the nonlinearity of RFs. We especially focused on the shift-invariance, the most well-studied nonlinearity in V1 complex cells[6]. Figure 5f,g show 20 of the 100 CNN RF images for two representative neurons. The predicted responses with respect to these RF images were all >99% of

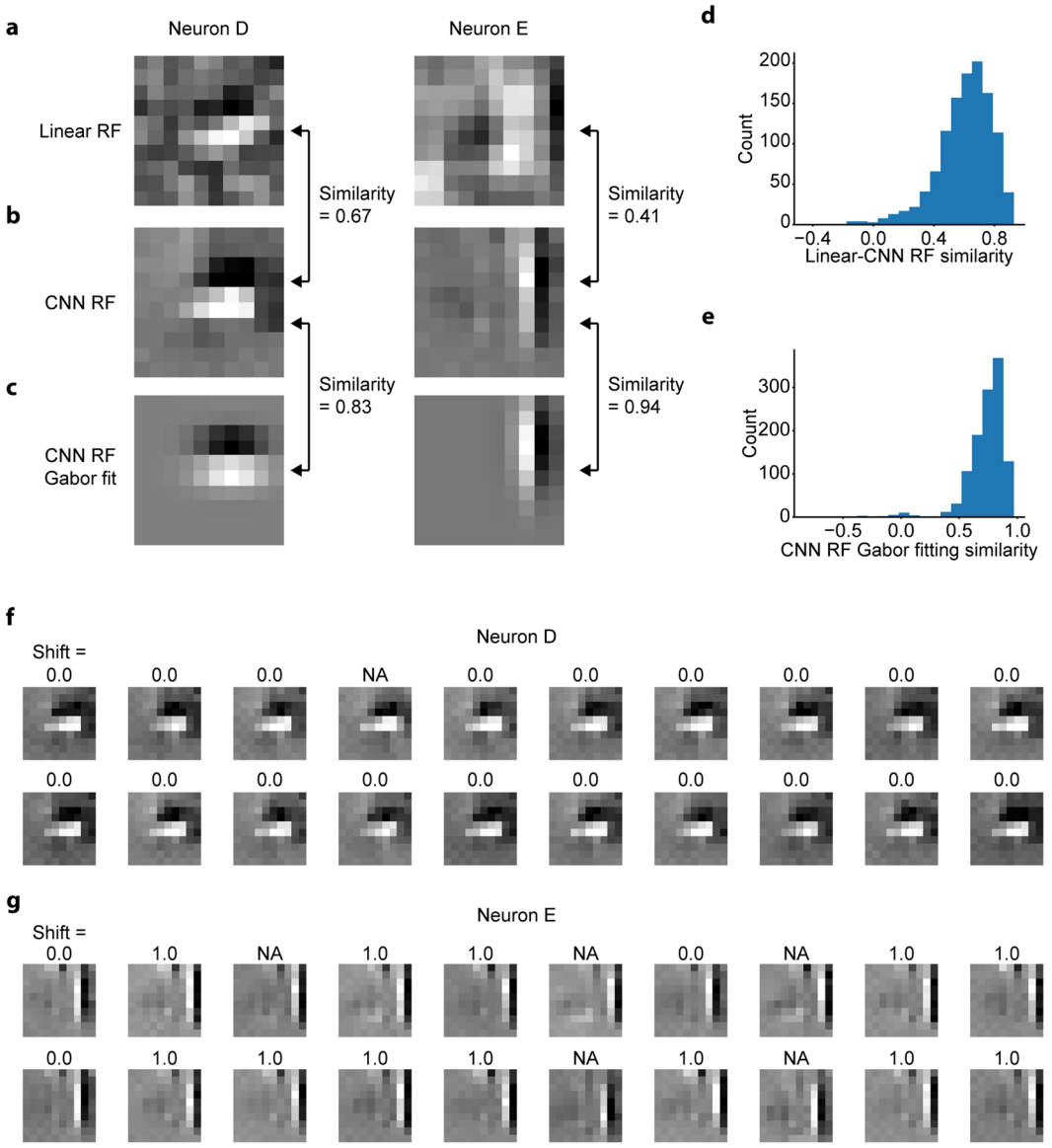

Figure 5. Estimating RFs of V1 neurons from trained CNNs. **(a)** Linearly estimated RFs (linear RFs) of two representative neurons (neurons D and E), using a regularised pseudoinverse method. **(b)** RFs estimated from the trained CNNs (CNN RFs) of the two representative neurons. **(c)** Gabor kernels fitted to CNN RFs of the two representative neurons. **(d)** Similarity between linear RFs and CNN RFs. Similarity was defined as the normalised pixelwise dot product between the linear RF and the CNN RF. **(e)** Gabor fitting similarity of CNN RFs, defined as the Pearson correlation coefficient between the CNN RF and the fitted Gabor kernel. Only neurons with a CNN prediction similarity >0.3 were analysed in (**d**,**e**) (N = 1160 neurons). **(f**,**g)** Results of iterative CNN RF estimations for neuron D (**f**) and neuron E (**g**). Only 20 out of the 100 generated RF images are shown in this figure. The number above each RF image indicates the shift pixel distance between the RF image and the top left RF image. The shift distance between the two images was calculated as the maximum distance of pixel shifts with which the zero-mean normalised cross correlation (ZNCC) > 0.95, projected orthogonally to the Gabor orientation. "NA" indicates that the ZNCC was not above 0.95 for any shift. While shift distances were zero or NA for RF images of neuron D, some RF images of neuron E were shifted to another by one pixel.

the maximum response in the actual data of each neuron, ensuring that the activations of the CNN output layers were indeed maximised. Importantly, RF images of neuron D had ON and OFF subregions in nearly identical positions (Fig. 5f). In contrast, some RF images of neuron E were horizontally shifted in relation to one another (Fig. 5g), suggesting that neuron E is shift-invariant and could be a complex cell.

When the iteratively visualised RF images were comprehensively compared, we found that while the RF images of neuron D were nearly identical (Supplementary Fig S5a left and Supplementary Fig S5b), the RF images of neuron E can be divided into several clusters (Supplementary Fig S5d left). However, we also discovered that the properties of these clusters of neuron E can be explained mostly by the parallel shifts, considering the high

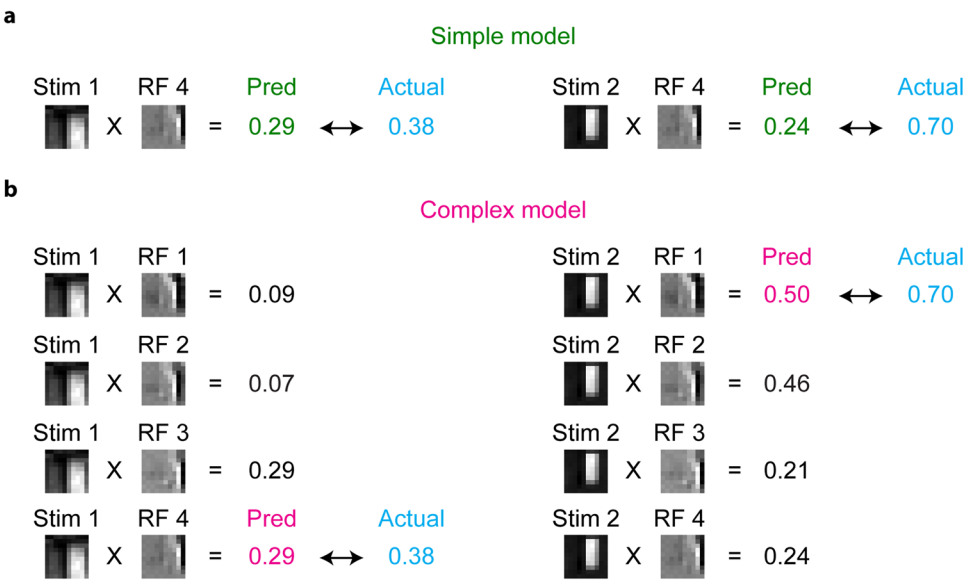

Figure 6. Schemes of the simple model and complex model. Schemes of the simple model and complex model are illustrated using RFs and actual responses of neuron E. **(a)** The simple model is a linear predictive model, which predicts the neuronal response as the normalised dot product between the stimulus image and one RF image (RF 4). **(b)** The complex model predicts the neuronal response as the maximum of the normalised dot products of the stimulus image and several RF images (RF 1–4). Note that the complex model predicted the neuronal response to Stim 2 better than the simple model for this neuron.

