SCIENCE CHINA
Technological Sciences

Special Topic: Brain-Inspired AI Research
•Article•

August 2024 Vol.67 No.8: 2310–2318
https://doi.org/10.1007/s11431-024-2748-3

A comparison of statistical learning of naturalistic textures between DCNNs and the human visual hierarchy

LU XinCheng[1], YUAN ZiQi[3], ZHANG YiChi[3], AI HaiLin[1], CHENG SiYuan[1], GE YiRan[1], FANG Fang[4,5,6,7*] & CHEN NiHong[8,1,2*]

[1] *Department of Psychological and Cognitive Sciences, Tsinghua University, Beijing 100084, China;*
[2] *IDG/McGovern Institute for Brain Research, Tsinghua University, Beijing 100084, China;*
[3] *Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China;*
[4] *School of Psychological and Cognitive Sciences and Beijing Key Laboratory of Behavior and Mental Health, Peking University, Beijing 100871, China;*
[5] *Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing 100871, China;*
[6] *Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China;*
[7] *IDG/McGovern Institute for Brain Research, Peking University, Beijing 100871, China;*
[8] *State Key Laboratory of Brain and Cognitive Science, Institute of Psychology, Chinese Academy of Sciences, Beijing 100101, China*

Received March 5, 2024; accepted July 16, 2024; published online July 30, 2024

The visual system continuously adapts to the statistical properties of the environment. Existing evidence shows a close resemblance between deep convolutional neural networks (CNNs) and primate visual stream in neural selectivity to naturalistic textures above the primary visual processing stage. This study delves into the mechanisms of perceptual learning in CNNs, focusing on how they assimilate the high-order statistics of natural textures. Our results show that a CNN model achieves a similar performance improvement as humans, as manifested in the learning pattern across different types of high-order image statistics. While L2 was the first stage exhibiting texture selectivity, we found that stages beyond L2 were critically involved in learning. The significant contribution of L4 to learning was manifested both in the modulations of texture-selective responses and in the consequences of training with frozen connection weights. Our findings highlight learning-dependent plasticity in the mid-to-high-level areas of the visual hierarchy. This research introduces an AI-inspired approach for studying learning-induced cortical plasticity, utilizing DCNNs as an experimental framework to formulate testable predictions for empirical brain studies.

CNN, perceptual learning, naturalistic texture, psychophysics

Citation: Lu X C, Yuan Z Q, Zhang Y C, et al. A comparison of statistical learning of naturalistic textures between DCNNs and the human visual hierarchy. Sci China Tech Sci, 2024, 67: 2310–2318, https://doi.org/10.1007/s11431-024-2748-3

1 Introduction

The environmental statistics shape our visual system during evolution and ontogeny [1,2]. Even after the critical period, enhancement in the sensory channel can be triggered by prolonged training with basic visual features [3,4], such as spatial frequency [5], orientation [6], and motion direction [7,8]. This type of training, known as perceptual learning, provides an ideal testbed for understanding plasticity in adults' brain.

Advances in deep neural networks (DNNs) have enabled the comparison of hierarchical feature representations between computational layers and primate visual cortical areas [9–12]. DNNs pre-trained on natural images exhibit a cor-

*Corresponding authors (email: ffang@pku.edu.cn; nihongch@tsinghua.edu.cn) respondence between the visual cortex and DNN layers, showing an increasing complexity of visual features along the ventral neural pathway. As a popular example of the CNN model, AlexNet has been shown to represent high-order statistical features of textures [13–15]. Notably, texture selectivity first emerged in Layer 2, consistent with the functional signature of naturalistic texture representation in macaque's and human's V2 [16].

In a recent psychophysical study, we demonstrated that the high-order statistical regularities embedded in the naturalistic textures can be acquired via perceptual learning, indicating a change in the neural populations above V1 [17]. While the neural substrate underlying statistical learning of naturalistic textures has not been empirically investigated, DNN provides a promising testbed to study learning-associated plasticity in the hierarchical visual system [18]. In this study, we tested whether training DNN with naturalistic textures produces comparable behavioral improvements for specific statistical components as in humans. The learning-induced changes in the DNN model yield important predictions to the neural substrate underlying statistical learning of naturalistic textures in the primate visual hierarchy.

2 Materials and methods

2.1 Stimuli

The stimuli were generated in the same way as our previous human psychophysical study [17]. In brief, we generated textures using a synthesis algorithm to capture the high-order statistical features [19]. Prototypes of textures are gray photographs (256 × 256 pixels) from www.textures.com. Each prototype texture is convolved with a bank of filters, which is analogous to the responses of V1 simple and complex cells, tuned to different orientations and spatial frequencies. Low-order statistics refer to the spatially averaged responses of filters selective for different orientations, positions and spatial scales. Then the model computed pairwise products across filter responses at different positions, and across different orientations and scales. The correlations, yielded by averaging all of these pairwise products across the spatial extent of the image, were categorized as three kinds of high-order statistics: (1) Linear which captures spectral features such as periodicity; (2) energy which captures junctions, corners, edges, lines, and contours; and (3) phase which distinguishes lines from edges and also captures gradients in intensity arising from shading.

