2023

Computational Mechanisms of Face Perception

Jinge Wang
West Virginia University, jnwang1@mix.wvu.edu

Computational Mechanisms of Face Perception

Jinge Wang

Dissertation submitted to the
Benjamin M. Statler College of Engineering and Mineral Resources
at West Virginia University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Computer Engineering

Xin Li, Ph.D., Chair
Shuo Wang, Ph.D.
Donald Adjeroh, Ph.D.
Katerina Goseva-Popstojanova, Ph.D.
Yuxin Liu, Ph.D.
Yong-Lak Park, Ph.D.

Lane Department of Computer Science and Electrical Engineering

Morgantown, West Virginia
2023

Keywords**:** Feature coding, Face perception, Deep learning, Critical period, Autism spectrum disorder (ASD), Identity coding, Identity-selective unit, Human single-neuron recordings

Abstract

Computational Mechanisms of Face Perception

Jinge Wang

The intertwined history of artificial intelligence and neuroscience has significantly impacted their development, with AI arising from and evolving alongside neuroscience. The remarkable performance of deep learning has inspired neuroscientists to investigate and utilize artificial neural networks as computational models to address biological issues. Studying the brain and its operational mechanisms can greatly enhance our understanding of neural networks, which has crucial implications for developing efficient AI algorithms. Many of the advanced perceptual and cognitive skills of biological systems are now possible to achieve through artificial intelligence systems, which is transforming our knowledge of brain function. Thus, the need for collaboration between the two disciplines demands emphasis. It's both intriguing and challenging to study the brain using computer science approaches, and this dissertation centers on exploring computational mechanisms related to face perception.

Face recognition, being the most active artificial intelligence research area, offers a wealth of data resources as well as a mature algorithm framework. From the perspective of neuroscience, face recognition is an important indicator of social cognitive formation and neural development. The ability to recognize faces is one of the most important cognitive functions. We first discuss the problem of how the brain encodes different face identities. By using DNNs to extract features from complex natural face images and project them into the feature space constructed by dimension reduction, we reveal a new face code in the human medial temporal lobe (MTL), where neurons encode visually similar identities. On this basis, we discover a subset of DNN units that are selective for facial identity. These identity-selective units exhibit a general ability to discriminate novel faces. By establishing coding similarities with real primate neurons, our study provides an important approach to understanding primate facial coding. Lastly, we discuss the impact of face learning during the critical period. We identify a critical period during DNN training and systematically discuss the use of facial information by the neural network both inside and outside the critical period. We further provide a computational explanation for the critical period influencing face learning through learning rate changes. In addition, we show an alternative method to partially recover the model outside the critical period by knowledge refinement and attention shifting.

Our current research not only highlights the importance of training orientation and visual experience in shaping neural responses to face features and reveals potential mechanisms for face recognition but also provides a practical set of ideas to test hypotheses and reconcile previous findings in neuroscience using computer methods.

Acknowledgments

I am deeply grateful to the many individuals who have contributed to and supported my Ph.D. studies. Completing this journey would not have been possible without their collective efforts.

First and foremost, I extend my profound appreciation to my advisors, Prof. Xin Li and Prof. Shuo Wang, for their unwavering guidance and support throughout my Ph.D. journey. Their patience, expertise, and valuable life advice have transformed me from an uncertain student into a confident researcher. I am immensely grateful for their mentorship in both my academic studies and personal growth.

Secondly, I wish to acknowledge the faculty and staff members of the Lane Department of Electrical Engineering and Computer Science. Throughout my years of study at this institution, they have consistently offered support and exhibited dedication to teaching. Their genuine desire to help has been instrumental in helping me reach this milestone.

My gratitude extends to my colleagues and friends Dr. Runnan Cao, Dr. Wentian Zhou, Dr. Siyuan Wu, Dr. Xuan Xu, Dr. Chuanbo Hu, Dr. Shan Jia, Dr. Xudong Liu, Xin Xiong, Mindi Ruan, Minglei Yin, and Na Zhang, whose kindness and assistance have enriched my Ph.D. experience.

I am also deeply thankful to my committee members Prof. Donald Adjeroh, Prof. Katerina Goseva-Popstojanova, Prof. Yong-Lak Park, and Prof. Yuxin Liu for their invaluable guidance, encouragement, and constructive feedback throughout my studies.

Lastly, I owe immense gratitude to my parents and my uncle's family for their unreserved support and love. Special thanks to my cherished wife, Qing Ye, for whom words alone cannot convey the depth of my gratitude and admiration. She is the cornerstone of my strength and a constant source of inspiration. Thanks for standing by my side with grace and patience throughout this journey.

Table of Contents

List of Figures

Chapter 1: Introduction

1.1 Motivation

The accelerated development of deep learning has positioned artificial intelligence at the forefront of academic research across various disciplines. Much like humans, AI derives inspiration and experience from nature and living organisms, continually expanding its knowledge and enhancing our world. Fundamentally, AI represents a bold endeavor by computer scientists to emulate and generalize human cognitive abilities. As such, the progression of AI is intimately connected to neuroscience research, with the two fields sharing a rich history of interaction [1]. A more profound comprehension of the biological brain has the potential to significantly contribute to the evolution of intelligent machines, while the rapid advancements in AI can offer invaluable support for neuroscience breakthroughs. This intersection between the two fields presents a promising area for future investigations, necessitating sustained collaboration. As AI researchers deepen their understanding of the brain, they can refine the development of intelligent machines. Conversely, as AI continues to evolve and progress, it can be employed to enhance our understanding of the brain, leading to further advancements in neuroscience. By collaborating, AI and neuroscience can catalyze each other's growth and propel progress in both domains.

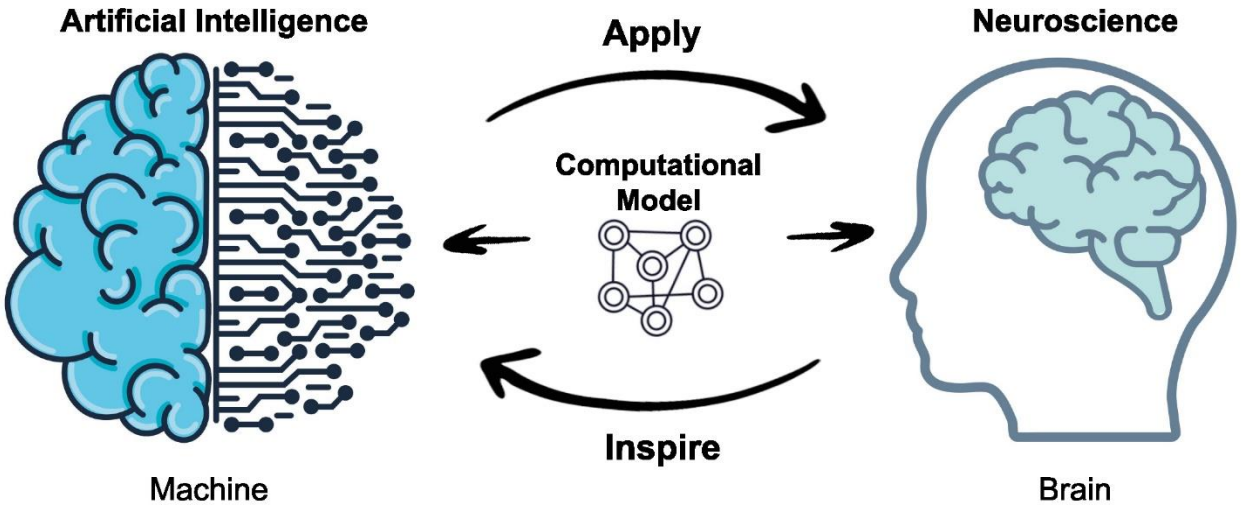


Figure 1.1 Relationship between AI and neuroscience

Artificial intelligence (AI) research aims to create computer systems capable of replicating or substituting biological or human intelligence in tasks encompassing perception, recognition,

decision-making, and control [2]. In contrast, neuroscience research seeks to delve deeper into the structure, function, and mechanisms of the biological brain, including its information processing, decision-making, and interactions with the environment [3]. AI can be viewed as a simulation of human intelligence, with numerous esteemed pioneers making significant contributions to the field. Alan Turing [4], Marvin Minsky and Seymour Papert [5], John McCarthy [6], and Geoffrey Hinton [7] are among those who aspired to not only construct machines that mimic the human brain but also to gain insights into how the brain computes and processes information using computational models. In the 1980s, neuroscientists started employing artificial neural networks as models for brain computation, leading to an increasing consensus in cognitive science that intelligence could be studied as a holistic phenomenon. By utilizing AI models, we can augment our understanding of the brain and develop superior AI systems, which can, in turn, facilitate further comprehension of the brain. Collaborative efforts between these two fields can unlock novel research and innovation opportunities.

Over the past decade, artificial neural networks have experienced remarkable progress, achieving new performance milestones. The extensive application of AI has facilitated the resolution of numerous challenges once deemed insurmountable by biological systems, giving rise to innovative ideas and methods. This advancement has inspired neuroscientists to utilize artificial neural networks to emulate biological brain networks, which was their original purpose. Recent technological breakthroughs in monitoring the activity of vast numbers of human brain neurons have significantly contributed to this success. The wealth of data has equipped researchers with the necessary groundwork to employ intricate artificial neural networks, allowing for a more direct comparison between artificial neural networks and real brains, and revealing striking parallels between artificial neural network representations and biological brain responses in pertinent tasks. This connection has been observed across various brain regions, encompassing low and high vision areas, language areas, motor areas, and prefrontal areas, in both feedforward and recurrent neural networks [8]. For example, researchers have demonstrated that the inferior temporal cortex, a critical region in primates responsible for object categorization, can be physically represented as an object space map with two axes that resemble those observed in the later layers of a deep network [9]. Concurrently, computational neuroscientists focused on the brain's learning function and plasticity have explored methods to train artificial neural networks [10]. These discoveries

have resulted in a significant uptick in studies employing AI to simulate animal behavior and cognition using techniques that hold theoretical relevance to the brain.

On the other hand, AI is rapidly evolving, generating new findings while simultaneously encountering obstacles in numerous domains. There is a growing expectation for neuroscience to offer additional insights that will guide the development of innovative artificial neural network approaches and further advance AI. In many areas, brain systems maintain a clear advantage over artificial neural networks. Investigating these gaps can yield a wealth of knowledge and resources for enhancing artificial neural network models. Over time, neuroscience inspirations have emerged across all facets of artificial neural network development. The research on convolutional neural networks and deep learning has been invigorated by the discovery of convolutional features and multilayer structures in the brain using electron detectors [11, 12]. Attentional mechanisms identified through positron emission tomography (PET) imaging systems [13] have had profound implications for artificial neural networks [14]. Functional magnetic resonance imaging (fMRI) techniques have inspired the study of memory modules in machine learning models [15], eventually leading to the development of the Long Short-Term Memory Model (LSTM) and other advancements. These remarkable accomplishments have prompted computer scientists to acknowledge the importance of neuroscience in guiding future AI advancements.

As artificial intelligence and neuroscience become increasingly intertwined, a critical issue is gaining prominence. A new direction has emerged at the intersection of these two fields: the interpretability of neural networks. The human brain, as the most complex instrument known, poses considerable challenges for direct experimentation aimed at analyzing existing issues or exploring unknown possibilities. Not only are the technological requirements exceptionally high, but the available resources are also limited. Consequently, large-scale, systematic research on the human brain has consistently faced difficulties. While brain modeling with artificial intelligence has introduced new opportunities, these studies often provide merely reference results for neuroscience research, lacking discussions regarding the rationale and validity of the findings. This is because the structure and operating mechanisms of the human brain are far more complex than those of artificial neural networks. The scientific community has criticized artificial neural networks as being a "black box," with interpretability posing a significant concern in computer science today. As the focus on interpretability intensifies, there has been a growing number of efforts to open the "black box." Various visualization tools are available to help observe the

working patterns of artificial neural networks, enabling AI to offer more robust support for neuroscience. In the pursuit of explanation, computer scientists have begun developing interpretive tools grounded in established biological concepts and hypotheses, combined with the inspiration derived from neuroscience.

Although neuroscience and artificial intelligence complement each other, the complexity of research in both fields has significantly increased over time. Boundaries have formed, and both neuroscience and artificial intelligence have established conventional research paradigms, making interdisciplinary research challenging. In particular, the field of computer science has recently focused on enhancing the performance of existing algorithms, adopting a routine that only concerns incremental applications of existing techniques. This has led to a tendency for new AI approaches to be highly homogeneous. Nonetheless, there is no denying that AI's accomplishments have profoundly influenced the field of neuroscience. AI has developed novel methods for high-dimensional data analysis and providing powerful new models to describe brain functions. Historically, neuroscience has attempted to use a "simple model" to explain as much brain activity as possible. However, the emergence of models with millions of parameters has rendered the issue of interpretation unavoidable, casting doubt on the conventional understanding of a 'simple explanation.' What are the benefits and implications of describing a brain area in terms of a neural network with numerous poorly understood units? I would argue that even if we consider this model a "black box," it remains beneficial if it can accurately predict or classify the tasks we propose, especially since we now have some ability to explain and analyze large-scale models. At the very least, we can use it to determine the optimal architecture and learning rules that the brain employs. Since these systems demonstrate in much lower dimensions, they are more intelligible than direct biological experimental data. Significant practical applications, such as brain-computer interfaces and sensory systems, may also be implemented. These achievements are already exciting. Even within the field of deep learning, there is a strong belief that these models are surprisingly and unreasonably effective. As a result, theoretical machine learning is currently garnering more attention, and a better understanding of powerful models is also being sought.

The intricate relationship between AI and neuroscience offers a multitude of opportunities for collaboration between the two fields. Our research focuses on utilizing existing artificial intelligence algorithms and interpretable analytical tools to analyze and explain neuroscience problems, specifically in relation to face perception. From a computer science perspective, we

have access to a large amount of face data, and algorithms related to face recognition are well-developed. In fact, deep neural networks have outperformed the human brain on certain face-related tasks. From a neuroscience perspective, the visual system is one of the most extensively studied brain regions. Furthermore, the ability to recognize and identify faces is critical for social communication. Previous successful attempts to use artificial intelligence in neuroscience, coupled with numerous biological experiments, have provided important insights into face perception analysis. By combining existing theories and hypotheses in neuroscience with established computer science techniques, we aim to uncover the computational mechanisms of face perception. Our approach involves an in-depth exploration of computational models of artificial neural networks, which we use to analyze and predict neural networks in the brain. The findings from this research provide valuable experience for further advancements in this field, fostering ongoing collaboration and innovation at the intersection of AI and neuroscience.

1.2 Research Approach and Contributions

1.2.1 Feature-based encoding of face identity by single neurons in the human medial temporal lobe

This study highlights the advantages of using DNNs to investigate neural encoding for face identity. By extracting features from complex natural face images using DNNs and projecting them onto the feature space constructed through DNN feature reduction, we uncovered a novel face code in the human medial temporal lobe (MTL). This code suggests that neurons encode visually similar identities based on shared features. We demonstrated that some MTL neurons are selective to multiple different face identities, based on shared features that form clusters in the representation of a deep neural network trained to recognize faces. Contrary to prevailing views, we found that these neurons represent an individual's face with feature-based encoding, rather than through association with concepts. The response of these feature neurons did not depend on face identity or face familiarity, and the region of feature space to which they are tuned predicted their response to new face stimuli. Our results provide critical evidence bridging the perception-driven representation of facial features in the higher visual cortex and the memory-driven representation of semantics in the MTL. This evidence may form the basis for declarative memory.

1.2.2 Face identity coding in the deep neural network and primate brain

A central challenge in cognitive neuroscience is understanding how the brain encodes faces. The impact of visual experience, learning, and memory on face perception and recognition remains largely unclear. In this study, we analyzed the response characteristics of a face-recognition DNN and discovered that identity-selective units in the DNN could generalize their discriminability to face identities displayed in various styles, as well as to face identities not involved in training. Visualization and manipulation of the DNN revealed the importance of identity-selective units in face recognition. By establishing coding similarities with real primate neurons, our study provided a valuable method for understanding face coding in primates. Moreover, by analyzing an artificial neural network dedicated to face recognition, we can formulate hypotheses that can be validated in the primate brain. Our study not only emphasizes the importance of the direction of training and visual experience in shaping neural responses to face identities but also offers a useful approach for testing these hypotheses and reconciling previous findings.

1.2.3 A critical period for developing face recognition

Face learning has critical periods during development, but the computational properties and mechanisms underlying these critical periods remain unknown. In this study, we demonstrated that similar to humans and animals, deep artificial neural networks exhibit critical periods during which a stimulus deficit could impair the development of face learning. We systematically investigated facial information integration during a critical period. Specifically, we identified a critical period during training when the DNN rapidly learned face identities. In the baseline condition, reduced facial information led to diminished model performance and a subsequent inability to use information from the corresponding facial parts. Full recovery could only be achieved when full-face information was provided within the critical period. Interestingly, complementary facial information provided within the critical period could even override the original model. We further provided a computational interpretation by analyzing the learning rate, which explained all the properties of critical periods. Lastly, we demonstrated an alternative approach using knowledge distillation and attention transfer to partially recover the model outside the critical period. Our results not only reveal computational mechanisms underlying face learning but also suggest strategies to recover from face processing deficits.

1.3 Background

1.3.1 The development of artificial neural networks inspired by brain science

Brain science is a highly interdisciplinary field that encompasses neuroscience, medicine, mathematics, computational science, and other disciplines. Its primary goals are to understand, protect, and simulate the brain, with simulation serving as both a vital source of inspiration and the ultimate objective of artificial intelligence. Brain simulation relies on neurobiology and computational neuroscience, with artificial intelligence playing a significant role in it. Deep learning, a biologically-inspired yet simplified model, is currently employed for this purpose. It offers the advantages of compatibility with existing computer architectures and a solid mathematical foundation. However, it also presents challenges such as high computational costs and suboptimal hardware integration. While advancements in deep learning and big data have allowed these models to surpass human performance in specific tasks, they struggle to address complex problems that the human brain can easily handle, necessitating significant computational and data resources. In contrast, the human brain is a highly optimized system, consuming only 25 watts of energy while housing nearly 100 billion neurons and an average of 7,000 synaptic connections per neuron. This vast network enables the efficient processing of complex problems with low energy consumption. Consequently, scientists seek to emulate the human brain's information processing capabilities, and the development of effective neural network models has become critical in the pursuit of brain simulation.

Early brain-like computing, also known as neural computation, utilized neuronal and synaptic models to perform real-world recognition tasks that emulated human brain functions, such as character and face recognition. This brings us to machine learning, the fundamental theory of artificial intelligence. Machine learning is grounded in statistical learning and optimization theory. Prior to the advent of big data and deep learning, these methods were primarily employed to mine and analyze relatively small datasets, utilizing techniques like logistic regression, support vector machines (SVM), principal component analysis (PCA), spanning trees, and boosting approaches. These methods enabled predictions or classifications of test data by obtaining model parameters based on the statistical features of training data through optimization techniques. By integrating machine learning and neuronal models, the first generation of neural networks, known as perceptrons, was introduced by American computer scientist F. Rosenblatt in 1958 [3]. A single-

layer perceptron is a feedforward network consisting of just one layer of neurons and a threshold activation function. By training the network weights, the perceptron can be made to respond to a set of input vectors with an output of either 0 or 1, achieving classification. Perceptrons have a robust mathematical foundation and can be regarded as a variant of polynomial regression.