peak cross-correlation of these RF images (Supplementary Fig S5e) and the representative RF image of each cluster (Supplementary Fig S5d right). We then showed that while the RF images of neuron D were shifted mostly parallelly to the Gabor orientation (Supplementary Fig S5c), the RF images of neuron E were shifted both parallelly and orthogonally to the Gabor orientation (Supplementary Fig S5f). When populations of neurons were analysed, among all RF pairs within each neuron, 96.5% ± 10.4% (mean ± SD) of pairs were shifted, further indicating that the properties of these visualised RFs can be explained mostly by the shifts. The maximum shift distance orthogonal to the Gabor orientation varied among all neurons (Supplementary Fig S5g), reflecting the distributed shift-invariance of V1 neurons.

Characterisation of shift invariance from iteratively estimated RF images. To quantitatively understand the shift-invariance, we then developed predictive models of visual responses for each simulated complex cell and V1 neuron, termed simple model and complex model, inspired by the stimulus-response properties of simple and complex cells. In the simple model, the response to a stimulus was predicted as the normalised dot product between the stimulus image and an RF image. The RF image that yielded the best prediction similarity was chosen and used for all stimulus images (Fig. 6a). In contrast, in the complex model, the response to each stimulus was predicted as the maximum of the normalised dot products between the stimulus image and several RF images (Fig. 6b). Here, RF images used in these models were selected from 100 RF images as ones that were shifted to one another. If there was no shifted RF image, the complex model was identical to the simple model (see Methods). Figure 6 shows examples of predictions from the simple and complex models for V1 neuron E. Although the response to one image (Stim 1) was predicted moderately well by both the simple model and complex model, the prediction for another image (Stim 2) by the simple model was far poorer than the prediction by the complex model. This difference is probably because the ON/OFF phase of the RF image used in the simple model (RF 4) did not match with that of Stim 2. On the other hand, the complex model had multiple RF images, and one RF image (RF 1) matched with Stim 2. These results suggest that the responses of this neuron are somewhat tolerant to phase shifts and that such complex cell-like properties were better captured by the complex model than by the simple model.

We then measured the prediction similarity of each model for all stimulus images by the Pearson correlation coefficient between the predicted responses and actual responses. As expected, the prediction of the complex model was better than that of the simple model for this neuron E (Fig. 7a), reflecting its shift-invariant property (Figs 5 and 6).

We compared the simple model and complex model for populations of V1 neurons (Fig. 7b), simulated simple cells, and simulated complex cells. We defined the complexness index for each cell by

$$Complexness = 1 - \frac{R_{simple}}{R_{complex}} \quad (1)$$

where R_{simple} and $R_{complex}$ are the response prediction similarity of the simple model and complex model, respectively. Cells with a Gabor fitting similarity (Figs 2h and 5e) ≤0.6 were omitted from this analysis, since we have

