frontiers | Frontiers in Computational Neuroscience

TYPE Original Research
PUBLISHED 03 November 2022
DOI 10.3389/fncom.2022.1030707

OPEN ACCESS

EDITED BY
Chang-Eop Kim,
Gachon University, South Korea

REVIEWED BY
Jeong-woo Sohn,
Catholic Kwandong University, South Korea
Yongseok Yoo,
Soongsil University, South Korea

*CORRESPONDENCE
Se-Bum Paik
sbpaik@kaist.ac.kr

†These authors have contributed equally to this work

RECEIVED 29 August 2022
ACCEPTED 13 October 2022
PUBLISHED 03 November 2022

CITATION
Cheon J, Baek S and Paik S-B (2022) Invariance of object detection in untrained deep neural networks. *Front. Comput. Neurosci.* 16:1030707. doi: 10.3389/fncom.2022.1030707

Invariance of object detection in untrained deep neural networks

Jeonghwan Cheon[1†], Seungdae Baek[1†] and Se-Bum Paik[1,2*]

[1]Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology, Daejeon, South Korea, [2]Program of Brain and Cognitive Engineering, Korea Advanced Institute of Science and Technology, Daejeon, South Korea

The ability to perceive visual objects with various types of transformations, such as rotation, translation, and scaling, is crucial for consistent object recognition. In machine learning, invariant object detection for a network is often implemented by augmentation with a massive number of training images, but the mechanism of invariant object detection in biological brains—how invariance arises initially and whether it requires visual experience—remains elusive. Here, using a model neural network of the hierarchical visual pathway of the brain, we show that invariance of object detection can emerge spontaneously in the complete absence of learning. First, we found that units selective to a particular object class arise in randomly initialized networks even before visual training. Intriguingly, these units show robust tuning to images of each object class under a wide range of image transformation types, such as viewpoint rotation. We confirmed that this "innate" invariance of object selectivity enables untrained networks to perform an object-detection task robustly, even with images that have been significantly modulated. Our computational model predicts that invariant object tuning originates from combinations of non-invariant units *via* random feedforward projections, and we confirmed that the predicted profile of feedforward projections is observed in untrained networks. Our results suggest that invariance of object detection is an innate characteristic that can emerge spontaneously in random feedforward networks.

KEYWORDS

object detection, invariant visual perception, deep neural network, random feedforward network, learning-free model, spontaneous emergence, biologically inspired neural network, visual pathway

Introduction

Visual object recognition is a crucial function for animal survival. Human and primates can detect objects robustly, despite huge variations in the position, size, and viewing angles (Logothetis et al., 1994; Tanaka, 1996; Connor et al., 2007; Pinto et al., 2008; DiCarlo et al., 2012; Poggio and Ullman, 2013). This challenging ability is thought to be based on invariant neural tuning in the brain—Neurons that selectively respond to

various objects have been observed in higher visual areas, and these neurons showed invariant object representation across various types of transformation (Perrett et al., 1991; Ito et al., 1995; Wallis and Rolls, 1997; Hung et al., 2005; Zoccolan et al., 2007; Li et al., 2009a; Freiwald and Tsao, 2010; Apurva Ratan Murty and Arun, 2015; Ratan Murty and Arun, 2017). Behavioral- and neural-level observations of this function have led many researchers to raise the important question of how this invariance of object detection emerges.

Often, this invariant neural tuning has been considered to develop from the learning of various types of visual transformations (Biederman, 1987; Li et al., 2009b). With the notion that visual experience of natural objects contains numerous variants that transform depending on the viewing conditions, it has been suggested that the capability to detect objects invariantly can develop gradually when observers repeatedly see objects with a wide range of variations (Földiák, 1991). Notably, in the machine learning field, invariant object recognition is also implemented via learning with a massive dataset. In this case, data augmentation, a specialized method to increase the dataset volume, is often applied (Simard et al., 2003; O'Gara and McGuinness, 2019; Shorten and Khoshgoftaar, 2019) to generate images through linear transformations such as rotation, positional shifting, and flipping, as in natural visual experience. Then, invariant object recognition is achieved from the training of the augmented dataset with computer vision models (Chen et al., 2019; O'Gara and McGuinness, 2019; Shorten and Khoshgoftaar, 2019). In contrast to the above scenario, observations in newborn animals suggest the possibility of its emergence without learning: Human infants show a preference to faces despite variations of the size and rotations in depth (Turati et al., 2008; Kobayashi et al., 2012; Ichikawa et al., 2019). In addition, newborn chicks can detect virtual objects from novel viewpoints (Wood, 2013). These findings imply that invariant object detection arises without visual experience, but the developmental mechanism of this invariance in biological brains—how object invariance arises innately in the complete absence of learning—remains elusive.

A model study using a biologically inspired deep neural network (DNN) (Krizhevsky et al., 2012; Simonyan and Zisserman, 2015) has been suggested as an effective approach to this problem (Paik and Ringach, 2011; DiCarlo et al., 2012; Yamins and DiCarlo, 2016; Sailamul et al., 2017; Baek et al., 2020, 2021; Jang et al., 2020; Kim et al., 2020, 2021; Park et al., 2021; Song et al., 2021). DNNs, which consist of a stack of feedforward projections inspired by the hierarchical structure of the visual pathway, can be used as simplified model to investigate various visual functions. For instance, it was reported that a DNN trained to natural images can predict neural responses in the primate visual pathways from an early visual area (e.g., primary visual cortex, V1) to a higher visual area (e.g., inferior temporal cortex, IT) (Cadieu et al., 2014; Yamins et al., 2014). Recent studies also provided insight into the origin of functional tuning in the brain, by showing that units that selectively respond to numerosity, faces, and various types of objects among visual stimuli can arise in a randomly initialized DNN without any learning (Baek et al., 2021; Kim et al., 2021).

By adopting a similar approach, here we show that object invariance can arise in completely untrained neural networks. Using AlexNet (Krizhevsky et al., 2012), a model designed along the structure of the visual stream, we found that units selective to various visual objects were observed in a randomly initialized DNN and that these units maintained selectivity across a wide range of variations, such as the viewpoint, even without any visual training. We observed that a certain proportion of the units show an invariant tuning to viewpoint, while other groups of units show tuning to a specific viewpoint. Preferred feature images obtained from the reverse-correlation method showed that each specific viewpoint unit encodes a shape from a particular view of an object, while invariant viewpoint units encode inclusive features from specific units with different preferred angles. We found that invariant units emerge by homogenous projections from specific units in the previous layer in a random feedforward network. Finally, we confirmed that this innate invariance enables the network to perform an object-detection task under an enormous range of variations of viewpoints. Overall, our results suggest that invariant object detection can emerge spontaneously from the random wiring of hierarchal feedforward projections in an untrained DNN.

Results

Emergence of object selectivity in untrained networks

To investigate the emergence of invariant object selectivity in an untrained model network, we used AlexNet (Krizhevsky et al., 2012), a biologically inspired DNN that models the structure of the ventral visual pathway. To find an object-selective response of an individual unit in the network, we investigated the responses of the final convolutional layer (Conv5), which is presumed to correspond to the IT domain of the brain. To simulate the condition of an untrained hierarchical network, we randomly initialized AlexNet using a standardized network initialization method (LeCun et al., 1998), by which the weights of the filters in every convolutional layer are randomly selected from a Gaussian distribution.

The stimulus set was designed to contain nine different object categories (e.g., Monitor, Bed, Chair, etc.) (**Figure 1A**). To define selective units for a specific target object, eight other class object sets and one scrambled set of the target object were used, following a previous experimental study (Stigliani et al., 2015) (see section "Materials and methods" for details). The images in each class were prepared by controlling the low-level features of the luminance, contrast, object location and object

size (**Supplementary Figure 1**). Specifically, the pixel value distribution of the object image and background image were calibrated using the same Gaussian distribution (mean = 127.5, s.d. = 51.0), and the intra-class image similarity was also controlled at statistically comparable level. As this stimulus was given as input for the randomly initialized networks, the responses were measured in the Conv5 layer and an analysis of object selectivity was conducted (**Figure 1B**).

We found object-selective units that show higher responses to a specific class of target images (e.g., Toilet) than to other non-target class images and scrambled images (two-sided rank-sum test, $P < 0.001$) (**Figure 1C**). Among the nine object categories, we observed that object-selective units emerge mostly in a few objects categories (**Figures 1D,E**). We observed toilet-selective units ($n = 565 \pm 55$ in 20 random networks, mean $\pm$ s.d.), sofa-selective units ($n = 339 \pm 68$), and monitor-selective units ($n = 294 \pm 54$) in the Conv5 layer (43,264 units; $13 \times 13 \times 256$, $N_{x-position} \times N_{y-position} \times N_{channel}$) (**Figure 1D**). In particular, the number of object-selective units was divided into large and small groups (**Figure 1E**, n = 20, two-sided rank-sum test, $^{*}P < 10^{-27}$). Large groups consisted of the toilet, sofa, and monitor groups ($n_{units} = 400 \pm 133$) and small groups were the dresser, desk, bed, chair, nightstand, and table groups ($n_{units} = 32 \pm 32$). Our previous study suggested that units selective to various visual objects can arise spontaneously from the simple configuration of the geometric components and that objects with a simple profile lead to a strong clustering of abstracted responses in the network, more likely to generate units selective to it (Baek et al., 2021). To validate this in the current result, we performed an analysis using a dimension reduction method (van der Maaten and Hinton, 2008). From the examination of a clustered representation of each object class in the latent space using the silhouette index (Kaufman and Rousseeuw, 2009), we found that classes in the large group with relatively simple configurations have higher silhouette indices than those in the small group (**Supplementary Figures 2A,B**). We also confirmed that there is a significant correlation between the degree of class clustering in the latent space and the number of selective units (**Supplementary Figure 2C**, Pearson correlation coefficient, $n_{Net} = 20$, $r = 0.62$, $P < 10^{-20}$). In the subsequent analyses, we investigated the results mostly for the three object classes which show a large number of selective units.