We used the Portilla-Simoncelli model to synthesize texture samples that had identical statistics based on each naturalistic prototype texture. One texture family consists of metametric texture samples from one particular prototype texture, which was generated with Gaussian white noise and iteratively adjusted until analysis of the synthetic image matches the original texture [16]. By manipulating the strength of the cross-filter correlation, a set of texture samples was generated for each family spanning a naturalness axis with nine values, equally spaced from 0 to 0.88. A naturalness level at 0 was referred to as noise, which only matched the spatially averaged filter responses, but not the spatially averaged correlations between filter responses.

To evaluate the transfer effect of learning, we generated additional families by replacing the linear/energy/phase component of the trained texture family with its counterpart in the untrained texture family (linear-/energy-/phase-).

2.2 Model

An AlexNet-based DNN was used to simulate visual learning of naturalistic textures. We briefly described the network architecture here (see ref. [13] for details). We adopted the first five layers of the original AlexNet, in which each unit is connected locally to a patch of units responding to a local visual field in the upstream layer or the input image. These layers feature a function known as convolution, corresponding to the cross-correlation between the inputs and the filters, enabling parameter sharing and sparsity of connections. A set of non-linear transformations, including ReLU (Rectified Linear) nonlinearity, max pooling, and local response normalization, were followed by each convolution layer. We took the layer outputs after these non-linear transformations, referring to them as L1 to L5 (Figure 1(a)). The fully connected layers in the original AlexNet were discarded to reduce model complexity. The network was tested on a three-alternative-forced choice (3AFC) oddity task for a direct comparison with human psychophysics. In a trial, each of the three images (two textures and one noise, or two noises and one texture) was processed by the same five convolutional layers, each yielding a vector readout in L5. The pairwise Euclidean distance between three vectors was calculated. We defined a probability for identifying the odd one, which is inversely related to the distance between the images from the same category (D_{same}), normalized by the distances across all image pairs according to eq. (1).

$$\frac{1}{p}=\sum_{i=1}^{3}\frac{D_{\text{same}}}{D_i}. \tag{1}$$

2.3 Training procedure

Network weights were initialized using the weights in the five convolutional layers of an AlexNet trained on naturalistic images (http://dl.caffe.berkeleyvisio-n.org/bvlc_alexnet.caffemodel).