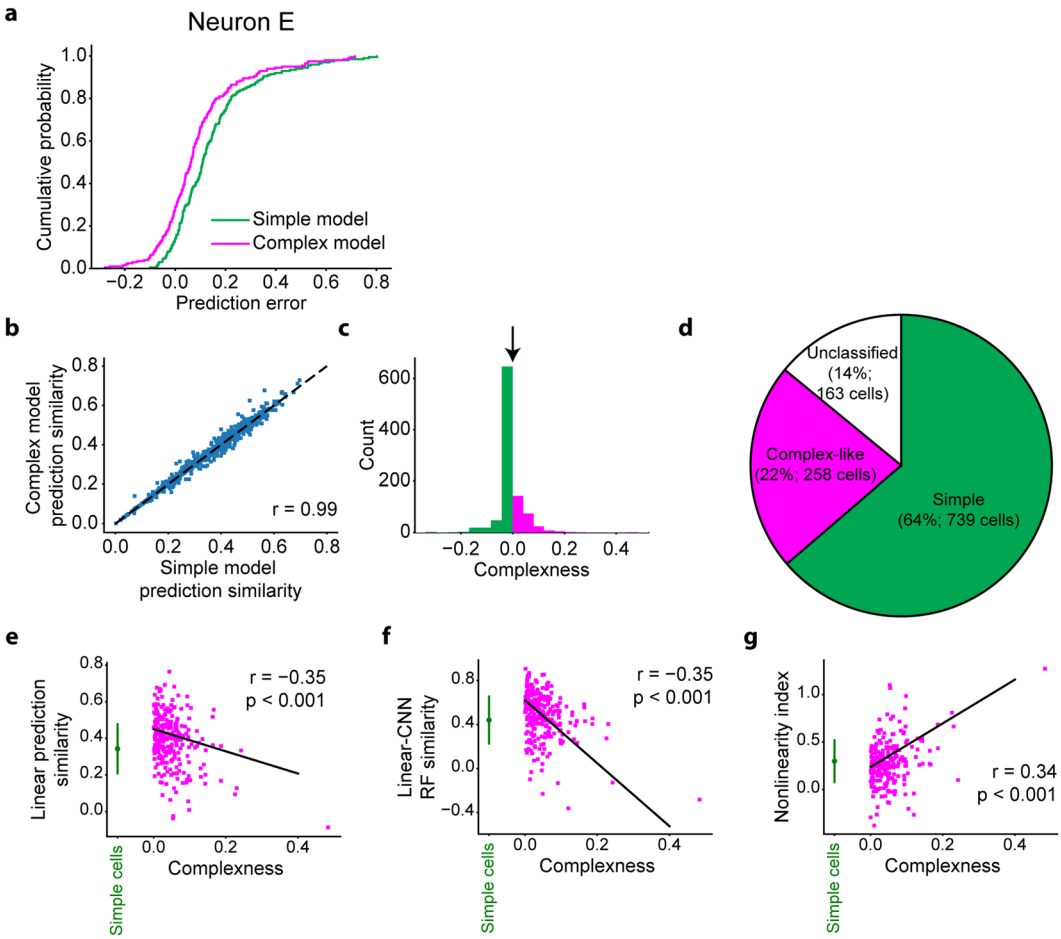

Figure 7. Simple cells and complex-like cells. **(a)** Cumulative distributions of prediction errors of the simple model (green) and the complex model (magenta) for neuron E. Prediction error was defined as the difference between the predicted response and actual response. **(b)** Relationship of similarities between the simple model and complex model (N = 997 neurons). Neurons with the Gabor fitting similarity ≤0.6, similarity of the simple model <0, or similarity of the complex model <0 were omitted from this analysis. **(c)** Distribution of complexness. Simple cells (green) and complex-like cells (magenta) were classified with threshold = 0 (black arrow). **(d)** Proportion of classified cells, simple cells, and complex-like cells among neurons with the CNN response prediction similarity >0.3. Classified cells were neurons with the Gabor fitting similarity >0.6, the response prediction similarity of the simple model >0, and the response prediction similarity of the complex model >0. Simple cells were neurons with complexness ≤0. Complex-like cells were neurons with complexness >0. **(e–g)** Relationships between complexness and linear (Lasso) prediction similarity (**e**), similarity between linear RFs and CNN RFs (**f**), and the nonlinearity index (**g**). Data of simple cells are presented as the mean ± s.d. (N = 739 neurons, green). Solid lines are the robust fit lines[74] for complex-like cells. Both linear prediction similarity and RF similarity of complex-like cells (magenta) negatively correlated with complexness (r = −0.35, p < 0.001, N = 258 neurons: e and r = −0.29, p < 0.001, N = 258 neurons: **f**), while the nonlinearity index of complex-like cells positively correlated with complexness (r = 0.34, p < 0.001, N = 258 neurons: **g**), suggesting that complexness defined here indeed reflected nonlinearity.

observed many non-Gabor-like RFs in this population. Cells with $R_{simple} < 0$, or $R_{complex} < 0$ were also omitted from this analysis. Then, we defined simple cells as cells with complexness ≤0 and complex-like cells as cells with complexness >0. The sensitivity (recall) of this classification for simulated data was 89% for simple cells and 85% for complex cells (Fig. 2l), ensuring the validity of this classification. In addition, the ratio of complex-like cells (26%, 258/997 neurons; Fig. 7c,d) among classified V1 neurons was consistent with that in a previous study[46]. When different thresholds for the cell selection were used, this ratio was 22.1−36.2% (median: 28.7%) (Supplementary Fig S3c).

We also compared complexness with other indices of linearity and nonlinearity using a dataset of V1 neurons. First, linear prediction similarity, measured as the prediction similarity of the L1-regularised linear regression model (Lasso), significantly anti-correlated with complexness for complex-like cells (Fig. 7e) (r = −0.35, N = 258, freedom = 256, t-value = −5.93, and p < 0.001; Student's t-test), suggesting that the linear regression models could not accurately predict the responses of neurons with high complexness. Similarity between linear RF images and CNN RF images also anti-correlated significantly with complexness (Fig. 7f) (r = −0.35, N = 258,

freedom = 256, t-value = −5.90, and $p < 0.001$; Student's t-test), suggesting that linear RFs could not accurately capture the RFs of neurons with high complexness. Furthermore, the nonlinearity index ((CNN prediction similarity – Lasso prediction similarity)/CNN prediction similarity; see Methods) significantly correlated with complexness (Fig. 7g) ($r = 0.34$, N = 258, freedom = 256, t-value = 5.83, and $p < 0.001$; Student's t-test), suggesting that the nonlinearity of V1 neurons was at least in part introduced by the nonlinearity of complex-like cells.

Simple cells and complex-like cells were not spatially clustered in V1. Finally, we tested whether simple cells and complex-like cells were spatially organised in the cortical space. We first investigated the spatial structure of complexness by comparing the difference in complexness with the cortical distance between all neuron pairs, which were sampled at individual planes (N = 129451 neuron pairs). We found no correlation between complexness and cortical distance ($r = -0.01$, freedom = 129449, and t-value = −3.87), suggesting no distinct spatial organisation of complexness (Fig. 8a left and 8b). We also calculated the cortical distances of all simple cell-simple cell pairs and complex-like cell-complex-like cell pairs. The cumulative distributions of these distances, normalised by the area, were both within the first and 99th percentiles of the position-permuted simulations (1,000 times for each plane; see Methods for the permutations), demonstrating no cluster organisation of simple cells or complex-like cells (Fig. 8a right and 8c).

Discussion

We first revealed that the CNN can well predict the responses to natural images for both simulated cells and V1 neurons (Figs 2c, 3b, 4b). This finding is not surprising in light of the recent successes of artificial neural networks, especially CNN, in computer vision[18–20]. Such successes could be attributed to the ability of CNN to acquire sophisticated statistics of high-dimensional data[47]. Likewise, the good prediction of CNN shown in this study is possibly due to its ability to capture higher-order nonlinearity between stimulus images and responses. Notably, the prediction of CNN was good even though the hyperparameters and layer structures of CNNs were identical for all types of cells, suggesting that CNN might be used as a general-purpose encoding model of visual neurons.

Using simulated cells, we showed that nonlinear RFs could be accurately estimated by CNN encoding models by the maximization-of-activation approach. In particular, various types of response nonlinearity could be visualised, including RFs with different phases for complex cells (Fig. 2d,f) and RFs with different orientations for rotation-invariant cells (Fig. 3c). One advantage of this RF estimation method is that it does not require an explicit assumption regarding the nonlinearities of RFs, whereas most methods for nonlinear RF estimation in previous studies do. Second-order Wiener kernel[12] and spike-triggered covariance[13–15] are capable of estimating RFs with second-order nonlinearity at most, and Fourier-based methods[48,49] estimate RFs that are linearised in the Fourier domain. The second advantage is that our method can directly visualise the image that is predicted to evoke the highest response of the target cell, in contrast to previously proposed RF estimations from artificial neural networks[29,31,32]. As suggested in[50], the disadvantage of the maximization-of-activation approach is that it may produce unrealistic images even if the maximisation of activation was successful because the candidate image space is extremely vast. To avoid this issue, we constrained the candidate image space to natural images by using L_p-norm and total variance regularisations. Although the hyperparameters of regularisations were fixed across all analysed cells, these regularisations worked well when considering the quality of the resultant RF images.

Although artificial neural networks and cortical neural networks have much in common[51], the former might not be an exact *in silico* implementation of the latter (e.g., the learning algorithms discussed in[52]). However, recent studies have suggested that the representations of CNNs and the activity of the visual cortex share hierarchical similarities[53–57]. These studies raise the possibility that the CNN encoding model could be applicable to neurons with complex nonlinearities, such as rotation-invariant neurons in the inferotemporal cortex[9]. Thus, the CNN encoding model and nonlinear RF characterisation proposed in this paper will contribute to future studies of neural computations not only in V1 but also in higher visual areas.

Methods

Acquisition of neural data. All experimental procedures were performed using C57BL/6 male mice (N = 3; Japan SLC, Hamamatsu, Shizuoka, Japan), which were approved by the Animal Care and Use Committee of Kyushu University and the University of Tokyo. All experiments were performed in accordance with approved guidelines and regulations. Anaesthesia was induced and maintained with isoflurane (5% for induction, 1.5% during surgery, and ~0.5% during imaging with a sedation of ~0.5 mg/kg chlorprothixene; Sigma-Aldrich, St Louis, MO, USA). After the skin was removed from the head, a custom-made metal head plate was attached to the skull with dental cement (Super Bond; Sun Medical, Moriyama, Shiga, Japan), and a craniotomy was made over V1 (centre position: 0–1 mm anterior from lambda, +2.5–3 mm lateral from midline). Then, 0.8 mM Oregon green BAPTA-1 (OGB-1; Life Technologies, Grand Island, NY, USA), dissolved with 10% Pluronic (Life Technologies) and 25 μM sulforhodamine 101 (SR101; Sigma-Aldrich) was pressure-injected using Picospritzer III (Parker Hannifin, Cleveland, OH, USA) approximately 400 μm below the cortical surface. The craniotomy was sealed with a coverslip and dental cement.