We investigated the number and the selective index of object units across the convolutional layers and found that the number of object units increases when the convolutional layers become deeper (**Supplementary Figure 3A**). The object-selective index for a single unit also shows a strong tendency to increase across convolution layers, demonstrating that object tuning becomes sharper through the network hierarchy (**Supplementary Figure 3B**). Furthermore, we found that the responses of an untrained network measured in the deep layer (Conv5) were clustered as object classes in the latent space, while raw images do not cluster in the latent space (**Figure 1F**).

Invariance of object-selective units in untrained networks

Next, to investigate whether the observed object-selective units show viewpoint-invariant representations of an object image, we measured the responses of object-selective units to target objects and non-target objects with various viewpoint angles. To do this, a viewpoint-variant stimulus set was generated (**Supplementary Figure 4**) by rotating the viewpoint of 3D objects on the horizontal plane (**Figure 2A**). For each object, we rendered 13 variant images at different viewpoints between $-90°$ and $90°$. Then, we measured the responses of selective units to target objects and non-target objects with various viewpoint angles (**Figure 2B**). We found that units show selective responses when an object image within a certain threshold is presented (**Figure 2C**, left, $n = 200$, one-sided rank-sum test, Toilet at 0° vs. Non-toilet, $^{*}P < 10^{-13}$; Toilet at 45° vs. Non-toilet, $^{**}P < 10^{-5}$), while the units did not show selectivity when an object image at a larger viewpoint angle was given (**Figure 2C**, left, Toilet at 90° vs. Non-toilet, NS, $P = 0.492$). Hence, the selectivity of object-selective units is maintained within a limited effective range (**Figure 2C**, right).

To investigate the effective range that maintains the selectivity of object units quantitatively, we investigated the responses of selective units with a viewpoint between $-90°$ and $90°$ and estimated the boundary of the viewpoint variation around which target-object tuning is lost. For example, we observed that object tuning of toilet units was retained when the viewpoint change was within 105° (**Figure 2D**, left, $n = 200$, one-sided rank-sum test, $P < 0.05$). Then, to verify whether the viewpoint invariance of an object-selective unit simply arises due to the similarity of the object shape upon a change of the viewpoint, we estimated the pixel-wise raw-image correlations between object images from a front view and a rotated view. We compared the effective ranges of viewpoint invariance between the selective responses and the image correlations (**Figure 2D**, right). For toilet units, we observed that the effective range of the selective responses is significantly wider than that of the image correlation (**Figure 2E**, Toilet units). Similarly, this tendency was commonly observed in other object-selective units (**Figures 2E,F**, $n = 20$, two-sided rank-sum test; Toilet unit, $^{*}P < 10^{-4}$; Sofa unit, $^{*}P < 10^{-4}$; Monitor unit, $^{*}P < 10^{-4}$). This result suggests that the observed invariance is not simply due to the similarity of the object images at different viewpoints but is a characteristic of object-selective units in untrained networks. To find the origin of the invariance in an untrained network, we also examined the single-unit-level characteristics of invariance. We found that each unit shows considerable variations in the response characteristics when a target object image with various viewpoints is given as the input. In particular, each unit shows various effective ranges (**Figures 3A,B**, left). Considering the definition of viewpoint invariance, we presumed that the top 30% of units were "viewpoint-invariant" units and the bottom

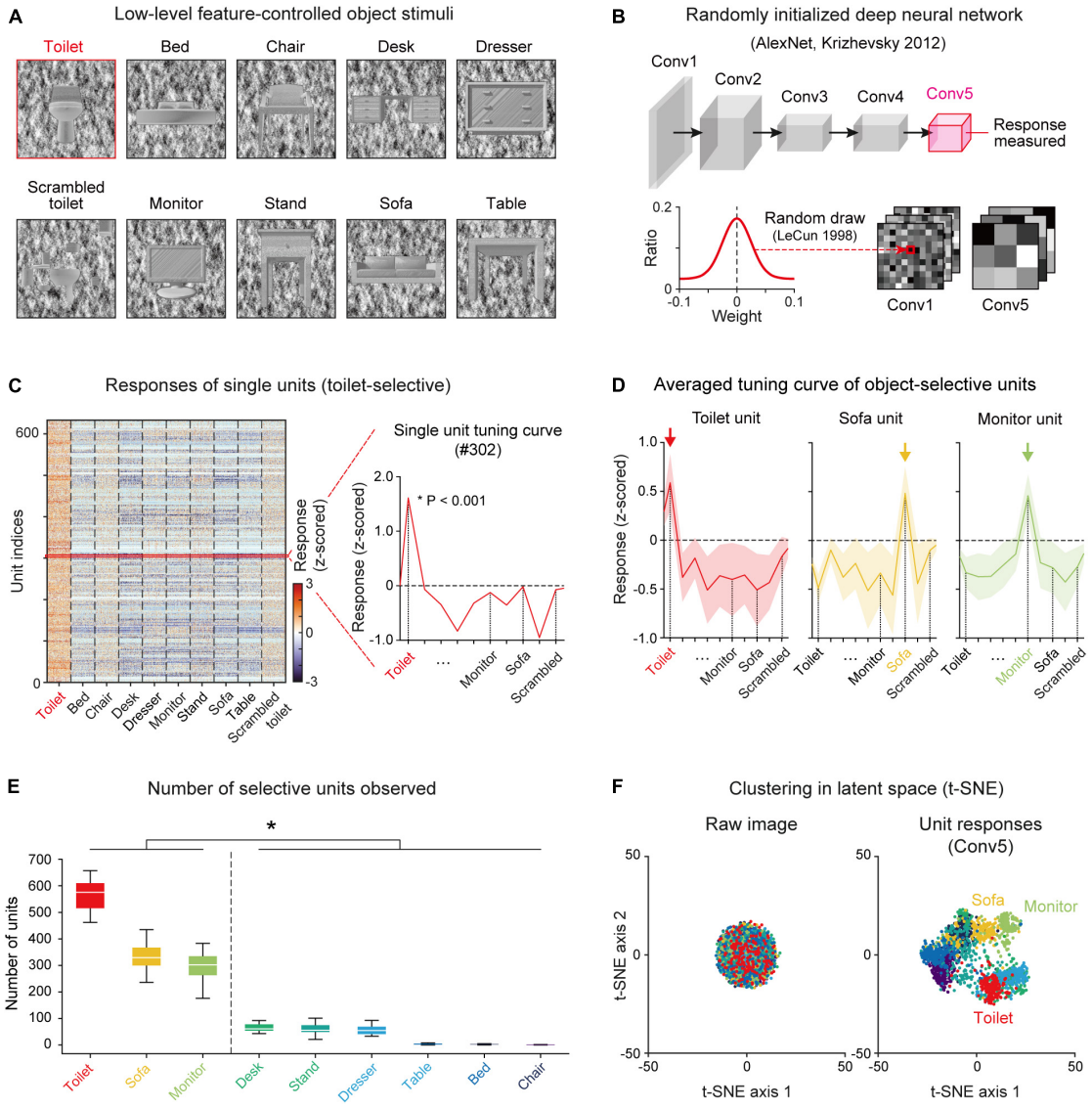

FIGURE 1
Emergence of selectivity to various objects in untrained networks: **(A)** The stimulus images were selected and modified from a publicly available CAD dataset (https://modelnet.cs.princeton.edu/) (see section "Materials and methods for details). The images contain nine different object categories. The low-level features of the luminance, contrast, object location and object size of images were calibrated equally across object classes. **(B)** The structure of a randomly initialized deep neural network. Five convolutional layers in AlexNet (Krizhevsky et al., 2012) were randomly drawn from a Gaussian distribution (LeCun et al., 1998). **(C)** Responses of each single toilet-selective unit (two-sided rank-sum test, $P < 0.001$). The red curve is an example tuning curve of a single unit. **(D)** Average responses of selective units for three object classes (Toilet, $n = 548$; Sofa, $n = 285$; Monitor, $n = 301$). Each arrow indicates the preferred object class. Shaded areas represent the standard deviation from the tuning curves of the target units. **(E)** The number of object-selective units for nine classes in untrained networks ($n = 20$). Box plots indicate the inter-quartile range (IQR between Q1 and Q3) of the dataset, the white line depicts the median and the whiskers correspond to the rest of the distribution (Q1–1.5*IQR, Q3 + 1.5*IQR). **(F)** Visualization of the latent space by the t-SNE method (van der Maaten and Hinton, 2008) from raw images and the responses of Conv5 units to each class. The raw images of each object class do not cluster in the latent space, but the responses of the untrained network collected in Conv5 were clustered in the latent space according to the class of the given image.