We chose honeycomb and grassland as two prototypes of

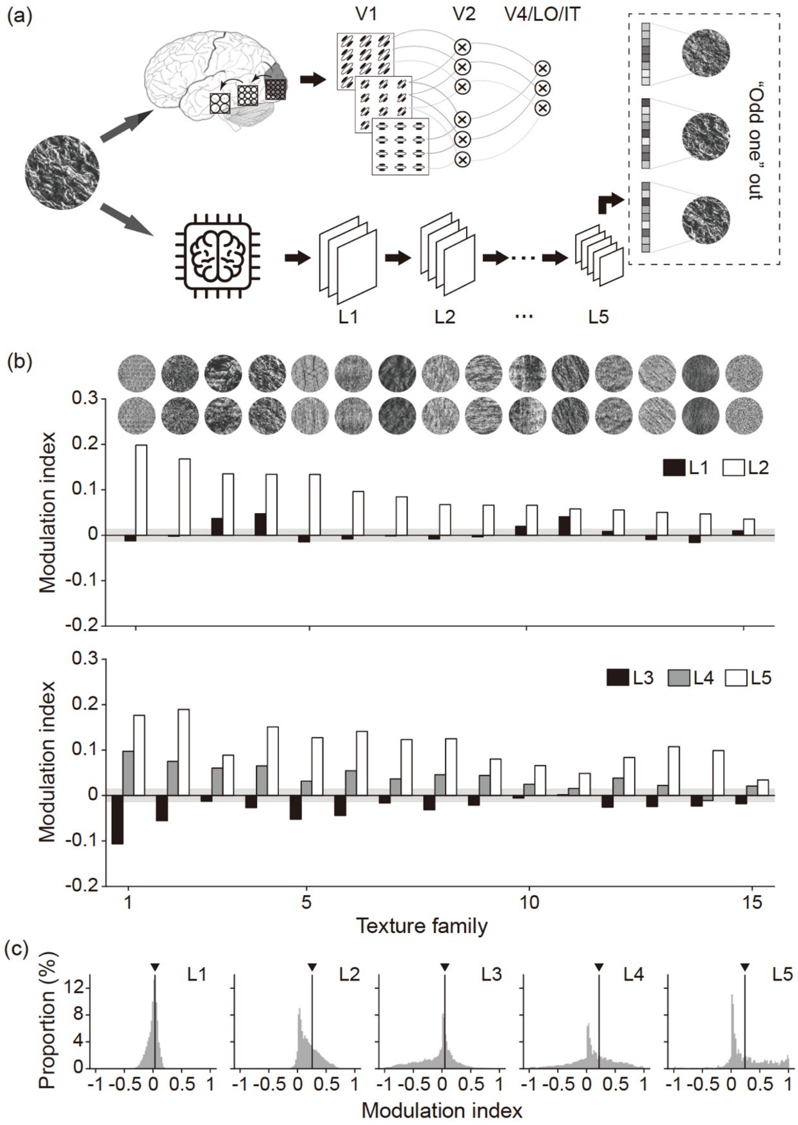


Figure 1 Model architecture (a) and texture modulation index in the pre-trained model ((b), (c)). (a) Schematic illustration of the DNN architecture alongside the primate visual hierarchy for processing naturalistic textures. Top: illustration of the texture analysis model [19] mapped to the primate visual hierarchy. The first stage partitioned images into subbands by convolving with a bank of filters tuned to a range of orientations and spatial frequencies. The linear response and local magnitude response were computed for each local filter, akin to V1 simple and complex cells. In the second stage, the model computed pairwise products across filter responses at different positions and across different orientations and scales for both sets of responses. Similar pairwise product computations are conjectured to be implemented in subsequent visual stages. Bottom: DNN architecture. To perform a 3AFC odd-one-out task, units from the last convolutional layer were read out. The decision was based on pairwise distance computed from the vector representation of each texture/noise image. (b) The modulation index at full naturalness, averaged across model units for various texture families. Corresponding textures and spectrally matched noise are shown at the top, sorted from high to low according to the modulation index in L2. The shaded area shows the expected modulation due to chance (2.5th and 97.5th percentiles of the null distribution). The first two texture families were used for the learning study. (c) Example histograms of modulation indices across units from L1 to L5. The vertical line denotes the criterion of top 30% modulation. Units with zero responses to both texture families in the learning study were excluded.

texture family and trained the network on each of them respectively, for direct comparison with previous human psychophysical results [17]. The network was trained in the 3AFC oddity task on stimulus triplets at all nine naturalness levels, which were randomly selected from 40 texture images and 40 spectrally matched noise images of each naturalness level. During training, a stochastic gradient descent (SGD) algorithm was used to optimize the Triplet Loss function [20], minimizing distance between two images of the same identity and maximizing distance between images of different identities. The learning rate and the momentum were set at 0.00005 and 0.9, respectively. Gradients were obtained by backpropagating the loss through layers [21]. The network was trained to simulate the learning process in human subjects, for 500 iterations of 18-image batches so that one iteration is analogous to a training block for human.

2.4 Behavioral performance and layer-dependent response

Before and after training, we tested the model's performance in discriminating the texture from their corresponding noise in the 3-AFC oddity task. The performance was evaluated for the trained texture family, the untrained family, and the trained family with its linear, energy, or phase component replaced by that of the untrained texture family [17] (Figure S1). Model performance at nine naturalness levels was used to construct the psychometric function. The model threshold was determined as the degree of naturalness at 66.7% accuracy. The learning effect at the performance level was defined as eq. (2).

$$\text{Improvement} = (\text{Threshold}_{\text{pre}} - \text{Threshold}_{\text{post}}) / \text{Threshold}_{\text{pre}} \times 100\%. \quad (2)$$

In addition to Pre and Post-tests, we tracked the model's performance during learning with the trained and the untrained texture family after every hundred iterations, respectively. A total of 30 triplets of texture/noise stimuli were used at each level of naturalness, which were randomly selected from 60 texture images and 60 spectrally matched noise images different from the training set.

As in previous DNN work, layer-dependent response to texture families was evaluated in the pre-trained model [15]. The test stimuli for the pre-trained model include synthesized textures from a set of 15 prototypes of naturalistic texture images according to ref. [19]. For each prototype, we generated 40 textures with a different random Gaussian white noise seed and 40 spectrally matched noise images. As depicted in eq. (3), Texture selectivity of each unit was quantified using the modulation index (MI) following neurophysiological studies [16].

$$MI = \frac{R_{\text{texture}} - R_{\text{noise}}}{R_{\text{texture}} + R_{\text{noise}}}, \quad (3)$$

where R denotes unit response to texture or noise. Similar to the ROI (regions of interest)-based approach in neuroimaging studies, for each layer, we identified units exhibiting the highest 30% modulation index to the trained texture family in the pre-test (Figure 1(c)). The layerwise learning effect was assessed by averaging changes across these units, quantified as eq. (4).

$$\text{Learning Index} = MI_{\text{post}} - MI_{\text{pre}}. \quad (4)$$

Weight change during learning was measured based on the difference from pre-trained values. The change in each layer was calculated at each iteration through eq. (5) [22].

$$d = \frac{\sum_i^N |\delta w_i|}{\sum_i^N |w_i|}, \quad (5)$$

where w_i is the ith element in the N-dimensional weight vector at each layer in the pre-trained model. δw indicates the change in the weight vector after training.