Neuronal activity was recorded using two-photon microscopy (A1R MP; Nikon, Minato-ku, Tokyo, Japan) with a 25× objective lens (NA = 1.1; PlanApo, Nikon) and Ti:Sapphire mode-locked laser (Mai Tai DeepSee; Spectra-Physics, Santa Clara, CA, USA). OGB-1 and SR101 were both excited at a wavelength of 920 nm, and their emissions were filtered at 525/50 nm and 629/56 nm, respectively. 507 × 507 μm or 338 × 338 μm images were obtained at 30 Hz using a resonant scanner with a 512 × 512-pixel resolution. We recorded data of four planes from three mice. The recording depths were 320 and 365 μm from one mouse, and 360 μm from the other two mice.

Visual stimuli were presented using PsychoPy[58] on a 32-inch LCD monitor (Samsung Electronics, Yeongtong, Suwon, South Korea) at a refresh rate of 60 Hz. Stimulus presentation was synchronised with imaging using

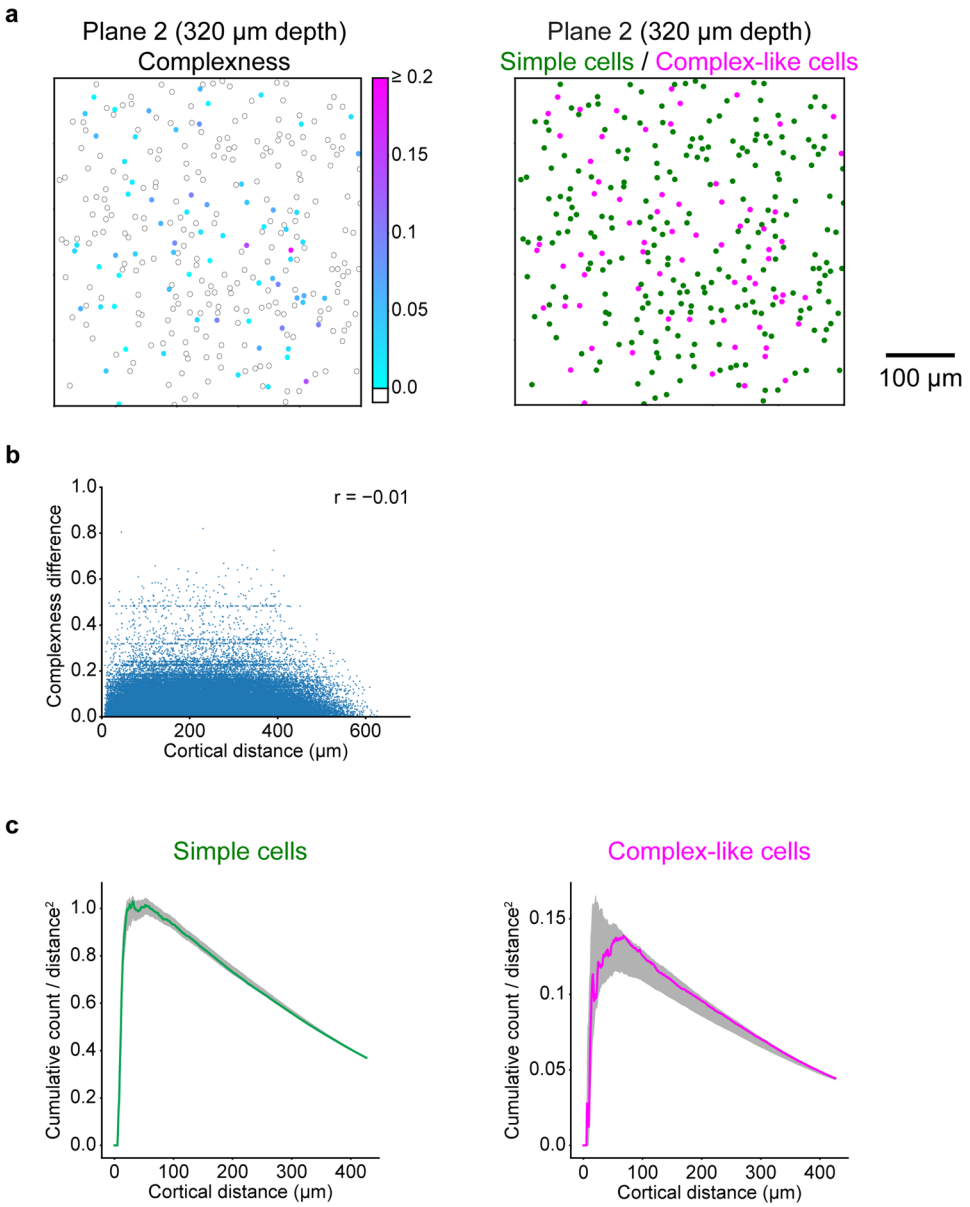

Figure 8. Spatial organisations of simple cells and complex-like cells. **(a)** Left: cortical distribution of complexness for the representative plane. The position of each neuron is represented as the circle annotated by the complexness (cyan to magenta for complex-like cells (complexness >0) and white for simple cells (complexness ≤0)). Right: cortical distribution of simple cells (N = 238 neurons, green) and complex-like cells (N = 70 neurons, magenta) for the representative plane. **(b)** Relationship between cortical distances and differences of complexness for all simple cells and complex-like cells. **(c)** Cumulative distributions of the number of simple cell-simple cell pairs (left) or complex-like cell-complex-like cell pairs (right) as a function of the cortical distance, normalised by the area. Dark shadows indicate the range from the first to 99th percentile of 1,000 position-permuted simulations for each plane. The cumulative distributions were both within the first and 99th percentiles of simulations, indicating no distinct spatial arrangements of simple cells or complex-like cells.

the transistor-transistor logic signal of image acquisition timing and its counter board (USB-6501, National Instruments, Austin, TX, USA).

First, the retinotopic position was determined using moving grating patches (contrast: 99.9%, spatial frequency: 0.04 cycles/degree, temporal frequency: 2 Hz). We first determined the coarse retinotopic position by presenting a grating patch with a 50-degree diameter at each 5 × 3 position covering the entire monitor. Then, a grating patch with a 20-degree diameter was presented at each 4 × 4 position covering an 80 × 80-degree space to fine-tune the position. The retinotopic position was defined as the position with the highest response.

Natural images (200, 1200, or 2200 images, 512 × 512 pixels) were obtained from the van Hateren Database[59] and McGill Calibrated Colour Image Database[60]. After each image was grey-scaled, it was preprocessed such that its contrast was 99.9% and its mean intensity across pixels was at an intensity level of approximately 50%, and then

masked with a circle with a 60-degree diameter. The stimulus presentation protocol consisted of 3−12 sessions. In one session, images were ordered pseudo-randomly, and each image was flashed three times in a row. Each flash was presented for 200 ms with 200-ms intervals between flashes in which a grey screen was presented.