30% units were "viewpoint-specific" units in the subsequent analyses. Indeed, we observed that each tentative viewpoint-specific unit has various preferred angles; i.e., they only respond to a particular view of an object (Figure 3B, right).

To verify our conjecture that "viewpoint-specific" and "viewpoint-invariant" units exist and can be classified according to the observed effective range of each unit (Figures 3A,B), we investigated the responses of object-selective units for object images with different viewpoints. Target object images with a viewpoint between −60° and 60° (five steps, 50 images per viewpoint class) were presented to the network, and the responses were measured. Indeed, we observed that there

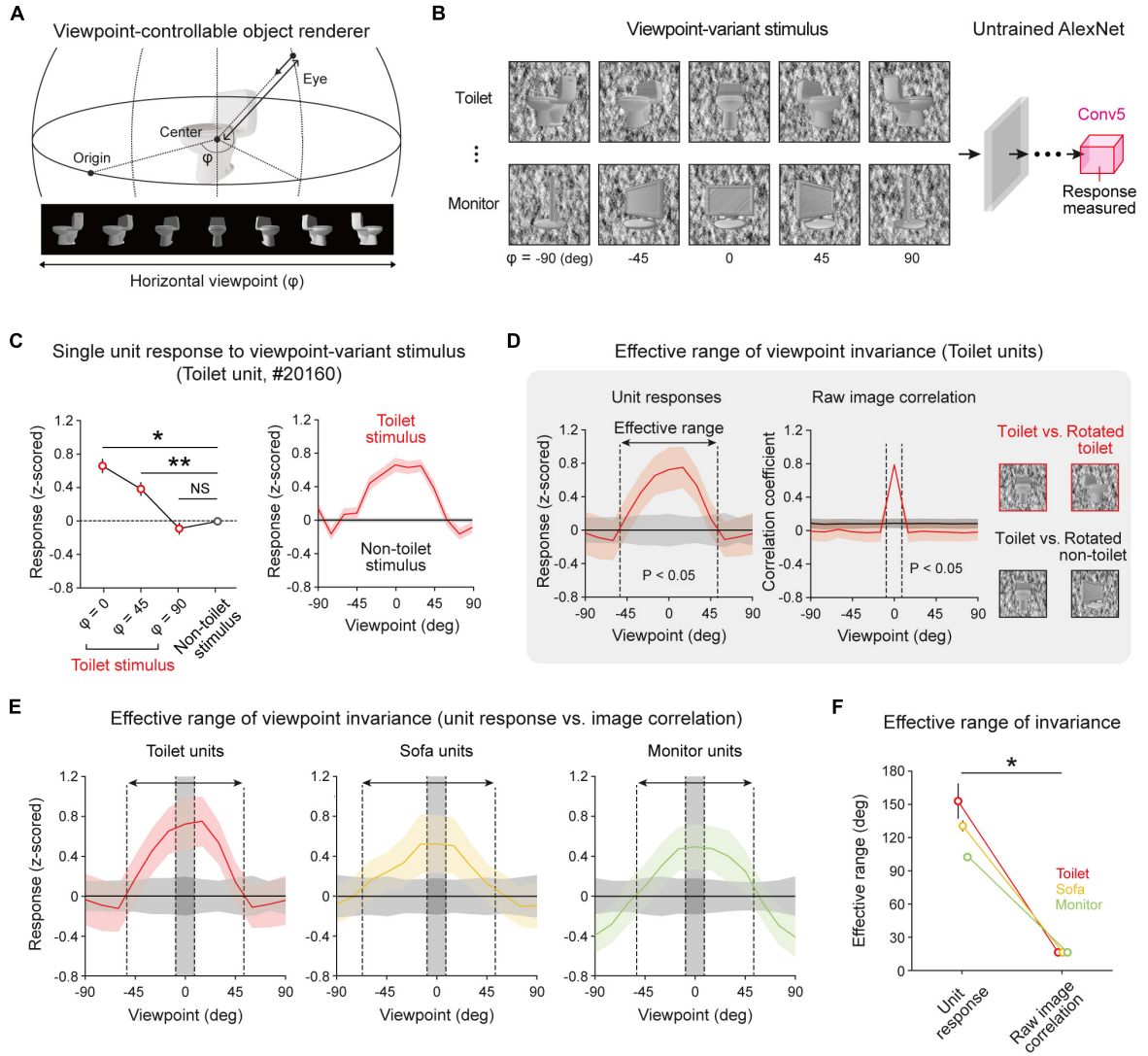

FIGURE 2
Viewpoint-invariant object selectivity observed in an untrained network: **(A)** An object renderer generates object images at various viewpoints rotating in a horizontal orbit. **(B)** The viewpoint-varying object stimulus was generated within the viewpoint range of −90° to +90° with 13 steps. The responses of the object-selective unit were measured in the final convolutional layer of an untrained AlexNet. **(C)** Viewpoint-invariant responses of a selective unit for viewpoint-rotated object stimulus; the tuning of a single toilet-selective unit shows a wide range of viewpoint invariance. Shaded areas and error bars represent the standard error of 200 images (n = 200, one-sided rank-sum test, Toilet at 0° vs. Non-toilet, $*P < 10^{-13}$; Toilet at 45° vs. Non-toilet, $**P < 10^{-5}$; NS, P = 0.492). **(D)** The effective range of the average responses of object units and the pixel-wise correlation of a raw image (n = 200, one-sided rank-sum test, $P < 0.05$). **(E)** Average response of object-selective units for an object stimulus at different viewpoint rotations. The arrow indicates the effective range of the selective response, and the shaded area between dashed lines indicates the effective range of the raw-image correlation. Shaded areas represent the standard deviation of 200 images. **(F)** Comparison of effective ranges between the selective response and raw-image correlation in each object-selective unit (n = 20, two-sided rank-sum test; Toilet unit, $*P < 10^{-4}$; Sofa unit, $*P < 10^{-4}$; Monitor unit, $*P < 10^{-4}$). Error bars indicate the standard deviation of 20 random networks.

are units that only respond to a particular viewpoint image (**Figure 3C**, Viewpoint-specific, one-way ANOVA with single peak filtering, $P < 0.05$) and units that respond invariantly to any viewpoint image (**Figure 3C**, Viewpoint-invariant, one-way ANOVA, $P > 0.05$). Viewpoint-specific units show highly tuned responses to one preferred viewpoint angle, while viewpoint-invariant units show a flat tuning curve to any viewpoint (**Figure 3D**).

Next, to visualize the distinct tuning features of viewpoint-specific and viewpoint-invariant units, we used a reverse-correlation method (Bonin et al., 2011; Baek et al., 2021) and obtained the preferred feature images (PFIs) of units (**Figure 3E**, see section "Materials and methods" for details). We found that each specific unit showed a PFI similar to an object image at the viewpoint angle of its preferred value. From this result, we confirmed that each specific unit encodes a shape from a