To assess the contribution of each DNN layer to learning, we trained a new (pre-trained) AlexNet model by successively freezing layers from L1 and compared the performance with the fully plastic network. The contribution of each layer was derived by calculating the accuracy drop from the successive addition of the downstream layer (e.g., the contribution of L2 was evaluated by subtracting the accuracy drop of L1 frozen condition from the L1 + L2 frozen condition).

3 Results

We incorporated all five convolutional layers from the AlexNet model (Figure 1(a)) with pre-trained weights for general object recognition. While this model was not tailored for naturalistic texture processing, we identified a selectivity to various texture families (Figure 1(b)). This selectivity was defined as enhanced responses to textures compared to spectrally matched noise (100% naturalness vs. 0% naturalness). Notably, texture selectivity was observed in L2, but not in L1. In addition to this established transformation from the initial to the secondary processing stage, we quantified the texture modulation index in stages beyond L2. We observed an averaged negative index in L3. In contrast, the modulation index turned positive and continued to increase from L4 to L5.

Next, we trained the network using a naturalistic texture discrimination task. We read out responses in the last convolutional layer and computed the distance between image pairs in a 3-AFC oddity task to distinguish texture from noise. The training was repeated for 36 conditions (2 texture families, 9 naturalness levels, 2 triples) with a stochastic gradient descent (SGD) algorithm to minimize the Triplet Loss function. We performed a two-way ANOVA on the model's discrimination threshold with texture family and iteration as two factors. The threshold, which is inversely related to sensitivity, gradually lowered during training (main effect of iterations: $F(5, 708) = 1167.74$, $p < 0.001$), resulting in a 55% and 44% performance improvement for the trained and the untrained family, respectively (Figure 2(a)). A significant interaction between iteration and texture family ($F(5, 708) = 15.28$, $p < 0.001$) was also found, which revealed the threshold of the trained family decreased more quickly than the untrained family. Post-hoc t-test showed that sensitivity to the trained texture saturated towards the end of training (400 iterations vs. 500 iterations for the trained texture: $t(59) = 3.16$, $p = 0.110$, Bonferroni corrected).

Similar to our human psychophysical study [17], we compared thresholds before and after training for five texture

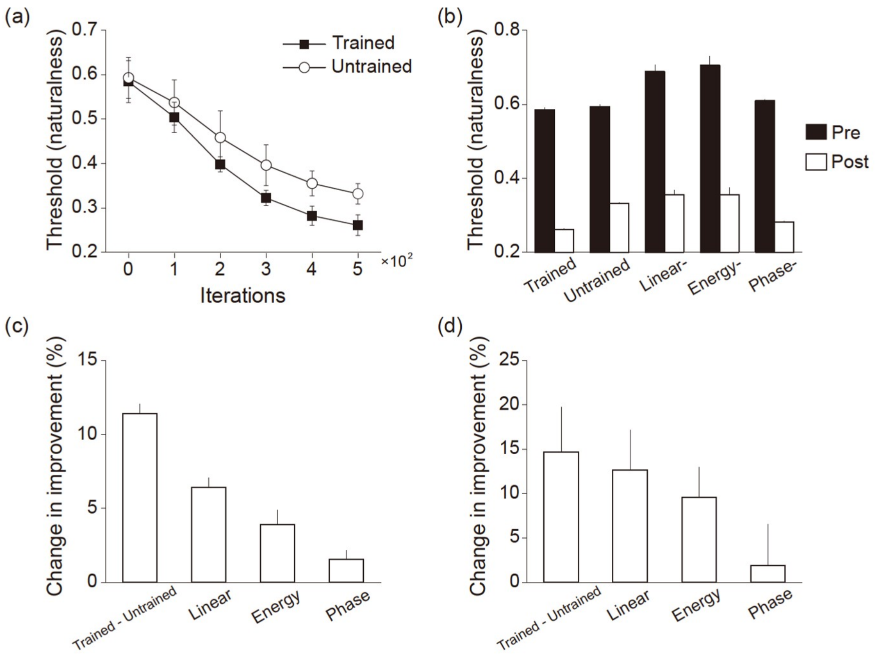
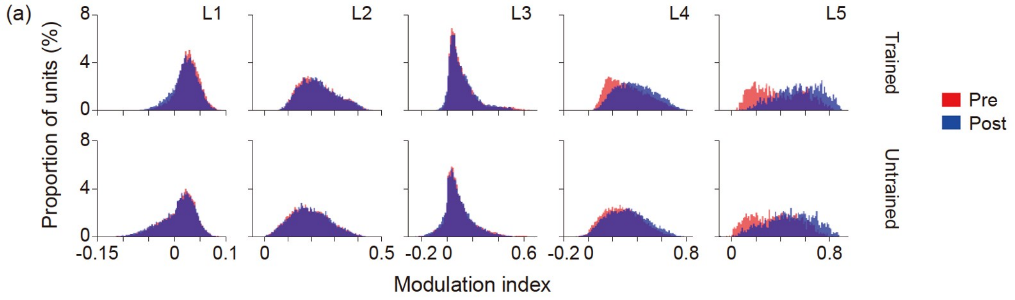