Acquisition of simulated data. The following types of artificial cells were simulated in this study: simple, complex, and rotation-invariant cells. A simple cell was modelled using a "linear-nonlinear" cascade formulated as shown below where the response to a stimulus was defined as the dot product between the stimulus image s and a Gabor-shaped filter f_1, followed by a rectifying nonlinearity[2] and a Gaussian noise (Fig. 2a).

$$R_{simple} = \max(s * f_1, 0) + noise \tag{2}$$

A complex cell was modelled using an energy model with two subunits[37,38]. In this model, each subunit calculated the dot product between the stimulus image s and a Gabor-shaped filter f_1, f_2. Then, the outputs of these two subunits were squared, summed together, and the square root was taken. Finally, a Gaussian noise was added to define the response (Fig. 2b). Here, the Gabor-shaped filters used in this model had identical amplitude, position, size, spatial frequency, and orientation; the phase was shifted by 90 degrees. Note that this procedure, formulated as follows, can also be viewed as a "linear-nonlinear-linear-nonlinear" cascade[30,61].

$$R_{complex} = \sqrt{(s * f_1)^2 + (s * f_2)^2} + noise \tag{3}$$

A rotation-invariant cell was modelled using 36 subunits. The i-th subunit ($1 \le i \le 36$) calculated the dot product between the stimulus image s and a Gabor-shaped filter f_i. After the maximum of the outputs of the subunits was taken, a Gaussian noise was added to define the response (Fig. 3a). Here, the Gabor-shaped filters used in this model f_i had identical amplitude, position, size, spatial frequency, and phase; the orientation of the i-th subunit was 5 (i - 1) degree.

$$R_{rotation-invariant} = \max(s * f_i) + noise \tag{4}$$

We simulated 30 simple cells, 70 complex cells, and 10 rotation-invariant cells. For each cell simulation, we performed 4 trials with a different random noise. The stimuli used in these three models were identical to the stimuli used in the acquisition of real neural data (2200 images), which were down-sampled to 10×10 pixels. The Gabor-shaped filter used in these models, a product of a two-dimensional Gaussian envelope and a sinusoidal wave, was formulated as follows:

$$G(x, y) = A \exp\left(-\left(\frac{x'^2}{2\sigma_1^2} + \frac{y'^2}{2\sigma_2^2}\right)\right)\cos(k_0 y' + \tau) \tag{5}$$

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}\begin{pmatrix} x - x_0 \\ y - y_0 \end{pmatrix} \tag{6}$$

where A is the amplitude, σ_1 and σ_2 are the standard deviations of the envelopes, k_0 is the frequency, τ is the phase, (x_0, y_0) is the centre coordinate, and θ is the orientation. The parameters for f_1 of simple cells and complex cells were sampled from a uniform distribution over the following range: $0.1 \le x_0/L_x \le 0.9$, $0.1 \le y_0/L_y \le 0.9$, $0 \le A \le 1$, $0.1 \le \sigma_1/L_x \le 0.2$, $0.1 \le \sigma_2/L_y \le 0.2$, $\pi/3 \le k_0 \le \pi$, $0 \le \theta \le 2\pi$, and $0 \le \tau \le 2\pi$, where L_x and L_y are the size of the stimulus image in the x and y dimension, respectively. The parameters for f_1 of rotation-invariant cells were sampled from a uniform distribution over the following range: $0 \le A \le 1$, $0.15 \le \sigma_1/L_x \le 0.2$, $0.15 \le \sigma_2/L_y \le 0.2$, $\pi/3 \le k_0 \le 2/3\ \pi$, and $0 \le \tau \le 2\pi$. x_0, y_0 and θ were set as $L_x/2$, $L_y/2$, and 0, respectively.

The noise was randomly sampled from a Gaussian distribution with a mean of zero and standard deviation of one, which resulted in trial-to-trial variability similar to that of real data.

Data preprocessing. Data analyses were performed using Matlab (Mathworks, Natick, MA, USA) and Python (2.7.13, 3.5.2, and 3.6.1). For real neural data, images were phase-corrected and aligned between frames[62]. To determine regions of interest (ROIs) for individual cells, images were averaged across frames, and slow spatial frequency components were removed from the frame-averaged image with a two-dimensional Gaussian filter whose standard deviation was approximately five times the diameter of the soma. ROIs were first automatically identified by template matching using a two-dimensional difference-of-Gaussian template and then corrected manually. SR101-positive cells, which were considered putative astrocytes[63], were removed from further analyses. The time course of the fluorescent signal of each cell was calculated by averaging the pixel intensities within an ROI. Out-of-focus fluorescence contamination was removed using a method described previously[64,65]. The neuronal response to each natural image was computed as the difference between averaged signals during the last 200 ms of presentation and averaged signals during the interval preceding the image presentation.

For both real data and simulated data, responses were averaged across all trials and scaled such that the values were between zero and one. Natural images used in further analyses were down-sampled to 10×10 pixels. We finally standardised the distribution of each pixel by subtracting the mean and then dividing it by the standard deviation.

Encoding models. Encoding models were developed for each cell. An L1-regularised linear regression model (Lasso), L2-regularised linear regression model (Ridge), and SVR with radius basis function kernel were

implemented using the Scikit-learn (0.18.1) framework[66]. The hyperparameters of these encoding models were optimised by exhaustive grid search with 5-fold cross-validation for data of 10 real V1 neurons. The optimised hyperparameters were as follows: the regularisation coefficients of Lasso and Ridge were 0.01 and 10^4, respectively, and the kernel coefficient and penalty parameter of SVR were both 0.01. The HSM was implemented as previously proposed[31] with hyperparameters identical to the ones used in the study.

CNNs were implemented using the Keras (2.0.3 and 2.0.6) and Tensorflow (1.1.0 and 1.2.1) framework[67]. A CNN consisted of the input layer, several hidden layers (convolutional layer, pooling layer, or fully connected layer), and the output layer. The activation of a convolutional layer was defined as the rectified linear (ReLU)[68] transformation of a two-dimensional convolution of the previous layer activation. Here, the number of convolutional filters in one layer was 32, the size of each filter was (3, 3), the stride size was (1, 1), and valid padding was used. The activation of a pooling layer was 2×2 max-pooling of the previous layer activation, and valid padding was also used. The activation of a fully connected layer was defined as the ReLU transformation of the weighted sum of the previous layer activation. If the previous layer had a two-dimensional shape, the activation was flattened to one dimension. The activation of the output layer was the sigmoidal transformation of the weighted sum of the previous layer. The size of the mini batch, dropout[69] rate, type of optimizer (stochastic gradient descent (SGD) or Adam[70]), learning rate decay coefficient of SGD, and number and types of hidden layers (convolutional, max-pooling, or fully connected) were optimised with 5-fold cross-validation for the data of 10 real V1 neurons that were sampled randomly. The optimised hyperparameters of CNN were as follows: the size of the mini batch was 5 or 30 (depending on the size of the dataset), the dropout rate of fully connected layers was 0.5, the optimizer was SGD, the learning rate decay coefficient was 5×10^{-5}, and the hidden layer structure was 4 successive convolutional layers and one pooling layer, followed by one fully connected layer (Fig. 1). Other hyperparameters were fixed. To investigate the robustness of the hyperparameter optimisation, we additionally performed a post hoc search of the best hyperparameter set using different ten neurons sampled randomly. The optimal hyperparameters were identical to the original ones, except for the learning rate decay coefficient, which was 5×10^{-9} in this case. However, since we observed that the difference of the learning rate decay coefficient had minimal influence, we concluded that the hyperparameter optimisation we incorporated in this study was robust.