10.3389/fncom.2022.1030707

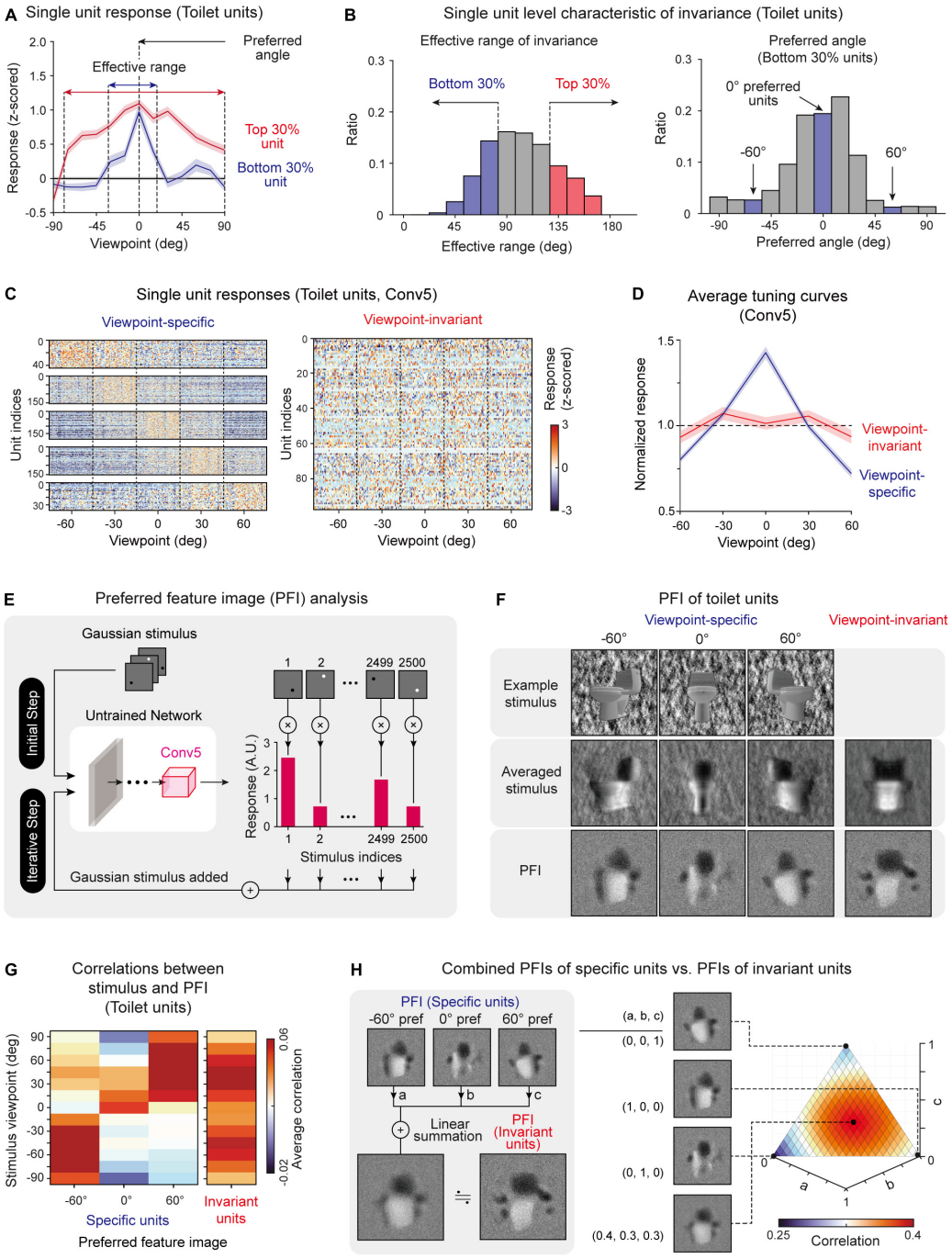

FIGURE 3

Single-unit-level analysis of invariance: **(A)** Viewpoint tuning curves of two sample toilet units: a unit with a wide effective range (Top 30%) and a unit with a narrow effective range (Bottom 30%). **(B)** Histogram of the invariance effective range of each unit and histogram of the preferred angle of units with a narrow effective range (Bottom 30%). **(C)** Responses of individual viewpoint-specific and viewpoint-invariant toilet-selective units in the Conv5 layer (Viewpoint-specific units, one-way ANOVA with single peak filtering, $P < 0.05$; Viewpoint-invariant units, one-way ANOVA, $P > 0.05$). **(D)** Average tuning curves of viewpoint-specific units ($n = 147$) and viewpoint-invariant units ($n = 95$) in an untrained network. Shaded areas represent the standard error of each type of unit. **(E)** Overall process of the preferred feature image (PFI) of target units in Conv5 of untrained networks using a reverse-correlation analysis (Bonin et al., 2011; Baek et al., 2021). The input stimulus was generated with a randomly positioned bright or dark dot blurred with a 2D Gaussian filter. The PFI was calculated as the response-weighted summation of the input stimulus. **(F)** Obtained preferred feature images of viewpoint-invariant and viewpoint-specific units with different preferred angles (−60°, 0°, 60°). An example stimulus and an average stimulus corresponding to preferred viewpoint angle. For invariant units, the average stimulus across all viewpoints is presented. **(G)** Correlation between the stimulus of various viewpoints and PFIs of various types of units. **(H)** Comparison between the PFI of an invariant unit and the weighted summation of PFIs of specific units. Here, "a," "b", and "c" represent the weight of each PFI for summation (a + b + c = 1). The 3D plot represents the pixel-wise correlations for different values of each weight pairs.

particular view of the object (Figure 3F, Specific). In contrast, the PFIs of viewpoint-invariant units were similar to the average stimulus image of various viewpoints (Figure 3F, Invariant). Notably, the calculation of the correlations between the stimulus of various viewpoints and the PFIs from each different type of unit reveals that the PFI of specific units shows a high correlation only with the stimulus image of the corresponding viewpoint, while that of invariant units shows high correlations with the stimulus images of various viewpoints (Figure 3G). From this observation, we hypothesized that the PFIs of invariant units can be expressed as a linear combination of the PFIs of specific units. We tested this scenario by searching for wiring coefficients that maximize the correlation between the PFIs of invariant units and a combined PFIs of specific units (Figure 3H, left). We observed a very high correlation when each PFI of a specific unit is linearly combined with fairly homogeneous coefficients (Figure 3H, right). The same tendency was observed in the PFIs of other object-selective units (Supplementary Figure 5). These results suggest that viewpoint-invariant units can originate from a homogenous combination of viewpoint-specific units.

The feedforward model can explain the spontaneous emergence of invariance

To validate the hypothesis that viewpoint-invariant units originate from the projection of viewpoint-specific units in the previous layer, we backtracked projections of the units from the source layer (Conv4) to the target layer (Conv5) and examined the weights of connected viewpoint-specific units. First, we confirmed that viewpoint-specific ($n = 765 \pm 102$) and viewpoint-invariant toilet-selective units ($n = 130 \pm 28$) exist in Conv4 as well as in Conv5 (Viewpoint-specific, $n = 504 \pm 81$, Viewpoint-invariant, $n = 96 \pm 16$). We confirmed that the viewpoint-specific units in Conv5 receive stronger input from units with the same object tuning than from other units in Conv4 (Figure 4A, left and middle, $n = 20$, two-sided rank-sum test, $^*P < 10^{-7}$). In more detail, the viewpoint-specific units in Conv5 receive inputs from Conv4 units strongly biased to a particular viewpoint angle (Figure 4A, right, $n = 20$, one-way ANOVA, $^*P < 10^{-11}$). This tendency of a strongly biased weight also appeared in other preferred viewpoints. The connectivity between viewpoint-specific units with the same preferred angle in the source and target layers showed significantly high weights compared to other projection directions (Figure 4B, n = 20, one-way ANOVA, $^*P < 0.05$).

We also found that viewpoint-invariant units in Conv5 are strongly connected to units with the same object selectivity in Conv4 (Figure 4C, left and middle, $n = 20$, two-sided rank-sum test, $^*P < 10^{-7}$), as in the case of viewpoint-specific units. However, the viewpoint-invariant units in Conv5 receive homogeneous inputs from specific units in Conv4 units with various preferred viewpoint angles (Figure 4C, right, $n = 20$, one-way ANOVA, NS, $P = 0.313$). To estimate the degree of homogeneity in the projection weights, we defined the homogenous index as the inverse of the standard deviation of the average weight connected to specific units with different viewpoints in the source layer. The index of the average weight connected to viewpoint-invariant units is significantly higher than that of the average weight connected to the viewpoint-specific units, indicating an unbiased input to the viewpoint-invariant units (Figure 4D, $n = 20$, two-sided rank-sum test, Toilet; Invariant units vs. Specific units, $^*P < 0.05$). This tendency was also observed in units with other object tunings (Figure 4D, $n = 20$, two-sided rank-sum test, Sofa and Monitor; Invariant units vs. Specific units, $^*P < 0.05$; Supplementary Figure 6). This implies that observed viewpoint invariance of object tuning can originate from hierarchical random feedforward projections.

To verify this developmental model further, we revisited earlier observations of invariant object tuning in the monkey IT which reported that neurons in the higher layer in the hierarchy show increased invariance (from the ML to the AM area) (Freiwald and Tsao, 2010). Our previous study (Baek et al., 2021) suggested that the viewpoint-invariant units in deeper layers emerges by receiving feedforward inputs from units with a fairly homogeneous distribution of viewpoint angles in previous layers. In this scenario, the degree of invariance is expected to increase in deeper layers because the chance of combined connectivity, which induces invariant responses, increases. To verify this scenario in the current result, we investigated the weight of invariant units connected to units in the previous layer (Supplementary Figure 7A). We found that viewpoint invariant units in each layer are strongly connected to specific units with various preferred viewpoint angle, or invariant units already exist in the previous layer, as expected. This tendency was observed consistently in units with various object selectivity (Supplementary Figure 7B, $n = 20$, two-sided rank-sum test, from invariant units, Conv4 to Conv5, Toilet: $^*P < 10^{-7}$, Sofa: $^*P < 10^{-7}$, Monitor: $^*P < 10^{-7}$; from specific units, Conv3 to Conv4, Toilet: $^*P < 10^{-7}$, Sofa: $^*P < 10^{-7}$, Monitor: $^*P < 10^{-6}$; Conv4 to Conv5, Toilet: $^{**}P < 10^{-7}$, Sofa: $^{**}P < 10^{-7}$, Monitor: $^{**}P < 10^{-7}$).

