Neural Networks 166 (2023) 692–703

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Visual information processing through the interplay between fine and coarse signal pathways

Xiaolong Zou [a,b,c,1], Zilong Ji [a,d,1], Tianqiu Zhang [a], Tiejun Huang [c,e], Si Wu [a,b,c,*]

[a] School of Psychological and Cognitive Sciences, IDG/McGovern Institute for Brain Research, Peking-Tsinghua Center for Life Sciences, Center of Quantitative Biology, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China
[b] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei, China
[c] Beijing Academy of Artificial Intelligence, Beijing, China
[d] Institue of Cognitive Neuroscience, University College London, London, UK
[e] School of Computer Science, Peking University, Beijing, China

ARTICLE INFO

Article history:
Received 23 November 2022
Received in revised form 19 July 2023
Accepted 30 July 2023
Available online 3 August 2023

Keywords:
Visual information processing
Two-pathway model
Convolution neural network
Imitation learning
Backward masking

ABSTRACT

Object recognition is often viewed as a feedforward, bottom-up process in machine learning, but in real neural systems, object recognition is a complicated process which involves the interplay between two signal pathways. One is the parvocellular pathway (P-pathway), which is slow and extracts fine features of objects; the other is the magnocellular pathway (M-pathway), which is fast and extracts coarse features of objects. It has been suggested that the interplay between the two pathways endows the neural system with the capacity of processing visual information rapidly, adaptively, and robustly. However, the underlying computational mechanism remains largely unknown. In this study, we build a two-pathway model to elucidate the computational properties associated with the interactions between two visual pathways. Specifically, we model two visual pathways using two convolution neural networks: one mimics the P-pathway, referred to as FineNet, which is deep, has small-size kernels, and receives detailed visual inputs; the other mimics the M-pathway, referred to as CoarseNet, which is shallow, has large-size kernels, and receives blurred visual inputs. We show that CoarseNet can learn from FineNet through imitation to improve its performance, FineNet can benefit from the feedback of CoarseNet to improve its robustness to noise; and the two pathways interact with each other to achieve rough-to-fine information processing. Using visual backward masking as an example, we further demonstrate that our model can explain visual cognitive behaviors that involve the interplay between two pathways. We hope that this study gives us insight into understanding the interaction principles between two visual pathways.

1. Introduction

Imagine you are driving a car on a highway and suddenly an object appears in your visual field, crossing the road. Your initial reaction is to slam on the brakes even before recognizing the object. This highlights a core difference between human vision and current machine learning strategies for object recognition. In machine learning, visual object recognition is often viewed as a feedforward, bottom up process, where object features are extracted from local to global in a hierarchical manner; whereas in human vision, we can capture the gist of a visual object at a glance without processing the details of it, a crucial ability for us (especially animals) to survive in competitive natural environments. This strategic difference has been demonstrated by a large volume of experimental data. For examples, Sugase, Yamane, Ueno, and Kawano (1999) found that neurons in the inferior temporal cortex (IT) of macaque monkeys convey the coarse information of an object much faster than the fine information of it; FMRI and MEG studies on humans showed that the activation of orbitofrontal cortex (OFC) precedes that of the temporal cortex when a blurred object was shown to the subject (Bar et al., 2006); Liu, Wang, Zhou, Ding, and Luo (2017) further demonstrated that the dorsal pathway extracts the coarse information of an object in less than 100 ms after the stimulus onset, and this coarse information guides the subsequent local information processing.

Indeed, the Reverse Hierarchy Theory for visual perception has proposed that although the representation of image features

[*] Corresponding author at: School of Psychological and Cognitive Sciences, IDG/McGovern Institute for Brain Research, Peking-Tsinghua Center for Life Sciences, Center of Quantitative Biology, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China.

E-mail addresses: xiaolz@pku.edu.cn (X. Zou), zilong.ji@ucl.ac.uk (Z. Ji), tianqiuakita@stu.pku.edu.cn (T. Zhang), tjhuang@pku.edu.cn (T. Huang), siwu@pku.edu.cn (S. Wu).

[1] These authors contribute equally to this work.

https://doi.org/10.1016/j.neunet.2023.07.048

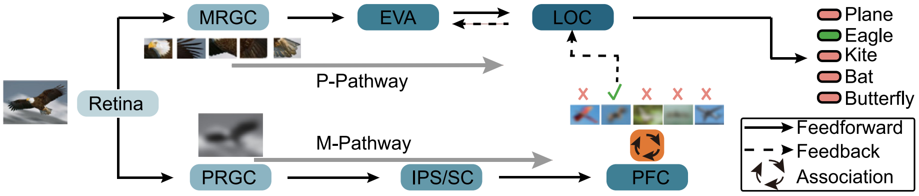

Fig. 1. Illustration of the two separated pathways for information processing in the visual system. An image of an eagle is processed through two pathways. Upper panel: the P-pathway processes the detailed information of the image. Lower panel: the M-pathway processes the coarse information of the image rapidly, generates predictions about the image (association), and modulates the information processing of the P-pathway (feedback). MRGC: midget retina ganglion cell. PRGC: parasol retina ganglion cells. EVA: early visual area. LOC: lateral occipital complex. IPS: intraparietal sulcus. SC: superior colliculus. PFC: prefrontal cortex.

along the ventral pathway goes from local to global, our perception of an object goes inversely from global to local (Hochstein & Ahissar, 2002). How does this happen in the brain? Experimental studies have revealed that there exist two anatomically and functionally separated signal pathways for visual information processing (see Fig. 1). One is called the parvocellular pathway (P-pathway), which starts from midget retina ganglion cells (MRGCs), projects to layers 3–6 in the lateral geniculate nucleus (LGN), and then primarily goes downstream along the ventral stream. The other is called the magnocellular pathway (M-pathway), which starts from parasol retina ganglion cells (PRGCs), projects to layers 1–2 of LGN, and then goes along the dorsal stream or the subcortical pathway (the superior colliculus and downstream areas). The two pathways have different neural response characteristics and complementary computational roles. Experimental findings have shown that the P-pathway is sensitive to colors and responds primarily to visual inputs of high spatial frequency; whereas the M-pathway is color blind and responds primarily to visual inputs of low spatial frequency (Derrington & Lennie, 1984). It has been suggested that the M-pathway serves as a short-cut to extract coarse information of images rapidly, while the P-pathway extracts fine features of images slowly, and the interplay between two pathways endows the neural system with the capacity of processing visual information rapidly, adaptively, and robustly (Bar, 2003; Bullier, 2001; Liu et al., 2017; Wang, Zhou, Zhuo, Chen, & Huang, 2020). For instance, by extracting the coarse information of an image, the M-pathway can generate predictions about what are expected in the visual field, and this knowledge subsequently modulate the fine information processing in the P-pathway (see Fig. 1).

Although the existence of separated P- and M- pathways is well known in the neuroscience field, exactly how they cooperate with each other to facilitate information processing remains poorly understood. The main difficulty comes from that to date, we still do not have much knowledge about the detailed structures of two pathways and the details of their interaction process, which prevent us from building a detailed biological model to elucidate the associated neural mechanisms. Recently, computational studies have demonstrated that deep neural networks can be useful models to describe visual information processing, e.g., it was shown that convolution neural networks (CNNs) can effectively mimic the neuronal response variability along the visual pathway (Kriegeskorte, 2015; Yamins, Hong, Cadieu, & DiCarlo, 2013). Inspired by these studies, in this work, we build up a two-pathway model using CNNs as building blocks to elucidate the computational properties of the interplay between two visual pathways (see Fig. 2). Specifically, we model the P-pathway using a relatively deep CNN, which has small-size kernels and receives detailed visual inputs, referred to as FineNet hereafter; and we model the M-pathway using a relatively shallow CNN, which has large-size kernels and receives blurred visual inputs, referred to as CoarseNet hereafter. Based on the proposed model, we investigate several computational issues associated with the interplay between two pathways, including how CoarseNet learns from FineNet via imitation, how FineNet benefits from CoarseNet via feedback to leverage its performance, and how they interact with each other to achieve rough-to-fine information processing. We also use the two-pathway model to reproduce the backward masking phenomenon observed in human psychophysic experiments. We hope that this modeling study, although it is still quite preliminary and misses many biological details, can give us some insight into understanding the interaction principles between two visual pathways.

2. The two-pathway model

The structure of our two-pathway model is illustrated in Fig. 2, where FineNet and CoarseNet mimic the P- and M- pathways, respectively. Notably, FineNet is deeper than CoarseNet, reflecting that the P-pathway goes through more feature analyzing relays (e.g., V1-V2-V4-IT along the ventral pathway) than the M-pathway. FineNet also has smaller convolutional kernels than CoarseNet, reflecting that MRGCs in the retina have much smaller receptive fields than PRGCs. Furthermore, we consider that FineNet receives detailed and colorful visual inputs, reflecting that MRGCs have small receptive fields and are color sensitive; while CoarseNet receives blurred and gray inputs, reflecting that PRGCs have large receptive fields and are color blind.

In the model, we consider that the two pathways interact with each other in three forms: (1) **Imitation learning**. Since CoarseNet has a shallow structure and receives blurred inputs, it is hard to train CoarseNet well for object recognition directly. Hence we consider that CoarseNet learns the feature representations of FineNet via an imitation process. Later we will argue that this has an important biological implication (see Section 3.1). (2) **Association**. It is supposed that the M-pathway generates predictions about what might be in the visual scene, which guides the information processing in the P-pathway. We model this by considering that CoarseNet predicts the representation of FineNet through a memory association process. (3) **Feedback**. It is known that coarse information can serve as a cognitive bias guiding the extraction of fine information of images. We model this by feeding the associated prediction back to an earlier layer of FineNet to enhance the fine feature extraction. The details of the two-pathway model are introduced below.