The training was formulated as follows:

$$W^* = \underset{W}{\operatorname{argmin}} \sum_{I,t} E(f(I;\ W),\ \ t) \tag{7}$$

where I is an image, t is the response, W is the parameters, and f is the model. E is the loss function defined as the mean squared error between the predicted responses and actual responses in the training dataset. The prediction similarity was defined as the Pearson correlation coefficient between the predicted responses and actual responses. The training procedures of CNNs were as follows. First, the training data were subdivided into data used to update the parameters (90% of training data) and data used to monitor generalisation performances (10% of training data: validation set). After the parameters were initialised by sampling from Glorot uniform distributions[71], they were updated iteratively by backpropagation[72], which was performed to minimise the loss function in either an SGD or Adam manner. SGD was formulated as follows:

$$v \leftarrow mv + \varepsilon \frac{\partial E(w)}{\partial w} \tag{8}$$

$$w \leftarrow w - v \tag{9}$$

where w is the parameter we want to update, m is the momentum coefficient (0.9), v is the momentum variable, ε is the learning rate (initial learning rate was 0.1), and $E(w)$ is the loss with respect to the batched data. Adam was formulated as previously suggested[70]. The training iterations were stopped upon saturation of the prediction similarity for the validation set.

The response prediction of each encoding model was evaluated in a 5-fold cross-validation manner for each cell not used for hyperparameter optimisations. To quantify the nonlinearity of each cell, we defined a nonlinearity index for each cell by comparing the response prediction similarity of Lasso and CNN in the following way:

$$nonlinearity\ index = 1 - \frac{R_{Lasso}}{R_{CNN}} \tag{10}$$

where R_{CNN} and R_{Lasso} are the response prediction similarity of CNN and Lasso, respectively.

RF estimation. Nonlinear RFs were estimated from trained CNNs using a regularised version of a maximization-of-activation approach[23,24]. Cells with a CNN prediction similarity ≤ 0.3 were omitted from this analysis. First, CNN was trained using all data for each cell. Then, starting with a randomly initialised image, an image I was updated iteratively by gradient ascent to maximise the following objective function $E(I)$:

$$E(I) = f(I;\ W^*) - \frac{\lambda_1}{M}\|I\|_\alpha^\alpha - \frac{\lambda_2}{M}\int\left(\left(\frac{\partial I}{\partial x}\right)^2 + \left(\frac{\partial I}{\partial y}\right)^2\right)^{\beta/2} dxdy \tag{11}$$

where f is the trained CNN model; W^* is the trained parameters, which is fixed in this procedure; λ_1, λ_2, α, and β are the regularisation parameters; and M is the size of the image. Here, α and β are fixed at 6 and 1, respectively,

following a previous research[26], and λ_1 and λ_2 are fixed at 10 and 2 after searching for the best combination of the parameters beforehand. The second and third terms are regularisation terms to minimise the α-norm and total variation[26] of the image, respectively. The RMSprop algorithm[73] was used as the gradient ascent formulated as follows:

$$I \leftarrow I + \frac{\alpha}{\sqrt{r + 10^{-7}}} \frac{\partial E(I)}{\partial I} \tag{12}$$

$$r \leftarrow \gamma r + (1 - \gamma)\left(\frac{\partial E(I)}{\partial I}\right)^2 \tag{13}$$

where γ is the decay coefficient (0.95) and α is the learning rate (1.0). The image I was updated ten times, which was enough for the convergence of the gradient ascent. The generated image was finally processed such that its mean was zero and standard deviation was one (RF image). To confirm that the generated RF image maximally activates the output layer, the whole process was repeated independently until we generated an image to which the predicted response was high (for most cells, >95% of the maximum response of the actual data of each cell). Note that for representative cells (Figs 2d,f, 3c and 5b), the predicted responses to the generated RF images were >99% of the maximum response of the actual data.

To quantitatively assess the generated RF images, we fitted each RF image with a Gabor kernel $G(x, y)$ using sequential least-squares programming implemented in Scipy (0.19.0). A Gabor kernel, a product of a two-dimensional Gaussian envelope and a sinusoidal wave, was formulated as follows:

$$G(x,\ y) = A \exp\left(-\left(\frac{x'^2}{2\sigma_1^2} + \frac{y'^2}{2\sigma_2^2}\right)\right)\cos(k_0 y' + \tau) \tag{14}$$

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} cos\ \theta & sin\ \theta \\ -sin\ \theta & cos\ \theta \end{pmatrix}\begin{pmatrix} x - x_0 \\ y - y_0 \end{pmatrix} \tag{15}$$

where A is the amplitude, σ_1 and σ_2 are the standard deviations of the envelopes, k_0 is the frequency, τ is the phase, (x_0, y_0) is the centre coordinate, and θ is the orientation. The goal of fitting was to minimise the pixelwise absolute error between the RF image and a Gabor kernel. This optimisation was started with seven different initial x_0 and seven different initial y_0 to ensure that the optimisation fell in the global minima. In addition, to create a reasonable Gabor kernel, we set bounds for some of the parameters: $0 \le x_0/L_x \le 1$, $0 \le y_0/L_y \le 1$, $0 \le \sigma_1/L_x \le 0.2$, $0 \le \sigma_2/L_y \le 0.2$, and $\pi/3 \le k_0 \le \pi$, where L_x and L_y are the size of the RF image in the x and y dimension, respectively. The quality of Gabor fitting was evaluated by the pixelwise Pearson correlation coefficient between the original RF image and the fitted Gabor kernel.

Linear RF images were created by a regularised pseudoinverse method described previously[40]. The regularisation parameter was optimised for each cell by exhaustive grid search in a 10-fold cross-validation manner. For each value in the grid, responses to the held-out test data were predicted using the created RF image. Prediction similarity was calculated as the Pearson correlation coefficient between the predicted responses and actual responses. The linear RF image was created using the value with the highest prediction similarity as the regularisation parameter.

Quantification of shift-invariance (complexness). To distinguish between simple cells and complex-like cells, we then created a "shifted image set", which contained CNN RF images that were shifted with respect to one another, selected from the 100 CNN RF images. For this purpose, a zero-mean normalised cross correlation (ZNCC) was calculated for every pair of RF images (I_1, I_2):

$$ZNCC(u,\ v) = \frac{\sum_y \sum_x (I_1(x + u,\ y + v) - \overline{I_1})(I_2(x,\ y) - \overline{I_2})}{\sqrt{\sum_y \sum_x (I_1(x + u,\ y + v) - \overline{I_1})^2}\sqrt{\sum_y \sum_x (I_2(x,\ y) - \overline{I_2})^2}} \tag{16}$$

where (u, v) is a pixel shift and $\overline{I_1}$ is the mean of I_1. If the ZNCC was above 0.95 for a (u, v) pair ($(u, v) \neq (0, 0)$), these two RF images were defined as shifted to each other by (u, v) pixels. Then, for each pair of shifted RF images, we calculated the shift distance as the maximum length of (u, v) vectors projected orthogonally to the Gabor orientation. Finally, starting with the two RF images with the largest shift distance, we iteratively collected RF images that were shifted from the already collected RF images to create the "shifted image set". If none of the 100 RF images were shifted to another, the "shifted image set" consisted of the RF image with the highest predicted response.