From this result, we investigated whether this connectivity profile induces an increased trend of invariance across layers in our model neural network and found that such layer-specific characteristics of viewpoint invariance also emerge in the untrained network we used. We observed that the level of invariance increased along the network hierarchy (Supplementary Figure 7C). To quantify these invariance characteristics, we introduced an invariance index of units, defined as the inverse of the standard deviation of responses across different viewpoints. We observed an increase in the invariance index of selective units higher up in the hierarchy in the untrained AlexNet (Supplementary Figure 7D). The viewpoint-invariance index in Conv4 is significantly higher than

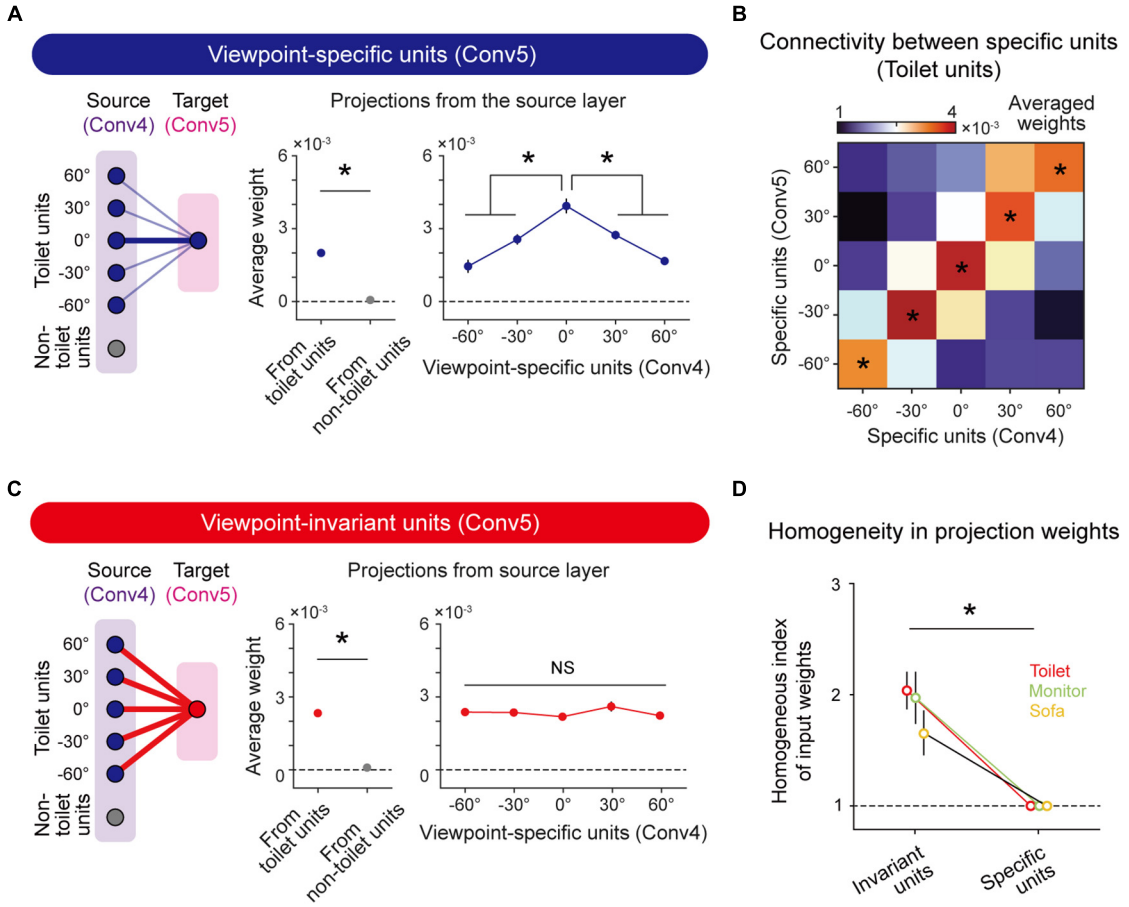

FIGURE 4
Emergence of viewpoint invariance based on unbiased projection from viewpoint-specific units: **(A)** Connectivity diagram (left); averaged weight from target object units and non-target object units (Conv4) to viewpoint-specific units (Conv5) (middle, $n = 20$, two-sided rank-sum test, $^{*}P < 10^{-7}$); averaged weight from viewpoint-specific units (Conv4) to viewpoint-specific units (Conv5) (right, $n = 20$, one-way ANOVA, $^{*}P < 10^{-11}$). Error bars indicate the standard error of 20 random networks. **(B)** Heatmap of weights between specific units from the source layer and specific units from the target layer. This heatmap shows biased input to specific units ($n = 20$, one-way ANOVA, $^{*}P < 0.05$). **(C)** Connectivity diagram (left); averaged weight from target object units and non-target object units (Conv4) to viewpoint-invariant units (Conv5) (middle; $n = 20$, two-sided rank-sum test, $^{*}P < 10^{-7}$); and averaged weight from viewpoint-specific units (Conv4) to viewpoint-invariant units (Conv5) (right, $n = 20$, one-way ANOVA, NS, $P = 0.313$). Error bars indicate the standard error of 20 random networks. **(D)** The homogeneous index of the input projection weight that connected viewpoint-specific units (Conv4) to viewpoint-invariant units (Conv5) and viewpoint-specific units (Conv5), respectively ($n = 20$, two-sided rank-sum test, $^{*}P < 0.05$). Error bars indicate the standard deviation of 20 random networks.

that in Conv3 ($n = 20$, two-sided rank-sum test, $^{*}P < 10^{-7}$). Also, the viewpoint-invariance index in Conv5 is significantly higher than that in Conv4 ($n = 20$, two-sided rank-sum test, $^{**}P < 10^{-7}$). This increasing tendency of the viewpoint-invariance index along the network hierarchy is also observed in other object-selective units ($n = 20$, two-sided rank-sum test; Sofa, $^{*}P < 10^{-7}$, $^{**}P < 10^{-7}$; Monitor, $^{*}P < 10^{-7}$, $^{**}P < 10^{-7}$). In addition, we confirmed that the same increasing tendency of the number of invariant units along the network hierarchy exists across various object tunings (**Supplementary Figure 7E**, $n = 20$, two-sided rank-sum test; Toilet, $^{*}P < 0.05$, $^{**}P < 0.001$; Sofa, $^{*}P < 0.05$; Monitor, $^{*}P < 10^{-5}$, $^{**}P < 10^{-6}$). These results suggest that our model provides a plausible scenario for understanding the spontaneous emergence of invariant object selectivity in untrained networks, which is supported by previous experimental observations of neural tunings.

Innate invariance enables invariant object detection without data-augmented learning

Next, we tested whether this innate invariance in untrained networks enables the network to perform the invariant object-detection task without learning. We expected that the information given by invariant object units is sufficient to detect an object while the viewpoint of the given object image varies, and in particular, that viewpoint-invariant units play a key