2.1. The inference process of the model

Denote the input to FineNet as $\boldsymbol{x}$ and the input to CoarseNet as $\hat{\boldsymbol{x}}$. $\hat{\boldsymbol{x}}$ is obtained by either filtering $\boldsymbol{x}$ with a 2D Gaussian filter or binarizing $\boldsymbol{x}$. Denote the output of CoarseNet to be $p^C(\hat{\boldsymbol{x}}) = \boldsymbol{f}^C\left[\mathbf{g}^C\left(\hat{\boldsymbol{x}}; \boldsymbol{\theta}^C\right); \boldsymbol{w}^C\right]$, where $\mathbf{g}^C(\cdot; \boldsymbol{\theta}^C)$ and $\boldsymbol{f}^C(\cdot; \boldsymbol{w}^C)$ represent, respectively, the feature extractor and the linear classifier of CoarseNet, and $\{\boldsymbol{\theta}^C, \boldsymbol{w}^C\}$ the trainable parameters. The output of FineNet is similarly denoted as $\boldsymbol{p}^F(\boldsymbol{x}) = \boldsymbol{f}^F\left\{\mathbf{g}^F\left[\boldsymbol{x}, \boldsymbol{O}(\hat{\boldsymbol{x}}, \boldsymbol{x}); \boldsymbol{\theta}^F\right]; \boldsymbol{w}^F\right\}$, where the feature extractor $\mathbf{g}^F(\cdot; \boldsymbol{\theta}^F)$ has an extra input component $\boldsymbol{O}(\hat{\boldsymbol{x}}; \boldsymbol{x})$, representing the feedback signal.

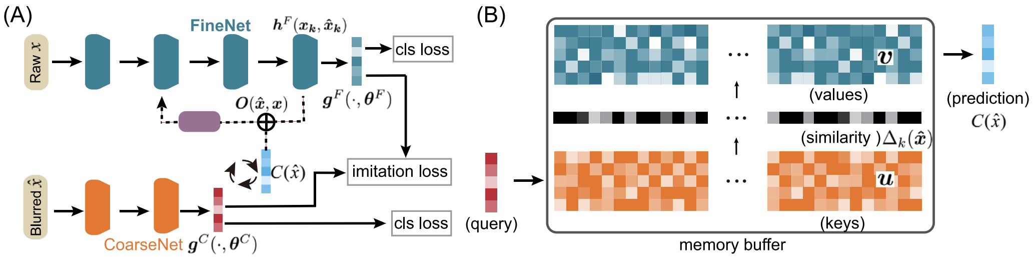

Fig. 2. Illustration of the two-pathway model. (A) The architecture of the model. The blue and orange blocks represent the feedforward convolutional layers in FineNet and CoarseNet, respectively. The purple one represents the feedback convolution block in FineNet. In inference, CoarseNet extracts coarse features $\boldsymbol{g}^C(\hat{\boldsymbol{x}})$ from a blurred image $\hat{\boldsymbol{x}}$, which serves as a cue to predict the fine features $\boldsymbol{C}(\hat{\boldsymbol{x}})$ of the image via association. The associated result is then combined with the deep representations $\boldsymbol{h}^F(\boldsymbol{x}, \hat{\boldsymbol{x}})$ to form a feedback signal $\boldsymbol{O}(\boldsymbol{x}, \hat{\boldsymbol{x}})$, and the latter modulates an early layer of FineNet. In training, FineNet is optimized by minimizing the classification loss, and CoarseNet by minimizing both classification and imitation losses. (B) Illustration of static memory association (SMA). A query of the coarse features $\boldsymbol{g}^C(\hat{\boldsymbol{x}})$ of the input $\hat{\boldsymbol{x}}$ is associated with a weighted summation of the fine features stored in the memory buffer, where the weighting coefficient $\Delta_k(\hat{\boldsymbol{x}})$ is the similarity between the coarse features and the key vector $\boldsymbol{u}_k$.

To generate the feedback signal $\boldsymbol{O}(\hat{\boldsymbol{x}}, \boldsymbol{x})$ in FineNet, we consider a memory association process. Two types of associations are exploited in this work, static memory association (SMA) and dynamic memory association (DMA) . They have the similar effect of using coarse features $\boldsymbol{g}^C(\cdot; \boldsymbol{\theta}^C)$ as a cue to predict fine features. SMA is simpler, but we also consider DMA, as it introduces temporal dynamics into our two-pathway model necessary for reproducing the backward masking experiment (see Section 4). For clearance, we only introduce SMA here (see Fig. 2B), and DMA is described in Appendix E. Specifically, we implement SMA with the cache memory model (Orhan, 2018), which performs a key–value association. The model stores a pair of a key matrix $\boldsymbol{u} \in R^{d\times K}$ and a value matrix $\boldsymbol{v} \in R^{c\times K}$ in the memory buffer, with K the number of memory items and d, c the dimensions of the key and value vectors, respectively. The columns $\boldsymbol{u}_k$ and $\boldsymbol{v}_k$ represent, respectively, the normalized $\boldsymbol{g}^C(\hat{\boldsymbol{x}}_k; \boldsymbol{\theta}^C)$ of CoarseNet and the flattened feature vector $\boldsymbol{h}^F(\boldsymbol{x_k}, \hat{\boldsymbol{x}}_k)$ of the last convolution layer of FineNet. When a specific query vector $\boldsymbol{g}^C(\hat{\boldsymbol{x}})$ of CoarseNet is presented, we first calculate its similarities with all key vectors stored in the memory buffer, which are given by $\Delta_k(\hat{\boldsymbol{x}}) = \exp\left[\beta \boldsymbol{g}^C(\hat{\boldsymbol{x}})^\top \boldsymbol{u}_k\right]$, for $k = 1, \dots K$, with β controlling the sharpness of similarity. After that, we calculate the associated result, i.e., the predicted fine features, which is given by $\boldsymbol{C}(\hat{\boldsymbol{x}}) = \sum_k \boldsymbol{v}_k \Delta_k(\hat{\boldsymbol{x}}) / \left[\sum_k \Delta_k(\hat{\boldsymbol{x}})\right]$. The inference of FineNet forms a continuous loop so that the feedback signal is updated iteratively (see Fig. 2A). At time step t, the feedback signal in FineNet is calculated by $\boldsymbol{O}_t(\hat{\boldsymbol{x}}, \boldsymbol{x}) = \boldsymbol{C}(\hat{\boldsymbol{x}}) + \boldsymbol{h}^F_{t-1}(\boldsymbol{x}, \hat{\boldsymbol{x}})$. Notably, at the first step $t = 1$, only the associated result from CoarseNet is available, which gives $\boldsymbol{O}_1(\hat{\boldsymbol{x}}, \boldsymbol{x}) = \boldsymbol{C}(\hat{\boldsymbol{x}})$. This reflects the fact that the M-pathway is much faster than the P-pathway, which generates the first feedback signal without interacting with high visual areas in the P-pathway.

In summary, the inference of the model involves interaction between two pathways: in response to an image, CoarseNet first generates its output and meanwhile predicts the fine features of FineNet through association; the predicted result is then combined with the deep representations of FineNet to form a feedback signal, which modulates the shallow layer of FineNet for feature extraction; this feedback loop can go on iteratively to continuously leverage the performance of FineNet.

2.2. The training of the model

During training, FineNet and CoarseNet are optimized jointly. To get the network output for an input, we run the feedback loop iteratively in FineNet for T steps. FineNet is optimized through minimizing the cross-entropy loss, which is given by

$$L_F = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} y_{i,j} \ln \boldsymbol{p}^F_j(\boldsymbol{x}_i), \tag{1}$$

where $\boldsymbol{p}^F_j$ is the jth element of $\boldsymbol{p}^F$, i.e., the likelihood of the jth class, and $y_{i,j}$ is the jth element of the one-hot label $\boldsymbol{y}_i$ for the image $\boldsymbol{x}_i$, which is 1 for the correct class and 0 otherwise. The summation runs over all images N and all classes K.

Since CoarseNet receives coarse inputs and has a shallow structure, we optimize it via a combination of classification and imitation losses, which is written as

$$L_C = \frac{1}{N} \times \sum_{i=1}^{N}\left[-\alpha \sum_{j=1}^{K} y_{i,j} \ln p^C_j(\hat{\boldsymbol{x}}_i) + \frac{1-\alpha}{2}\|\boldsymbol{g}^C(\hat{\boldsymbol{x}}_i) - \boldsymbol{g}^F(\boldsymbol{x}_i, \hat{\boldsymbol{x}}_i)\|^2\right], \tag{2}$$

where the symbol $\|\cdot\|$ denotes L_2 normal, and α is a hyperparameter balancing the cross-entropy loss and the imitation loss.

Since SMA aims to store the long-term correlation (association) between the feature representations of CoarseNet and FineNet, we update its key and value matrices after every N ($N = 2$ is used in this study) training epochs .

3. Interplay between two pathways

In our two-pathway model, FineNet and CoarseNet interact with each other in three forms, including imitation learning, association and feedback. We explore how these three interactions affect the computational properties of the two-pathway model.