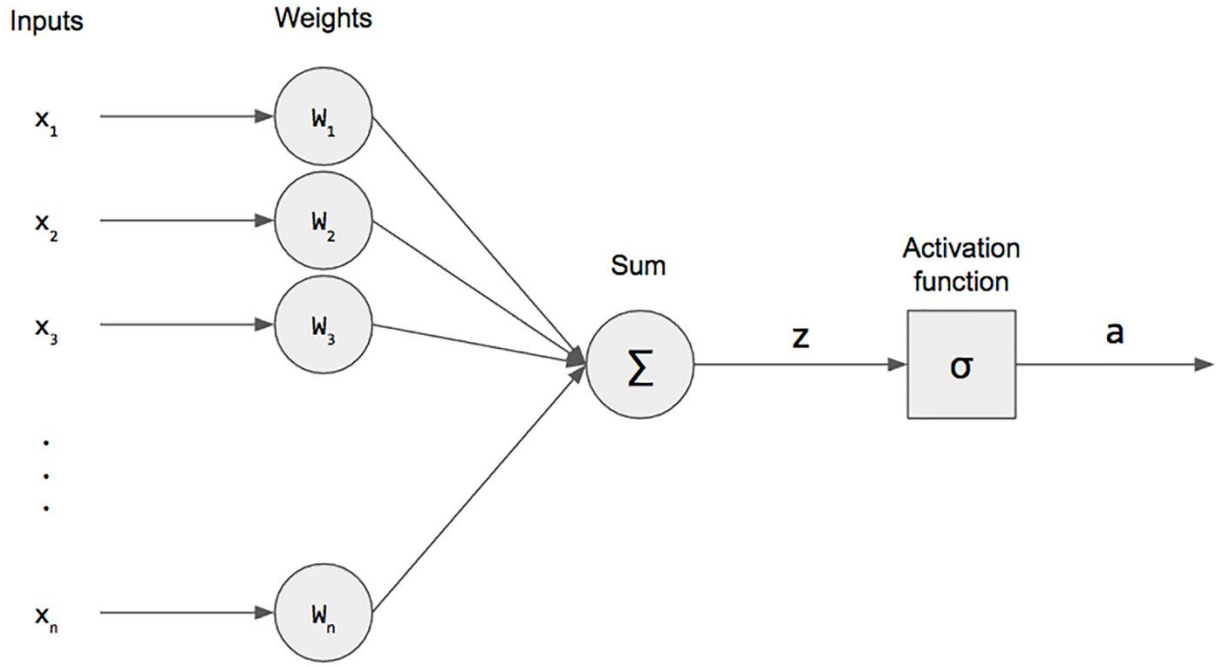


Figure 1.2 Perceptron

The perceptron is limited to linear classification and regression tasks within its single layer. However, with the introduction of the multi-layer perceptron (MLP), a differentiable sigmoid function replaced the activation function. Initially, the MLP could not be trained until the development of the backpropagation method. This breakthrough allowed neural networks to dynamically modify their connections by calculating the network's error function and feeding it back through the constituent layers. As a result, a new generation of artificial intelligence emerged, capable of advanced skills such as image and speech recognition [5, 7, 16]. Today, the backpropagation method remains widely used for training deep neural networks [10, 17]. The MLP's network structure, which allows for autonomous learning, represents the second generation of neural networks and now serves as the foundational theoretical model for deep learning.

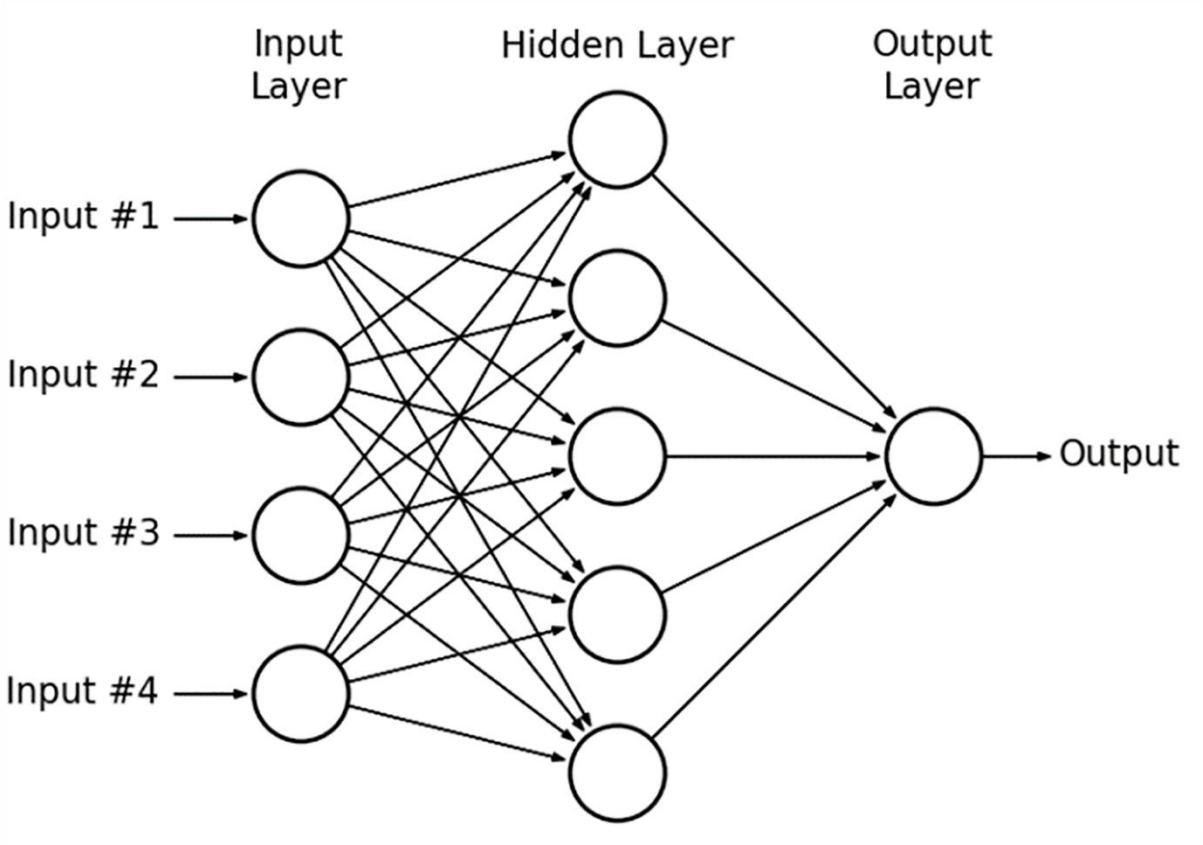


Figure 1.3 Multi-layer perceptron

The first two generations of neural networks were developed in the context of data science, where the input and output of neurons were abstracted into vectors, and the function of neurons was primarily to do weighted summation operations of matrices. However, because biological neurons process information as bioelectrical signals in the form of pulses rather than analog the computation of digital matrices, the first two generations of neural networks preserve just the neural network structure while greatly simplifying the neuronal model in the network. Spiking Neural Network (SNN) was first proposed by W. Maass in 1997 [18], and its underlying layer uses impulse functions to mimic bioelectrical signals as a way to transfer information between neurons, which can be considered as a third generation neural network. On the one hand, the SNN structure is more bio-interpretable and can be utilized as a fundamental tool for computational neuroscience simulation of biological brain activities; on the other hand, the SNN structure is easier to implement in hardware due to its information transfer via pulses. However, because the impulse function is not derivable, the SNN cannot be trained using t backpropagation algorithm and

stochastic gradient descent directly. The learning algorithm for SNN has been a major research challenge in recent years [19-21].

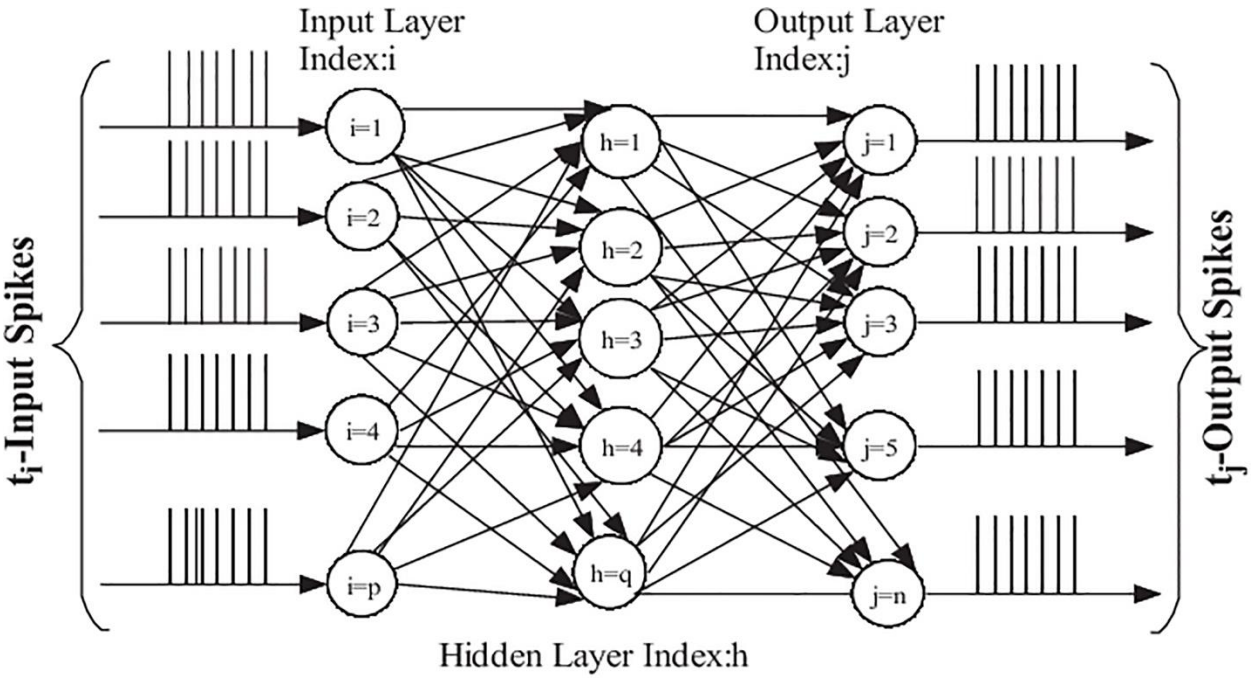


Figure 1.4 Spiking neural network

1.3.2 Brain regions and visual pathway

The brain, the most complex organ in the human body, weighs approximately three pounds and serves as the center of intelligence, the interpreter of sensory input, the initiator of body movement, and the regulator of behavior. Composed of numerous distinct parts, the brain contains specific regions, known as lobes, that perform specialized types of work. These lobes collaborate to enable the brain's overall functioning and contribute to our cognitive abilities and behaviors.

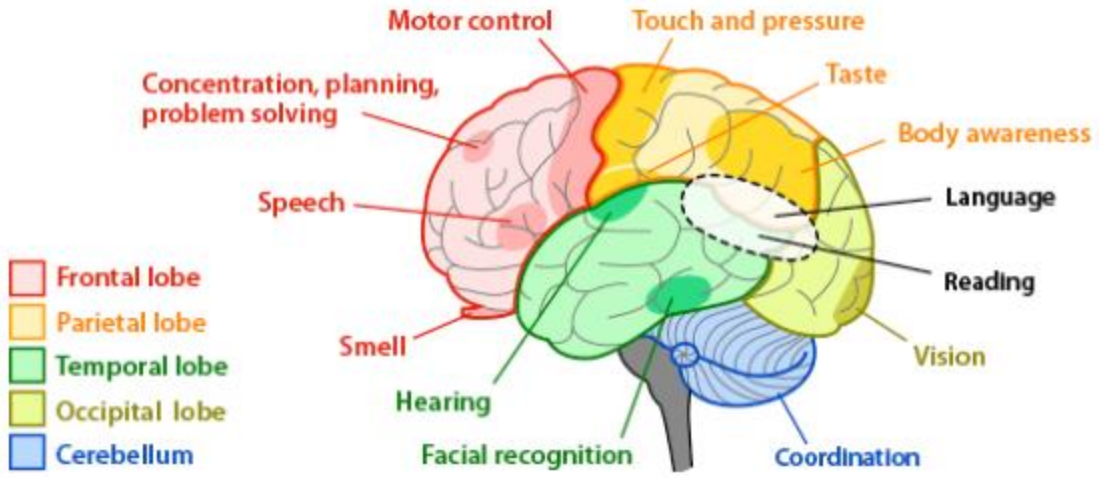


Figure 1.5 Brain regions

In our research, we paid more attention to the Temporal lobe, which is serve for hearing, recognizing, emotion and long-term memory.

The visual pathway refers to the anatomical structures responsible for transforming light into bioelectrical signals that the brain can interpret. Starting at the retina, the pathway concludes at the primary visual cortex. In computer science, convolutional neural networks (CNNs) are often utilized as models for biological vision due to their ability to perform visual tasks at a level comparable to humans. CNNs share a basic architecture with the visual system and extract features at different depths, which directly correlates with the activity in various regions of the visual system. CNNs possess two critical components that parallel the visual hierarchy in terms of properties and architecture. Firstly, as input progresses through the network layers, the receptive field size of individual units increases, mirroring the signal processing that occurs from V1 to IT within the visual system. Secondly, with increasing layer depth, CNNs learn more complex feature patterns, akin to the biological signal tuning that evolves from simple lines in V1 to semantic components in IT. This escalation in feature complexity can be directly observed through visualization techniques employed in CNNs.

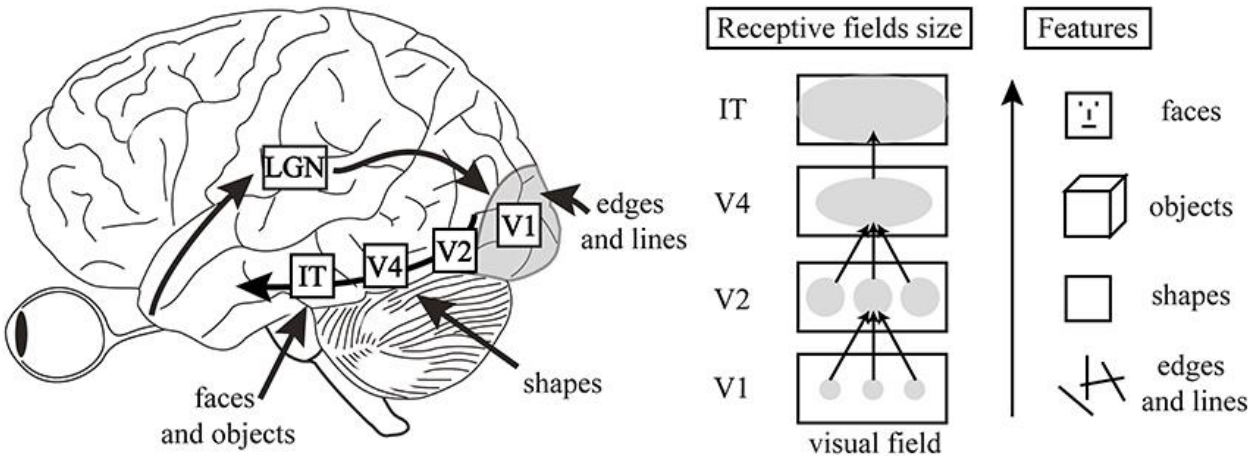


Figure 1.6 Visual pathway

1.3.3 Brain Data Acquisition

The availability of brain activity data further provides the possibility for computer analysis to participate in neuroscience. There are several methods to capture information of brain activities:

1.3.3.1 Electroencephalography (EEG)

EEG records the electrical activity generated by neural firing. It is a referential recording and does not represent absolute voltage; rather, it indicates a relative increase or decrease in electrical activity at specific brain locations. EEG devices allow for mobile testing, enabling brain processes to be recorded outside of laboratory environments. EEG has good temporal resolution since it captures information on when brain processes are active.

1.3.3.2 Magnetoencephalography (MEG)

MEG measures the magnetic fields generated by neural activity. MEG devices are stationary and sensitive to movement, requiring participants to lie or sit nearly motionless and minimize movements. The primary advantage of MEG is its combination of high temporal resolution, similar to EEG, with high spatial resolution.

1.3.3.3 Functional Magnetic Resonance Imaging (fMRI)

fMRI is a functional neuroimaging technique that measures brain function by detecting changes in blood flow associated with neural activity. fMRI has a trade-off between spatial and temporal resolution and is relatively slow compared to methods such as EEG or MEG. However, fMRI has good spatial resolution, capturing information on where brain processes occur.

1.3.3.4 Single-unit recording

Unlike the methods mentioned above, single-unit recording is an invasive data collection approach. It measures the electrophysiological responses of a single neuron using a microelectrode system. A microelectrode is inserted into the brain, where it records the rate of voltage change over time. Single-unit recordings are typically acquired during brain surgery, making such data highly valuable

In our study, we obtained brain activity data from single neuron recordings of human neurosurgical patients. To acquire the neuronal response from humans, we conducted single-neuron recordings from five neurosurgical patients (16 sessions in total). All participants provided written informed consent using procedures approved by the Institutional Review Board of West Virginia University (WVU).

1.3.4 Interpretive and statistical tools

1.3.4.1 Feature map visualization

Feature maps, or activation maps, capture the result of applying filters to input, such as an input image or another feature map. Visualizing a feature map for a specific input image helps understand what features of the input are detected or preserved in the feature maps. This allows us to see how the input data is decomposed in the network and what aspects of the original image are focused on by different filters. We want to visualize the feature map in three dimensions: width, height, and depth (channels). Each channel corresponds to an independent feature. The proper way to visualize these feature maps is to convert the content of each channel into a two-dimensional image separately or average the feature maps across channels to get the global information of the layer.

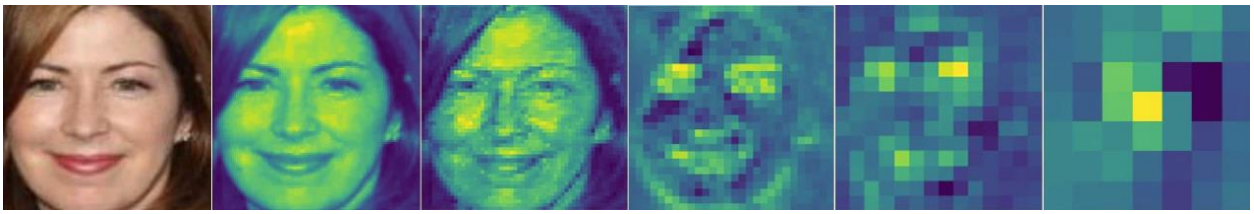


Figure 1.7 Feature map visualization

1.3.4.2 t-SNE

T-distributed Stochastic Neighbor Embedding (t-SNE) is a variation of stochastic neighbor embedding (SNE). It is a manifold learning algorithm for dimensionality reduction and is commonly used for high-dimensional data visualization with multiple classes. The basic principle

is to project each data point to the corresponding probability distribution through a mapping transformation. Specifically, a Gaussian distribution is used to transform the distance into a probability distribution in the high-dimensional space, and a long-tailed distribution is used to transform the distance into a probability distribution in the low-dimensional space. This allows lower distances in the high-dimensional space to have a larger distance after mapping, avoiding focusing too much on local features and ignoring global features when reducing dimensionality.