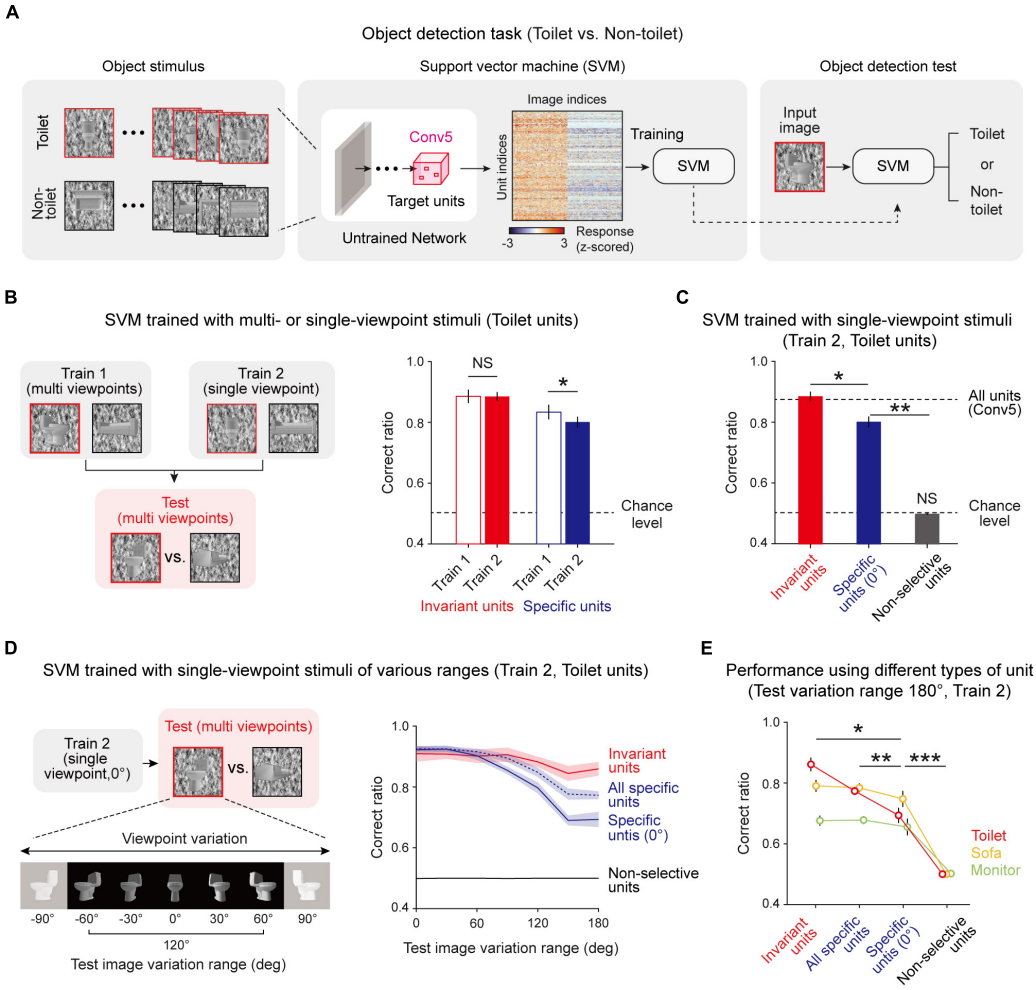

FIGURE 5
Invariantly tuned unit responses enable invariant object detection: **(A)** Overall process of the object-detection task and SVM classifier using the responses of object-selective units. To train the SVM classifier, 30 images of a target object and 30 images of a non-target object were used. Among the 60 images, 40 images were randomly sampled for training and the remaining 20 were used to test the task performance. The responses of untrained network units for these images are obtained and used to train and test the SVM to classify whether or not the given image is the target object. **(B)** Train 1 uses object images with various viewpoints to train the SVM, while Train 2 uses object images only with a center-fixed viewpoint. For the test SVM, object images with various viewpoints are used (n = 20, two-sided rank-sum test, Invariant, NS, P = 0.735; Specific, *$P < 0.001$). **(C)** Performance of the Train 2 method using invariant units, specific units, and non-selective units (n = 20, two-sided rank-sum test, Invariant vs. Specific, *$P < 10^{-7}$; Specific vs. Chance level, **$P < 10^{-7}$; Non-selective vs. Chance level, NS, P = 0.116). The upper dashed line represents the performance when all units in Conv5 are used, and the lower dashed line indicates the chance level of the task. **(D)** The task in various test-image-variation-range conditions. The test images were randomly sampled within the given viewpoint variation range. The Train 2 performances of invariant units, all specific units, specific units only with 0° preferred, and non-selective units were assessed. **(E)** Comparison of performances across the types of object units. The performance for each case was measured using test images with a 180° variation range (n = 20, two-sided rank-sum test, Invariant vs. Specific with center view, *$P < 0.05$; All specific vs. Specific with center view, **$P < 0.01$; Specific with center view vs. Non-selective, ***$P < 10^{-6}$). Shaded areas and error bars indicate the standard deviation of 20 random networks.

role in enabling invariant object detection. To confirm this hypothesis, we designed two different methods to train an SVM which classifies whether or not a given image is a target object, using unit responses to stimulus given (Figure 5A). In the first case (Train 1), the SVM is trained using an object image with various viewpoints to train the SVM, while it is trained using the object image only with a center-fixed viewpoint in the second condition (Train 2). After training, object images with various viewpoints were used for the test session (Figure 5B, left). We performed this process using both invariant units and specific units.

We found that the performances of the SVM using invariant units only and those of the SVM using specific units only are noticeably different (Figure 5B, right, Invariant, n = 20, two-sided rank-sum test, NS, P = 0.735; Specific, n = 20, two-sided rank-sum test, *$P < 0.001$). Invariant units show the same level

of performance regardless of training with various viewpoints, implying that the information given by invariant units is sufficient to detect images with varying viewpoints. In contrast, specific units show significantly lower performance outcomes when trained only with a fixed viewpoint (Train 2). Hence, we investigated in more depth the performances of SVMs trained in the second condition (Train 2) using different groups of units (**Figure 5C**, n = 20, two-sided rank-sum test, Invariant vs. Specific, $^{*}P < 10^{-7}$; Specific vs. Chance level, $^{**}P < 10^{-7}$; Non-selective vs. Chance level, NS, $P = 0.116$). First, the SVM using the responses of invariant units shows significantly high performance compared to the SVM using the responses of specific units. Second, the SVM using invariant units shows the same level of performance compared to when it is trained with all units in the Conv5 layer, implying that the invariance of the untrained network mainly relies on invariant units. To confirm that invariance enables invariant object detection in a wide variation range, we tested this concept in different viewpoint-variation ranges (**Figure 5D**, left). When the viewpoint variation range in the test set becomes wider, the SVM using specific units rapidly loses its performance capabilities, whereas the SVM using invariant units maintains its high-performance outcomes (**Figure 5D**, right). Interestingly, the SVM using all specific units with different preferred angles outperforms the SVM using specific units only with the same preferred angle. This trend was also observed using other object-selective units (**Figure 5E**, n = 20, two-sided rank-sum test, Invariant vs. Specific with center view, $^{*}P < 0.05$; All specific vs. Specific with center view, $^{**}P < 0.01$; Specific with center view vs. Non-selective, $^{***}P < 10^{-6}$). These results demonstrate that invariance in an untrained network enables an object-detection task with images with various viewpoints for a wide variation range, even without data-augmented learning.

Discussion

We showed that selectivity to various object emerges in randomly initialized networks and that this selectivity is robustly preserved even as the viewpoint changes significantly in the complete absence of learning. Furthermore, we found that the invariant tuning property can arise solely from the distribution of weights in feedforward projections. These results suggest that the statistical complexity of hierarchical neural network circuits allows the initial development of selectivity as well as invariance to various objects across a wide range of transformations.

Our results imply that innate invariance of object selectivity can arise from random feedforward projection, but this does not mean that there is no effect of experience on the development of this function. In fact, observations in various animals support the contention that this invariant function is affected by visual experience. In pigeons, the ability to detect objects across different variations in the viewing conditions is enhanced gradually during the visual training process (Watanabe, 1999). In the monkey, size-invariant object representation is reshaped by unsupervised visual experience (Li and DiCarlo, 2010). Considering the above neurological and behavioral evidence, at an early developmental stage, the innate invariance of object selectivity arises from the structure of neural circuits, and this function can be refined by visual experience during the subsequent developmental process. Specifically, repeated experience with a particular object under various viewpoints will further strengthen the existing selectivity and synaptic weights according to a biologically observed learning rule, such as Hebbian learning. In our model, the invariance of object selectivity can also be more robust and have a widened effective range by visual experience.

Although the current study investigated only viewpoint invariance, we anticipate that invariance to other types of image transformations, such as position, size, and rotation, can also emerge spontaneously in untrained neural networks. Previous studies using an untrained DNN provide supporting evidence. Kim et al. (2021) showed that number selectivity spontaneously emerge in a randomly initialized DNN. Here, number selective units, defined as units that selectivity respond to only numbers of dots in images and respond invariantly to other visual features (e.g., locations, sizes, and convex hulls of dots), contain invariant characteristics of neural tuning. This implies that invariant tuning to various image transformations can arise in untrained neural networks. In addition, Baek et al. (2021) found that face-selectivity can emerge initially and that this tuning shows invariant representation to position, size, rotation, and viewpoint variations to face images. Based on the above results, we expect that various types of invariance of selectivity to objects as well as faces can spontaneously arise in completely untrained neural networks.

We proposed a method of generating invariance without learning, in contrast to previous approaches that implement the same function by relying on a massive training process. In the machine learning field, invariant object recognition has been implemented by learning a great many images. To learn invariant object features, the data-augmentation method is often applied (Simard et al., 2003; O'Gara and McGuinness, 2019; Shorten and Khoshgoftaar, 2019), which generates images with variations through linear transformation, such as positional shifts, rotation, and flipping. However, data augmentation is inefficient in terms of the computational cost. One study that examined changes of the accuracy and training time by data augmentation (O'Gara and McGuinness, 2019) found that twice the learning time is required to slightly improve the accuracy by introducing data augmentation. Thus, our findings can provide clues for addressing the limitations of the data augmentation method to implement invariant functions. By finding selective units in initially randomized networks and applying a training algorithm (Zhuang et al., 2021) toward strengthening innate selectivity and invariance, we expect to

reduce the computational cost of implementing invariant object recognition.

In summary, we conclude that invariance of object selectivity can arise from the statistical variance of randomly wired bottom-up projections in untrained hierarchical neural networks. Our findings may provide new insight into the developmental mechanism of innate cognitive functions in biological and artificial neural networks.

Materials and methods

Untrained AlexNet

Currently, DNN models, which have a biologically inspired hierarchical structure, provide an effective approach for investigating functions in the brain (Paik and Ringach, 2011; DiCarlo et al., 2012; Yamins and DiCarlo, 2016; Sailamul et al., 2017; Baek et al., 2020, 2021; Jang et al., 2020; Kim et al., 2020, 2021; Park et al., 2021; Song et al., 2021). Several studies have reported that DNNs trained to natural images can predict the neural responses of the monkey inferior temporal cortex (IT) (Cadieu et al., 2014; Yamins et al., 2014), known as the area for object recognition. Furthermore, a previous study by the authors found that face-selectivity can arise without experience using a randomly initialized DNN (Baek et al., 2021).

Following earlier work, we used a randomly initialized (untrained) AlexNet (Krizhevsky et al., 2012) consisting of feature extraction and classification layers. AlexNet extracts the features of the input image from five convolutional layers and a pooling layer. It uses a rectified linear unit (ReLU) as an activation function. This activation function allows us to investigate non-linear activity of the type that similarly occurs in the human brain. To randomly initialize the AlexNet, we used standard randomizing method (LeCun et al., 1998). For each filter, each weight was randomly drawn from a Gaussian distribution with a zero mean and the standard deviation set to the square root of the unit number of the previous layer. With this method, we can generate an untrained state of a neural network that balances the strength of the input signal across the layers.