3.1. Imitation learning from FineNet improves CoarseNet's performance

In the two-pathway model, CoarseNet is supposed to generate a good initial guess of the image, which further serves as a cognitive bias to facilitate the performance of FineNet. However, since CoarseNet is shallow, has large convolution kernels, and receives coarse inputs, it is hard to train CoarseNet well independently. Therefore, we consider that CoarseNet learns the feature representations of FineNet via imitation learning. This is an interesting issue and may have some far-reaching implications to brain functions (see discussions below). We therefore carry out a separate experiment to study the imitation learning effect. Specifically, we focus on exploring how CoarseNet learns from FineNet via imitation, without considering other interactions between two pathways. To evaluate the model performance, we use Pascalvoc-mask and CIFAR-10 datasets (see Appendix A.1). To generate blurred inputs $\hat{\boldsymbol{x}}$ to CoarseNet, we either adopt low-pass filter $\boldsymbol{x}$ using a 2D Gaussian filter with $std = 2$ or binarizing $\boldsymbol{x}$ using a shape mask (see examples in Fig. 3A), mimicking the

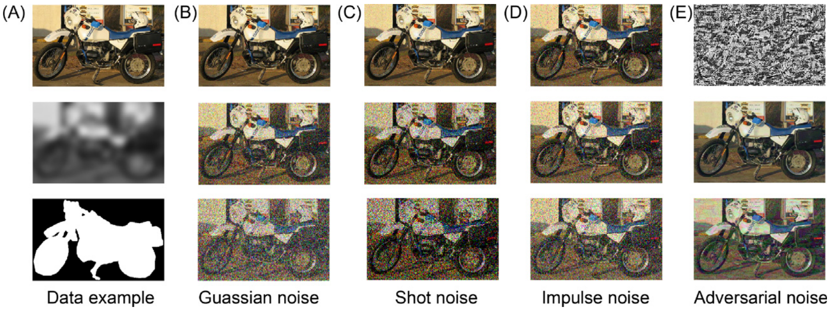

Fig. 3. Examples of visual inputs used in the experiments. (A) Examples of visual inputs used for training FineNet and CoarseNet. From up to down, a raw image to FineNet, the corresponding low-pass filtered image (blurred) to CoarseNet, and the corresponding binarized image (mask data) to CoarseNet. (B-E) Different kinds of noise disrupted inputs. (B) Examples of Gaussian noise with $std = 0.04, 0.3, 0.6$, respectively. (C) Examples of shot noise with $c = 100, 3, 1$, respectively. (D) Examples of impulse noise with $p = 0.07, 0.15, 0.3$, respectively. (E) Adversarial noise. Up: the adversarial noise of the example image in (A)-up, obtained by the Fast Gradient Sign Method (Goodfellow, Shlens, & Szegedy, 2014); Middle and down: the adversarial examples with the noise levels of 0.1 and 0.5, respectively.

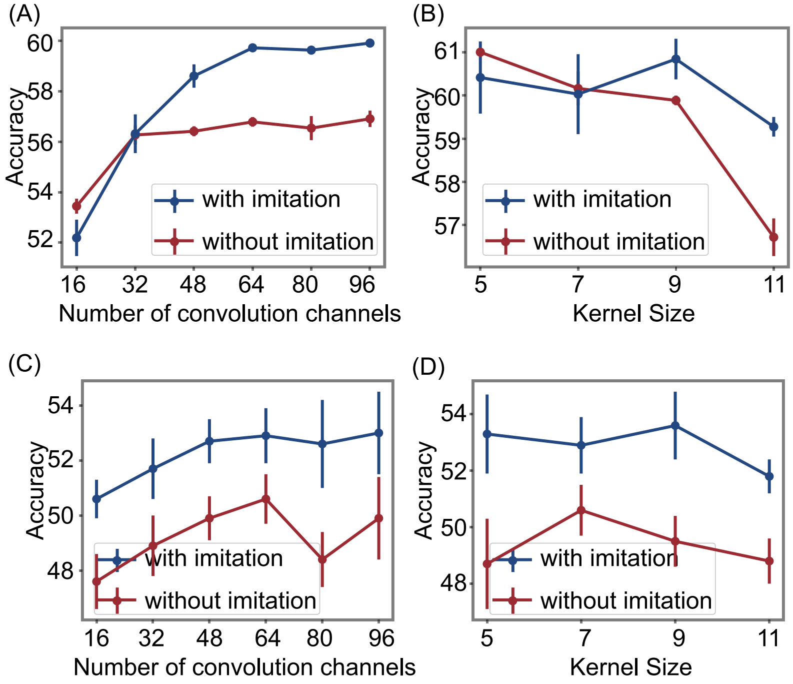

Fig. 4. Imitation learning from FineNet improves the performance of CoarseNet. (A-B): performances of CoarseNet trained on low-pass filtered images from CIFAR-10. (A) Performances vs. the number of convolution channels. (B) Performances vs. the size of convolution kernel. (C-D): performances of CoarseNet trained on the binarized Pascalvoc-mask. (C) Performance vs. the number of convolution channels. (D) Performance vs. the size of convolution kernel. The number of convolution channels and the size of convolutional kernel refer to that in the first layer in CoarseNet. Mean and std are obtained by averaging over 5 trials with random network initialization. See Appendix A for the details of the training and testing data.

coarse input to M-pathway. The detailed implementations of the model are presented in Appendix C.

Fig. 4 present the results, which demonstrate that over a wide range of parameters, the classification accuracy of CoarseNet with imitation learning is improved considerably compared to that without imitation learning. Specifically, with respect to the number of convolution kernels in CoarseNet, the improvement is significant when the number of kernels is large (Fig. 4A for low-pass filtered inputs; Fig. 4C for binarized inputs); with respect to the size of kernels in CoarseNet, the improvement is also significant (Fig. 4B for low-pass filtered inputs; Fig. 4D for binarized inputs). The fact that the effect of imitation learning also depends on the network parameters (see Fig. 4A) indicates that in reality there is a trade-off between the simplicity of the M-pathway structure and the effect of the M-pathway learning from the P-pathway.

3.1.1. Biological implications of imitation learning

From the computational point of view, the brain faces a difficulty of "designing" properly the M-pathway. On one hand, the M-pathway needs to be shallow and process coarse visual inputs in order to generate quick responses (which is important in a dangerous environment); on the other hand, the M-pathway needs to efficiently generate approximated, if not accurate, recognition of an object, serving as a good initial guess for further processing. However, it is a well-known fact that a shallow neural network alone is unable to achieve good object recognition (this has actually motivated the development of deep neural networks). So, how does the brain resolve this dilemma? Here, our study suggests that the strategy of imitation learning proposed in machine learning (Hinton, Vinyals, & Dean, 2015) may provide a solution to this challenge, that is, the shallow M-pathway learns the representations of the deep P-pathway through imitation

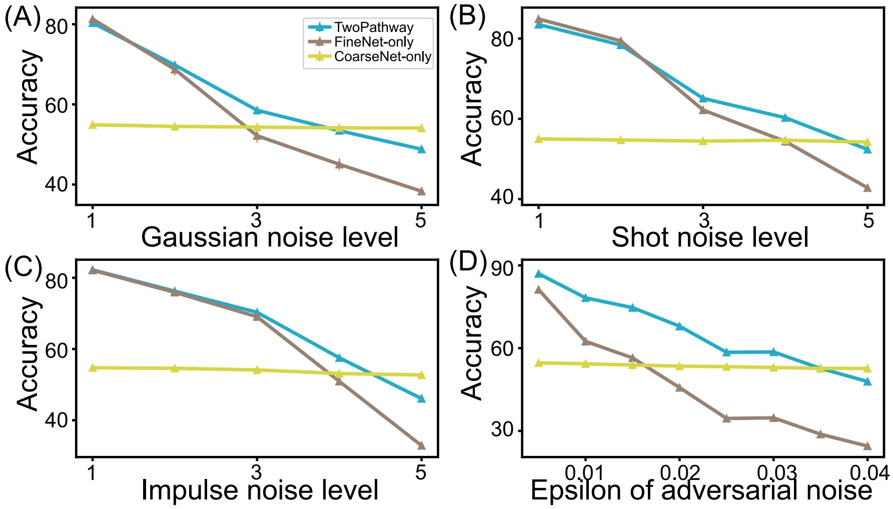

Fig. 5. Model performances against noises and adversarial noise perturbations. The loop of feedback interactions is 3 for TwoPathway model. CoarseNet in these models takes grayed and low-pass filtered inputs with $std = 2$. Adversarial noises are generated by attacking FineNet using the Fast Gradient Sign Method (Goodfellow et al., 2014). In (A-C), model performances against noise perturbations. FineNet and CoarseNet are trained independently on the dataset. (D) Model performances against adversarial noise perturbations. Mean and std are obtained by averaging over 4 trials with random network initialization and seed.

to improve its performance. Imitation learning may also be involved in other brain functions, such as for knowledge transfer and memory consolidation between hippocampus and neocortex (Alvarez & Squire, 1994; Sirota, Csicsvari, Buhl, & Buzsáki, 2003). During the acquisition of motor skills, it has been observed that neural activities gradually shift from the prefrontal cortex to the premotor, posteriorparietal, and cerebellar areas (the so-called scaffolding-storage proposed by Petersen, Van Mier, Fiez, & Raichle, 1998), indicating that imitation learning may occur across cortical regions. Notably, the brain also has resources to implement imitation learning, e.g., the widely observed synchronized oscillations between cortical regions (Buzsáki & Draguhn, 2004) can modify neuron connections via Hebbian plasticity to support the transfer of neural representations. It will be interesting to explore how imitation learning is realized in real neural systems.