1.3.4.3 Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) [22] is a popular technique for visualizing which regions in the original image contribute to the final output. It uses the gradients of the target category flowing into a certain convolution layer, usually the last one, to produce a coarse localization map highlighting the important pixels/regions in the image for predicting the category. This approach reveals the model's implicit attention, making the real contributor of features in the input image distinguishable. Grad-CAM is an improvement over the previous approach, CAM [23], for both versatility and accuracy.

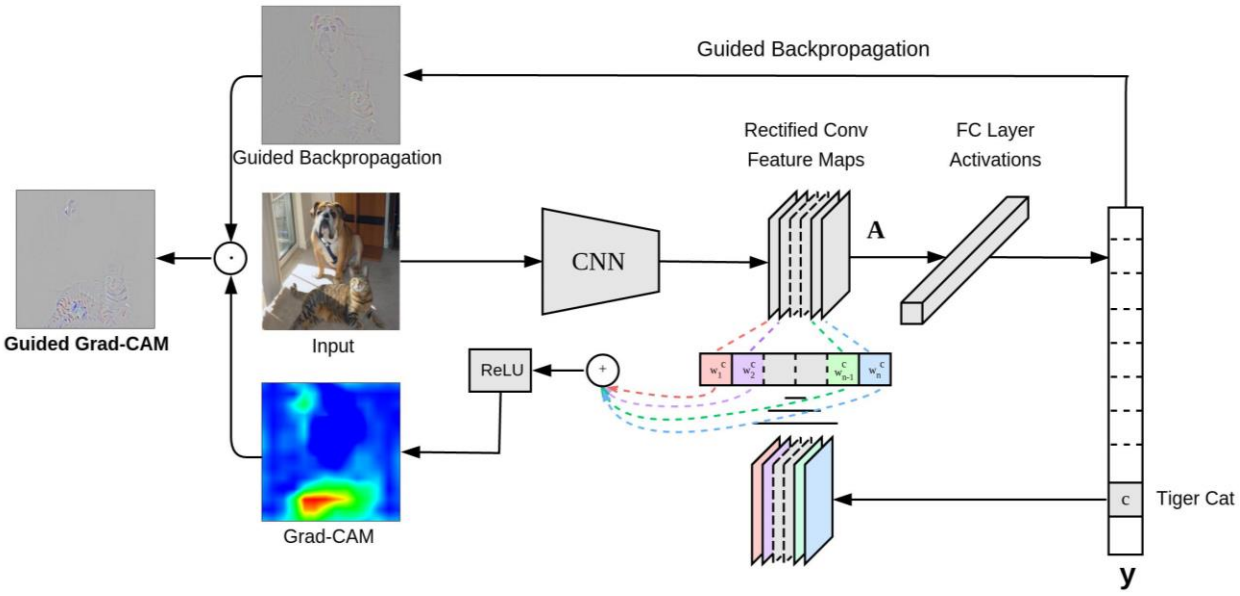


Figure 1.8 Grad-CAM

1.4 Related Work

1.4.1 Feature-based encoding of face identity by single neurons in the human medial temporal lobe

How the brain encodes different face identities is one of the most fundamental and intriguing questions in neuroscience. There are two extreme hypotheses. The feature-based model posits that face representations are encoded over a broad and distributed population of neurons [24] (1-4).

Under this model, recognizing a particular individual requires access to many neurons, with each neuron responding to many different faces that share specific visual features such as shape and skin texture (e.g., (5) and (6)). Conclusive evidence for feature-based coding, in particular axis-based feature coding (i.e., neurons parametrically correlate with facial features along specific axes in face space), has recently been revealed in the non-human primate inferotemporal cortex (IT) (7-10). In contrast, on the other extreme, the exemplar-based model posits that explicit facial representations in the brain are formed by highly selective (sparse) but at the same time highly visually invariant neurons (11-14). Identity neurons that selectively respond to many different images showing a specific person's face embody the exemplar-based coding and are common in the human hippocampus and other parts of the medial temporal lobe (MTL) (13, 14). Recent studies have shown that the responses of identity neurons are clustered by high-level conceptual or semantic relatedness (e.g., Bill Clinton and Hillary Clinton) rather than by lower-level facial features (15, 16). Feature-based and exemplar-based models are not mutually exclusive given that both types of neurons have been observed in different brain regions; but there appears to be an abrupt transition from a distributed axis-coding model in the higher visual cortex to a sparse exemplar-based model in the MTL. The neural computations achieving this transformation remain little understood. Here, we ask the critical question of how the brain transitions from the representation of facial features processed in the higher visual cortex to the representation of identities in the MTL. We hypothesize that there are traces of feature-based encoding in the MTL and these remaining feature-based responses will enable the transformation from feature-based coding to exemplar-based coding.

To test this hypothesis, we recorded from 578 neurons in the amygdala and hippocampus (MTL areas) of 5 neurosurgical patients. At the meanwhile, we extracted features from the images

shown to the patients using a pre-trained deep neural network (DNN) VGG-16 trained to recognize faces. We then constructed a two-dimensional stimulus feature space using t-distributed stochastic neighbor embedding (t-SNE) feature reduction for each DNN layer. The feature space demonstrated an organized structure. DNN had no access to semantic information about the faces (e.g., gender, ethnicity, social traits), and therefore, the representation of each face in the feature space was entirely driven by visual features. Thus, faces were distributed in the feature space purely based on their visual appearance, regardless of any semantic information or conceptual association with each other. We next projected the neuronal responses of a given neuron to each face onto this visual feature space. Strikingly, this revealed that some neurons were selective to different identities that were clustered in the visual feature space, which suggests that these neurons responded to face identities that were in fact visually similar.

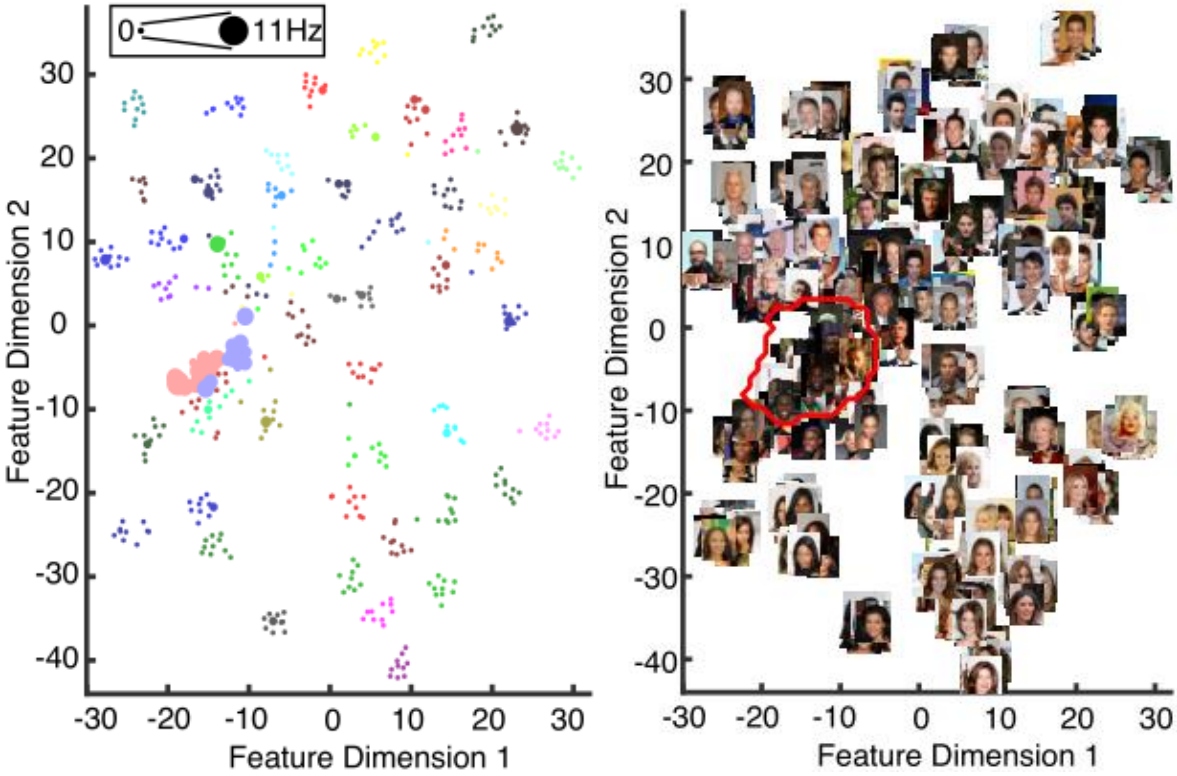


Figure 1.9 Identity Neuron

We next conduct experiment to reveal that the response of identity neurons in the human MTL can encode identities that are related visually rather than conceptually. We further identified feature neurons in the MTL that exhibited region-based feature coding. We showed that feature

neurons were not dependent on identity selectivity nor face familiarity, and their tuning regions could be validated by new face stimuli.

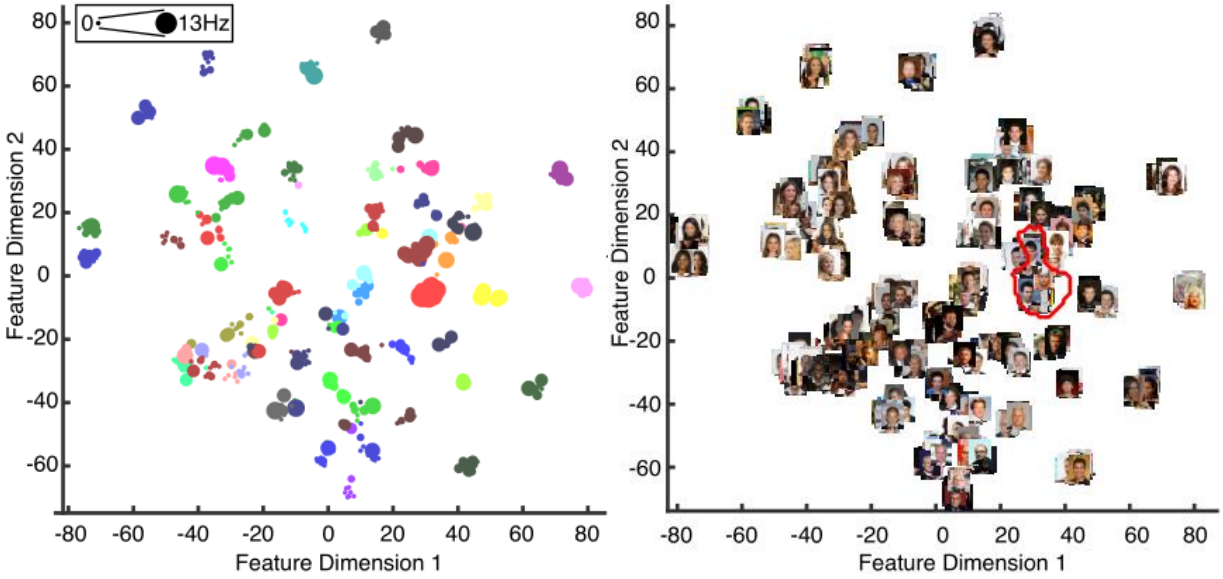


Figure 1.10 Feature Neuron

The findings bridge the two extreme hypotheses by illustrating region-based feature coding in the MTL, which may form the basis for feature-invariant exemplar-based coding and semantic memory.This study reiterates the advantages of using DNNs to study neural encoding for face identity: by extracting features from complex natural face images using DNNs and projecting them onto the feature space constructed by DNN feature reduction, It revealed a novel face code in the human MTL that neurons encode visually similar identities.

1.4.2 The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks

Starting from the 1990s, Yan Lecun et al [25] first proposed the idea of neural network pruning, i.e., pruning certain parameters in the network that do not contribute much to the output results.This method can greatly improve the speed of the model run, but at the same time it will have some impact on the model accuracy. After nearly 30 years of research, neural network pruning techniques can now reduce more than 90% of the parameters used in training the network to reduce storage requirements and improve the inferential computational performance of the model. The usual network pruning steps are shown below, first training a large, over-parameterized model, then pruning the parameters of the trained model according to certain guidelines, and finally fine-tuning the pruned model to obtain the lost accuracy.

Figure 1.11 Network pruning steps

However, it has been found through long-term experiments that the sparser the network is, the harder it is to train and the slower it is to learn, so the sparse architecture network generated by pruning is difficult to train from the beginning. In this artical, the lottery ticket theory is articulated by the authors: Among all possible connection in a dense, randomly-initialized, feed-forward architectures, there is a sparse subnetwork called the winning ticket that may achieve test accuracy comparable to the original network in a similar number of epochs.

Algorithm to attention the winning ticket:

1. randomly initialize the neural network
2. iterate several times to train the neural network and obtain the network parameters.
3. prune a certain percentage of the post-training parameters
4. initialize the remaining parameter values with the initial values at the beginning of the network, and construct the winning lottery network (i.e., the pruned and reinitialized network).

In which, this process uses iterative pruning, i.e., if we want to prune the weights of percent p through n iterations, then each iteration prunes the weights of one nth of percent p. This method can obtain a smaller size sub-network compared to one-time pruning and achieve the accuracy of the original network. However, repeated training also means that the finding cost is expensive.
The initialization of the winning ticket is as important as the structure of the winning ticket. A randomly initialized network is slower to learn than initializing with the original network and loses test accuracy when the pruning rate is very small.

The authors conducted the same experiments with the fully connected network Lenet on the MNIST dataset, and with the convolutional network on the Cifar-10 dataset trained by the iterative pruning method. The accuracy of most of the winning tickets was found to be significantly higher than that of the original network. With the pruning of the network, winning tickets learn faster, test with higher accuracy and generalize better compared to the original network. It was also demonstrated that Dropout - proposed by Hinton et al. in 2012 [26] to improve test accuracy

by randomly disabling a small portion of each training (i.e., randomly sampling a sub-network) - is equally applicable in the training of winning tickets. and pointed out that Dropout can improve the initial test accuracy but slow down the learning speed. So iterative pruning strategy and Dropout need to be used simultaneously in a complementary way to find the winning ticket faster. Finally, to demonstrate the validity of winning tickets in deeper networks, the authors look for winning tickets in VGG and ResNet networks. Unlike the previous pruning of each layer separately in the same proportion, ResNet-18 and VGG-19 are pruned in all convolutional layers with lower contributing weights (Global Pruning). This is because for these deeper networks, some layers have far more parameters than others. If all layers were pruned at the same ratio, these smaller layers would become bottlenecks and prevent us from finding the winning ticket with the smallest structure.

The final results, however, show that different learning rates and random reinitialization affect whether the winning ticket can be found. Still, these experiments can give us some inspiration, such as how to design better network structures and further improve model performance.

1.4.3 Critical Learning Periods in Deep Networks

In this paper, the authors propose the concept that for deep neural networks, similar to the learning process in animals and humans, there is a "critical learning period" for skill learning. From a biological perspective, critical periods are the early postnatal developmental windows during which perceptual deficits can lead to permanent skill impairment. Researchers in the field of biology have identified and documented critical periods affecting a range of species and systems, including vision in kittens, song learning in birds, and more. For humans, uncorrected eye defects (e.g., strabismus, cataracts) during critical periods of visual development can lead to adult amblyopia.

Researchers in the field of biology have established that the existence of critical periods in humans or animals is due to the biochemical modulation of windows of neuronal plasticity [27]. The earliest origins of neural networks were expected to mimic the working patterns of human brain neurons, and Achille et al. demonstrated in the literature that deep neural networks respond to sensory deficits in a manner similar to that observed in human and animal models. However, in neural networks, defects do not affect low-level statistical features, such as vertical flipping of

images, while having no lasting impact on performance, as well as can eventually be overcome with further training. This finding triggered the authors to think that the presence of "critical periods" in deep neural network learning may arise from information processing rather than biochemical phenomena.

A very well-known example of a critical period defect affecting humans is amblyopia (loss of vision in one eye) caused by cataracts in humans during infancy or childhood [28]. Even after surgical correction of cataracts, the ability of patients to regain normal vision in the affected eye depends on the duration of the visual defect and the age of onset, with early and long-term visual defects causing more severe effects. The goal of this paper is to investigate the impact of similar defects in DNN. To this end, the authors trained a standard All-CNN architecture to classify objects in images from the CIFAR-10 database. SGD was used for training in the experiments. To simulate the effect of cataracts, the authors used downsampling and then interpolation to obtain blurred images, corrupting small-scale image details. Training was continued to ensure that the network converged and that it could obtain the same number of uncorrupted images as in the control experiment.

The presence of a critical period can be easily observed in the performance of the network. If the blur is not removed at the early phase of the network training, the final performance is severely degraded (up to three times more error) compared to the baseline approach. This decrease in performance follows a trend commonly observed in animals, such as the loss of visual acuity observed in kittens deprived of a single eye after birth in relation to the length of the defect, as demonstrated in an earlier study [29].

From the results given by the experiment, it is natural to ask: Is there a corresponding critical period of learning for any change in the input data distribution? The authors say that this is not the case for neuronal networks, which are plastic enough to adapt to high-level changes in sensory processing. For example, adult humans can quickly adapt to some drastic changes, such as the inversion of the visual field. The authors also present corresponding results demonstrating that DNNs are largely unaffected by high-level defects, such as vertical flipping of images or random arrangement of output labels. After defect correction, the network quickly recovers its baseline performance. This implies a more fine-grained interaction between the structure of the data distribution and the optimization algorithm, which leads to a critical period.

Next, the authors imposed a more drastic defect attack on the network, causing each image to be replaced by white noise. Surprisingly this extreme defect shows a significantly lighter effect than the one obtained by blurring the images. Training the network with white noise does not provide any information about the natural images, so the effect of white noise is milder compared to other defects (e.g., image blurring). However, white noise contains some information that causes the network to (incorrectly) learn fine structures that are not present in the image.

When defects affect only high-level features (vertical flipping of images) or the last layer of the CNN (label swapping), the network does not exhibit critical periods (test accuracy remains essentially flat). On the other hand, defects similar to perceptual deprivation (images replaced by random noise) do lead to defects, but their impact is not as severe as in the case of image blurring. It is also noted that adding more convolutional layers increases the impact of critical period defects. Fully connected networks trained on the MNIST library also suffer from image blurring defects during the critical period. Therefore, the authors argue that the convolutional structure is not necessary for reproducing the critical period of model training, nor is the use of natural images. Similarly, ResNet-18 trained on CIFAR-10 has a critical period that is clearer than in the standard convolutional network. In the authors' analysis, ResNets allow gradients to be more easily back-propagated to lower layers, and the presence of their critical period could suggest that the critical period is not caused by gradient disappearance. The existence of critical periods does critically depend on the depth of the network. The authors confirm that a critical period exists even when the network is trained at a constant learning rate. When using Adam as the optimizer, a critical period similar to the standard SGD can still be observed by renormalizing the gradient using the running average of its first two moments. At the same time, changing the hyperparameters of the optimization can change the shape of the critical period, increasing the weight decay to make the critical period longer and less sharp. This can be interpreted as both slowing down the convergence of the network and limiting the ability of the high-level to change to overcome defects, thus encouraging the low-level to learn new features as well.