Figure 2 Learning-induced performance improvement in CNN and humans. (a) Learning curve of CNN. Since 1 SEM error bars are hardly visible, error bars indicate 1 SD across trials. (b) Texture discrimination performance before and after training. Error bars indicate 1 SEM across trials. (c) Learning-induced improvements across texture families in CNN. Error bars indicate 1 SEM across trials. (d) Learning-induced improvements across texture families in human psychophysics. Figure 2(c) is adapted from ref. [17]. Error bars indicate 1 SEM across subjects.

families: the trained texture, untrained texture, and the trained texture with its high-order statistics (linear, energy, or phase) substituted by that of the untrained texture family (Figure 2(b)). Training significantly lowered the thresholds for all five texture families (all $t(59) > 15.05$, $p < 0.001$, Bonferroni corrected). To assess the contribution of each statistical component, we quantified the percent improvement for each family using the formula ($\text{Threshold}_{\text{pre}}$ − $\text{Threshold}_{\text{post}}$)/$\text{Threshold}_{\text{pre}}$ × 100%, and then compared it to the improvement of the trained texture. Our findings revealed that substituting the statistics led to a significant decline in the model's improvement (one-sample *t*-test: untrained, linear, energy: all $t(59) > 4.06$, $p < 0.001$, phase: $t(59) = 2.63$, $p < 0.05$, Bonferroni corrected; Figure 2(c)). The drop in improvement, ranked from highest to lowest, was linear (6%), energy (4%), and phase (2%). This pattern mirrored the behavioral patterns observed in human psychophysics, showing the same order of improvement drop among statistics-substituting conditions: linear (13%), energy (10%), and phase (2%) (Figure 2(d)).

To characterize the learning effect across layers, we analyzed the distributions of modulation indices of units that exhibited texture selectivity in the pre-trained model (Figure 3(a)). A learning index was defined by contrasting the averaged modulation index before and after training (Figure 3(b)). An index above/below zero indicates that training increased/decreased the response to the texture. Positive learning indices with the greatest magnitudes were observed in L4 and L5 for both trained and untrained textures. A two-way ANOVA on the learning index showed a significant interaction between training and layers ($F(4, 10790) = 71.41$, $p < 0.001$). Notably, the learning index of the trained texture was higher than that of the untrained texture (post-hoc *t*-test: both $t(1079) > 12.93$, $p < 0.001$, Bonferroni corrected). In addition, a representational similarity analysis (RSA) showed that learning enhanced the representational similarity between the trained and untrained texture families (Figure S2).

To quantify the contributions of different layers to learning, we measured improvement drop when specific layers were frozen (Figure 4). A two-way ANOVA on accumulative improvement showed a main effect of frozen layers ($F(4, 590) = 3951.15$, $p < 0.001$; Figure 4(a)). In general, successively freezing downstream layers led to a step-by-step performance drop above L2 (post-hoc *t*-test: L2 vs L3, L3 vs L4: both $t(119) > 43.08$, $p < 0.001$, Bonferroni corrected; Figure 4(b)). A two-way ANOVA on layerwise improvement drop showed a main effect of frozen layers ($F(4, 590) = 2018.96$, $p < 0.001$; Figure 4(c)). Post-hoc *t*-test revealed significant difference between L3 and L4 ($t(119) = 12.36$, $p < 0.001$, Bonferroni corrected), as well as between L4 and L5 ($t(119) = 39.20$, $p < 0.001$, Bonferroni corrected). The most substantial drop (23%) was identified when freezing L3, i.e., the connection weights between L2 and L3. In addition, freezing L4 led to the second largest drop (19%), followed by freezing L5 (6%). These findings indicate that compensatory



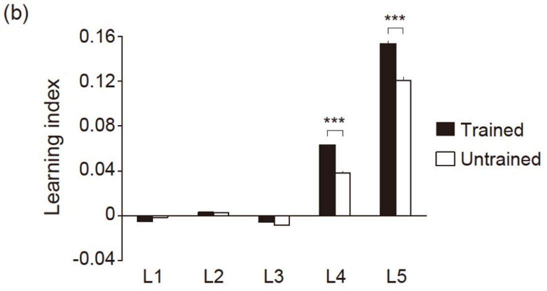


Figure 3 (Color online) Learning-induced changes in modulation indices across CNN layers. (a) Distributions of unit modulation indices for trained and untrained textures before and after training. (b) Learning effect on modulation index. Learning index is defined by contrasting the modulation index before and after training. Error bars indicate 1 SEM across trials.