3.2. CoarseNet improves FineNet's robustness to noise

A deep CNN trained for image classification is known to overly rely on local textures rather than the global shape of objects (Baker, Lu, Erlikhman, & Kellman, 2018; Geirhos et al., 2018, 2018), which is sensitive to unseen noises. In our model, since CoarseNet processes blurred visual inputs, whereby the local texture information is no longer the main cue supporting object classification, we expect that CoarseNet is robust to noise corruptions. Furthermore, through association and feedback, we expect that the robustness of FineNet to noises is also leveraged. We carry out simulations to test this hypothesis. The implementation details are presented in Appendix D.

Fig. 5 presents the results, which compares the performance of the two-pathway model with those of FineNet and CoarseNet only without interaction. The models were trained on the clean CIFAR-10 dataset and tested by adding various noise perturbations, including Gaussian, shot, impulse and adversarial noises (for details, see Appendix A.2). We see that the noise robustness of the two-pathway model is improved significantly compared to that of FineNet only without the feedback of CoarseNet. Notably, although CoarseNet has much lower accuracy compared to FineNet, it is robust to all kinds of noises. This indicates that CoarseNet can generate a robust association, which help to improve the robustness of FineNet to noises.

To exclude the possibility that noise robustness of our model comes from the feedback interactions in FineNet itself rather than the feedback from CoarseNet, we carry out simulations by including feedback loops between higher and lower layers in FineNet without considering the feedback from CoarseNet. As shown in Table 1, with the loop of feedback interactions increases, the performances of FineNet in both clean and noisy datasets increase, but they are still inferior to the two-pathway model. These results confirm that the interplay between two pathways does contribute to improving the noise robustness of the model.

Finally, we carry out ablation study to analyze the contributions of different elements of the model, including FineNet, CoarseNet, the association module (SMA), and the feedback loop, and confirm that when any one of them is missing or modified, the robustness of the model to noise is degraded dramatically (see Fig. 6).

3.3. The two-pathway model implements rough-to-fine information processing

In the above sections, we have considered that the two pathways recognize the same categorical information of images. In reality, the two pathways may also process different levels of categorical information of images, and object recognition goes from rough to fine. For instance, CoarseNet may recognize the higher category of an object (e.g., reptile), and FineNet recognizes the lower category of the object (e.g., turtle). In such a case, the result of CoarseNet can serve as a cognitive bias to facilitate the performance of FineNet. We carry out experiments to test this hypothesis.

We construct a rough-to-fine recognition task using the CIFAR-100 dataset. We randomly choose 15 000 images from CIFAR-100, with each image having a super- and a sub-class labels, e.g., an image belongs to a sub-class turtle and super-class reptile (see Fig. 7A). There are totally 5 super-classes (people, reptiles, small mammals, trees, vehicles) and each of which further contains 5 sub-classes (e.g., for reptiles, the five sub-classes are crocodile, dinosaur, lizard, snake, and turtle) (see Appendix A.1). We train CoarseNet and FineNet to recognize the super- and sub-classes of each image, respectively. After training, we test the performances of the two-pathway model on a test dataset with Gaussian noise (see Appendix A.2).

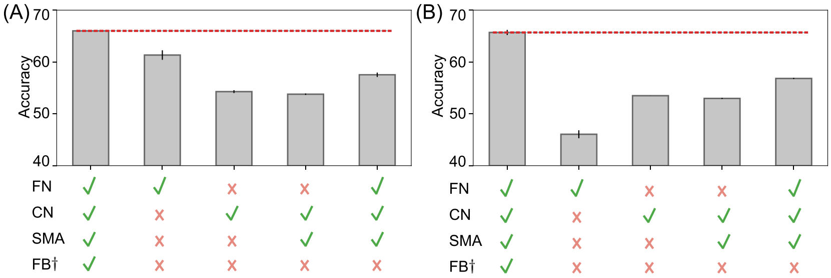

Fig. 6. Component analysis of the two-pathway model. FN: FineNet. CN: CoarseNet. SMA: static associative memory. FB†: the long-range feedback from CoarseNet and the higher layer of FineNet to the second layer of FineNet; without FB† means a short-range feedback to the third layer of FineNet. Red dashed line: the performance of the two-pathway model. (A) Model performances with respect to Gaussian, impulse and shot noises. (B) Model performances with respect to adversarial noise. Mean and std are obtained by averaging over 4 trials with random network initialization. Experimental details are the same as in Table 1.

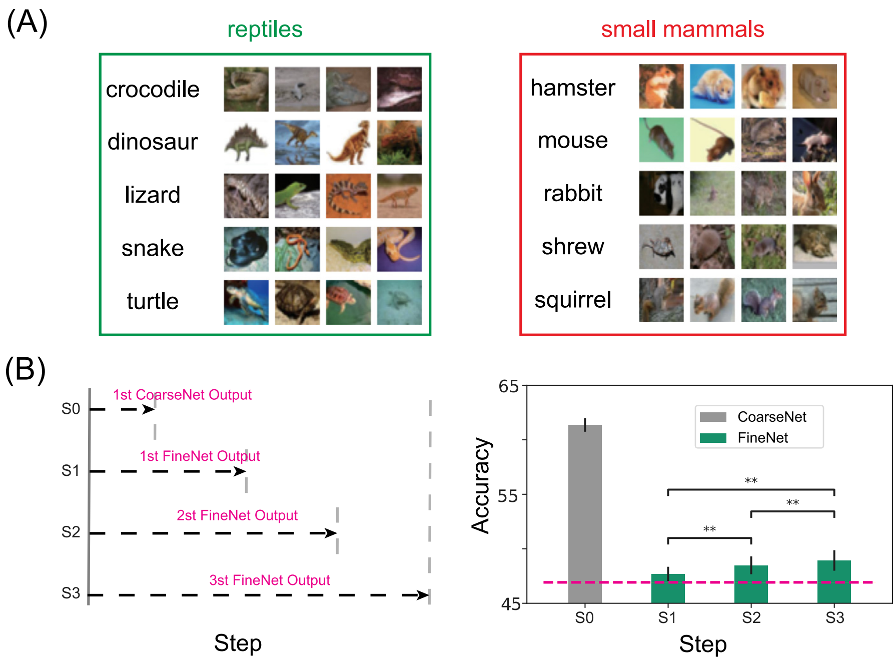

Fig. 7. Coarse to Fine information processing in the two-pathway model, where CoarseNet and FineNet perform super- and sub-class classifications, respectively. (A) Example images with super- and sub-class labels. Left panel: five sub-classes image samples of the super-class reptiles. Right panel: five sub-classes image samples of the super-class small mammals. (B) Left panel: the inference of the two-pathway model is divided into four steps. S0: CoarseNet recognizes the super-class label of the image; S1: FineNet recognizes the sub-class label of the image through the 1st-round of feedback interaction; S2 and S3: FineNet recognizes the sub-class label of the image through the 2nd and third rounds of feedback interaction. Right panel: the performances of the model over steps. The dashed red line denotes the accuracy of FineNet-only without interaction with CoarseNet. The model performances are evaluated on a test dataset with Gaussian noise and *severity* $= 2$. A paired t-test is conducted, and statistical significance is denoted as: *** indicates a significance level of $P < 0.001$, ** indicates $P < 0.01$, and * indicates $P < 0.05$. Mean and std are obtained by averaging over 8 trials with random network initialization. The other experimental settings are the same as in Table 1.

Table 1
Comparing the performance of the two-pathway model with that of FineNet with internal feedback loops. FineNet-only with $n_{fd} = 0, 1, 3$, refer to FineNet without feedback connection, with 1 loop of feedback interaction, and with 3 loops of feedback interaction, respectively. FineNet takes clean RGB images and CoarseNet the grayed, low-pass filtered images. The network performances for Gaussian, shot, and impulse noises are obtained by averaging over 5 different noise perturbation levels, and the results for adversarial noises are obtained by averaging over 8 different noise perturbation levels. Mean and std are obtained by averaging over 4 trials.

Models	Clean	Gaussian noise	Shot noise	Impulse noise
FineNet-only ($n_{fd} = 0$)	$86.9_{\pm 0.1}$	$50.0_{\pm 0.5}$	$57.8_{\pm 0.8}$	$59.0_{\pm 0.4}$
FineNet-only ($n_{fd} = 1$)	$88.4_{\pm 0.0}$	$56.6_{\pm 1.2}$	$64.4_{\pm 1.2}$	$61.4_{\pm 1.4}$
FineNet-only ($n_{fd} = 3$)	$88.0_{\pm 0.0}$	$59.1_{\pm 1.4}$	$65.9_{\pm 1.2}$	$63.0_{\pm 0.1}$
Two-pathway model	$86.7_{\pm 0.2}$	$\mathbf{62.2}_{\pm 0.2}$	$\mathbf{68.0}_{\pm 0.1}$	$\mathbf{66.5}_{\pm 0.5}$

Notably, the networks (CNNs) we use do not have temporal dynamics. To reflect the temporal dynamics of the real neural system, we decompose the model's outputs into multiple steps: Step 1: CoarseNet generates its output, predicting the super-class label of the image; Step 2: CoarseNet feedbacks to FineNet, and FineNet predicts the sub-class label of the image; Step 3 and 4: FineNet receives the feedback from CoarseNet and outputs the sub-class label of the image iteratively. The results are shown in Fig. 7B. We see that indeed through the feedback from CoarseNet, the accuracy of FineNet recognizing the sub-class label of an image increases over time, manifesting a characteristic of rough-to-fine information processing.