1.4.4 Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer

Inspired by the way humans observe objects, the concept of attention was introduced into neural networks, and has been successfully applied in computer vision and nature language processing to improve the performance of network models. The main idea of knowledge distillation is to transfer the "knowledge" from a complex teacher model with high learning ability to a simple student model. It also has significant effects on the performance enhancement of the model through optimization strategies such as mutual learning and self-learning of neural networks and data resources such as label-free and cross-modal. The authors argue that in knowledge distillation, the attentional map learned by the teacher's network can be used as a kind of knowledge that can be distilled to the student's network, so that the learning network can try to generate an attentional map similar to the teacher's network, thus improving the performance of the student's network.

By distilling the knowledge of the attention mechanism, the network is expected to automatically learn what needs to be noticed in the picture, so that the network learns to ignore irrelevant information and focus on the key information. That is, decide which part of the input needs to be attended to. Allocate limited information processing resources to the important parts.

In terms of the attentional map generation, through comparison, the activation-based method has better performance. Assume that A is the feature map after activation of a layer in CNN.

$$\mathrm{A} \in R^{C \times H \times W}$$

In which C is the number of feature map channels, H and W are the height and width of the feature map.

$$\mathrm{F}: R^{C \times H \times W} \rightarrow R^{H \times W}$$

Based on the activation-based attention transfer, the feature map A after activation layer is used as input, and the spatial attention map F is output.

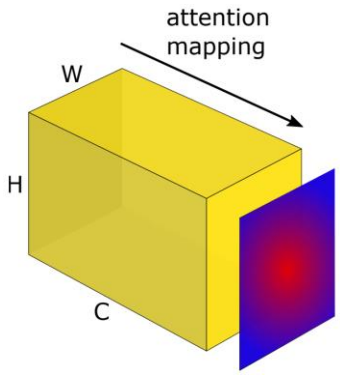


Figure 1.12 Attention mapping over feature dimension.

3 ways of activation-based attention map generation:

Sum of absolute values: $F_{sum}(A) = \sum_{i=1}^{C}|A_i|$

Sum of absolute values raised to the power of p (where p > 1): $F_{sum}^{p}(A) = \sum_{i=1}^{C}|A_i|^p$

Max of absolute values raised to the power of p (where p > 1): $F_{max}^{p}(A) = max_{i=1,C}|A_i|^p$

The activation of neurons are not only spatially correlated, but these correlations are strongly related to accuracy. Better performing and more powerful networks tend to have more accurate attention areas.

Different layers in the network have different attention areas and generate different attention maps. For example, the middle layer has the highest activation for the most discriminative regions, while the highest layer responds to the overall feature regions of the target

Compared with $F_{sum}(A)$, $F_{sum}^{p}(A)$ assigns a larger weight to the spatial location corresponding to the neuron with the highest activation level. For example, assigning more weights to the most discriminative part among all activated neurons at the same spatial location, $F_{max}^{p}(A)$ selects only the largest of them to assign weights, while $F_{sum}^{p}(A)$ prefers regions where multiple neurons are activated.

Implementation process of knowledge distillation using attention shifting

The attention map of the teacher's network is generated first, which can be obtained in any of the three ways mentioned above. The goal of knowledge distillation is to make the attention map generated by the student network as similar as possible to the attention map generated by the teacher network, so as to improve the accuracy of the student network.

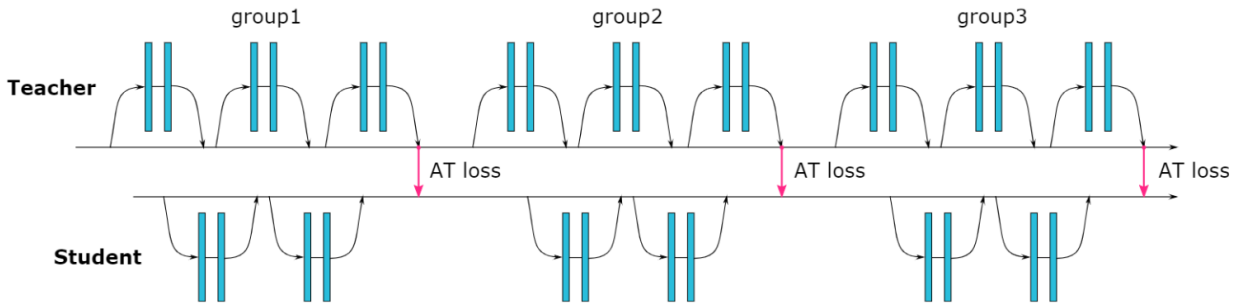


Figure 1.13 Schematics of teacher-student attention

Loss function design

$$Q_s^j = vec\left(F\left(A_S^j\right)\right)$$

$$Q_T^j = vec\left(F\left(A_T^j\right)\right)$$

$$L_{AT} = L(W_S, x) + \frac{\beta}{2}\sum_{j\in I}\left\|\frac{Q_s^j}{\left\|Q_s^j\right\|_2} - \frac{Q_T^j}{\left\|Q_T^j\right\|_2}\right\|_p$$

On the CIFAR-10 dataset, attention maps were generated using method 2, the sum of absolute values raised to the power of p. Combined with knowledge distillation, the attention transfer effectively reduced the recognition error rate.

This paper introduces attention mechanisms into the knowledge distillation domain to improve the performance of student networks. Different attention graph generation approaches are explored, each with a different focus. It inspires the research direction of knowledge distillation and innovates the knowledge distillation method. It also provides valuable experience for my future experiments.

Chapter 2: Face identity coding in the deep neural network and primate brain

2.1 Introduction

The ability to identify and recognize faces is one of the most important cognitive functions in social communications. Primates have a dedicated neural system to process faces. Neurons that are selectively responsive to faces (i.e., face-selective neurons) have been observed in a distributed network of brain areas, notably including the face patches in the temporal cortex [30] (shown in monkeys) as well as the amygdala and hippocampus [31, 32] (shown in both monkeys and humans). In particular, there are two extreme hypotheses about how neurons encode and represent faces. The exemplar-based model posits that faces are represented by highly selective, sparse, but visually invariant neurons. This model has been supposed by single-neuron recordings in the human amygdala, hippocampus, and other parts of the medial temporal lobe (MTL). The feature-based model posits that faces are represented by simultaneous activation of a broad and distributed population of neurons and each neuron responds to many pictures with similar basic features [33, 34]. This model has been supported by single-neuron recordings in the monkey inferotemporal (IT) cortex. Although exemplar-based and feature-based models are not mutually exclusive because both types of neurons have been observed in different brain regions, it remains unclear how to bridge these prior findings and how the brain transitions from one model to the other, which is largely due to the lack of simultaneous access to the activity of the entire face processing neural network (i.e., the whole population of neurons from all brain regions involved in face processing) [35].

To address this limitation, in this study, we exploit the opportunity to analyze the activity of all units from an artificial neural network (ANN) dedicated to face recognition. Importantly, we were able to compare the results directly with recordings from the primate neurons, which will shed light on how the primate brain encode face identities. Rapid advances in computer vision and development of deep neural networks (DNNs) have provided an unprecedented opportunity to study face recognition and representation (see [36] for a review). DNNs can help researchers to understand the functional architecture of the primate brain and test the computational benefits of fundamental organizational features for the visual system. For example, within a class of

biologically plausible hierarchical neural network models, there is a strong correlation between a model's categorization performance and its ability to predict single-neuron responses from the IT cortex [8]. In addition, both human neuroimaging studies [37] and intracranial electroencephalogram (EEG) studies [38] have shown that features in DNNs can be represented in the human brain, which in turn can explain our ability to recognize individual faces. Recent studies in monkeys have shown that images synthesized by DNNs can control neural population activity. Using natural face stimuli and face features extracted from DNNs, our recent work has shown that neurons in the human MTL encode visually similar identities [39].

The present study continues from our recent work showing feature-based encoding of face identities in the human MTL using DNN-extracted visual features [39]. The motivation for the present study is largely two-fold. First, the sparse coding hypothesis for face recognition has been inspected recently under the framework of DNN. It has been shown that identity, gender, and viewpoint information all contributes to individual unit responses [40, 41], similar to the neuronal coding of facial attributes in the primate brain. Experimentally, it is difficult to comprehensively characterize the tuning properties of neurons from the human brain to a large number of facial attributes; therefore, in this study, we will conduct *in silico* experiments to probe the tuning properties of DNN units using a diverse array of stimuli, which will in turn shed light on the tuning properties of human neurons. Second, several lines of research in DNN (e.g., dropout [42], neural architecture search [43]) have shown supporting evidence about the so-called *lottery ticket hypothesis* [44] (i.e., among all possible fully-connected and feedforward architectures, the winning ticket is a sparse subnetwork). This hypothesis has important biological implications into both energy efficiency and generalization properties, but there is still a missing parallel and convergent connection between the sparse coding hypothesis in neuroscience [45] and the lottery ticket hypothesis in deep learning [44]. To fill in this gap, in this study we used a DNN as a proxy model for studying the tuning properties of neurons that encode face identities. For the first time, our present study will directly compare and link the sparse coding of face identities between human neurons and DNN units.

2.2 Results

2.2.1 Identity-selective DNN units

We used 500 natural face images of 50 celebrities (10 faces per identity) to elicit response from a pre-trained deep neural network (DNN) VGG-16 trained to recognize faces (see **Figure 2.1a** for DNN architecture and **Supplementary Figure 1** for stimuli). We performed a fine-tuning on the top/output layer FC8 to ensure that the network was able to discriminate the identities used in the present study (see **Methods** for details). The pre-trained DNN had an accuracy of 94.2%±2.3% (mean±SD across identities) in identity recognition after fine-tuning. Note that to make our results more comparable to the literature, we used the original VGG network without any fine-tuning for further analyses.

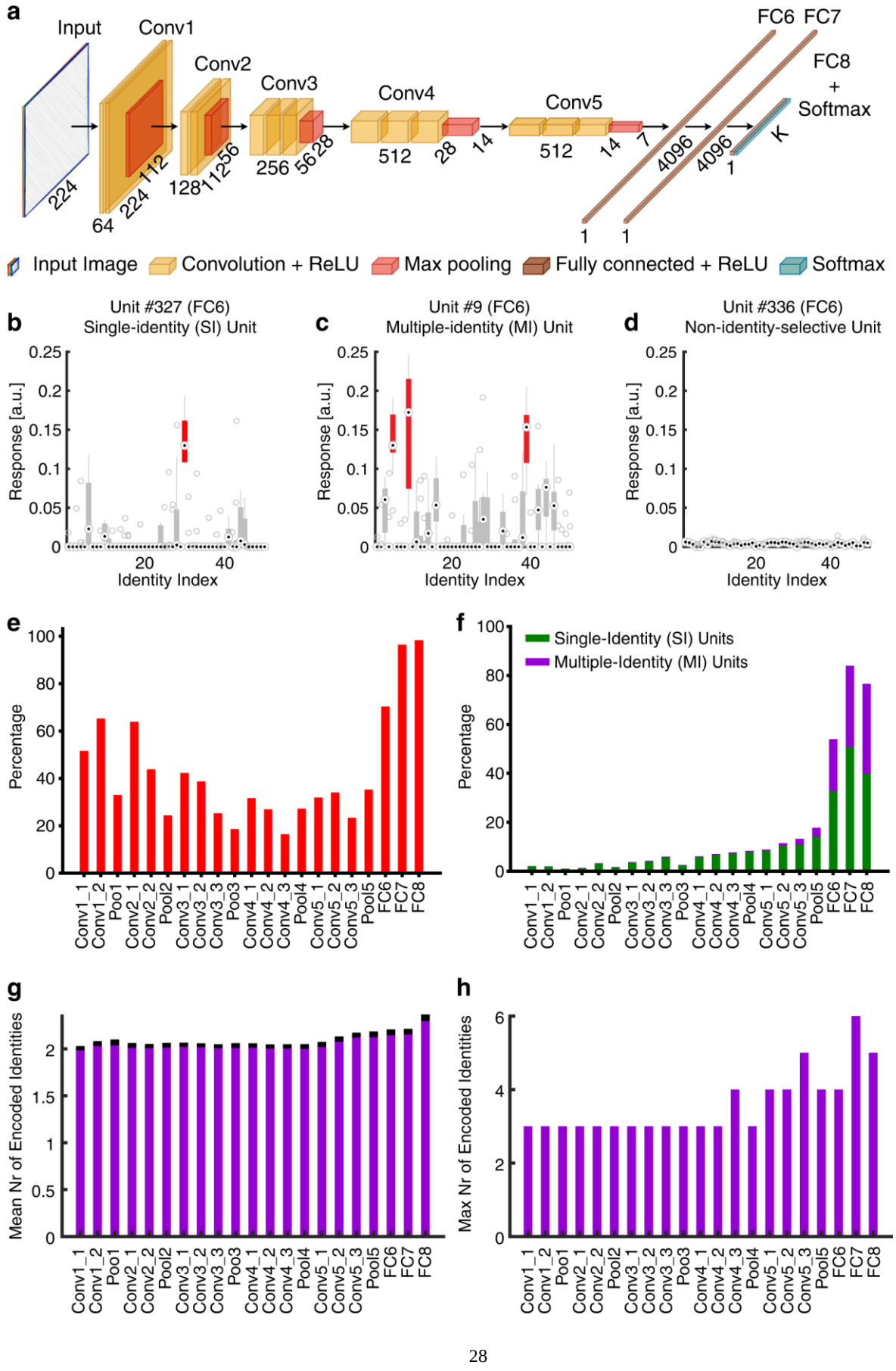

a
Input
Conv1
Conv2
Conv3
Conv4
Conv5
FC6
FC7
FC8
+
Softmax
224
64
224
112
128
112
56
256
56
28
512
28
14
512
14
7
4096
4096
1
K
1
1
Input Image
Convolution + ReLU
Max pooling
Fully connected + ReLU
Softmax
b
Unit #327 (FC6)
Single-identity (SI) Unit
c
Unit #9 (FC6)
Multiple-identity (MI) Unit
d
Unit #336 (FC6)
Non-identity-selective Unit
Response [a.u.]
Identity Index
e
Percentage
f
Single-Identity (SI) Units
Multiple-Identity (MI) Units
Percentage
g
Mean Nr of Encoded Identities
h
Max Nr of Encoded Identities
Conv1_1
Conv1_2
Pool1
Conv2_1
Conv2_2
Pool2
Conv3_1
Conv3_2
Conv3_3
Pool3
Conv4_1
Conv4_2
Conv4_3
Pool4
Conv5_1
Conv5_2
Conv5_3
Pool5
FC6
FC7
FC8

Figure 2.1 Identity-selective units in a pre-trained VGG-16 deep neural network (DNN).
a Structure of the VGG-16 DNN. The convolutional neural network (CNN) consisted of a feature extraction section (13 convolutional layers) and a classification section (three fully connected (FC) layers). The feature extraction section was consistent with the typical architecture of a CNN. A 3 × 3 filter with 1-pixel padding and 1-pixel stride was applied to each convolutional layer, which followed by rectified linear unit (ReLU) operation. Every convolutional block was followed by a max-pooling operation with a stride of two pixels. There were three FC layers in each classification section: the first two had 4096 channels each, and the third performed a K-way classification. Each FC layer was followed by a ReLU and 50% dropout to avoid overfitting. A nonlinear Softmax operation was applied to the final output of VGG-16 network to make the classification prediction of 50 identities. **b** An example of a single-identity (SI) unit. **c** An example of a multiple-identity (MI) unit. **d** An example of a non-identity-selective unit (i.e., the unit did not encode any particular identities). Shown are responses of DNN units to 50 identities (500 faces in total; 10 faces per identity) in arbitrary units (a.u.). On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, and the whiskers extend to the most extreme data points the algorithm considers to be not outliers. Encoded identities are shown in red. **e** Percentage of identity-selective units for each DNN layer. **f** Percentage of single-identity (SI; green) and multiple-identity (MI; purple) units for each DNN layer. **g** The average number of identities encoded by MI units. **h** The maximum number of identities encoded by MI units.

We identified a subset of DNN units that showed a significantly unequal response to different identities (one-way ANOVA of activation for each DNN unit: $P < 0.01$; **Supplementary Figure 2a**), and we refer to this population of units as identity-selective units (see **Figure 2.1b-d** for examples and **Figure 2.1e-g** for summary). There were identity-selective units in every layer (**Figure 2.1e**). On the one hand, the high proportion of identity-selective units in earlier DNN layers (**Figure 2.1e**) suggested that simple facial features (e.g., higher/lower contrast, more curves, more wrinkles, or more colorful makeup) could discriminate face identities. On the other hand, the higher percentage of identity-selective units in the later DNN layers (**Figure 2.1e**) was likely because the later DNN layers were closer to the output (note that the output was the identity of the input face). Interestingly, we observed that a substantial amount of units were selective to multiple identities (referred to here as multiple-identity [MI] units; **Figure 2.1c, f, g, h**; see **Methods** and **Supplementary Figure 2.2a** for selection procedure), consistent with prior studies with direct recordings from single neurons in the human brain. Compared to units that were selective to a single identity (referred to here as single-identity [SI] units; **Figure 2.1b, f**), the percentage of MI units increased in the later DNN layers (**Figure 2.1f**). Furthermore, the average number (**Figure 2.1g**) and maximum number (**Figure 2.1h**) of identities encoded by MI units showed that the "receptive fields" of the MI units increased from earlier layers to later layers. Lastly, some identities were encoded by more SI and MI units than the other identities (**Supplementary Figure 2b**).

2.2.2 Identity-selective DNN units demonstrated generalized selectivity to face identities

We first analyzed the selectivity properties of identity-selective units. We used a support vector machine (SVM) to assess to what extent units from a specific DNN layer could distinguish the input stimuli. First, with the original stimuli used to select identity-selective units, we found that the discriminability of all DNN units for face identities increased in the later DNN layers (**Figure 2.2a**). This was expected because the later DNN layers were closer to the output and contained more information about face identities. Notably, such discriminability was primarily driven by identity-selective units; and non-identity-selective units alone could not discriminate different identities (**Figure 2.2a**). Interestingly, identity-selective units had even better discriminability than all units in the earlier DNN layers (**Figure 2.2a**); and SI and MI units had even better discriminability than all identity-selective units (**Figure 2.2a**; also note that SI units had slightly better discriminability than MI units). It is worth noting that although the last five layers (Conv5_3, Pool5, FC6, FC7, and FC8) had a very different percentage of identity-selective units (**Figure 2.1e**), the identity discrimination performance was similar across these layers (**Figure 2.2a**; similarly for other stimuli as shown in **Figure 2.2**), indicating that the layer Conv5_3 might have already contained sufficient information for identity discrimination. The similar accuracy across all last five layers and unit types (all vs. identity-selective vs. SI vs. MI) indicated that the information for identity discrimination was already saturated.