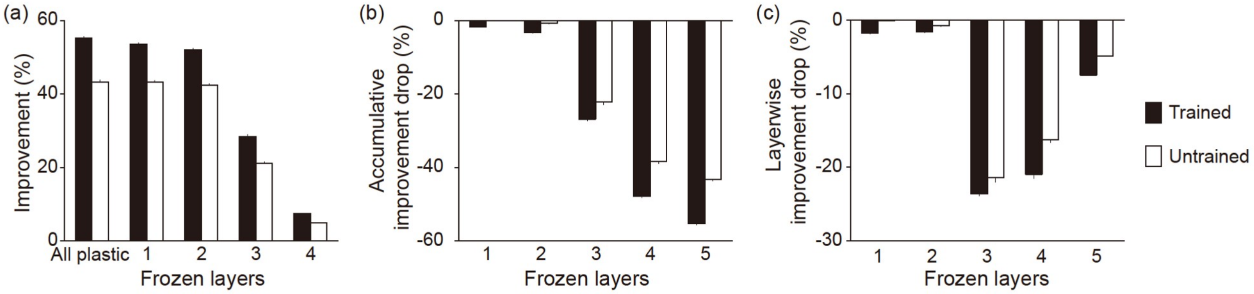


Figure 4 The impact of freezing layers on learning. (a) Learning-induced improvement as successive layers were frozen. (b) Drop in improvement obtained by subtracting the frozen condition from the all-plastic condition. As more layers were frozen, the magnitude of the drop increased. (c) Layerwise contribution in performance, quantified by isolating the drop in each added layer. For instance, the y-value at $x = 2$ in (c) was derived by subtracting the y-value at $x = 1$ from the y-value at $x = 2$ in (b). Error bars indicate 1 SEM across trials.

changes may occur in downstream layers when L1 is frozen. In contrast, higher-level processing beyond the secondary visual stage plays a crucial role in the statistical learning of naturalistic textures, which may be difficult to compensate for with changes in other layers.

To characterize the time course of learning in the weight space across layers, we calculated the weight change relative to the pre-trained weight among 500 iterations in the fully plastic model (Figure 5(a)). Overall, weights changed faster in the first half of training, as the change rate of layers beyond L1 peaked at approximately 200 iterations (Figure 5(b)). Notably, L1 exhibited a sustained increase in weight change throughout learning. Its magnitude of final weight change was ranked second only to that of L3 (Figure 5(c)).

4 Discussion

We found that training DCNN models to differentiate naturalistic textures from noise led to enhanced sensitivity to high-order statistics, a pattern that mirrors learning in humans. Alongside this texture-specific learning effect, we observed an increase in texture selectivity in the units' responses, starting in layers beyond L3. Moreover, training the model with frozen weights in these layers resulted in a significant decline in performance. These findings underscore the plasticity in the mid-to-high levels of the visual hierarchy, induced by the statistical learning of naturalistic textures.

Firstly, we observed that the pre-trained model exhibited selectivity in response to textures beyond its initial proces-

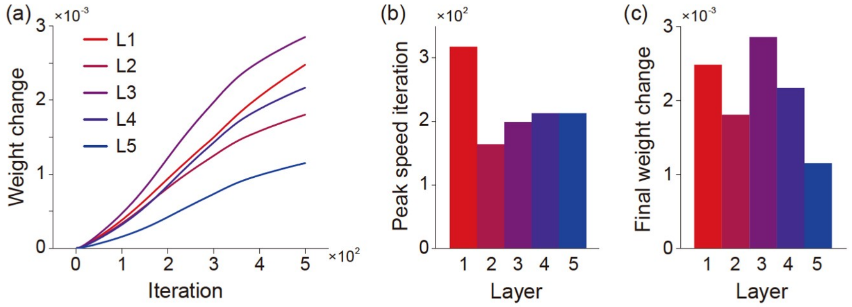


Figure 5 (Color online) Learning in the weight space. (a) Layerwise weight change trajectories during learning; (b) iteration at which the rate of change peaked; (c) final layer change.

sing stage. Our perceptual training was carried out on the CNN model pre-trained for general object recognition, without specific tuning for textures. Nevertheless, the texture selectivity in our CNN was developed in a manner consistent with previous computational and empirical neural evidence. In humans and macaques, high-order statistics embedded in naturalistic textures have been found to modulate neural responses in V2, but not in V1 [16]. Consistently, L2 was the first stage to exhibit texture selectivity. Further downstream, we observed increasing magnitudes of texture selectivity, which aligns with findings from previous CNN work [15,23]. The increment in texture selectivity along the ventral visual hierarchy is supported by fMRI and neurophysiological studies, with V4 exhibiting higher texture selectivity than V2 [24–26]. These results demonstrate that the layer-dependent CNN processing mirrors the human visual processing of naturalistic textures.