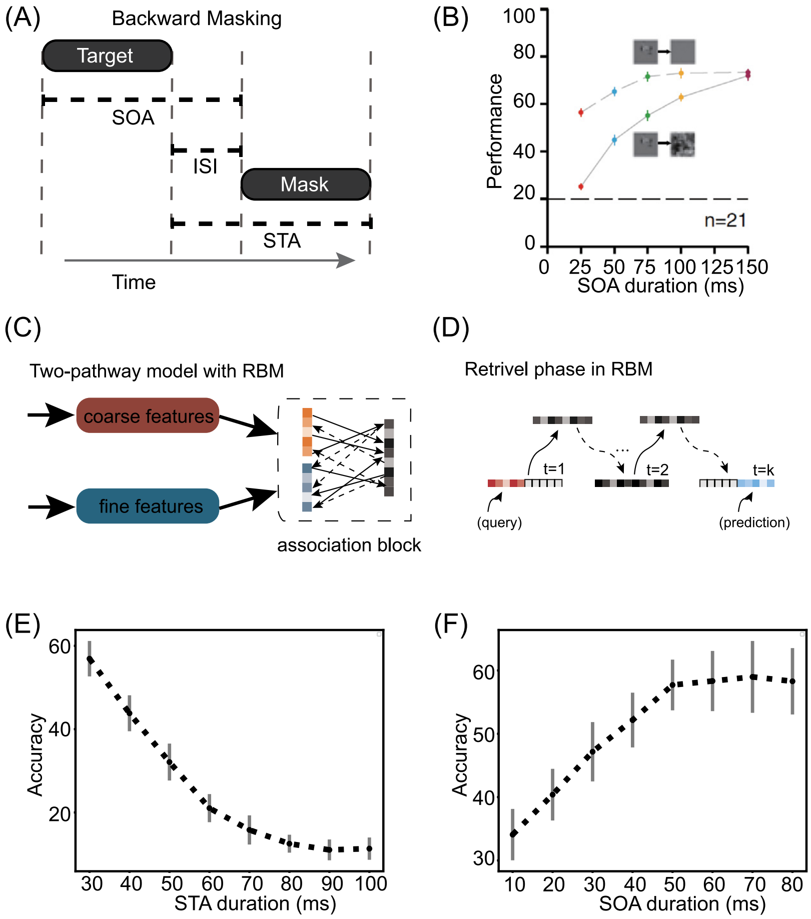

Fig. 8. The two-pathway model explains visual backward masking. (A) The paradigm of the backward masking experiment, adapted from Macknik and Martinez-Conde (2007). SOA: the time interval between the onsets of target and mask; ISI: the interval between the termination of target and the onset of mask; STA: the interval between the terminations of target and mask. (B) The experimental result, adapted from Tang et al. (2018). (C) The two-pathway model with RBM. RBM plays the role of a dynamical associative memory, which consists of a hidden layer and a visible layer, and the visible layer receives concatenated features from CoarseNet and FineNet. (D) The retrieval phase in the two-pathway model. The coarse features from CoarseNet are fed into the visible layer of RBM and the prediction is generated through the dynamics of RBM. (E) The recognition accuracy of the two-pathway model vs. different values of STA. ISI = 0. (F) The recognition accuracy of the two-pathway model vs. different values of SOA. Mean and std are obtained by averaging over 10 trials with random network initialization.

4. Two-pathway processing accounts for visual backward masking

Visual backward masking is a classic experiment widely used in cognitive psychology to investigate attention, awareness and dyslexia. In the cognitive experiment, a masking stimulus is presented after the target stimulus with a brief delay (usually 30–70 ms), which incurs a failure of the subject to consciously perceive the target (see Fig. 8A). The important factors affecting the masking effect are the target duration and the mask duration, referred to as stimulus onset asynchrony (SOA) and stimulus termination asynchrony (STA), respectively. An example study is display in Fig. 8B (Tang et al., 2018), in which subjects were required to perform a 5-classes recognition task involving objects that were either partially or fully visible, and object images were followed by either a gray screen (without masking) or a spatially overlapping noise pattern (with masking). SOA varies from 25 to 150 ms in randomly ordered trials. The experiment found that subjects' performances were dramatically disturbed when SOA is very small, and they were improved gradually when SOA increases. To explain visual backward masking, both feedback and feedforward mechanisms were proposed in the literature. The feedback mechanism suggests that the feedback from higher visual areas to V1 leads to the invisibility of the target stimulus (Lamme, Zipser, & Spekreijse, 2002). However, many experimental findings do not support that feedback plays a crucial role in visual masking (Macknik & Martinez-Conde, 2004, 2007; Martinez-Conde, Macknik, & Hubel, 2004), rather they suggest that visual masking is primarily driven by feedforward (non-reverberatory) lateral interactions between the target and mask (Macknik & Martinez-Conde, 2007). One well-known feedforward mechanism is the conceptual two-channel framework proposed by Breitmeyer et al. which comprises a fast transient channel and a slow sustained channel (Breitmeyer & Ganz, 1976), and our study proposes a neural network model to implement this framework.

Our two-pathway model naturally implements the two-channel processing idea. To capture the temporal effect in the experiment, we consider a dynamical memory association (DMA) process implemented by a restrict Boltzmann machine (RBM), which holds the same idea of using coarse features as a cue to predict fine features as SMA (see Fig. 8C, note that we use RBM instead of SMA to model visual masking is only because RBM involves iterative memory association, which allows us to model the time delay between the M and P pathways; while SMA performs one-shot computation and is not suitable to model this process. There is no extra interaction induced between SMA and RBM). The visible part of RBM is composed of the concatenated features from CoarseNet and FineNet, which are associated with each other through hidden variables, as shown in Fig. 8D. In the simulation, one iteration in RBM equals to a time step of 10 ms. As suggested by the neural data (Bar et al., 2006; Liu et al., 2017; Sugase et al., 1999), information processing in the M-pathway proceeds that in the P-pathway for about 50 ms, thus RBM will only receive coarse features as the input at the first 5 iterations. When the target information from FineNet arrives at the visible layer of RBM, it will interact with the features from CoarseNet. Note that at this moment the features can be the mask or the coarse feature of the target, depending on the SOA value. The network evolves for another period of time (500 ms), and the final features in the visible layer are used for recognition. For the training of DMA implemented by RBM, please refer to Appendix E.

Because of the time lag between two pathways, the coarse information of the mask is confounded with the fine information of the target, leading to wrong association that interferes the perception. The results are presented in Fig. 8E,F, which shows that the larger the STA or the shorter the SOA, the stronger the interference of the mask. Our model successfully reproduces the backward masking effect as observed in the experiment. As shown in Fig. 8F, the classification accuracy of the model increases with SOA, agreeing with the experimental findings (Tang et al., 2018).

5. Conclusion and discussion

In the present study, we have built a two-pathway model based on CNNs to study the interplay between two visual pathways. The model is composed of FineNet and CoarseNet, with the former extracting fine information of visual inputs and the latter extracting coarse information of inputs. CoarseNet processes information rapidly, whose result serves as a feedback to facilitate the performance of FineNet. Our study demonstrates several appealing properties associated with the interplay between two pathways, which are: (1) through imitation, CoarseNet can learn from FineNet to improve its prediction of inputs; (2) through association and feedback from CoarseNet, the robustness to noise of FineNet is improved significantly; (3) the prediction of CoarseNet can serve as a cognitive bias to leverage the performance of FineNet, achieving rough-to-fine information processing. Furthermore, we show that the two-pathway model can explain the visual backward masking phenomenon as observed in the experiment.

While there is currently no direct biological evidence for the presence of imitation learning in the M-pathway, numerous experimental findings and computational requirement strongly suggest that imitation learning should occur in the M-pathway. Firstly, imitation learning between different pathways has been suggested to be the cause of cognitive automaticity (Alvarez & Squire, 1994; Ashby, Ennis, & Spiering, 2007; Hélie, Roeder, & Ashby, 2010; Kawai et al., 2015; Murray & Escola, 2019; Pollmann & Maertens, 2005), enabling our brain to perform cognitive tasks rapidly, efficiently, and effortlessly after sufficient practice (Haith & Krakauer, 2018), especially in memory consolidation and motor skill learning. In addition, experiments have provided evidence that the M-pathway exhibits faster processing of image information compared to the P-pathway (Bar et al., 2006) and can process visual information in an automatic manner (Tamietto & de Gelder, 2010). Thus, the M-pathway has been proposed to play a significant role in cognitive automaticity, such as automatic categorization judgment (Ashby & Maddox, 2011), supporting the potential role of imitation learning in the M-pathway. Secondly, from the computational point of view, the M-pathway receives coarse visual input and is characterized by a shallow hierarchy, which limit its performance in recognition tasks. To enhance task performance, imitation learning from the P-pathway seems to be a natural solution.

Recently, Bakhtiari et al. also proposed a deep network model with two parallel pathways to simulate the visual cortex (Bakhtiari, Mineault, Lillicrap, Pack, & Richards, 2021). They found that when trained on a video dataset using a self-supervised predictive loss function, the model can capture the properties of both the ventral and dorsal visual pathways. Although both are studying visual information processing using CNNs, their model is different from ours on two key issues. First, their model consists of two CNNs of similar structures mimicking the ventral and dorsal pathways, while our model consists of two CNNs of very different structures, mimicking the difference between the P- and M- pathways. Second, their model focuses on the information fusion between two pathways, while our model focuses on the interplay between two pathways. Overall, the two models are studying the different parts (with some overlap) and different functions of the visual system, it will be interesting to integrate them together in future work.