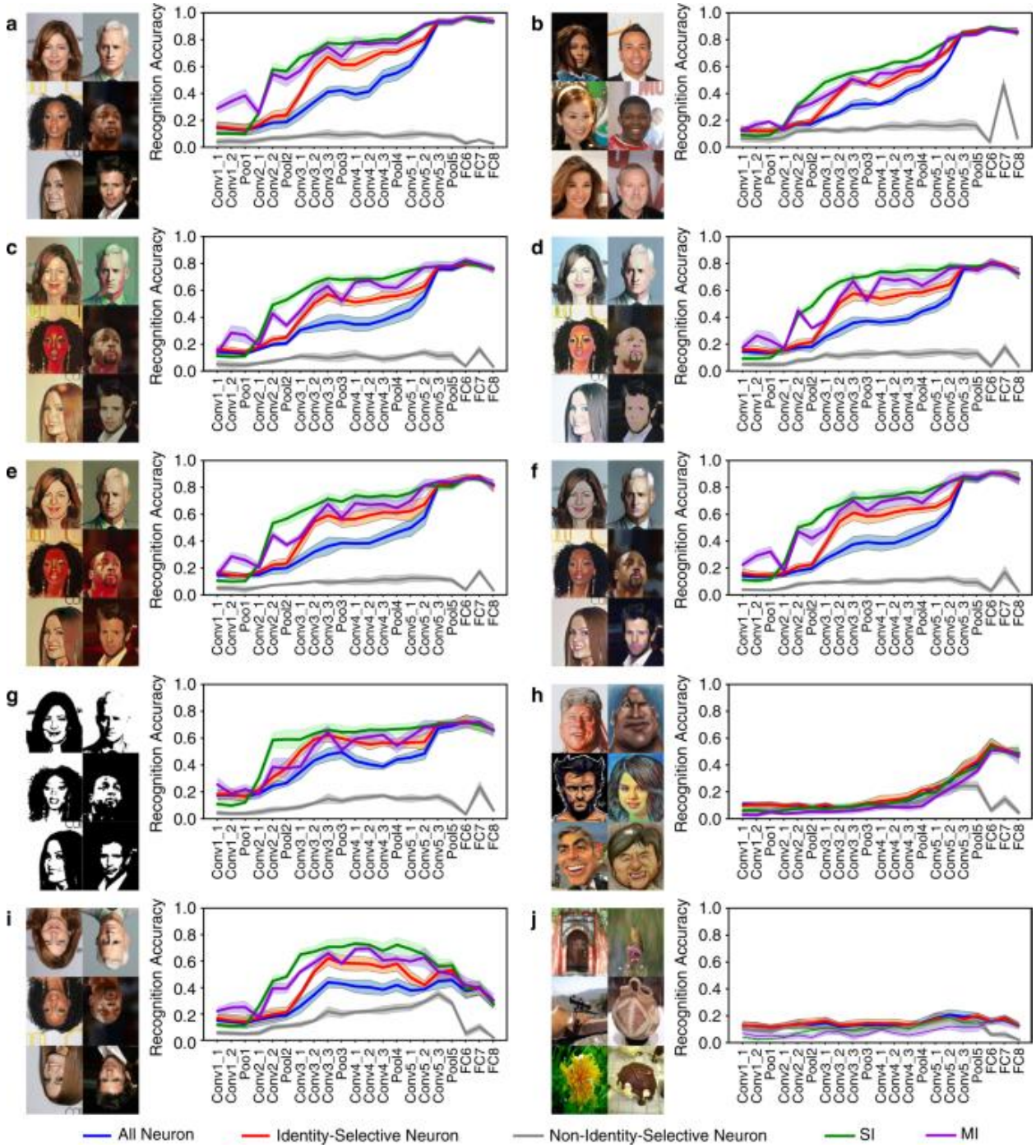


Figure 2.2 Selectivity properties of identity-selective units.
a Original faces used to identify identity-selective units. **b** Faces from a different set of 50 identities randomly selected from the CelebA database. **c** Original faces in the cartoon style Hayao. **d** Original faces in the cartoon style Hosoda. **e** Original faces in the cartoon style Paprika. **f** Original faces in the cartoon style Shinkai. **g** Original faces in the Mooney style. **h** A different set of celebrity caricature faces. **i** Original faces in inversion. **j** A set of non-face objects selected from the ImageNet stimuli. Identity recognition accuracy is shown for each deep neural network (DNN) layer. Error shade denotes one standard deviation across five-fold cross-validation. Blue: all units from each DNN layer. Red: identity-selective units. Gray: non-identity-selective units. Green: single-identity (SI) unit. Purple: multiple-identity (MI) units.

Second, we found that the DNN could discriminate identities in a different set of celebrity faces (**Figure 2.2b**) as well as the original celebrity faces transformed to various cartoon styles (**Figure 2.2c-f**), and the response profile was similar to that with the original stimuli (although cartoon faces had an overall reduced discrimination accuracy): identity-selective units primarily drove the discrimination and had even better discriminability than all units in the earlier DNN layers whereas non-identity-selective units could not discriminate identities in all these tests. Within identity-selective units, SI and MI units primarily drove the discrimination.

Third, we found that identity-selective units could still discriminate face identities even with very limited information in faces (**Figure 2.2g**). Although the accuracy was reduced to discriminate these low-information two-tone Mooney faces (generated from the original celebrity faces), a similar pattern of response was observed with identity-selective units (especially SI and MI units) primarily driving the discrimination.

Fourth, we found that identity-selective units could discriminate a different set of celebrity caricature faces in an exaggerated cartoon style (**Figure 2.2h**). Although the overall accuracy was reduced, identity-selective units still played the dominant role in discriminating the identities and non-identity-selective units did not contribute to the discrimination. It is also worth noting that identity-selective units (including SI and MI units) as well as all units could not discriminate face identities any more in earlier layers (**Figure 2.2h**).

Fifth, we found that the DNN could still discriminate inverted faces (**Figure 2.2i**), although the accuracy was reduced, consistent with impaired discrimination of inverted face in humans [46]. Such discrimination was again driven by identity-selective units (especially SI and MI units). However, we found that the DNN (all units, identity-selective units, SI units, MI units, and non-identity-selective units) could barely discriminate non-face object categories (**Figure 2.2j**). Therefore, the response could only be generalized within faces.

Lastly, we conducted the following control analyses. (1) We derived similar results when we equated the number of units per layer when comparing identity-selective and non-identity-selective units (**Supplementary Figure 3**), since classification performance could depend on the number of units (features). (2) We found that DNN units selective to identities from one race (e.g., Caucasian) could also discriminate face identities from other races, suggesting that the DNN and identity-selective units had cross-race generalizability. (3) We found that DNN units selective to identities from one gender (e.g., male) could also discriminate face identities from the other gender,

suggesting that the DNN and identity-selective units had cross-gender generalizability. (4) We found that combined SI and MI units (i.e., a subset of identity-selective units selected by an additional criterion that the response for certain identities stood out from the global mean response; see **Methods** and **Supplementary Figure 2a**) demonstrated even stronger discriminability of face identities (**Supplementary Figure 2c**). (5) Although most of our stimuli were frontal faces (**Supplementary Figure 1**), we confirmed that the DNN could well discriminate profile faces as well (**Supplementary Figure 4a, b**). (6) Using a VGG network pre-trained for ImageNet object stimuli [47, 48], we confirmed that the VGG network could discriminate the non-face object categories (**Supplementary Figure 4c**; see **Figure 2.2j** for a comparison). (7) We employed different DNNs and found that our findings could generalize to other DNNs (**Supplementary Figure 5**; note that as expected face recognition performance was reduced in some DNNs that were not trained for face recognition).

Together, our results showed that identity-selective units played a general and critical role in discriminating face identities under various circumstances, whereas non-identity-selective units could discriminate face identities in none of the circumstances. Therefore, our results suggested that although the discriminability varied as a function of the level of information contained in the stimuli, a subset of DNN units were consistently involved in the face identity discrimination and these units demonstrated generalized selectivity to face identities.

2.2.3 DNN visualization explained the role of identity-selective units in face recognition

We next visualized the response of identity-selective vs. non-identity-selective units in order to understand why identity-selective units but not non-identity-selective units played a general and critical role in face recognition. First, we found that identity-selective units indeed corresponded to the critical visual features of the stimuli such as the eyes, nose, and mouth (**Figure 2.3a**). Second, when we constructed a two-dimensional stimulus feature space using t-distributed stochastic neighbor embedding (t-SNE) feature reduction for each DNN layer, we found that face identities clustered in the feature space constructed by identity-selective units but not non-identity-selective units (**Figure 2.3b**), confirming that identity-selective units could discriminate face identities. Similar results were derived if we constructed a three-dimensional feature space or used different perplexity parameters for t-SNE (balance between local and global aspects of the data).

We could also replicate our results in the full dimensional space of the DNN. Together, DNN visualization revealed that identity-selective units encoded critical stimulus features so that they embodied a general discriminability of face identities.

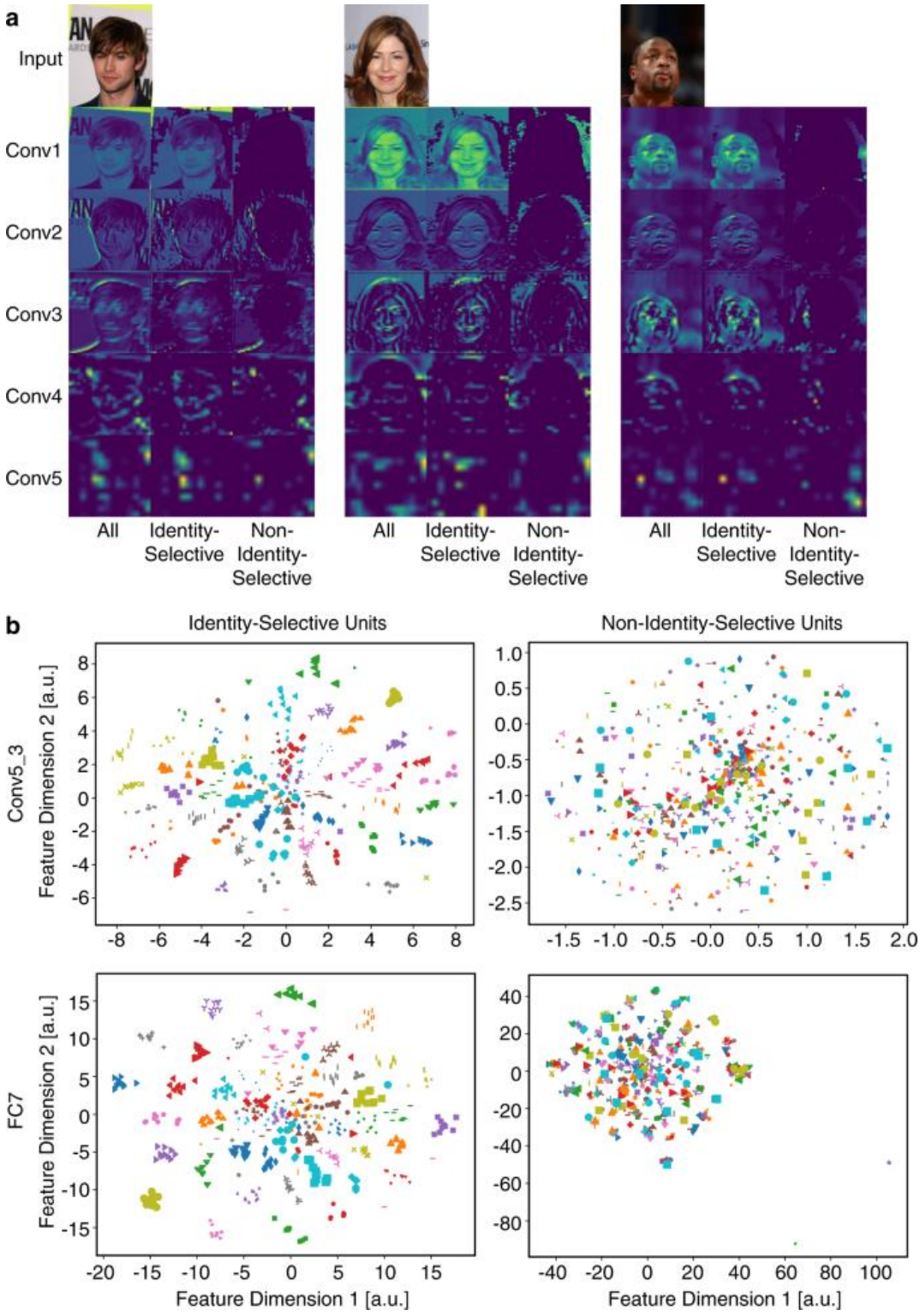


Figure 2.3 Visualization of the deep neural network (DNN).
a Example activation maps for a subset of DNN layers. Left: all units. Middle: identity-selective units. Right: non-identity-selective units. **b** Visualization in the t-SNE space for DNN layers Conv5_3 (upper row) and FC7 (lower row). Each color-shape combination represents a face identity (ten faces per identity). The feature dimensions are in arbitrary units (a.u.). Left column: identity-selective units. Right column: non-identity-selective units.

2.2.4 Lesion and perturbation of the network

We next investigated how critical the identity-selective units as well as the trained network weight structure were to discriminate face identities by lesioning and perturbing the network.

First, following each convolutional layer, we added a "RandDrop" layer (i.e., a binary mask applied to the preceding layer; **Figure 2.4a**) to randomly set a subset of DNN units to be 0, which partially lesioned the network. Indeed, we found that identity recognition accuracy decreased as a function of increasing lesion amount (**Figure 2.4b-f**). With a small amount (10%) of information loss (**Figure 2.4b**), the network could still well discriminate face identities and only had a small decrease in performance compared to the intact network (**Figure 2.2a**; the decrease in performance was primarily in later layers). When 30%-50% of DNN units were dropped (**Figure 2.4c, d**), only earlier layers had a comparable performance as the intact network but later layers had a significant decrease of performance. When 70%-90% of DNN units were dropped (**Figure 2.4e, f**), the network could not perform identity discrimination any more. Notably, in all these cases, identity-selective units still played the dominant role in discriminating the identities and non-identity-selective units did not contribute to the discrimination. It is also worth noting that we dropped the same percentage of units for every layer, so information loss would accumulate in later layers. Interestingly, we found that dropping units from a single layer had only limited impact on recognition performance in subsequent layers (**Supplementary Figure 6**; 30% of units were dropped; only dropping units from the very first layer Conv1_1 impacted performance in subsequent layers; no re-training was involved), suggesting that the network had great plasticity.

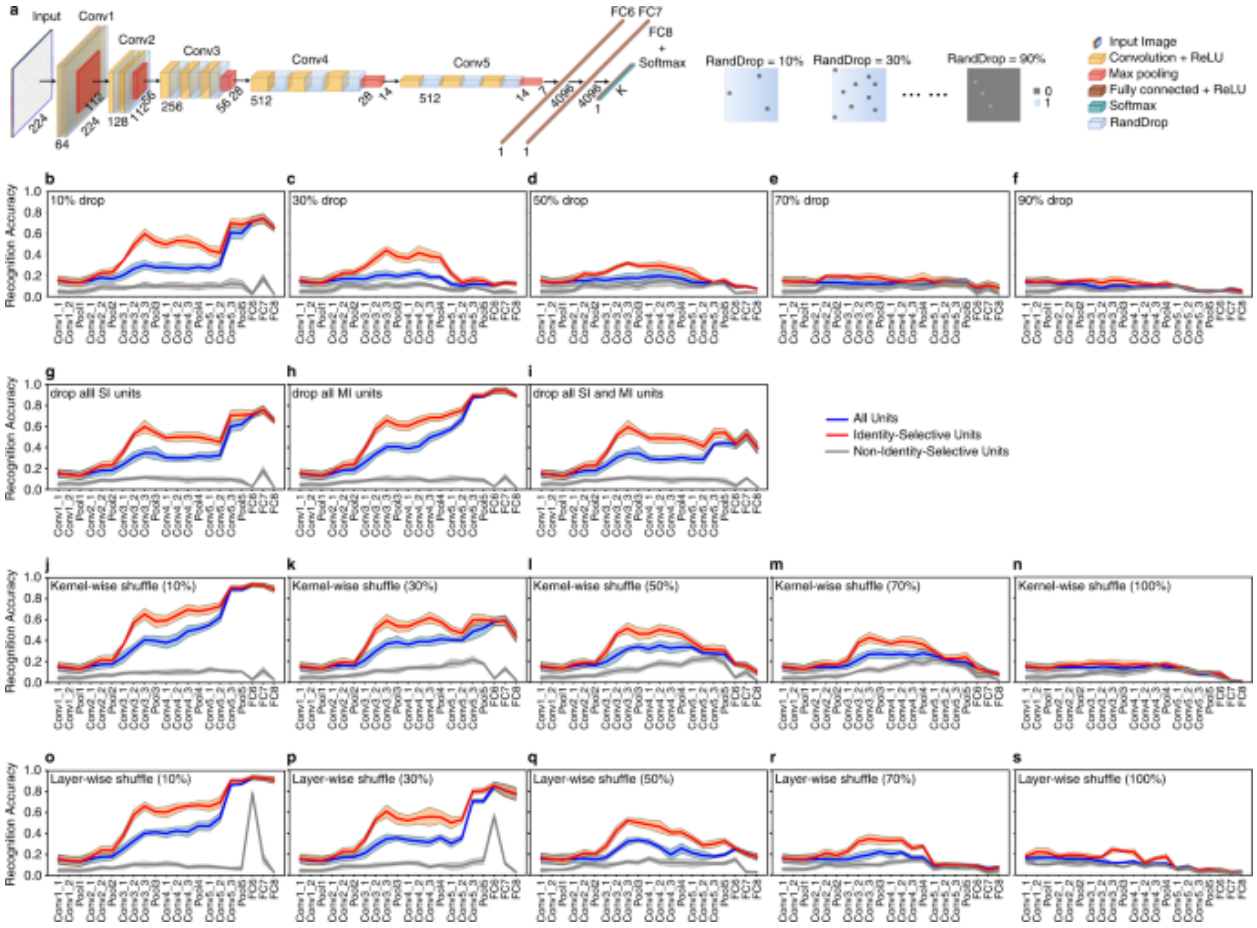


Figure 2.4 Manipulation of the deep neural network (DNN).
a Illustration of how random dropout of DNN units was performed. We added a binary mask following every convolutional layer in the original DNN architecture to randomly deactivate a subset of DNN units. The percentage of dropped DNN units was controlled by the percentage of zeros in the binary mask and varied from 10 to 90%. **b**–**f** Recognition accuracy following a random dropout of DNN units. **g**–**i** Recognition accuracy following a complete dropout of (**g**) SI units, (**h**) MI units, and (**i**) combined SI and MI units. **j**–**n** Recognition accuracy following kernel-wise shuffle. **o**–**s** Recognition accuracy following layer-wise shuffle. Legend conventions as in Figure. 2.2.

Furthermore, we investigated the impact of SI and MI units on identity discrimination by dropping these units (**Figure 2.4g-i**). We found that when dropping SI units (**Figure 2.4g**) and MI units (**Figure 2.4h**) alone, the network could still well discriminate face identities, indicating that SI and MI units could well complement each other. The decrease in performance was primarily in the later layers (likely due to accumulation of information loss, as shown above), and the earlier layers had a comparable performance as the intact network (**Figure 2.2a**). Dropping all SI units (**Figure 2.4g**) had a similar impact as dropping 10% of all units (**Figure 2.4b**), whereas dropping all MI units (**Figure 2.4h**) had a less impact and led to a performance comparable to the intact network (**Figure 2.2a**), indicating that MI units were less important compared to SI units in identity

discrimination (see also **Figure 2.2** where SI units had a slightly better identity discriminability compared to MI units). As expected, when we dropped both SI and MI units (**Figure 2.4i**), the performance further decreased, leading to a performance comparable to dropping 30% of all units (**Figure 2.4c**).