Next, we characterized the discrimination performance of the DCNN before and after training with specific texture families. The learning curve follows a pattern similar to that observed in human perceptual learning studies, where repetitive training trials lead to a gradual lowering and eventual saturation of the threshold for the trained texture family. To determine if this enhanced performance was specific to the high-order statistical components embedded in the texture, we substituted different types of statistics in the trained texture. We discovered that replacing these statistical components weakened the learning effect, which mirrors the learning pattern observed in humans, as reported by Cheng et al. [17], suggesting a learning specificity in the statistics of naturalistic textures.

Considering the learning effect at the model performance level, we characterized the improvement before and after training on layerwise response across units. We found that learning enhanced texture selectivity specific to the trained texture in higher stages L4 and L5. These results suggest a multi-stage plasticity induced by naturalistic texture learning. Supporting this hypothesis, our human psychophysical research reported a partial transfer of learning across visual hemifields [17]. By simulating the learning process in DCNN, the present study yields critical predictions about the neural substrates underlying texture perceptual learning.

To further assess the causal contribution of specific layers to learning, we trained the model with step-by-step weight freezing. We observed a decline in performance when freezing the connection weights above L2. Notably, the most substantial decrease in learning performance occurred when the weights in L2-L3 connection were fixed. Although compensatory changes may take place in other layers under freezing conditions, they cannot recover the pivotal role of the downstream visual processing stages beyond L2. Taken together with the learning effect on layerwise texture selectivity, these freezing findings support the idea of multiple high-level stages involved in the statistical learning of naturalistic textures.

It may seem unexpected that the learning-induced enhancement in texture selectivity was not observed in L2, the initial stage known for its preference for naturalistic textures over their spectrally matched noise. This suggests that the specific learning effects on the trained texture family might not necessarily enhance the texture selectivity at this early stage. Previous studies have indicated that in primates' visual cortex, texture-selective processing progresses from V2 to V4 [25,27]. It is possible that V4 develops a specialized computational module, as different image dimensions have been shown to modulate neuronal response in this area [28]. In humans, our recent 7T fMRI work revealed no columnar organization for texture processing in V2. Meanwhile, we found enhanced texture selectivity in V4 and significant feedback connectivity from V4 to V2 [26]. These results highlight the critical role of mid-to-high-level visual areas, above V2, in the experience-dependent acquisition of texture selectivity.

In the present model findings, the learning effect beyond L2 was also reflected in the weight space, as weights in L2-L3, and L3-L4 were substantially enhanced after learning. These findings were in line with the role of V3 [29] and V4 [24–26,28,30] in texture processing in human and primate visual cortex. The learning-induced changes in feedforward connectivity generate important predictions for future studies. In human, layer-dependent fMRI allows for interrogating feedforward and feedback connectivity in the cortex [26,31], thus providing a promising method to testify the observed changes in weight space.

While substantial and irreplaceable changes were identified in the mid-to-high-level processing stages, we also observed changes in the weight space at the early stage. Specifically, L1 exhibited a continual weight change which saturated the latest across layers during learning. In addition, the final weight change in L1 was ranked the second, only to L3 in a fully plastic model. These results suggest that learning-induced rerouting of information may start early, even in areas that do not exhibit a texture-selective representation. However, weight changes at this early stage might not be crucial to learning, as indicated by the negligible changes in model performance when L1 was frozen. This underlines the need for further research into the relationship between weight changes across different layers and their impact on layer-dependent information representation.

Our findings help bridge the gap in our understanding of neural plasticity beyond the secondary visual cortical processing stage, setting a foundation for future electrophysiological and fMRI studies. Future works on CNNs should examine the layerwise representation of various dimensions of statistical information (e.g., coarseness, directionality, and regularity) to establish a closer link between DCNN and primate mid-level visual processing [24,28,30]. Furthermore, the emergence of texture selectivity can serve as a criterion for testifying the biological plausibility in neural networks, whether achieved with a backpropagation algorithm with minimal parameter intervention [22,32], or through shallow neural network model employing self-supervised training to match the layerwise complexity [33]. Finally, our work paves the way for studying dynamic texture perception and material attributes [34,35] in more complex, naturalistic scenes.

In sum, the current study demonstrates human-like learning in the DCNN for the acquisition of high-order statistics embedded in naturalistic textures, highlighting an AI-inspired approach to studying learning-induced cortical plasticity.

This work was supported by the National Natural Science Foundation of China (Grant Nos. 31971031, 31930053, and 32171039), and the STI2030-Major Projects (Grant Nos. 2021ZD0203600, 2022ZD0204802, and 2022ZD0204804).