It is believed that our visual system primarily learns and organizes itself in an unsupervised manner during the developmental period (Zhuang et al., 2020). Here, we training the model using supervised learning has two gains: (1) it brings us some insight into the computational principles of the two-pathway model; (2) supervised learning may be seen as partially modeling the evolutionary process of the neural system. Our model can be extended to accommodate unsupervised learning straightforwardly. For example, we can train our two-pathway model using some unsupervised methods, such as deep clustering method (Oord, Li, & Vinyals, 2018), BYOL (Grill et al., 2020), SimCLR (Chen, Kornblith, Norouzi, & Hinton, 2020) and so on. Moreover, we can leverage the dynamic interplay characteristic of our two-pathway model to enhance unsupervised learning on video data. For example, the coarse M-pathway can be trained in a unsupervised manner to predict spatial fine representations of the P-pathway, while the slow P-pathway can be trained in a unsupervised manner to predict future coarse features of the M-pathway.

Notably, the present study focuses on visual information processing, but the interplay between two pathways also exists in auditory information processing. It has been found that in the primate auditory cortex, two anatomically distinct streams exit, one is the ventral auditory pathway for analyzing the semantic information of sound-emitting objects and the other the dorsal auditory pathway for locating these objects (Santoro et al., 2014). Studies of the human auditory cortex further suggest that the ventral auditory pathway preferably encodes the fine-grained spectral information of sound in a slow processing manner, while the dorsal pathway encodes the coarse spectral information in a fast processing manner (Santoro et al., 2014). The separation of two pathways is proposed to support flexible auditory cognition (Santoro et al., 2014; Zulfiqar, Moerel, & Formisano, 2020). It will be interesting to extend our two-pathway model to the auditory system and explore whether the interaction principles found in this work are applicable to the auditory information processing.

We would like to point out that the current work is still a very preliminary modeling study of the dynamical interaction between two visual pathways, and there are a lot of space to improve. First, the current model considers that CoarseNet receives low-pass filtered or binarizing inputs mimicking the property of PRGCs, while in reality, the retina is not a simple prefilter of visual input, but has diverse ganglion cell types extracting different features of visual scene (Gollisch & Meister, 2010). Also, many inhibitory neurons, such as horizontal and amarcrine cells, exist in the retina, which interact with ganglion cells to execute complicated information processing (Gollisch & Meister, 2010). Incorporating these retinal functions in our model will improve the simulation

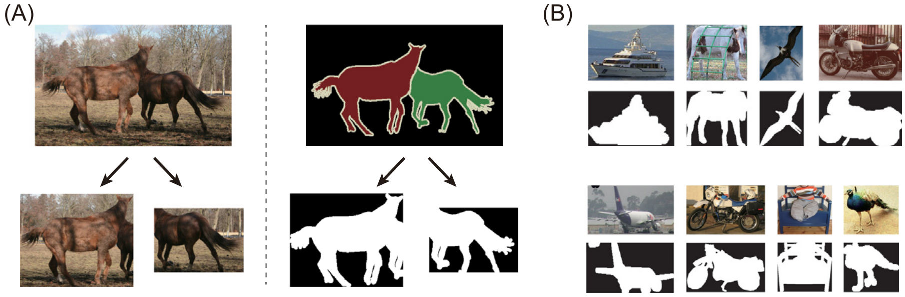

Fig. A.9. Data examples of the dataset Pascalvoc-mask. (A) left panel: a raw image and the corresponding objects; right panel: the segment of the raw image and the corresponding object segments. (B) Examples of Pascalvoc-mask. Images in the RGB channel and their masked counterparts.

of the M-pathway and P-pathway. Second, the current model considers a simple memory module, SMA or DMA to mediate the interaction between FineNet and CoarseNet, while the real memory association process in the brain is much more complicated and efficient. In future work, we will include the biologically more plausible memory module in the model, such as to combine the synaptic plasticity and addressing mechanism (Tyulmankov, Yang, & Abbott, 2022) and the hierarchical associative memory process (Krotov, 2021). Third, the current model employs CNNs as the building block to construct FineNet and CoarseNet, which miss the lateral connections between neurons that widely exist in the visual cortex. These lateral connections are known to play important roles in dynamical visual information processing (Gilbert & Li, 2013), and should be included in our future work. Nevertheless, we hope that this preliminary modeling study can give us some insights into understanding the interaction principles between two visual pathways and may inspire us to develop new object recognition architectures.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

We use publicly published standard datasets, and when the article is published, we will release the code.

Acknowledgments

This work was supported by National Key R&D Program of China (No. 2020AAA0105200), Science and Technology Innovation 2030-Brain Science and Brain-inspired Intelligence Project (No. 2021ZD0200204).

Appendix A. Datasets and manipulation of the input

A.1. Three datasets

We use three datasets, Pascalvoc-mask, CIFAR-10 and CIFAR-100 to evaluate our model. Pascalvoc-mask and CIFAR-10 are used to demonstrate the effect of imitation learning in Fig. 4. CIFAR-10 and CIFAR-100 are used to test the noise robustness and the rough-to-fine processing property of our model.

Pascalvoc-mask is a new dataset we created from the Pascalvoc2012 dataset (Everingham et al., 2015), which contains 20 foreground object classes. The goal of the original Pascalvoc2012 dataset is to recognize objects from a number of visual object classes in realistic scenes. There are two main tasks (classification and detection) and two additional competitions (segmentation and action classification). In the current study, we are only interested in those objects with precise annotated segments. Totally, there are 2913 images with 6866 objects having annotated segments. To create the Pascal-mask dataset, firstly, we extract each object image from the raw image and each object segment from the corresponding "SegmentationObject" image set according to the bounding box information (see Fig. A.9A). Secondly, we remove object images with low resolution (the number of pixels in width or height is less than 50) or large aspect ratio (width/height or height/width is more than 3). In this way, the number of remaining objects is 4887. Thirdly, to obtain the masked counterpart of each object, we gray the object segment by setting the pixel values for objects to be 1 and backgrounds to be 0 (see Fig. A.9B). All object images are resized to 64 × 64 to fit the input of CoarseNet. The new dataset consists of 4887 object images with masks. We split the dataset into 4512 training and 375 testing images, where the testing set is all from the Pascalvoc2007 testing set. See Table A.2 for the details of Pascalvoc-mask. The dataset can be found at https://drive.google.com/file/d/1TP0QsFBtVwXaCENGTwuk9ZhDlkGMyTOj/view?usp=sharing.

CIFAR-10 consists of 60 000 32 × 32 color images for 10 classes, with 6000 images per class, and they are split into 5000 training and 1000 test images in each class.

CIFAR-100 is a harder version of the CIFAR-10 dataset, which has 100 classes, with 600 images per class, and they are split into 500 training and 100 testing images in each class. The 100 classes in the CIFAR-100 are further grouped into 20 superclasses, so that each image has a pair of sub-class and super-class labels (this information is used to test the rough-to-fine processing property). In the rough-to-fine task, 5 superclasses are randomly selected, containing 15 000 images, 12 500 for training and 2500 for testing. The detailed superclasses and subclasses are shown in the Table A.3.

A.2. Manipulating the input with different noises

Evaluating the model performances under different kinds of noise disruption is a main task in the current study. Here we describe the details of manipulating inputs with various forms of noise. Four types of noises are used, Gaussian, shot, impulse, and adversarial noises. Please see Fig. 3 for details.

We obtain the performances of models on Gaussian noise, shot noise, and impulse noise dataset by averaging over 5 amplitude levels (see Fig. 5). For Gaussian noise, the 5 levels correspond to the noise variance $std = [0.04, 0.06, 0.08, 0.09, 0.10]$. For shot noise, the 5 levels correspond to the multiplication parameter $c = [500, 250, 100, 75, 50]$. For impulse noise, the 5 levels correspond to the probability $p = [0.01, 0.02, 0.03, 0.05, 0.07]$. For adversarial noise, we average 9 different levels with $\epsilon = [0.005, 0.01, 0.015, 0.02, 0.025, 0.03, 0.035, 0.04]$.

Table A.2
The number of samples in each class of Pascalvoc-mask. Digits in each column mean the training/testing numbers. The number of classes is 20 and the total number of training examples/testing examples is 4512/375.

Aeroplane	Bicycle	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow
104/10	147/15	215/11	126/9	92/6	174/12	206/16	261/17	398/33	204/10
Diningtable	Dog	Horse	Motorbike	Person	Pottedplant	Sheep	Sofa	Train	Tvmonitor
115/14	267/16	176/11	162/18	1017/100	186/14	188/21	178/14	138/12	158/16

Table A.3
The detailed superclasses and subclasses used in the rough-to-fine task.

Superclasses	Subclasses
People	Baby, boy, girl, man, woman
Reptiles	Crocodile, dinosaur, lizard, snake, turtle
Small mammals	Hamster, mouse, rabbit, shrew, squirrel
Trees	Maple, oak, palm, pine, willow
Vehicles 1	Bicycle, bus, motorcycle, pickup truck, train

Appendix B. Implementation details of the SMA buffer

The SMA buffer is initialized by using the feature representations obtained from FineNet and CoarseNet, which are trained independently for several optimization iterations. The SMA buffer size, denoted as K, is set to be equal to the number of training samples, e.g., $K = 40\,000$ in Fig. 5. The inverse temperature parameter β is set to be 100. During each SMA update, the key and value matrices in the SMA are replaced by the feature vectors from the newly trained CoarseNet and FineNet. Using different parameter settings will not change our results qualitatively, but the optimal values of parameters vary with the task.