Viewpoint-controllable object stimulus renderer

There are a few well-known objects image datasets, such as ImageNet (Russakovsky et al., 2015), which are often used in DNN studies. However, this image dataset is not sufficient for investigating the effects of viewpoint variance quantitatively. Also, generally used image datasets do not control for low-level features such as luminance, contrast, position, and intra-class image similarity. For this reason, we developed a viewpoint-controllable and low-level feature-controlled object stimulus renderer.

ModelNet10 (Wu et al., 2015), a publicly available 3D object dataset which contains 10 different object classes with aligned orientations, was used to render the stimulus in our study (we used only nine object classes due to an insufficient number of CAD files). Each CAD file is converted to an image at a given horizontal viewpoint using the object render. After capturing the object, the renderer generates a phase-scrambled background image. Using a sample natural image, it scrambles the phase of the given natural image in the Fourier domain and returns it to the original space. These phase-scrambled backgrounds are often used in human fMRI studies to exclude the effects of the background in visual processing (Stigliani et al., 2015). For the object images and phase-scrambled backgrounds, the overall pixel intensity is normalized in each case to have an identical intensity distribution ($Pixel_{mean}$ = 127.5, $Pixel_{std}$ = 51.0). Using this renderer, we generated various viewpoint object stimulus sets in which low-level features were properly calibrated.

Stimulus dataset

We prepared three types of visual stimulus datasets specialized to each task. (1) Object dataset (**Supplementary Figure 1A**): This set was used to find units that selectively respond to a particular object class. It contains nine object classes (bed, chair, desk, dresser, nightstand, monitor, sofa, table, and toilet), and 200 images are prepared in each object class. To render the images of the object dataset, the viewpoint variation angle was randomly set between −30° and +30°. In the object dataset, brightness and contrast of the images are precisely controlled to be equal across object classes (**Supplementary Figures 1B,C**). In addition, the intra-class similarity of the images in each object category was calibrated at a statistically comparable level (**Supplementary Figure 1D**). (2) Viewpoint dataset (**Supplementary Figure 4A**): This set was used to test the viewpoint-invariant characteristics of the object-selective units. This dataset consists of 13 subsets which have different viewpoints from −180° to +180° on a linear scale. It contains 250 different object identities in an object class. Among them, 200 object identities are identical to those used in the object dataset. They were used to analyze the viewpoint-invariant characteristics of the object-selective units quantitatively. The remaining 50 object identities were used not to find object-selective units but to distinguish object-selective units with or without viewpoint invariance. In the viewpoint dataset, the luminance and contrast are also controlled (**Supplementary Figures 4B,C**). (3) SVM dataset: This set was used to train and test the SVM that performs the object-detection task. It contains 60 different object identities in an object class, which were not

used for finding object-selective units. Specifically, it consists of 18 subsets with different viewpoint variations ranging from 0° to 180°. For example, a subset with a 180° viewpoint variation range contains images that show different viewpoints of objects within −90° and +90°.

Analysis of responses of the network units

Using the totally untrained AlexNet, we measured the responses of the target layer for each designed stimulus. For each response from the target convolution layers, each unit of an activation map was separately recorded for different classes of the stimulus. Based on our previous study, object-selective units were defined as units that showed a significantly greater mean response to target object images compared to those of non-target object images ($P < 0.001$, two-sided rank-sum test). To analyze the responses of each unit, it was necessary to regularize the raw response. To normalize the raw response, we used the z-scoring method. Furthermore, we used a trick in the z-score in that we subtracted $\overline{R}_{second\ max}$ from $\overline{R}_{target-object}$. $\overline{R}_{second\ max}$ indicates the response for an object class that leads to the second maximum response for that unit. Therefore, if the z-scored response is higher than zero, our unit shows a higher raw response to the target object than to the second maximum object, indicating selectivity.

$$Response\left(z-scored\right) = \frac{\overline{R}_{target-object} - \overline{R}_{second\ max}}{\sigma_{all\ object}}$$

To quantify the degree of tuning, an object selectivity index (OSI) of a single unit was defined using the follow formula. This index is modified from the face-selective index (FSI), which defined in previous experimental research (Aparicio et al., 2016).

$$Object\ Selectiveity\ Index\ (OSI) = \frac{(\overline{R}_{target-object} - \overline{R}_{non-target-object})}{\sqrt{(\sigma^2_{target-object} + \sigma^2_{non-target-object})/2}}$$

$\overline{R}_{target-object}$ is the average response to target-object images and $\overline{R}_{non-target-object}$ is the average response to all non-target-object images. A higher OSI indicates fine tuning and an OSI of zero indicates equal responses to target and non-target object images.

Among the object-selective units, we defined a viewpoint-invariant unit as a unit for which the response was not significantly different (one-way ANOVA, $P > 0.05$) for all viewpoint classes. Similarly, viewpoint-specific units are defined as a unit for which the response was significantly high for one preferred viewpoint class (one-way ANOVA with single peak filtering, $P < 0.05$). For this, we detected a peak by thresholding the value of the average signal plus the standard deviation, as often done in the field of signal processing.

To measure the invariant index quantitatively, we calculated the inverse of the standard deviation of the average responses for images within each viewpoint class.

$$Viewpoint\ Invariance\ Index = \frac{1}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\overline{R}_{viewpoint\ i} - \mu)^2}}$$

$\overline{R}_{viewpoint\ i}$ is the average response to a viewpoint class and μ is average response for all viewpoint classes. n is total number of viewpoint classes.

Preferred feature image analysis

To achieve the preferred input features of each target unit, we estimated the receptive field of units using the reverse correlation method (Bonin et al., 2011). For this, the initial stimulus set was prepared using 2,500 random local 2D Gaussian filters and the corresponding responses were measured. An initial preferred feature image was achieved from the weighted sum of these responses. In the next iteration, the PFI was re-estimated using a stimulus set consisting of the summation of the previous PFI and the random Gaussian filters. These iterations were repeated 100 times to obtain the final PFI.

Connectivity analysis

To investigate the connectivity between object-selective units across convolutional layers, we backtracked projections of the units from the source layer (Conv4) to the projection layer (Conv5). This backtracking process is opposite of the group convolution process. To backtrack the origin of a unit in the projection layer, we investigated all connected weights and units in the source layers.

To measure the degree of homogeneity in the input projection weight to a single target unit, the homogeneous index was defined as

$$Homogeneous\ Index = \frac{1}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\overline{W}_{specific\ unit\ i} - \mu)^2}},$$

where $\overline{W}_{viewpoint\ i}$ is the average weight from specific units in the source layer to a unit in the projection layer and μ is the average weight from all specific units. n is the total number of viewpoint-specific units with different preferred angles. To compare the unbiased properties of specific and invariant units, we normalized the homogenous index so that the average index value of viewpoint-specific units reaches unity.

Object-detection task

To validate viewpoint-invariant object-selectivity that spontaneously emerges in an untrained DNN, we trained a support vector machine (SVM) using the responses of object-selective units with two types of training. For Train 1, target-object (n = 40) or non-target-object (n = 40) images, which shows different viewpoints of objects within a range of −60° and +60° were randomly presented to the networks, and the observed responses of the Conv5 layer were used to train the SVM. For Train 2, most of the processes are nearly identical compared Train 1, but the only difference is in how the train images are presented. We prepared target-object and non-target object images without viewpoint variation (front-view only). After training the SVM, we investigated the performance with the responses of object-selective units for a stimulus with viewpoint variation. Here, target-object (n = 20) or non-target-object (n = 20) images were also randomly presented to the networks, and the responses from the Conv5 layer was used to test the SVM.

Data availability statement

The stimulus datasets and the MATLAB codes for this study are available at https://github.com/vsnnlab/Invariance.

Author contributions

S-BP conceived of the project. JC, SB, and S-BP designed the model and wrote the manuscript. JC performed the simulations. JC and SB analyzed the data. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by a grant from the National Research Foundation of Korea (NRF) funded by the Korean government (MSIT) (Nos. NRF-2022R1A2C3008991, NRF-2021M3E5D2A01019544, and NRF-2019M3E5D2A01058328), the Singularity Professor Research Project of KAIST, and the KAIST Undergraduate Research Participation (URP) program (to S-BP).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fncom.2022.1030707/full#supplementary-material

References

Aparicio, P. L., Issa, E. B., and DiCarlo, J. J. (2016). Neurophysiological organization of the middle face patch in macaque inferior temporal cortex. *J. Neurosci.* 36, 12729–12745. doi: 10.1523/JNEUROSCI.0237-16.2016

Apurva Ratan Murty, N., and Arun, S. P. (2015). Dynamics of 3D view invariance in monkey inferotemporal cortex. *J. Neurophysiol.* 113, 2180–2194. doi: 10.1152/jn.00810.2014

Baek, S., Park, Y., and Paik, S.-B. (2020). Sparse long-range connections in visual cortex for cost-efficient small-world networks. *bioRxiv* [Preprint]. doi: 10.1101/2020.03.19.998468

Baek, S., Song, M., Jang, J., Kim, G., and Paik, S.-B. (2021). Face detection in untrained deep neural networks. *Nat. Commun.* 12, 1–15. doi: 10.1038/s41467-021-27606-9