1 Olshausen B A, Field D J. Natural image statistics and efficient coding. Network-Comput Neural Syst, 1996, 7: 333–339

2 Simoncelli E P, Olshausen B A. Natural image statistics and neural representation. Annu Rev Neurosci, 2001, 24: 1193–1216

3 Fahle M, Poggio T A. Perceptual Learning. Cambridge: MIT Press, 2002

4 Sagi D. Perceptual learning in vision research. Vision Res, 2011, 51: 1552–1566

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

6 Dosher B A, Lu Z L. Perceptual learning reflects external noise filtering and internal noise reduction through channel reweighting. Proc Natl Acad Sci USA, 1998, 95: 13988–13993

7 Ball K, Sekuler R. A specific and enduring improvement in visual motion discrimination. Science, 1982, 218: 697–698

8 Watanabe T, Náñez J E, Sasaki Y. Perceptual learning without perception. Nature, 2001, 413: 844–848

9 Guclu U, van Gerven M A J. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. J Neurosci, 2015, 35: 10005–10014

10 Yamins D L K, DiCarlo J J. Using goal-driven deep learning models to understand sensory cortex. Nat Neurosci, 2016, 19: 356–365

11 McClure P, Kriegeskorte N. Representational distance learning for deep neural networks. Front Comput Neurosci, 2016, 10: 131

12 Horikawa T, Kamitani Y. Generic decoding of seen and imagined objects using hierarchical visual features. Nat Commun, 2017, 8: 15037

13 Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In: Proceedings of the Advances in Neural Information Processing Systems (NIPS). Lake Tahoe, 2012. 1097–1105

14 Zeiler M D, Fergus R. Visualizing and Understanding Convolutional Networks. Cham: Springer International Publishing, 2014. 818–833

15 Laskar M N U, Sanchez Giraldo L G, Schwartz O. Deep neural networks capture texture sensitivity in V2. J Vision, 2020, 20: 21

16 Freeman J, Ziemba C M, Heeger D J, et al. A functional and perceptual signature of the second visual area in primates. Nat Neurosci, 2013, 16: 974–981

17 Cheng S, Ai H, Ge Y, et al. Visual statistical learning of naturalistic textures.. J Exp Psychol-Hum Percept Perform, 2023, 49: 1579–1590

18 Saxe A, Nelli S, Summerfield C. If deep learning is the answer, what is the question? Nat Rev Neurosci, 2021, 22: 55–67

19 Portilla J, Simoncelli E P. A parametric texture model based on joint statistics of complex wavelet coefficients. Int J Comput Vision, 2000, 40: 49–70

20 Schroff F, Kalenichenko D, Philbin J, et al. FaceNet: A unified embedding for face recognition and clustering. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2015. 815–823

21 Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors. Nature, 1986, 323: 533–536

22 Wenliang L K, Seitz A R. Deep neural networks for modeling visual perceptual learning. J Neurosci, 2018, 38: 6028–6044

23 Zhuang C, Wang Y, Yamins D, et al. Deep learning predicts correlation between a functional signature of higher visual areas and sparse firing of neurons. Front Comput Neurosci, 2017, 11: 100

24 Okazawa G, Tajima S, Komatsu H. Image statistics underlying natural texture selectivity of neurons in macaque V4. Proc Natl Acad Sci USA, 2015, 112: E351–360

25 Okazawa G, Tajima S, Komatsu H. Gradual development of visual texture-selective properties between macaque areas V2 and V4. Cereb Cortex, 2017, 27: 4867

26 Ai H, Lin W, Liu C, et al. Mesoscale functional organization and connectivity of color, disparity, and naturalistic texture in human second visual area. Elife, 2024, 13, doi: 10.7554/eLife.93171.1

27 Ziemba C M, Freeman J, Simoncelli E P, et al. Contextual modulation of sensitivity to naturalistic image structure in macaque V2. J Neu-

rophysiol, 2018, 120: 409–420

28 Kim T, Bair W, Pasupathy A. Perceptual texture dimensions modulate neuronal response dynamics in visual cortical area V4. J Neurosci, 2022, 42: 631–642

29 Kohler P J, Clarke A, Yakovleva A, et al. Representation of maximally regular textures in human visual cortex. J Neurosci, 2016, 36: 714–729

30 Hatanaka G, Inagaki M, Takeuchi R F, et al. Processing of visual statistics of naturalistic videos in macaque visual areas V1 and V4. Brain Struct Funct, 2022, 227: 1385–1403

31 Jia K, Zamboni E, Kemper V, et al. Recurrent processing drives perceptual plasticity. Curr Biol, 2020, 30: 4177–4187.e4

32 Cheng Y A, Sanayei M, Chen X, et al. Noise reduction as a unified mechanism of perceptual learning in humans, macaques, and convolutional neural networks. bioRxiv, 2023, doi:10.1101/2023.11.13.566963

33 Parthasarathy N, Simoncelli E P. Self-supervised learning of a biologically-inspired visual texture model.. arXiv: 2006.16976

34 Morgenstern Y, Kersten D J. The perceptual dimensions of natural dynamic flow. J Vision, 2017, 17: 7

35 Bi W, Jin P, Nienborg H, et al. Estimating mechanical properties of cloth from videos using dense motion trajectories: Human psychophysics and machine learning. J Vision, 2018, 18: 12