Appendix C. Implementation details of imitation learning

In Section 3.1, we illustrate the effect of imitation learning to CoarseNet. Here, we will introduce the implementation details of FineNet and CoarseNet in imitation learning task. FineNet used in this task consists of three stacked layers, each of which comprises a 128-filter 3 × 3 convolution, followed by a batch normalization, a ReLU nonlinearity, and 2 × 2 max-pooling. CoarseNet has two stacked layers with the same composition as in FineNet, except that it comprises 64-filter 11 × 11 convolution in the first layer and 128-filter 9 × 9 convolution in the second layer. The balancing term $\alpha = 0.4$ is used when training CoarseNet. Both FineNet and CoarseNet have a fully-connected layer of 1000 units before the readout layer. Except for normalizing with the channel-wise mean and standard deviation of the whole dataset, no other pre-processing strategies are adopted. Both FineNet and CoarseNet share the same training settings: the total number of training epochs is 150, SGD with a momentum term 0.9 is used to optimize parameters, and the initial learning rate is 0.05 which is multiplied with 0.1 after 100 and 125 epochs.

Appendix D. Implementation details of noise robustness task with our two-pathway model

In Section 3.2, we illustrate the computational property of two-pathway model in noise robustness task. Here, we will introduce the implementation details of our two-pathway model in the task (the implementation is also adopt in Section 3.3). Without loss of generality, FineNet adopt slightly deeper structures than that used in the imitation learning task (see examples in Fig. 3A). In the experiments, FineNet consists of four convolutional layers (see Fig. 2A), each of which comprises a 3 × 3 convolution, followed by a group normalization, a ReLU nonlinearity. The numbers of convolutional filters in 4 layers are [64, 128, 256, 512]. CoarseNet consists of two convolutional layers with the same composition as in FineNet, except that it comprises 128-filter 7 × 7 convolution kernels in the first layer and 512-filter 5 × 5 convolution kernels in the second layer. The balancing term $\alpha = 0.0$ is used when training CoarseNet. Both FineNet and CoarseNet have a global pooling layer before the readout layer (for generating $\boldsymbol{g}^C(\hat{\boldsymbol{x}}; \boldsymbol{\theta}^C)$ and $\boldsymbol{g}^F(\boldsymbol{x}, \boldsymbol{O}(\hat{\boldsymbol{x}}, \boldsymbol{x}); \boldsymbol{\theta}^F)$, respectively). The feedback kernel consists of an upsample layer and an 1 × 1 convolutional layer, followed by group normalization and sigmoid nonlinearity. It takes $\boldsymbol{O}(\hat{\boldsymbol{x}}, \boldsymbol{x})$ as the input and output a weighting term to modulate the representations in the second convolutional layer of FineNet via element-wise multiplication. During training, the memory buffer in SMA is updated after every two epochs, and $\beta = 100$. Both FineNet and CoarseNet share the same training settings as that in imitation learning task.

Appendix E. Restrict boltzmann machine (RBM) as a dynamical memory association model

To investigate the effect of different factors on the target visibility, e.g., the task duration (SOA), the mask duration (STA), we modified the similarity-based association phase into RBM, which introduces dynamics in the association phase. RBM is a simplified version of Boltzmann Machine (BM), with the latter being an extension of the Hopfield model with stochastic dynamics (Hinton, Osindero, & Teh, 2006). Both BM (Ackley, Hinton, & Sejnowski, 1985) and the Hopfield model (Hopfield, 1982) can be used to capture how memory patterns are stored as stationary states of neural circuits via recurrent connections between neurons. RBM consists of a visible and a hidden layers with no within-layer connections. Denote the input to the visible layer as $\boldsymbol{v}$, activities at the hidden layer as $\boldsymbol{h}$ and the connection matrix between two layers is $\boldsymbol{W}$. The energy function of a RBM is written as

$$E(\boldsymbol{v}, \boldsymbol{h}) = -\boldsymbol{a}\boldsymbol{v}^T - \boldsymbol{b}\boldsymbol{h}^T - \boldsymbol{v}\boldsymbol{W}\boldsymbol{h}^T, \tag{E.1}$$

where $\boldsymbol{a}$ and $\boldsymbol{b}$ represent the bias vectors in the visible and the hidden layers, respectively. The joint probability of a configuration $(\boldsymbol{v}, \boldsymbol{h})$ is written as

$$P(\boldsymbol{v}, \boldsymbol{h}) = \frac{e^{-E(\boldsymbol{v},\boldsymbol{h})}}{Z}, \tag{E.2}$$

where Z is the partition function given by $Z = \sum_{\boldsymbol{h}} \sum_{\boldsymbol{v}} e^{-E(\boldsymbol{v},\boldsymbol{h})}$. The probability of a specific $\boldsymbol{v}$ is

$$P(\boldsymbol{v}) = \frac{1}{Z} \sum_{\boldsymbol{h}} e^{-E(\boldsymbol{v},\boldsymbol{h})}. \tag{E.3}$$

The appealing property of the bipartite graph structure of RBM is that the conditional distributions $P(\boldsymbol{h}|\boldsymbol{v})$ and $P(\boldsymbol{v}|\boldsymbol{h})$ are factorial, i.e.,

$$P(\boldsymbol{v}|\boldsymbol{h}) = \prod_{i}^{n_v} P(v_i|\boldsymbol{h}), \quad P(v_i = 1|\boldsymbol{h}) = \sigma(a_i + \sum_{j=1}^{n_h} w_{ij} h_j), \tag{E.4}$$

$$P(\boldsymbol{h}|\boldsymbol{v}) = \prod_{i}^{n_h} P(h_i|\boldsymbol{v}), \quad P(h_i = 1|\boldsymbol{v}) = \sigma(b_i + \sum_{j=1}^{n_v} w_{ji} v_j), \tag{E.5}$$

where $\sigma(x) = 1/(1 + e^{-x/T})$ is a sigmoid function, with T the temperature.

To implement the association phase, we construct $\boldsymbol{v}$ by concatenating the features from both CoarseNet and FineNet, e.g., $\boldsymbol{v} =$

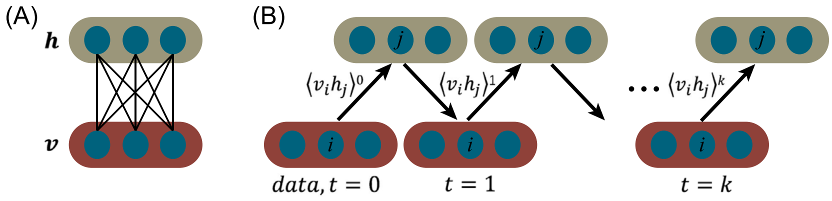
(A)
$\boldsymbol{h}$
$\boldsymbol{v}$
(B)
j
i
$\langle v_i h_j\rangle^0$
$\langle v_i h_j\rangle^1$
$\cdots$ $\langle v_i h_j\rangle^k$
$data, t = 0$
$t = 1$
$t = k$

Fig. E.10. Learning in RBM. (A) Diagram of a RBM with three visible and three hidden units. There are only connections between layers. (B) Illustrating the Gibbs sampling process in RBM during training. At time $t = 0$, the visible units $\boldsymbol{v}$ are initialized and the hidden units are updated according to $\boldsymbol{h} \sim P(\boldsymbol{h}|\boldsymbol{v})$. At time $t = 1$, the visible units are updated according to $\boldsymbol{v} \sim P(\boldsymbol{v}|\boldsymbol{h})$ and the correlations $\langle v_i h_j\rangle$ are the statistics used for contrastive learning in the RBM. The number of units in the visible layer is 1000, with 500 units for coarse features and 500 units for fine features.

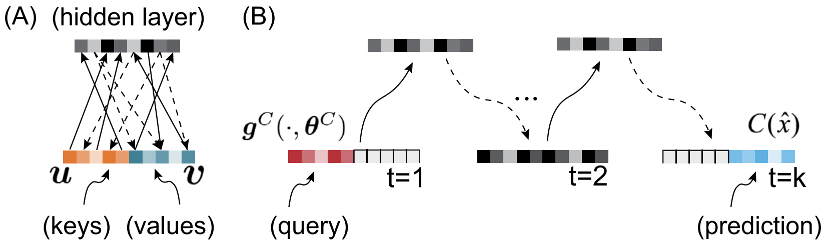

Fig. E.11. DMA implemented by RBM. (A) The training phase (see Fig. E.10 for the details). (B) The retrieval phase. The coarse probe $g^C(\hat{\boldsymbol{x}})$ of a visual object is fed into the visible layer of the trained RBM and the prediction $\boldsymbol{O}(\hat{\boldsymbol{x}})$ is retrieved through the dynamics of RBM.