Second, we perturbed the network by breaking the optimal weight structure (i.e., connection between DNN units) derived from training and generated a random permutation of weights (i.e., connections between DNN units) to evaluate the impact of the model training on identity recognition. We employed two approaches. Kernel-wise shuffle randomly permuted the weights in a single kernel. Similar to the above dropout results, the DNN gradually lost the ability to discriminate face identities with increasing levels of kernel-wise shuffle (**Figure 2.4j-n**). Unlike kernel-wise shuffle that only rearranged the weights within one kernel, layer-wise shuffle pooled the weights of all kernels from a layer and reorganized the weights to form new kernels. Again, we found that with increasing levels of layer-wise shuffle, the DNN's ability to discriminate face identities decreased and was eventually abolished (**Figure 2.4o-s**).

Together, our model manipulation suggested that identity-selective units as well as the optimal DNN unit connections derived from training were critical to face identity discrimination.

2.2.5 Establishing the relationship between artificial DNN units and real monkey neurons

It has been suggested that DNNs share similarities with the primate visual cortex and can therefore help us better understand the sensory cortex [49]. Here, we explored the relationship between artificial DNN units and real primate neurons. We first analyzed whether the DNN had a similar encoding as monkey inferotemporal (IT) neurons. We used the same stimuli (500 natural face images of 50 celebrities) and recorded neuronal activity using two Utah arrays in the anterior and central IT cortex (see **Methods**) while the monkey performed a passive viewing task (**Figure 2.5a**). We identified 53 multi-unit activity (MUA) channels that showed sufficient internal consistency and we focused on these channels for further analysis.

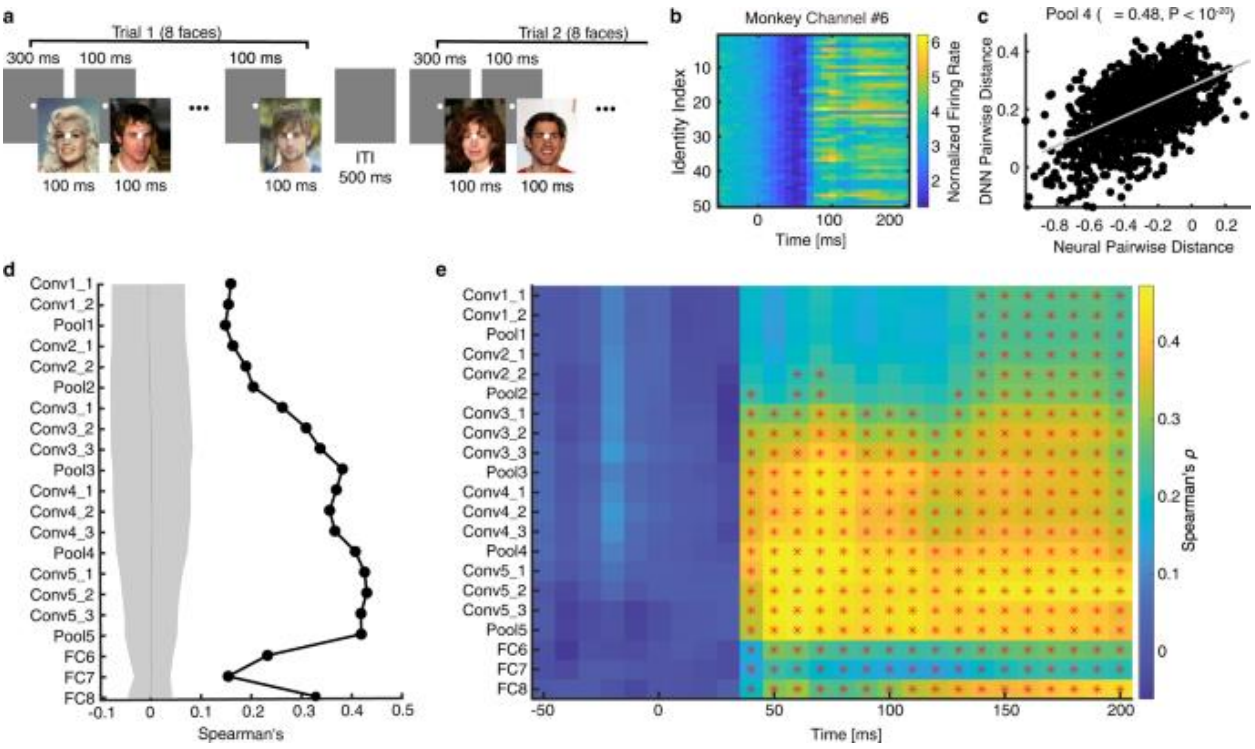


Figure 2.5 Comparison between the deep neural network (DNN) units and (real) monkey inferotemporal (IT) cortical neurons.

a Task used to acquire neural responses from a monkey. In each trial, eight faces were presented for 100 ms each, followed by a fixed interstimulus interval (ISI) of 100 ms. There was a central fixation point of 300 ms at the beginning of each trial and there was an intertrial interval (ITI) of at least 500 ms following each trial. The central fixation point persisted through the trial. **b**, **c** An example multiunit-activity (MUA) channel showing a significant correspondence with the DNN feature space. **b** MUA to 50 identities, shown in 10 ms time bins. Time 0 denotes the stimulus onset. Firing rate was normalized to the average of the gray images (i.e., control stimuli). **c** Correlation between MUA pairwise distance and DNN layer Pool4 feature pairwise distance (see "Methods"). Each dot represents a face pair, and the gray line denotes the linear fit. **d** Correlation between pairwise distance in the monkey inferotemporal (IT) neuronal face space and pairwise distance in the DNN face space. Here, we used the mean firing rate in a time window 70 ms to 180 ms after stimulus onset as the response to each face, and we averaged the responses to ten faces for each face identity. We calculated the correlation using all identities. Solid circles represent a significant correlation (permutation test: $P < 0.05$, Bonferroni correction across layers). The shaded area denotes ±SD across permutation runs. **e** Temporal dynamics of correlation of pairwise distance between monkey neurons and DNN units (bin size = 40 ms, step size = 10 ms). Color coding indicates the magnitude of Spearman's correlation. Asterisks (*) indicate a significant correlation in that bin (permutation test: $P < 0.05$, Bonferroni correction across time bins for each layer).

We found that IT MUA not only showed face responsiveness (**Figure 2.5b**) but also encoded the geometry of DNN layers (**Figure 2.5c**). To formally quantify this result at the group level, we used DNN units to construct a face space and correlated that with the IT neuronal face

space using a distance metric (see **Methods**). Using pairwise activation similarities [50], we found that the pairwise distance from the intermediate to later DNN layers significantly correlated with the neuronal pairwise distance from the monkey IT cortex (**Figure 2.5d**; non-parametric permutation test: $P < 0.05$, Bonferroni correction), suggesting that the population of DNN units encoded the geometry of the face space similarly as monkey IT neurons. We further investigated the temporal dynamics of the correspondence between face spaces and found a strong correlation starting from ~50 ms after stimulus onset (**Figure 2.5e**; permutation test: $P < 0.05$, Bonferroni correction across time bins for each DNN layer; note that a moving time window was used), consistent with the response latency of monkey IT neurons. Therefore, with our direct recordings from the monkey IT cortex using the same stimuli, we showed that the DNN shared a similar encoding of faces as the monkey IT cortex.

2.2.6 Establishing the relationship between artificial DNN units and real human neurons

The DNN performs the face recognition task similarly as humans. Does the ensemble of DNN units share representational similarity with the ensemble of human neurons? In order to answer this question, we used the same stimuli (500 natural face images of 50 celebrities) and recorded from 667 neurons in the MTL (340 neurons from the amygdala, 222 neurons form the anterior hippocampus, and 105 neurons from the posterior hippocampus; firing rate > 0.15 Hz) of 8 neurosurgical patients (23 sessions in total) [39]. Patients performed a one-back task (**Figure 2.6a**; accuracy = 77.38%±4.94% [mean±SD across sessions]) and they could well recognize the faces [39]. The responses of 76/667 neurons (11.39%) differed between different face identities in a window 250-1250 ms following stimulus onset and these neurons were the real human identity-selective neurons (see **Figure 2.6b-d** for example neurons). We grouped amygdala and hippocampal neurons as a single neuronal population (i.e., MTL neurons) for further analysis because they show very similar identity selectivity responses.

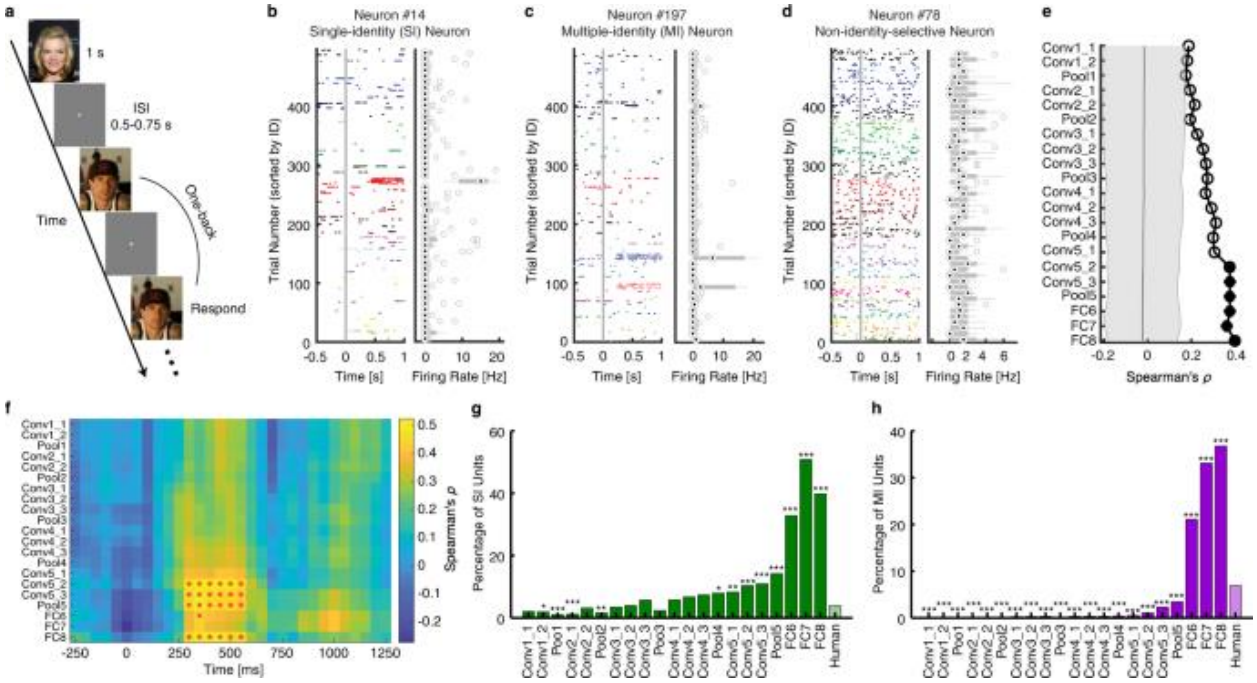


Figure 2.6 Comparison between the deep neural network (DNN) units and (real) human medial temporal lobe (MTL) neurons.

a Task used to acquire single-neuron responses from humans. We employed a one-back task, in which patients responded whenever an identical famous face was repeated. Each face was presented for 1 s, followed by a jittered interstimulus interval (ISI) of 0.5–0.75 s. **b**–**d** Neuronal responses to 500 faces (50 identities) shown in raster plots. **b** An example of a single-identity (SI) neuron. **c** An example of a multiple-identity (MI) neuron. **d** An example of a non-identity-selective neuron (i.e., the neuron did not encode any particular identities). Trials are aligned to face stimulus onset (gray line) and are grouped by individual identity. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, and the whiskers extend to the most extreme data points the algorithm considers to be not outliers. **e** Correlation between pairwise distance in the human MTL neuronal face space and pairwise distance in the DNN face space. Here, we used the mean firing rate in a time window 250–1000 ms after stimulus onset as the response to each face, and we averaged the responses to 10 faces for each face identity. We calculated the correlation using the top ten identities that were most frequently encoded by MTL neurons. Solid circles represent a significant correlation (permutation test: $P < 0.05$, corrected by false discovery rate (FDR)35 for $Q < 0.05$) and open circles represent a non-significant correlation. The shaded area denotes ±SD across permutation runs. **f** Temporal dynamics of correlation of pairwise distance between human neurons and DNN units (bin size = 500 ms, step size = 50 ms). Color coding indicates Spearman's correlation coefficient. Asterisks (*) indicate a significant correlation in that bin (permutation test: $P < 0.05$, FDR corrected across time bins for each layer). **g** Percentage of SI units in each DNN layer and comparison with SI neurons from the human MTL. **h** Percentage of MI units in each DNN layer and comparison with MI neurons from the human MTL. Asterisks indicate a significant difference in the percentage using χ^2-test with Bonferroni correction for multiple comparisons. $*P < 0.05$, $**P < 0.01$, and $***P < 0.001$.

Similar to our analysis of IT MUA, we used DNN units to construct a face space and correlated that with the MTL neuronal face space using a distance metric (see **Methods**). We found that the pairwise distance from the later/top DNN layers significantly correlated with the neuronal

pairwise distance from the human MTL (**Figure 2.6e**; non-parametric permutation test: $P < 0.05$, corrected by false discovery rate [FDR] [51] for $Q < 0.05$), suggesting that the population of DNN units encoded the geometry of the face space similarly as human MTL neurons. Notably, compared to IT neurons that had a drop in correspondence with DNN features from the later layers (**Figure 2.5d**), MTL neurons had a smooth increase of correspondence and peaked at the top/output layer (**Figure 2.6e**), consistent with the different processing stages along the ventral visual pathway. Here, we calculated the correlation between DNN units and human neurons using the 10 identities that were most frequently encoded by MTL neurons, given the sparseness of MTL responses; but we derived a similar pattern of results using all 50 identities. We further investigated the temporal dynamics of the correspondence between face spaces and found a significant correlation in later DNN layers starting from 250 ms after stimulus onset (**Figure 2.6f**; permutation test: $P < 0.05$, FDR corrected across time bins for each DNN layer), consistent with the response latency of human MTL neurons [52].

In addition, we found that compared to the human MTL, the DNN had a significantly higher percentage of SI units in later layers (starting from the layer Pool4; **Figure 2.6g**; χ^2-test with Bonferroni correction for multiple comparisons). The DNN also had a significantly higher percentage of MI units in the layers FC6, FC7, and FC8 (**Figure 2.6h**) but a significantly lower percentage of MI units in all other layers, where we did not expect to observe MI units because faces of the same identity were not yet clustered.

Together, for the first time, using human single-neuron recordings we directly compared identity selectivity between artificial units and real human neurons. Our results have revealed a systematic correspondence between the two face recognition systems.

2.2.7 Region-based feature coding in DNN units and a mechanism underlying face recognition

How does the DNN transition from representing visual features (in earlier and intermediate layers) to representing identities (in later and output layers)? Inspired by the primate visual system, one possible mechanism is that earlier DNN layers encode the axes of a face space and provide information to later DNN layers, which encode a region in the high-level feature space and are selective to identities that fall in this region. This mechanism has been instantiated in the primate brain: IT neurons encode visual features and axes of the feature space (notably, the axes of our

CelebA face space in the current study has been shown [39]) whereas MTL neurons are selective to specific face identities [53]. Our recent study has supported such transition by revealing a novel *region-based feature coding* by real human MTL neurons [39], i.e., human MTL neurons encode a specific region in the feature space and are selective to identities that are clustered in this region. Here, we further compared between artificial and primate neural systems and explored whether DNN units also demonstrated region-based feature coding, which will provide critical insights into the computational mechanisms underlying face recognition.

We focused on the DNN layers FC6 and FC7, where faces of the same identity was clustered; and we primarily observed human MTL neurons demonstrating region-based feature coding in the face feature spaces from these layers [39]. Notably, the face feature spaces showed an organized structure: for example, Feature Dimension 2 represented a gender dichotomy, and darker skinned faces were clustered at the bottom left corner of the feature space (**Figure 2.7a**). Indeed, we found that a large number of MI units in these layers (24.9% for FC6 and 36.5% for FC7) encoded identities that were adjacent in the feature space (see **Figure 2.7b, c**), demonstrating region-based feature coding. We refer to this subpopulation of MI units as feature MI units. On the other hand, non-feature MI units did not have selective identities clustered in the feature space (**Figure 2.7d**), and non-identity-selective units did not encode any particular identities (**Figure 2.7e**).

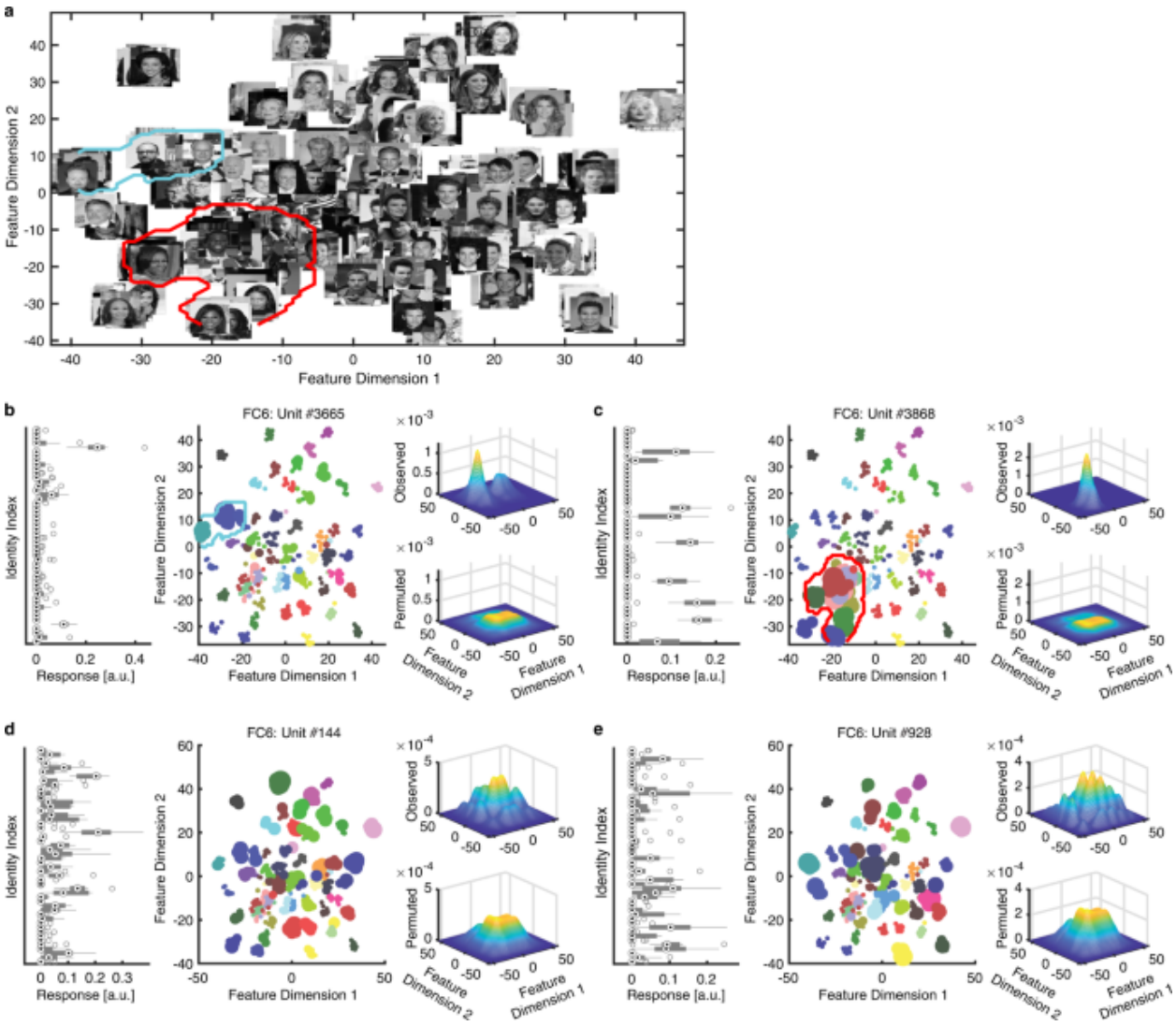


Figure 2.7 Region-based feature coding in DNN units.
a The face feature space constructed by t-distributed stochastic neighbor embedding (t-SNE) for the DNN layer FC6. All stimuli are shown in this space in grayscale. **b**–**e** Example DNN units. **b**, **c** Two examples of MI units that demonstrated region-based feature coding (i.e., the units encoded a region in the feature space). **d** An example MI unit that did not demonstrate region-based feature coding (i.e., the encoded identities were not adjacent to each other in the feature space). **e** A non-identity-selective unit (i.e., the unit did not encode any particular identities). All examples were from the DNN layer FC6. Left: Response of DNN units to 50 identities (500 faces in total; ten faces per identity) in arbitrary units (a.u.). On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, and the whiskers extend to the most extreme data points the algorithm considers to be not outliers. Middle: Projection of the DNN activation onto the feature space. Each color represents a different identity. The size of the dot indicates the level of activation. Right: Estimate of the spike density in the feature space. By comparing observed (upper) vs. permuted (lower) responses, we could identify a region where the observed response was significantly higher in the feature space. This region was defined as the tuning region of a unit (delineated by the red/cyan outlines).