A simple model and complex model were created for each cell as follows (Fig. 6). In the simple model, the response to a stimulus image was predicted as the normalised dot product between the stimulus image and one RF image selected from the "shifted image set". The RF image that yielded the best prediction similarity was chosen and used for all stimulus images. In the complex model, the response to a single stimulus image was predicted as the maximum of the normalised dot products between the stimulus image and RF images in the "shifted image set". The RF image with the maximal dot product was selected for each stimulus image separately. The prediction similarity for each model was quantified as the Pearson correlation coefficient between the predicted responses and actual responses among all stimulus-response datasets. Finally, the complexness index for each cell was defined by

$$Complexness = 1 - \frac{R_{simple}}{R_{complex}} \quad (17)$$

where R_{simple} and $R_{complex}$ are the response prediction similarity of the simple model and complex model, respectively. Cells with the Gabor fitting similarity ≤ 0.6, $R_{simple} < 0$, or $R_{complex} < 0$ were omitted from this analysis.

Spatial organisations of simple cells and complex-like cells. The spatial organisations of simple cells and complex-like cells were evaluated in two ways. First, for each pair of neurons, we calculated the in-between cortical distance and the difference in complexness. A relationship between the cortical distances and the complexness differences is indicative of a spatial organisation[62]. Second, we calculated the cumulative distributions of the in-between cortical distances for all pairs of simple cells and for all pairs of complex-like cells. To statistically evaluate the cumulative distributions, we permuted the cell positions 1,000 times independently for each plane. For each permutation, cell positions of simple cells were randomly sampled from original cell positions of simple and complex-like cells. Other positions were allocated for complex-like cells. After the cell positions were determined, the cumulative distributions of the in-between cortical distances were calculated. After repeating this procedure independently 1,000 times for each plane, the first and 99th percentiles of the permuted cumulative distributions were calculated for the significance levels.

Data Availability

The datasets and codes are available from the corresponding authors on reasonable request.

References

1. Hubel, D. H. & Wiesel, T. N. Receptive fields of single neurones in the cat's striate cortex. *J. Physiol.* **148**, 574–591 (1959).
2. Movshon, J. A., Thompson, I. D. & Tolhurst, D. J. Spatial summation in the receptive fields of simple cells in the cat's striate cortex. *J. Physiol.* **283**, 53–77 (1978).
3. Dean, A. F. & Tolhurst, D. J. On the distinctness of simple and complex cells in the visual cortex of the cat. *J. Physiol.* **344**, 305–325 (1983).
4. Tolhurst, D. J. & Dean, A. F. Spatial summation by simple cells in the striate cortex of the cat. *Exp. Brain Res.* **66**, 607–620 (1987).
5. DeAngelis, G. C., Ohzawa, I. & Freeman, R. D. Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. II. Linearity of temporal and spatial summation. *J. Neurophysiol.* **69**, 1118–1135 (1993).
6. Hubel, D. H. & Wiesel, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* **160**, 106–154 (1962).
7. Ito, M., Tamura, H., Fujita, I. & Tanaka, K. Size and position invariance of neuronal responses in monkey inferotemporal cortex. *J. Neurophysiol.* **73**, 218–226 (1995).
8. Brincat, S. L. & Connor, C. E. Underlying principles of visual shape selectivity in posterior inferotemporal cortex. *Nat. Neurosci.* **7**, 880–886 (2004).
9. Ratan Murty, N. A. & Arun, S. P. A Balanced Comparison of Object Invariances in Monkey IT Neurons. *eNeuro* **4** (2017).
10. Freiwald, W. A. & Tsao, D. Y. Functional Compartmentalization and Viewpoint Generalization Within the Macaque Face-Processing System. *Science* **330**, 845–851 (2010).
11. Jones, J. P. & Palmer, L. A. The two-dimensional spatial structure of simple receptive fields in cat striate cortex. *J. Neurophysiol.* **58**, 1187–1211 (1987).
12. Emerson, R. C., Citron, M. C., Vaughn, W. J. & Klein, S. A. Nonlinear directionally selective subunits in complex cells of cat striate cortex. *J. Neurophysiol.* **58**, 33–65 (1987).
13. Touryan, J., Lau, B. & Dan, Y. Isolation of relevant visual features from random stimuli for cortical complex cells. *J. Neurosci.* **22**, 10811–10818 (2002).
14. Touryan, J., Felsen, G. & Dan, Y. Spatial structure of complex cell receptive fields measured with natural images. *Neuron* **45**, 781–791 (2005).
15. Rust, N. C., Schwartz, O., Movshon, J. A. & Simoncelli, E. P. Spatiotemporal elements of macaque V1 receptive fields. *Neuron* **46**, 945–956 (2005).
16. Carandini, M. *et al*. Do we know what the early visual system does? *J. Neurosci.* **25**, 10577–10597 (2005).
17. Hornik, K., Stinchcombe, M. & White, H. Multilayer feedforward networks are universal approximators. *Neural Networks* **2**, 359–366 (1989).
18. Hinton, G. E. & Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science* **313**, 504–507 (2006).
19. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25* (eds Pereira, F., Burges, C. J. C., Bottou, L. & Weinberger, K. Q.) 1097–1105 (Curran Associates, Inc., 2012).
20. Radford, A., Metz, L. & Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In *International Conference on Learning Representations (ICLR)* (2016).
21. Fukushima, K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* **36**, 193–202 (1980).
22. LeCun, Y. *et al*. Handwritten Digit Recognition with a Back-Propagation Network. In *Advances in Neural Information Processing Systems 2* (ed. Touretzky, D. S.) 396–404 (Morgan-Kaufmann, 1990).
23. Erhan, D., Bengio, Y., Courville, A. & Vincent, P. Visualizing higher-layer features of a deep network. *Tech. report, Univ. Montr.* 1–13 (2009).
24. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In *International Conference on Learning Representations (ICLR) Workshop* (2014).
25. Zeiler, M. D. & Fergus, R. Visualizing and Understanding Convolutional Networks. In *European Conference on Computer Vision (ECCV)* (2014).
26. Mahendran, A. & Vedaldi, A. Understanding deep image representations by inverting them. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015).
27. Lehky, S. R., Sejnowski, T. J. & Desimone, R. Predicting responses of nonlinear neurons in monkey striate cortex to complex patterns. *J. Neurosci.* **12**, 3568–3581 (1992).
28. Lau, B., Stanley, G. B. & Dan, Y. Computational subunits of visual cortical neurons revealed by artificial neural networks. *Proc. Natl. Acad. Sci. USA* **99**, 8974–8979 (2002).