$[\boldsymbol{v}^C, \boldsymbol{v}^F]$, with $\boldsymbol{v}^{\boldsymbol{C}} = g^C(\hat{x})$ and $\boldsymbol{v}^{\boldsymbol{F}} = g^F(x)$. Given the training examples (features of input images), RBM is optimized through minimizing the negative log-likelihood:

$$L_{RBM} = -\frac{1}{N}\sum_{i=1}^{N}\log\left[P(\boldsymbol{v}_i)\right] = -\frac{1}{N}\sum_{i=1}^{N}\log\left[P(\boldsymbol{v}_i^F, \boldsymbol{v}_i^C)\right]. \tag{E.6}$$

The derivative of the log likelihood with respect to a connection weight is calculated to be,

$$\frac{-\partial \log P(\boldsymbol{v})}{\partial W_{ij}} = \langle v_i h_j\rangle_{data} - \langle v_i h_j\rangle_{model}, \tag{E.7}$$

where the first and the second terms in the right hand of the equation denote expectations over the distributions of data and the model, respectively. The first expectation is tractable. For the second expectation, we apply the strategy of contrastive divergence (CD) gradient (Hinton, 2002), which approximates the expectation over the model distribution by a sample generated via a number of Gibbs sampling iterations, with the initial state of the visible units being the training sample, as illustrated in Fig. E.10B. More specifically, we use the correlation statistics $\langle v_i h_j\rangle^k$ after k step Gibbs sampling to replace the $\langle v_i h_j\rangle_{model}$ to update the connection weights, i.e.,

$$\Delta W_{ij} = \epsilon(\langle v_i h_j\rangle^0 - \langle v_i h_j\rangle^k), \tag{E.8}$$

where ϵ is the learning rate. During the training, $\boldsymbol{v}$ and $\boldsymbol{h}$ are sampled from $P(\boldsymbol{h}|\boldsymbol{v})$ and $P(\boldsymbol{v}|\boldsymbol{h})$ alternatively. The total number of training epochs is 2000. We use SGD to optimize the RBM with an initial learning rate of 0.1, which is multiplied with 0.1 after 500 and 1000 epochs.

Once the training is finished, we can feed a partial feature to the visible layer of RBM, and retrieve the complete one. For example, given a partial feature $\boldsymbol{v}_0$ at time 0, the hidden representations of RBM is $\boldsymbol{h}_0 = P(\boldsymbol{h} = 1|\boldsymbol{v}_0 = \sigma(\boldsymbol{a} + \boldsymbol{W}\boldsymbol{v}_0)$ and the updated activation in the visible layer is $\boldsymbol{v}_1 = \sigma(\boldsymbol{b} + \boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{h}_1)$. After k iterations, we can get a $\boldsymbol{v}_{\boldsymbol{k}}$ which is a complete feature corresponding to $\boldsymbol{v}_0$ (see Fig. E.11B).

References

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science*, *9*(1), 147–169.

Alvarez, P., & Squire, L. R. (1994). Memory consolidation and the medial temporal lobe: a simple network model. *Proceedings of the National Academy of Sciences*, *91*(15), 7041–7045.

Ashby, F. G., Ennis, J. M., & Spiering, B. J. (2007). A neurobiological theory of automaticity in perceptual categorization. *Psychological Review*, *114 3*, 632–656.

Ashby, F. G., & Maddox, W. T. (2011). Human category learning 2.0. *Annals of the New York Academy of Sciences*, *1224*.

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLoS Computational Biology*, *14*(12), Article e1006613.

Bakhtiari, S., Mineault, P., Lillicrap, T., Pack, C., & Richards, B. (2021). The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning. *Advances in Neural Information Processing Systems*, *34*, 25164–25178.

Bar, M. (2003). A cortical mechanism for triggering top-down facilitation in visual object recognition. *Journal of Cognitive Neuroscience*, *15*(4), 600–609.

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmid, A. M., Dale, A. M., et al. (2006). Top-down facilitation of visual recognition. *Proceedings of the National Academy of Sciences*, *103*(2), 449–454.

Breitmeyer, B. G., & Ganz, L. (1976). Implications of sustained and transient channels for theories of visual pattern masking, saccadic suppression, and information processing. *Psychological Review*, *83*(1), 1.

Bullier, J. (2001). Integrated model of visual processing. *Brain Research Reviews*, *36*(2–3), 96–107.

Buzsáki, G., & Draguhn, A. (2004). Neuronal oscillations in cortical networks. *science*, *304*(5679), 1926–1929.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. E. (2020). A simple framework for contrastive learning of visual representations. ArXiv abs/2002.05709.

Derrington, A., & Lennie, P. (1984). Spatial and temporal contrast sensitivities of neurones in lateral geniculate nucleus of macaque. *The Journal of Physiology*, *357*(1), 219–240.

Everingham, M., Eslami, S. A., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2015). The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, *111*(1), 98–136.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint arXiv:1811.12231.

Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. In *Advances in neural information processing systems* (pp. 7538–7550).

Gilbert, C. D., & Li, W. (2013). Top-down influences on visual processing. *Nature Reviews Neuroscience*, *14*(5), 350–363.

Gollisch, T., & Meister, M. (2010). Eye smarter than scientists believed: neural computations in circuits of the retina. *Neuron*, *65*(2), 150–164.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., et al. (2020). Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, *33*, 21271–21284.

Haith, A. M., & Krakauer, J. W. (2018). The multiple effects of practice: skill, habit and reduced cognitive load. *Current Opinion in Behavioral Sciences*, *20*, 196–201.

Hélie, S., Roeder, J. L., & Ashby, F. G. (2010). Evidence for cortical automaticity in rule-based categorization. *The Journal of Neuroscience*, *30*, 14225–14234.

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Computation*, *14*(8), 1771–1800.

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*(7), 1527–1554.

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

Hochstein, S., & Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron*, *36*(5), 791–804.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79*(8), 2554–2558.

Kawai, R., Markman, T., Poddar, R., Ko, R., Fantana, A. L., Dhawale, A. K., et al. (2015). Motor cortex is required for learning but not for executing a motor skill. *Neuron*, *86*, 800–812.

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*, 417–446.

Krotov, D. (2021). Hierarchical associative memory. arXiv preprint arXiv:2107.06446.

Lamme, V. A., Zipser, K., & Spekreijse, H. (2002). Masking interrupts figure-ground signals in V1. *Journal of Cognitive Neuroscience*, *14*(7), 1044–1053.

Liu, L., Wang, F., Zhou, K., Ding, N., & Luo, H. (2017). Perceptual integration rapidly activates dorsal visual pathway to guide local processing in early visual areas. *Plos Biology*, *15*(11), Article e2003646.

Macknik, S. L., & Martinez-Conde, S. (2004). Dichoptic visual masking reveals that early binocular neurons exhibit weak interocular suppression: Implications for binocular vision and visual awareness. *Journal of Cognitive Neuroscience*, *16*(6), 1049–1059.

Macknik, S. L., & Martinez-Conde, S. (2007). The role of feedback in visual masking and visual processing. *Advances in Cognitive Psychology*, *3*(1–2), 125.

Martinez-Conde, S., Macknik, S. L., & Hubel, D. H. (2004). The role of fixational eye movements in visual perception. *Nature Reviews Neuroscience*, *5*(3), 229–240.

Murray, J. M., & Escola, G. S. (2019). Remembrance of things practiced with fast and slow learning in cortical and subcortical pathways. *Nature Communications*, *11*.

Oord, A. v. d., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748.

Orhan, E. (2018). A simple cache model for image recognition. In *Advances in neural information processing systems* (pp. 10107–10116).

Petersen, S. E., Van Mier, H., Fiez, J. A., & Raichle, M. E. (1998). The effects of practice on the functional anatomy of task performance. *Proceedings of the National Academy of Sciences*, *95*(3), 853–860.

Pollmann, S., & Maertens, M. (2005). Shift of activity from attention to motor-related brain areas during visual learning. *Nature Neuroscience*, *8*, 1494–1496.

Santoro, R., Moerel, M., De Martino, F., Goebel, R., Ugurbil, K., Yacoub, E., et al. (2014). Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. *PLoS Computational Biology*, *10*(1), Article e1003412.

Sirota, A., Csicsvari, J., Buhl, D., & Buzsáki, G. (2003). Communication between neocortex and hippocampus during sleep in rodents. *Proceedings of the National Academy of Sciences*, *100*(4), 2065–2069.

Sugase, Y., Yamane, S., Ueno, S., & Kawano, K. (1999). Global and fine information coded by single neurons in the temporal visual cortex. *Nature*, *400*(6747), 869–873.

Tamietto, M., & de Gelder, B. (2010). Neural bases of the non-conscious perception of emotional signals. *Nature Reviews Neuroscience*, *11*, 697–709.

Tang, H., Schrimpf, M., Lotter, W., Moerman, C., Paredes, A., Caro, J. O., et al. (2018). Recurrent computations for visual pattern completion. *Proceedings of the National Academy of Sciences*, *115*(35), 8835–8840.

Tyulmankov, D., Yang, G. R., & Abbott, L. (2022). Meta-learning synaptic plasticity and memory addressing for continual familiarity detection. *Neuron*, *110*(3), 544–557.

Wang, W., Zhou, T., Zhuo, Y., Chen, L., & Huang, Y. (2020). Subcortical magnocellular visual system facilities object recognition by processing topological property. *BioRxiv*.

Yamins, D. L., Hong, H., Cadieu, C., & DiCarlo, J. J. (2013). Hierarchical modular optimization of convolutional networks achieves representations similar to macaque IT and human ventral stream. In *Advances in neural information processing systems* (pp. 3093–3101).

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., et al. (2020). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences of the United States of America*, *118*.

Zulfiqar, I., Moerel, M., & Formisano, E. (2020). Spectro-temporal processing in a two-stream computational model of auditory cortex. *Frontiers in Computational Neuroscience*, *13*, 95.