At the population level, we found that the tuning region of an individual feature MI unit covered approximately 5-6% of the 2D feature space (**Figure 2.8a**; note that when we calculated the tuning region, we adjusted the kernel size to be proportional to the feature dimensions such that the percentage of space coverage was not subject to the actual size of the feature space). In contrast, the response of an individual SI or non-feature MI unit covered a significantly smaller region in the feature space (**Figure 2.8a**; two-tailed unpaired *t*-test: $P < 0.001$ for all comparisons). As expected, the distance in the face space between encoded identities was smaller for feature MI units compared with non-feature MI units (**Figure 2.8b**). As a whole, the entire population of DNN units covered approximately 55-60% of the feature space (**Figure 2.8c**; some areas were encoded by multiple units), and the covered areas were similar for SI, feature MI, and non-feature MI units (**Figure 2.8c**).

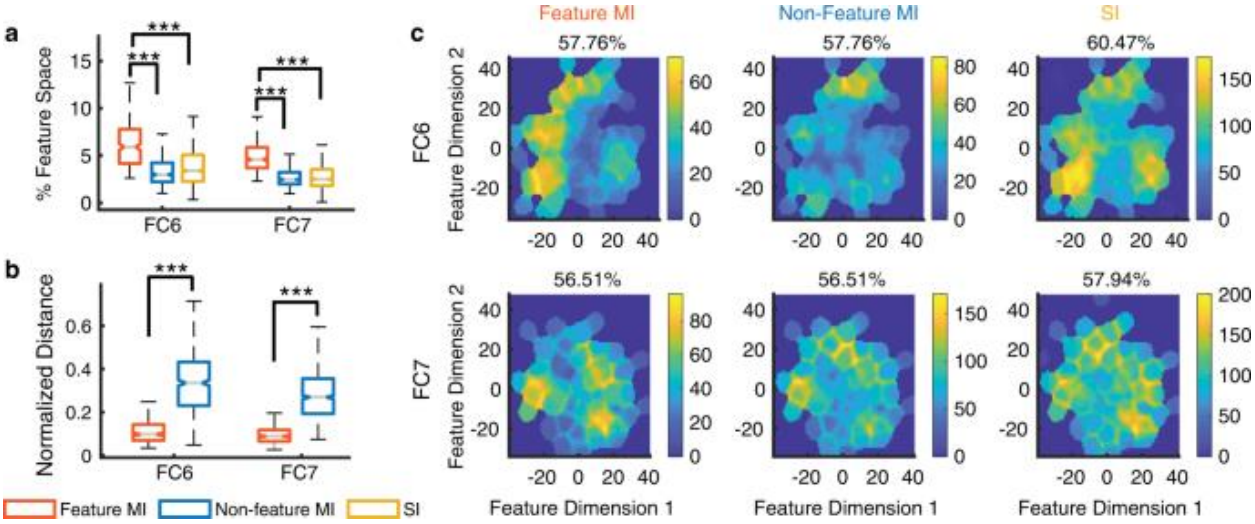


Figure 2.8 Summary of region-based feature coding for SI and MI DNN units.
a Percentage of feature space covered by tuning regions of SI and MI units. Note that here we did not apply the threshold for minimal cluster size for SI and non-feature MI units in order to compare between different types of units. **b** Normalized distance between MI unit's selective identities in the feature space. To be comparable for different layers, Euclidean distance was normalized by the maximum distance (i.e., diagonal line) of the feature space. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, and the whiskers extend to the most extreme data points the algorithm considers to be not outliers. Asterisks indicate a significant difference between feature MI units and non-feature MI units using two-tailed unpaired *t* test. ***$P < 0.001$. **c** The aggregated tuning regions of the population of units. Color bars show the counts of overlap between individual tuning regions. The numbers above the density map show the percentage of feature space covered by the tuning regions of the population of units.

We conducted several control analyses to ensure that our findings were robust in regard to the construction of the feature space. (1) We derived similar results if we constructed a three-dimensional feature space, or used different perplexity parameters for t-SNE or kernel/cluster size

parameters to detect a tuning region. (2) We derived similar results if we constructed the feature space using uniform manifold approximation and projection (UMAP) or principal component analysis (PCA). (3) We could replicate our findings using full DNN features, where the Euclidian distance between encoded identities was significantly smaller than that of non-encoded identities.

Together, we found that as neurons in the human MTL, DNN units also demonstrated region-based feature coding. Together with the default axis-based coding (i.e., DNN units encode a linear combination of features from a previous layer) and exemplar-based coding (i.e., DNN units output a face identity with visual invariance of the input images) in the DNN, our findings may provide an important mechanism that explains how the DNN transitions from representing visual features to representing identities and thus performs face recognition tasks. Importantly, our results suggest that artificial and primate neural systems share similar computational mechanisms for face recognition.

2.3 Discussion

In this study, we analyzed the response characteristics of a face recognition DNN and found that identity-selective units in the DNN could generalize their discriminability to face identities shown in various styles as well as face identities that were not involved in the training. Visualization and manipulation of the DNN showed the importance of identity-selective units in face recognition. By establishing the coding similarity with real primate neurons, our study provided an important method to understand face coding in primates. Furthermore, by analyzing an artificial neural network dedicated to face recognition, we will be able to formulate hypotheses that can be validated in the primate brain.

2.3.1 Possible caveats

In this study, we focused on analyzing the VGG-face model, which was pre-trained for face recognition and expected to contain identity-selective units. Using this functioning face recognition model that is highly capable of processing face identity information enabled us to (1) test the generalizability of identity response to different categories of stimuli, (2) compare with primate visual systems, and (3) visualize and perturb the network to reveal the critical features for face identity discrimination. It is worth noting that here we also explored other DNN models (**Supplementary Figure 5**) and found that our findings could generalize to other DNNs (although

as expected face recognition performance was reduced in some DNNs that were not trained for face recognition), consistent with a previous report surveying a large class of DNN models for face representation [54]. Interestingly, a recent study has even shown that face-selective units can emerge from an untrained DNN [55]. A future study will need to investigate whether identity-selective units can emerge from an untrained DNN.

We found that although cartoon faces had a decreased discriminability in general, they had a similar pattern of response as natural faces across DNN layers (**Figure 2.2a, c-f**). Furthermore, we found that inverted faces elicited a similar response as upright faces in the early and middle DNN layers (**Figure 2.2a, i**), suggesting that the DNN used similar information for both inverted and upright faces, a form of viewpoint invariance. However, in contrast to all upright faces that had increasing discriminability across layers, the discriminability decreased in later layers for inverted faces. Therefore, the impaired discriminability of inverted faces in humans may stem from neurons downstream in the visual processing stream [46]. Furthermore, consistent with our DNN lesion results (**Supplementary Figure 6**), it has been shown that DNN units demonstrate distributed and sparse codes to represent different face attributes [41]. Lastly, our present results may depend on the hyper-parameters used in the study (batch-norm vs. drop-out, pooling, architecture, dataset size, etc) and our results should be interpreted in the context of our DNN architecture and set of hyper-parameters.

Although the SVM had substantially more features (i.e., DNN units used for classification) than observations (i.e., training faces), the high recognition accuracy of identity-selective units in testing suggested that our results could not be simply explained by overfitting. Furthermore, the low recognition accuracy of non-identity-selective units in testing provided specificity of our approach; and notably, later DNN layers had fewer units but a higher recognition accuracy. Interestingly, such "overfitting" also appears in the human brain as a large number of neurons are often simultaneously activated by a single percept; and the theory of backward feature correction can well explain such "overfitting" [56]. It is also worth noting that we derived similar results when we input the same number of identity-selective units and non-identity-selective units to the SVM (**Supplementary Figure 3**), suggesting that the difference in encoded information between unit groups could not be attributed to different number of input units.

2.3.2 Identity selectivity in the DNN

We identified the DNN units critical for identity recognition, and the evolution of identity-selective units across DNN layers (**Figure 2.1e**). The identity-selective units in the earlier layers primarily corresponded to image pixels containing information about faces, but with the increase of kernel size in later layers, identity-selective units encoded more holistic information about face identities. In particular, the fully connected layers utilized information from all units from the previous layers. Therefore, identity selectivity could not be solely attributed to the receptive field of the DNN units. It is worth noting that our present results were not about *face* selectivity (i.e., contrasting response between faces vs. objects) but *identity* selectivity (i.e., contrasting response between face identities, which does not require face selectivity [39]). However, we found that the response of face identity-selective units could well generalize within faces but barely generalize to non-face objects, consistent with a dedicated and specialized face perception system. It is also worth noting that most monkey MUA channels showed strong face responsiveness (i.e., modulation by face onset; e.g., **Figure 2.5b**), consistent with previous studies [30].

We observed both identity-selective units that were selective to a single identity (SI units) and identity-selective units that were selective to multiple identities (MI units), analogous to the SI neurons and MI neurons from the human brain. The SI units and MI units identified in the present study are also reminiscent of the concept cells of the exemplar-based model [57]. Concept cells primarily appear in the human medial temporal lobe, respond in a remarkably selective and abstract manner to particular persons or objects, may be crucial for memory formation [57]. Given that we here show computational similarities between the primate visual system and the DNN, future studies can benefit from the computational architecture of the DNNs and address two important questions: (1) how concept neurons arise computationally (i.e., transition from representation of visual features to representation of concepts), and (2) how concept neurons create associations and transits between related concepts to form episodic memories.

2.3.3 Comparing artificial versus primate visual systems for face identity coding

We found that the DNN shared a similar coding with both monkey IT cortex and human MTL (see also [58]). Specifically, the intermediate to later DNN layers corresponded to the IT neuronal space whereas the later/top DNN layers corresponded to the MTL neuronal space, consistent with the ventral visual processing pathway in the primate brain [59, 60]. In particular,

in addition to the axis-based coding (i.e., DNN units encode a linear combination of features from a previous layer) and exemplar-based coding (i.e., DNN units output a face identity with visual invariance of the input images) in the DNN as observed in the IT cortex and MTL [53], respectively, we confirmed the region-based coding in the DNN, which is an important mechanism that bridges the representation of visual features and the representation of identities. It is worthing that the correlation strength with DNN layers was different between the monkey (**Figure 2.5**) and human (**Figure 2.6**) visual systems, which was likely due to differences in recording method (Utah array vs. microwire), noise level in recordings, repetition (repeated vs. single) and duration (100 ms vs. 1 s) of stimulus presentation, recording location (cortical vs. subcortical), and spike sorting (multi-unit vs. single-unit), so the correlation was not directly comparable between the IT cortex and the MTL. However, for each visual system, we found a significant correlation; and importantly, we found that the most strongly encoded DNN layer differed between brain areas in accordance with the ventral visual stream.

We found that identity-selective units had a general discriminability to face identities shown in various styles, consistent with feature-invariant coding of face identities by neurons in the human medial temporal lobe (MTL). Furthermore, identity selectivity could be generalized to face identities that were not involved in the training, similar to how memory is formed in the human brain [61]. Consistent with our prior findings that only a small proportion (~20%) of human neurons are involved in coding a certain task aspect, such as emotion content [62], emotion subjective judgment [63], attention, task sequence [64], visual selectivity [65], eye movement [66], social judgment [67], as well as face identity [39], in the present study we found a large population non-identity-selective DNN units that did not contribute to coding face identities. On the other hand, we quantitatively compared the proportion of SI and MI units/neurons between the DNN and human brain and found that the DNN in general had a higher proportion of identity-selective units than the human MTL (**Figure 2.6f, g**). This was likely because the human MTL is involved in many aspects of cognitive functions whereas the DNN has only been optimized to recognize face identities. Notably, the distribution of identity-selective and non-identity-selective DNN units across layers may provide new insights into understanding the human visual processing stream, where our currently available technology does not allow simultaneous sampling of neurons along the entire visual processing stream.

2.3.4 Contribution of DNNs to understanding primate visual processing

Rapid advances in DNNs have offered new opportunities for studying face perception by providing computational proxy models. State-of-the-art DNNs such as the VGG-face [68] and DeepFace [69] have achieved excellent face recognition performance and even outperformed humans. These DNNs are biologically inspired and therefore have the potential to successfully provide insight into the underlying mechanisms of brain functions, especially with respect to the perception and recognition of visual stimuli such as faces. Existing work at the intersection of DNNs and face perception can be broadly classified into two categories: face reconstruction (decoding models) and face recognition (encoding models). The former includes the reconstruction of faces from fMRI patterns (see [70], [71] for more general natural image reconstruction). The latter includes a flurry of literature on the convergent evolution of face spaces across DNN layers and human face-selective brain areas [50], the neurally plausible efficient inverse graphics model for face processing [72], and spontaneous generation of face recognition in untrained DNNs [55]. Our recent study showing feature-based encoding of face identities in the human MTL using DNN-extracted visual features [39] also employed an encoding model.

DNNs have the following advantages to help us better understand visual processing in primates [49]. First, most previous studies of face space had to use computer-generated faces in order to parametrically vary the faces but DNNs are able to extract features from real human faces and subsequently manipulate these features to generate new unique faces while providing well controlled stimuli to investigate differences in neural responses to feature changes. Second, DNNs have simultaneous access to the activity of the entire face processing neural network (i.e., the whole population of neurons from all brain regions involved in face processing) [35], which is particularly useful to study visual processing pathways (e.g., how the brain transitions from one face coding model to another). Third, we are readily able to selectively manipulate DNN units to study a causal effect. These advantages have been embodied in our present study.

2.3.5 Future directions

Our present study points to several future directions. First, most faces in the present study had a frontal view and were primarily emotionally neutral. A future study will further test faces from different angles (e.g., profile faces) and/or emotional faces. Second, an interesting future study will be to compare DNN lesion results with behavior of human brain lesion patients, who

demonstrate impaired face perception (e.g., [62]). The lesion approach can also help us better understand the functional segregation of face recognition [73]. Third, only one DNN architecture was examined in the present work; but more DNN architectures will need to be considered in order to further generalize our results. Lastly, region-based feature coding may also provide an account for object recognition (e.g., using the AlexNet) and visual selectivity (also observed in human MTL neurons): objects falling within the coding region of a neuron/unit may elicit an elevated response. Again, this mechanism applies to both artificial units and human neurons and needs to be tested in a future study.

A central challenge in cognitive neuroscience is to understand how the brain encodes faces. In particular, it remains largely unclear how visual experience, learning, and memory shape face perception and recognition. There are two competing hypotheses about the emergence of face-selective neurons. One hypothesis argues that face-selective neurons require visual experience to develop, and this hypothesis has been supported by fMRI studies in the monkey fusiform face areas. The other hypothesis argues that face-selective neurons have an innate origin, and this hypothesis has been supported by studies from human infants and adults without visual experience of faces. Along this line, future studies will be needed to explore if identity-selective units can spontaneously emerge from untrained neural network (but see also the lottery ticket hypothesis [44]) or neural networks trained with other image databases (e.g., ImageNet). Although the DNN used in the present study was a pre-trained artificial neural network, it demonstrated strong ability to generalize to new faces, comparable to the primate visual system. Importantly, *in silico* experiments allow us to test a large set of parameters and better control experimental conditions, which is often not feasible when directly working with human or animal subjects. Therefore, our present study not only highlights the importance of the direction of training and visual experience in shaping the neural response to face identities, but also provides a useful approach to test these hypotheses and reconcile previous findings.

2.4 Methods

2.4.1 Stimuli

We employed the following stimuli in this study (**Figure 2.2**).

(1) For the original stimuli, we used faces of celebrities from the CelebA dataset [74], and we selected 50 identities with 10 images for each identity, totaling 500 face images. The identities were selected to include both genders and multiple races (see also **Supplementary Figure 1** and **Figure 2.7a**).

(2) We selected another 500 faces from 50 different identities (10 images per identity) from the CelebA dataset as a testing set.

(3) We generated four versions of cartoon faces (Hayao, Hosoda, Paprika, Shinkai) of the original stimuli using CartoonGAN [75].

(4) We generated Mooney faces by first transforming the original images into gray scale. We then filtered the images with a two-dimensional Gaussian smoothing kernel with standard deviation of 0.5. We lastly thresholded the images using a threshold determined for each individual face based on its luminance (threshold = mean luminance of the cropped image center − 0.03).

(5) We randomly selected 500 caricature faces of 50 identities (10 images per identity) from the IIIT-CFW dataset [76].

(6) We randomly selected 500 non-face objects from 50 categories (10 objects per category) from the ImageNet database [77].

2.4.2 Deep neural network (DNN)

We used the well-known deep neural network (DNN) implementation based on the VGG-16 convolutional neural network (CNN) architecture [68] (**Figure 2.1a**; https://www.robots.ox.ac.uk/~vgg/software/vgg_face/). The inputs to the first convolutional layer were RGB images of fixed size (224×224 pixels). The images were passed through a stack of convolutional layers, where the filters were used with a very small receptive field (3×3 pixels, which was the smallest size to capture the notion of left/right, up/down, and center). The convolution stride was fixed to 1 pixel; and the spatial padding of the convolutional layer input was 1-pixel for 3×3 convolutional layers such that the spatial resolution was preserved after convolution. Spatial pooling was carried out by five max-pooling layers, which followed some of the convolutional layers. Max pooling was performed over a 2×2 pixel window, with a stride of 2 pixels. Three fully connected (FC) layers followed a stack of convolutional layers: the first two had 4096 channels each, the third performed 50-way classification and thus contained 50 channels (one for each identity). The final layer was the soft-max layer.