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychol. Rev.* 94:115. doi: 10.1037/0033-295X.94.2.115

Bonin, V., Histed, M. H., Yurgenson, S., and Clay Reid, R. (2011). Local diversity and fine-scale organization of receptive fields in mouse visual cortex. *J. Neurosci.* 31, 18506–18521. doi: 10.1523/JNEUROSCI.2974-11.2011

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., et al. (2014). Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS Comput. Biol.* 10:e1003963. doi: 10.1371/journal.pcbi.1003963

Chen, W., Tian, L., Fan, L., and Wang, Y. (2019). "Augmentation invariant training," in *Proceedings of the 2019 IEEE/CVF international conference on computer vision workshop (ICCVW)*, (Seoul: IEEE), 2963–2971. doi: 10.1109/ICCVW.2019.00358

Connor, C. E., Brincat, S. L., and Pasupathy, A. (2007). Transformation of shape information in the ventral pathway. *Curr. Opin. Neurobiol.* 17, 140–147. doi: 10.1016/j.conb.2007.03.002

DiCarlo, J. J., Zoccolan, D., and Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron* 73, 415–434. doi: 10.1016/j.neuron.2012.01.010

Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Comput.* 3, 194–200. doi: 10.1162/neco.1991.3.2.194

Freiwald, W. A., and Tsao, D. Y. (2010). Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science* 330, 845–851. doi: 10.1126/science.1194908

Hung, C. P., Kreiman, G., Poggio, T., and DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science* 310, 863–866. doi: 10.1126/science.1117593

Ichikawa, H., Nakato, E., Igarashi, Y., Okada, M., Kanazawa, S., Yamaguchi, M. K., et al. (2019). A longitudinal study of infant view-invariant face processing during the first 3–8 months of life. *Neuroimage* 186, 817–824. doi: 10.1016/j.neuroimage.2018.11.031

Ito, M., Tamura, H., Fujita, I., and Tanaka, K. (1995). Size and position invariance of neuronal responses in monkey inferotemporal cortex. *J. Neurophysiol.* 73, 218–226. doi: 10.1152/jn.1995.73.1.218

Jang, J., Song, M., and Paik, S. B. (2020). Retino-cortical mapping ratio predicts columnar and salt-and-pepper organization in mammalian visual cortex. *Cell Rep.* 30, 3270.e–3279.e. doi: 10.1016/j.celrep.2020.02.038

Kaufman, L., and Rousseeuw, P. J. (2009). *Finding groups in data: An introduction to cluster analysis*. Hoboken, NJ: John Wiley & Sons.

Kim, G., Jang, J., Baek, S., Song, M., and Paik, S.-B. (2021). Visual number sense in untrained deep neural networks. *Sci. Adv.* 7:eabd6127. doi: 10.1126/sciadv.abd6127

Kim, J., Song, M., Jang, J., and Paik, S. B. (2020). Spontaneous retinal waves can generate long-range horizontal connectivity in visual cortex. *J. Neurosci.* 40, 6584–6599. doi: 10.1523/JNEUROSCI.0649-20.2020

Kobayashi, M., Otsuka, Y., Kanazawa, S., Yamaguchi, M. K., and Kakigi, R. (2012). Size-invariant representation of face in infant brain: An fNIRS-adaptation study. *Neuroreport* 23, 984–988. doi: 10.1097/WNR.0b013e32835a4b86

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* 25, 1097–1105.

LeCun, Y., Bottou, L., Orr, G., and Muller, K.-R. (1998). *Ef?cient backprop. Neural networks tricks trade*. New York, NY: Springer. doi: 10.1007/3-540-49430-8_2

Li, N., Cox, D. D., Zoccolan, D., and DiCarlo, J. J. (2009a). What response properties do individual neurons need to underlie position and clutter "invariant" object recognition? *J. Neurophysiol.* 102, 360–376. doi: 10.1152/jn.90745.2008

Li, N., and DiCarlo, J. J. (2010). Unsupervised natural visual experience rapidly reshapes size-invariant object representation in inferior temporal cortex. *Neuron* 67, 1062–1075. doi: 10.1016/j.neuron.2010.08.029

Li, Y., Pizlo, Z., and Steinman, R. M. (2009b). A computational model that recovers the 3D shape of an object from a single 2D retinal representation. *Vision Res.* 49, 979–991. doi: 10.1016/j.visres.2008.05.013

Logothetis, N. K., Pauls, J., Bülthoff, H. H., and Poggio, T. (1994). View-dependent object recognition by monkeys. *Curr. Biol.* 4, 401–414. doi: 10.1016/S0960-9822(00)00089-0

O'Gara, S., and McGuinness, K. (2019). "Comparing data augmentation strategies for deep image classification," in *Proceedings of the irish machine vision & image processing conference*, (Dublin: Technological University Dublin).

Paik, S. B., and Ringach, D. L. (2011). Retinal origin of orientation maps in visual cortex. *Nat. Neurosci.* 14, 919–925. doi: 10.1038/nn.2824

Park, Y., Baek, S., and Paik, S. B. (2021). A brain-inspired network architecture for cost-efficient object recognition in shallow hierarchical neural networks. *Neural Netw.* 134, 76–85. doi: 10.1016/j.neunet.2020.11.013

Perrett, D. I., Oram, M. W., Harries, M. H., Bevan, R., Hietanen, J. K., Benson, P. J., et al. (1991). Viewer-centred and object-centred coding of heads in the macaque temporal cortex. *Exp. Brain Res.* 86, 159–173. doi: 10.1007/BF00231050

Pinto, N., Cox, D. D., and DiCarlo, J. J. (2008). Why is real-world visual object recognition hard? *PLoS Comput. Biol.* 4:0151–0156. doi: 10.1371/journal.pcbi.0040027

Poggio, T., and Ullman, S. (2013). Vision: Are models of object recognition catching up with the brain? *Ann. N.Y. Acad. Sci.* 1305, 72–82. doi: 10.1111/nyas.12148

Ratan Murty, N. A., and Arun, S. P. (2017). A balanced comparison of object invariances in monkey IT neurons. *eNeuro* 4, 1–10. doi: 10.1523/ENEURO.0333-16.2017

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* 115, 211–252. doi: 10.1007/s11263-015-0816-y

Sailamul, P., Jang, J., and Paik, S. B. (2017). Synaptic convergence regulates synchronization-dependent spike transfer in feedforward neural networks. *J. Comput. Neurosci.* 43, 189–202. doi: 10.1007/s10827-017-0657-5

Shorten, C., and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *J. Big Data* 6, 1–48. doi: 10.1186/s40537-019-0197-0

Simard, P. Y., Steinkraus, D., and Platt, J. C. (2003). "Best practices for convolutional neural networks applied to visual document analysis," in *Proceedings of the international conference on document analysis and recognition, ICDAR 2003-Janua*, (Edinburgh: IEEE), 958–963. doi: 10.1109/ICDAR.2003.1227801

Simonyan, K., and Zisserman, A. (2015). "Very deep convolutional networks for large-scale image recognition," in *Proceedings of the 3rd international conference learning representations ICLR 2015, San Diego - Conf.* 1–14.

Song, M., Jang, J., Kim, G., and Paik, S. B. (2021). Projection of orthogonal tiling from the retina to the visual cortex. *Cell Rep.* 34:108581. doi: 10.1016/j.celrep.2020.108581

Stigliani, A., Weiner, K. S., and Grill-Spector, K. (2015). Temporal processing capacity in high-level visual cortex is domain specific. *J. Neurosci.* 35, 12412–12424. doi: 10.1523/JNEUROSCI.4822-14.2015

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annu. Rev. Neurosci.* 19, 109–139. doi: 10.1146/annurev.ne.19.030196.000545

Turati, C., Bulf, H., and Simion, F. (2008). Newborns' face recognition over changes in viewpoint. *Cognition* 106, 1300–1321. doi: 10.1016/j.cognition.2007.06.005

van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.

Wallis, G., and Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Prog. Neurobiol.* 51, 167–194. doi: 10.1016/S0301-0082(96)00054-8

Watanabe, S. (1999). Enhancement of viewpoint invariance by experience in pigeons. *Cah. Psychol. Cogn.* 18, 321–335.

Wood, J. N. (2013). Newborn chickens generate invariant object representations at the onset of visual object experience. *Proc. Natl. Acad. Sci. U.S.A.* 110, 14000–14005. doi: 10.1073/pnas.1308246110

Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., et al. (2015). "3d shapenets: A deep representation for volumetric shapes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, CVPR2015, Boston, 1912–1920.

Yamins, D. L. K., and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* 19, 356–365. doi: 10.1038/nn.4244

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* 111, 8619–8624. doi: 10.1073/pnas.1403112111

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., et al. (2021). Unsupervised neural network models of the ventral visual stream. *Proc. Natl. Acad. Sci. U.S.A.* 118:e2014196118. doi: 10.1073/pnas.2014196118

Zoccolan, D., Kouh, M., Poggio, T., and DiCarlo, J. J. (2007). Trade-off between object selectivity and tolerance in monkey inferotemporal cortex. *J. Neurosci.* 27, 12292–12307. doi: 10.1523/JNEUROSCI.1897-07.2007