29. Prenger, R., Wu, M. C. K., David, S. V. & Gallant, J. L. Nonlinear V1 responses to natural scenes revealed by neural network analysis. *Neural Networks* **17**, 663–679 (2004).
30. Vintch, B., Movshon, J. A. & Simoncelli, E. P. A Convolutional Subunit Model for Neuronal Responses in Macaque V1. *J. Neurosci.* **35**, 14829–14841 (2015).
31. Antolík, J., Hofer, S. B., Bednar, J. A. & Mrsic-Flogel, T. D. Model Constrained by Visual Hierarchy Improves Prediction of Neural Responses to Natural Scenes. *PLoS Comput. Biol.* **12**, e1004927 (2016).
32. Kindel, W. F., Christensen, E. D. & Zylberberg, J. Using deep learning to reveal the neural code for images in primary visual cortex. *arXiv:1706.06208* (2017).
33. Cadena, S. A. *et al*. Deep convolutional models improve predictions of macaque V1 responses to natural images. *bioRxiv* (2017).
34. Klindt, D., Ecker, A. S., Euler, T. & Bethge, M. Neural system identification for large populations separating what and where. In *Advances in Neural Information Processing Systems 30* (eds Guyon, I. *et al*.) 3506–3516 (Curran Associates, Inc., 2017).
35. Zhang, Y., Lee, T. S., Li, M., Liu, F. & Tang, S. Convolutional neural network models of V1 responses to complex patterns. *bioRxiv* (2018).
36. Ecker, A. S. *et al*. A rotation-equivariant convolutional neural network model of primary visual cortex. *arXiv:1809.10504* (2018).
37. Adelson, E. H. & Bergen, J. R. Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A* **2**, 284-299 (1985).
38. Körding, K. P., Kayser, C., Einhäuser, W. & König, P. How are complex cell properties adapted to the statistics of natural stimuli? *J. Neurophysiol.* **91**, 206–212 (2004).
39. Glaser, J. I., Chowdhury, R. H., Perich, M. G., Miller, L. E. & Kording, K. P. Machine learning for neural decoding. *arXiv:1708.00909* (2017).
40. Smyth, D., Willmore, B., Baker, G. E., Thompson, I. D. & Tolhurst, D. J. The receptive-field organization of simple cells in primary visual cortex of ferrets under natural scene stimulation. *J. Neurosci.* **23**, 4746–4759 (2003).
41. McIntosh, L., Maheswaranathan, N., Nayebi, A., Ganguli, S. & Baccus, S. Deep Learning Models of the Retinal Response to Natural Scenes. In *Advances in Neural Information Processing Systems 29* (eds Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I. & Garnett, R.) 1369–1377 (Curran Associates, Inc., 2016).
42. Reid, R. C., Victor, J. D. & Shapley, R. M. The use of m-sequences in the analysis of visual neurons: linear receptive field properties. *Vis. Neurosci.* **14**, 1015–1027 (1997).
43. Jammalamadaka, S. R. & SenGupta, A. *Topics in Circular Statistics*. (World Scientific, 2001).
44. Jones, J. P. & Palmer, L. A. An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *J. Neurophysiol.* **58**, 1233–1258 (1987).
45. Pnevmatikakis, E. A. *et al*. Simultaneous Denoising, Deconvolution, and Demixing of Calcium Imaging Data. *Neuron* **89**, 285–299 (2016).
46. Niell, C. M. & Stryker, M. P. Highly Selective Receptive Fields in Mouse Visual Cortex. *J. Neurosci.* **28**, 7520–7536 (2008).
47. LeCun, Y., Bengio, Y. & Hinton, G. E. Deep learning. *Nature* **521**, 436–444 (2015).
48. David, S. V., Vinje, W. E. & Gallant, J. L. Natural stimulus statistics alter the receptive field structure of V1 neurons. *J. Neurosci.* **24**, 6991–7006 (2004).
49. David, S. V. & Gallant, J. L. Predicting neuronal responses during natural vision. *Netw. Comput. Neural Syst.* **16**, 239–260 (2005).
50. Nguyen, A., Yosinski, J. & Clune, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015).
51. Hassabis, D., Kumaran, D., Summerfield, C. & Botvinick, M. Neuroscience-Inspired Artificial Intelligence. *Neuron* **95**, 245–258 (2017).
52. Bengio, Y., Lee, D. H., Bornschein, J., Mesnard, T. & Lin, Z. Towards Biologically Plausible Deep Learning. *arXiv:1502.04156* (2015).
53. Yamins, D. L. K. *et al*. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. USA* **111**, 8619–8624 (2014).
54. Khaligh-Razavi, S. M. & Kriegeskorte, N. Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Comput. Biol.* **10**, e1003915 (2014).
55. Cadieu, C. F. *et al*. Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Comput. Biol.* **10**, e1003963 (2014).
56. Güçlü, U. & van Gerven, M. A. J. Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *J. Neurosci.* **35**, 10005–10014 (2015).
57. Horikawa, T. & Kamitani, Y. Generic decoding of seen and imagined objects using hierarchical visual features. *Nat. Commun.* **8**, 15037 (2017).
58. Peirce, J. W. Generating stimuli for neuroscience using PsychoPy. *Front. Neuroinform.* **2**, 1–8 (2009).
59. van Hateren, J. H. & van der Schaaf, A. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc. R. Soc. B Biol. Sci.* **265**, 359–366 (1998).
60. Olmos, A. & Kingdom, F. A. A. A biologically inspired algorithm for the recovery of shading and reflectance images. *Perception* **33**, 1463–1473 (2004).
61. McFarland, J. M., Cui, Y. & Butts, D. A. Inferring Nonlinear Neuronal Computation Based on Physiologically Plausible Inputs. *PLoS Comput. Biol.* **9**, e1003143 (2013).
62. Ohki, K., Chung, S., Ch'ng, Y. H., Kara, P. & Reid, R. C. Functional imaging with cellular resolution reveals precise micro-architecture in visual cortex. *Nature* **433**, 597–603 (2005).
63. Nimmerjahn, A., Kirchhoff, F., Kerr, J. N. D. & Helmchen, F. Sulforhodamine 101 as a specific marker of astroglia in the neocortex *in vivo*. *Nat. Methods* **1**, 31–37 (2004).
64. Kerlin, A. M., Andermann, M. L., Berezovskii, V. K. & Reid, R. C. Broadly Tuned Response Properties of Diverse Inhibitory Neuron Subtypes in Mouse Visual Cortex. *Neuron* **67**, 858–871 (2010).
65. Hagihara, K. M., Murakami, T., Yoshida, T., Tagawa, Y. & Ohki, K. Neuronal activity is not required for the initial formation and maturation of visual selectivity. *Nat. Neurosci.* **18**, 1780–1788 (2015).
66. Pedregosa, F. *et al*. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
67. Abadi, M. *et al*. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv:1603.04467* (2016).
68. Nair, V. & Hinton, G. E. Rectified Linear Units Improve Restricted Boltzmann Machines. In *International Conference on Machine Learning (ICML)* (2010).
69. Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. R. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv:1207.0580* (2012).
70. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)* (2015).
71. Glorot, X. & Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Artificial Intelligence and Statistics (AISTATS)* **9**, 249–256 (2010).
72. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323**, 533–536 (1986).
73. Hinton, G. E., Srivastava, N. & Swersky, K. Lecture 6e-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA Neural Networks Mach. Learn* (2012).
74. Fischler, M. A. & Bolles, R. C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* **24**, 381–395 (1981).

Acknowledgements

We thank all members of the Ohki laboratory, especially Ms. T. Inoue, Y. Sono, A. Ohmori, A. Honda, and M. Nakamichi for animal care. This work was supported by grants from Brain Mapping by Integrated Neurotechnologies for Disease Studies (Brain/MINDS), Japan Agency for Medical Research and Development (AMED) (to K.O.); International Research Center for Neurointelligence (WPI-IRCN), Japan Society for the Promotion of Sciences (JSPS) (to K.O.); Core Research for Evolutionary Science and Technology (CREST), AMED (to K.O.); Strategic International Research Cooperative Program (SICP), AMED (to K.O.); JSPS KAKENHI (Grant Number 25221001 and 25117004 to K.O. and 15K16573 and 17K13276 to T.Y.); the Ichiro Kanehara Foundation for the Promotion of Medical Sciences and Medical Care (to T.Y.); and the Uehara Memorial Foundation (to T.Y.). J.U. was supported by the Takeda Science Foundation and Masayoshi Son Foundation.

Author Contributions

J.U., T.Y. and K.O. designed the study; T.Y. performed the experiments; J.U., T.Y. and K.O. analysed the data; and J.U., T.Y. and K.O. wrote the paper.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-40535-4.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.