We fine-tuned the FC8 layer with the original CelebA stimuli to confirm that this pre-trained model was able to discriminate the identities and ensure that the pre-trained model was suitable as a feature extractor. Specifically, we modified the output layer to 50 units for our model. Two-thirds of the original CelebA stimuli were used as the training set and the remaining stimuli were used as the testing set. We used the Adam optimizer with an initial learning rate of 5×10^{-4} and we had 10 epochs in total. A learning rate scheduler was applied after each epoch with the gamma value set to 0.9 to facilitate the convergence of the loss function. To update the weights during fine-tuning, we computed the cross-entropy loss on random batches of four face images (scaled to 224×224 pixels) for back propagation. We used 5-fold cross validation, which reached an accuracy of approximately 95%. Note that only the FC8 layer was fine-tuned and all the other layers were frozen. It is also worth noting that the VGG-16 was originally trained with 23 of the 50 identities involved in the present study (including Adam Levine, Bahar Soomekh, Betty White, Dana Delany, Dean Geyer, Eduardo Noriega, Hugh Jackman, Isla Fisher, John Slattery, Katherine Bailess, Kevin Hart, Logan Marshall-Green, Mario Lopez, Missi Pyle, Natalie Zea, Olivia Palermo, Rachel McAdams, Ron Perlman, Shawn Ashmore, Steven Soderbergh, Tim Gunn, Treat Williams, Zac Efron). However, we derived similar results when we excluded the identities involved in the original VGG-16 training (see also our generalization results with new identities in **Figure 2.2b**).

To visualize the DNN response, we subsequently applied a t-distributed stochastic neighbor embedding (t-SNE) method to convert high-dimensional features into a two-dimensional feature space. t-SNE is a variation of stochastic neighbor embedding (SNE) [78], a commonly used method for multiple class high-dimensional data visualization [79]. We applied t-SNE for each layer, with the cost function parameter (Prep) of t-SNE, representing the perplexity of the conditional probability distribution induced by a Gaussian kernel, set individually for each layer. We also used t-SNE to construct a face feature space so that we were able to investigate region-based feature coding for both DNN units and primate neurons [39].

2.4.4 Selection of identity-selective DNN units and primate neurons

To select identity-selective units, we used a one-way ANOVA to identify identity-selective units that had a significantly unequal response to different identities ($P < 0.01$; **Supplementary Figure 2a**). We further imposed an *additional criterion* to identify a *subset* of identity-selective

units with selective identities (**Supplementary Figure 2a**): the response of an identity was 2 standard deviations (SD) above the mean of responses from all identities. The rationale for implementing the criterion rather than utilizing the post hoc test of ANOVA is that the post hoc tests can only determine which pairs exhibit significant differences. However, our objective is to ascertain which identity or identities are encoded by each individual neuron. Furthermore, this criterion has been applied in primate neural analyses as well. In order to maintain consistency, we finally adopted the same criterion to identify the subsets. These identified identities whose response stood out from the global mean were the encoded identities. We refer to the units that encoded a single identity as single-identity (SI) units and we refer to the units that encoded multiple identities as multiple-identity (MI) units.

We followed the identical selection procedure for primate neurons. We used the mean firing rate in a time window 250-1000 ms after stimulus onset as the response to each face for primate neurons. Note that we also used this response to study the correlation between DNN units and primate neurons.

2.4.5 Assessment of face recognition accuracy

We used a support vector machine (SVM) to assess face recognition accuracy for each group of DNN units from a specific DNN layer. We employed a 5-fold cross-validation: we randomly partitioned the stimuli into 5 equal portions, and in each run, 4 portions of the stimuli were used as the training dataset and the remaining portion of the stimuli was used as the test dataset. We used the python package Scikit-learn [80] to build our radial basis function (RBF) kernel SVM classifier.

2.4.6 DNN lesion and perturbation

By analogy with brain lesions, we designed a "random-drop model" to lesion the original VGG-16 network in order to understand how many DNN units were needed to discriminate face identities. We conducted two experiments. In the first experiment, following *every* convolutional layer, we included a binary mask for the preceding layer that randomly set a subset of DNN units to be 0. In the second experiment, we applied the binary mask to a specific layer only. Note that in both experiments, both identity-selective and non-identity-selective units were dropped, according to their proportions.

In addition, we perturbed the network by rearranging the weights in the model. We conducted two experiments. In the first experiment, kernel-wise shuffle, we randomly permuted the weights in a single kernel. Since the kernel size of the network was 3 by 3 for all layers, kernel-wise shuffle permuted the 9 weight values for each kernel. In the second experiment, layer-wise shuffle, we pooled the weights of all kernels from a layer and reorganized the weights to form new kernels. Note that in both experiments, both identity-selective and non-identity-selective units were shuffled.

2.4.7 Neural recordings from a monkey

One male rhesus macaque (Macaca mulatta) was used in this study. All procedures conformed to local and U.S. National Institutes of Health guidelines, including the U.S. National Institutes of Health Guide for Care and Use of Laboratory Animals. All experiments were performed with the approval of the MIT Institutional Animal Care and Use Committee (IACUC). The monkey passively viewed the original CelebA stimuli (**Figure 2.5a**). In each trial, the monkey first viewed a white central fixation point (0.2 degrees of visual angle [DVA]) on a gray background for 300 ms to initiate a trial. Then, 8 faces were presented for 100 ms each, each followed by a blank (gray) screen for an inter-stimulus-interval (ISI) of 100 ms. The central fixation point persisted through the trial, and fluid reward was given if the monkey successfully fixated through the entire trial. The inter-trial-interval (ITI) of blank gray screen was at least 500 ms. We recorded 4155 trials in total, and we rejected 666 trials where the monkey broke the fixation (±2 DVA). For each round of presentation, we generated a random sequence for the 500 faces; and we used different sequences for different rounds of presentation. On average, each face was presented 55.7±1.49 (mean±SD) times. Note that we randomly inserted one gray image in each round of presentation as a control stimulus for baseline normalization.

The monkey was chronically implanted with two Utah arrays (Blackrock Microsystems) in the anterior and central inferotemporal (IT) cortex (see [81, 82] for details). Each array consisted of one 10-by-10 electrode grid with 96 active iridium oxide electrodes. Each electrode was 1.5 mm long with an inter-electrode distance of 400 μm. During each recording session, band-pass filtered (0.1 Hz to 7.5 kHz) neural activity was recorded continuously at a sampling rate of 20 kHz using Intan Recording Controller (Intan Technologies, LLC). We detected the multi-unit spikes after the raw data were zero-phase band-pass filtered between 300-6000 Hz (Matlab ellip function, fourth

order with 0.1 decibel pass-band ripple and 40 dB stop-band attenuation), and we used multi-unit activity (MUA) for analyses. A multiunit spike event was defined as the threshold crossing when voltage (falling edge) deviated by more than three times the standard deviation of the raw voltage values. We estimated internal consistency for each channel using a standardized image set that was run before the recording session on the same day and we accepted 53 MUA channels (from two arrays) that showed sufficient internal consistency (> 0.6). Consistent with previous studies [81, 82], we used the mean firing rate in a time window 70 ms to 180 ms after stimulus onset as the response to each face. We averaged the response from repeated presentations for each face.

2.4.8 Single-neuron recordings in human neurosurgical patients

To acquire the neuronal response from humans, we conducted single-neuron recordings from 5 neurosurgical patients (16 sessions in total). All participants provided written informed consent using procedures approved by the Institutional Review Board of West Virginia University (WVU). The detailed procedure has been described in our previous study [39]. Briefly, we employed a 1-back task for the original CelebA stimuli (**Figure 2.6a**). In each trial, a single face was presented at the center of the screen for a fixed duration of 1 second, with uniformly jittered inter trial interval (ITI) of 0.5-0.75 seconds. Patients pressed a button if the present face image was *identical* to the immediately previous image. Each face was shown once unless repeated in one-back trials; and we excluded responses from one-back trials to have an equal number of responses for each face.

We recorded from implanted depth electrodes in the amygdala and hippocampus from patients with pharmacologically intractable epilepsy. Bipolar wide-band recordings (0.1-9000 Hz), using one of the eight microwires as reference, were sampled at 32 kHz and stored continuously for off-line analysis with a Neuralynx system. The raw signal was filtered with a zero-phase lag 300-3000 Hz bandpass filter and spikes were sorted using a semi-automatic template matching algorithm as described previously [83]. Units were carefully isolated and recording and spike sorting quality were assessed quantitatively. Only units with an average firing rate of at least 0.15 Hz (entire task) were considered. Only single units were considered. Trials were aligned to stimulus onset and we used the mean firing rate in a time window 250 ms to 1000 ms after stimulus onset as the response to each face.

2.4.9 Pairwise distances in the face space

We employed a pairwise distance metric [50] to compare neural coding of face identities between primate neurons and DNN units. For each pair of identities, we used the dissimilarity value (1−Pearson's *r*) [84] as a distance metric. The primate neuronal distance metric was calculated between firing rates of all recorded neurons and the DNN distance metric was calculated between feature weights of all DNN units. In more precise terms, the computation of the primate neuronal distance metric involves treating the firing rate of all recorded brain neurons in response to an image as a feature vector representing that face. Given that there are 500 stimuli in total, this process results in 500 distinct feature vectors. Subsequently, we calculate the dissimilarity between each pair of the 500 faces to generate the neuronal distance metric for primates. The same procedure is applied to the DNN units, with the only difference being that the feature vector for the DNN distance metric is derived from the activations of all units in a particular layer in response to an image. Importantly, the input images for both the brain and DNN systems are identical. We then correlated the primate neuronal distance metric and the DNN distance metric. To determine statistical significance, we used a non-parametric permutation test with 1000 runs. In each run, we randomly shuffled the face labels and calculated the correlation between the primate neuronal distance metric and the DNN distance metric. The distribution of correlation coefficients computed *with* shuffling (i.e., null distribution) was eventually compared to the one *without* shuffling (i.e., observed response). If the correlation coefficient of the observed response was greater than 95% of the correlation coefficients from the null distribution, it was considered *significant*. A significant correlation indicated that the DNN face space corresponded to the primate neuronal face space [50]. We computed the correlation for each DNN layer so that we could determine the specific layer that the neuronal population encoded. For each face identity, we averaged the response of all faces of that identity to get a single mean firing rate. To get temporal dynamics, for human neurons, we used a moving window with a bin size of 500 ms and a step size of 50 ms. The first bin started −300 ms relative to trial onset (bin center was thus 50 ms before trial onset), and we tested 19 consecutive bins (the last bin was thus from 600 ms to 1100 ms after trial onset). We used false discovery rate (FDR) [51] to correct for multiple comparisons across DNN layers or time bins. For monkey neurons, we used a moving window with a bin size of 40 ms and a step size of 10 ms. The first bin started −70 ms relative to stimulus onset (bin center was thus 50 ms before stimulus onset), and we tested 26 consecutive bins (the last bin was thus from 180 ms to 220 ms after stimulus

onset). We used Bonferroni correction to correct for multiple comparisons across DNN layers or time bins.

2.4.10 Selection of DNN feature units

We employed the same procedure to select DNN feature units as we did with human neurons [39]. We first estimated a continuous spike density map in the feature space by smoothing the discrete activation map using a 2D Gaussian kernel (kernel size = feature dimension range * 0.2, SD = 4). We then estimated statistical significance for each pixel by permutation testing: in each of the 1000 runs, we randomly shuffled the labels of faces. We calculated the p-value for each pixel by comparing the observed spike density value to those from the null distribution derived from permutation. We lastly selected the region with significant pixels (permutation $P < 0.01$, cluster size > 0.23 * pixel number in the whole space). We also applied a mask to exclude pixels from the edges and corners of the spike density map where there were no faces because these regions were susceptible to false positives given our procedure. If a unit had a region with significant pixels, the unit was defined as a "feature unit" and demonstrated "region-based feature coding". We selected feature units for each individual DNN layer.

2.4.11 Data availability

All data and statistical analysis code are available on GitHub (https://github.com/JingeW/ID_selective).

Chapter 3: A critical period for developing face recognition

3.1 Introduction

A critical period is a time window during the development when some particular experience must be undergone for the complete development of language and sensory systems to occur [85]. The critical period hypothesis was originally proposed for the acquisition of a second language [86] and visual perception [87]. In children born with opacity or deviation of the eyes, the deprived eye will suffer from lacking cortical response despite a healthy retina. The consequence of such sensory deprivation can lead to lifelong amblyopia (due to ocular dominance plasticity) [88, 89]. Similarly, it has been hypothesized that there is a critical period for the development of the fusiform face area (FFA), which has an intriguing connection with the difference in face processing by individuals with autism [27, 90-96]. However, it has remained an open question whether there are critical periods in the development of face processing [93], what the computational mechanisms of critical periods are, and what the developmental trajectory of facial feature selection is.

It has been argued that the neural coding of visual stimuli can change over development [97]. A study using functional MRI to examine the development of several functionally defined regions including object, face, and place-selective cortices in different age groups (children, adolescents, and adults) has shown that development occurred by the expansion of the FFA and parahippocampal place area (PPA) into the surrounding cortex is correlated with improved recognition memory for faces and places, respectively [98]. Furthermore, microstructural proliferation in the human cortex is coupled with the development of face processing [99]. In addition to the developmental trajectory, face learning plays an important role in the computational mechanisms of critical periods. Our recent study has provided a neuronal mechanism for face learning: neuronal distance between face identities increases as a function of exposure, suggesting that faces become more neurally distinct after learning [96]. Notably, the core components of face processing and their neuromaturational time course in typical development (TD) may facilitate our understanding of face processing deficits in autism as well as developing clinical tools for early diagnosis and remediation [100].

Given the challenges of creating a critical period in the physical world, developing computational surrogate models [101] has become an appealing alternative. Deep neural networks (DNN) such as VGG-Face [102] and FaceNet [103] have achieved comparable or even superior face recognition performance compared to human observers. These DNN-based surrogate models have made it convenient to conduct experiments with deprived stimuli or perturbation of network architectures [36, 104]. For example, it has been experimentally shown the existence of a "critical period ''(usually the first few epochs) in the DNN [101], suggests that critical periods are not restricted to biological systems, but can emerge naturally in learning systems, whether biological or artificial, due to fundamental constraints arising from learning dynamics and information processing. It has been shown that the use of unsupervised learning in DNNs can provide a quantitative model of the ventral visual processing system and serve as a biologically plausible computational model of primate sensory learning [105]. More broadly, DNNs provide an important approach to testing the computational benefits of fundamental organizational features of the visual system [106]. DNNs create a highly organized face similarity structure where natural image variation is organized hierarchically, offering an important theoretical framework to understand identity coding [107]. Furthermore, it has been shown that brain-like functional specialization emerges spontaneously in DNNs and reflects a computational optimization for face recognition [108]. In sum, *in silico* experiments with DNNs have provided unprecedented opportunities to understand face coding and learning, especially when artificial models show correspondence with brain models [8, 39, 50, 104, 109].

In this study, we hypothesize that similar to humans and animals, deep artificial neural networks exhibit critical periods during which a stimulus deficit can impair the development of face learning. We further hypothesize that face learning can only be restored when providing information *within* the critical period but not *outside* the critical period. We seek for a computational account for critical periods and explore possible ways to restore face learning. We hypothesize that learning rate is a key factor for critical periods. We lastly explore the correspondence with primate visual systems, which may in turn explain the recovery mechanism from the critical periods.

3.2 Results

3.2.1 A critical period for developing face recognition

We first trained a DNN based on the ResNet50 architecture (**Figure 3.2A**) using full-face images (**Figure 3.1A**). We observed a rapid increase of performance in early training epochs (**Figure 3.2B**), which reached the plateau after 30 epochs. We thus defined the first 30 epochs as the critical period (see **Methods**; see also **Supplementary Figure 7**; note that this definition is based on established studies of critical periods [101]).

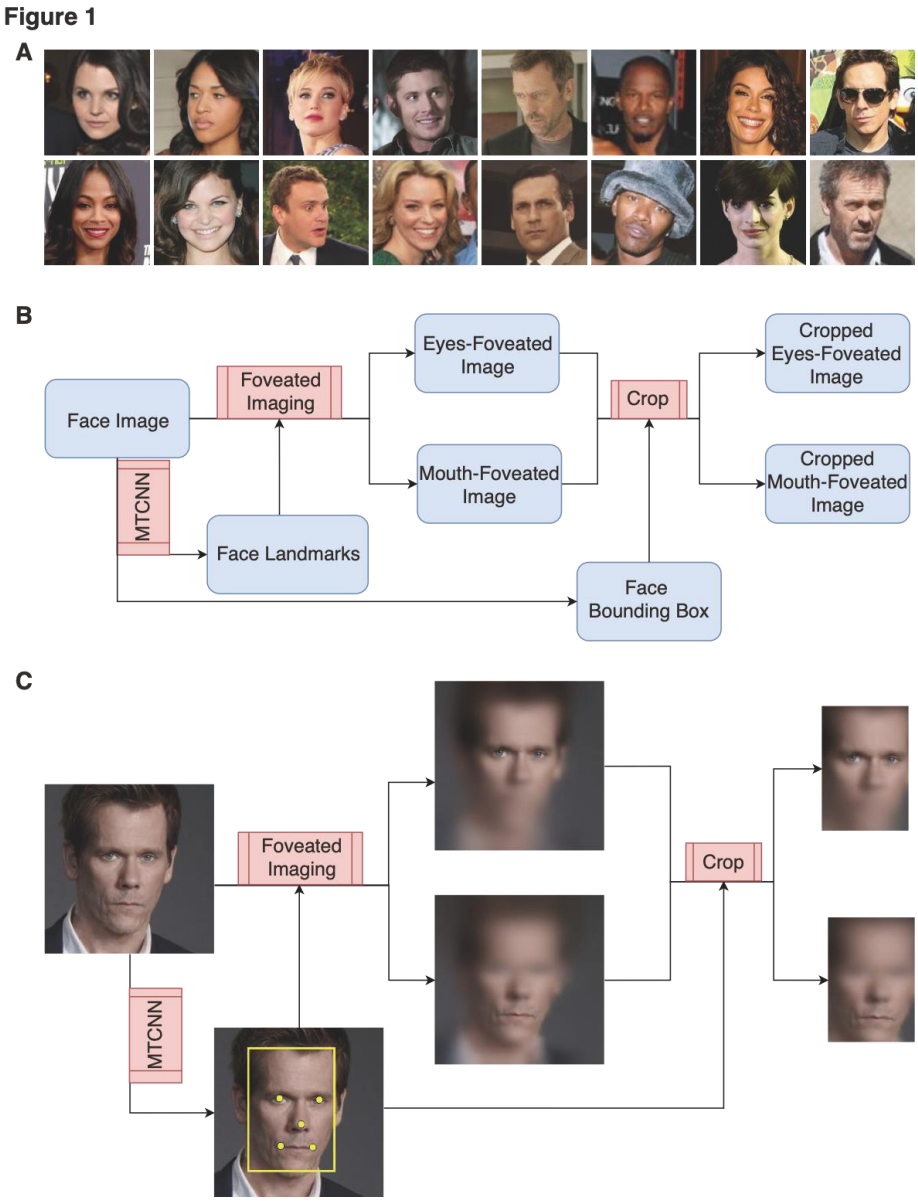


Figure 3.1 Sample stimuli and image processing pipeline.

(A) Sample stimuli from the CASIA_WebFace dataset. **(B)** Image processing pipeline. MTCNN was applied to the original face images to detect a tight bounding box outlining the face area and facial landmarks (centers of the eyes, nose tip, and corners of the mouth). Foveation imaging was applied to the original face images to derive two sets of foveated images (eyes-foveated and mouth-foveated). We lastly cropped the images based on the bounding box derived using the MTCNN. **(C)** An example showing the image processing pipeline.

Figure 2

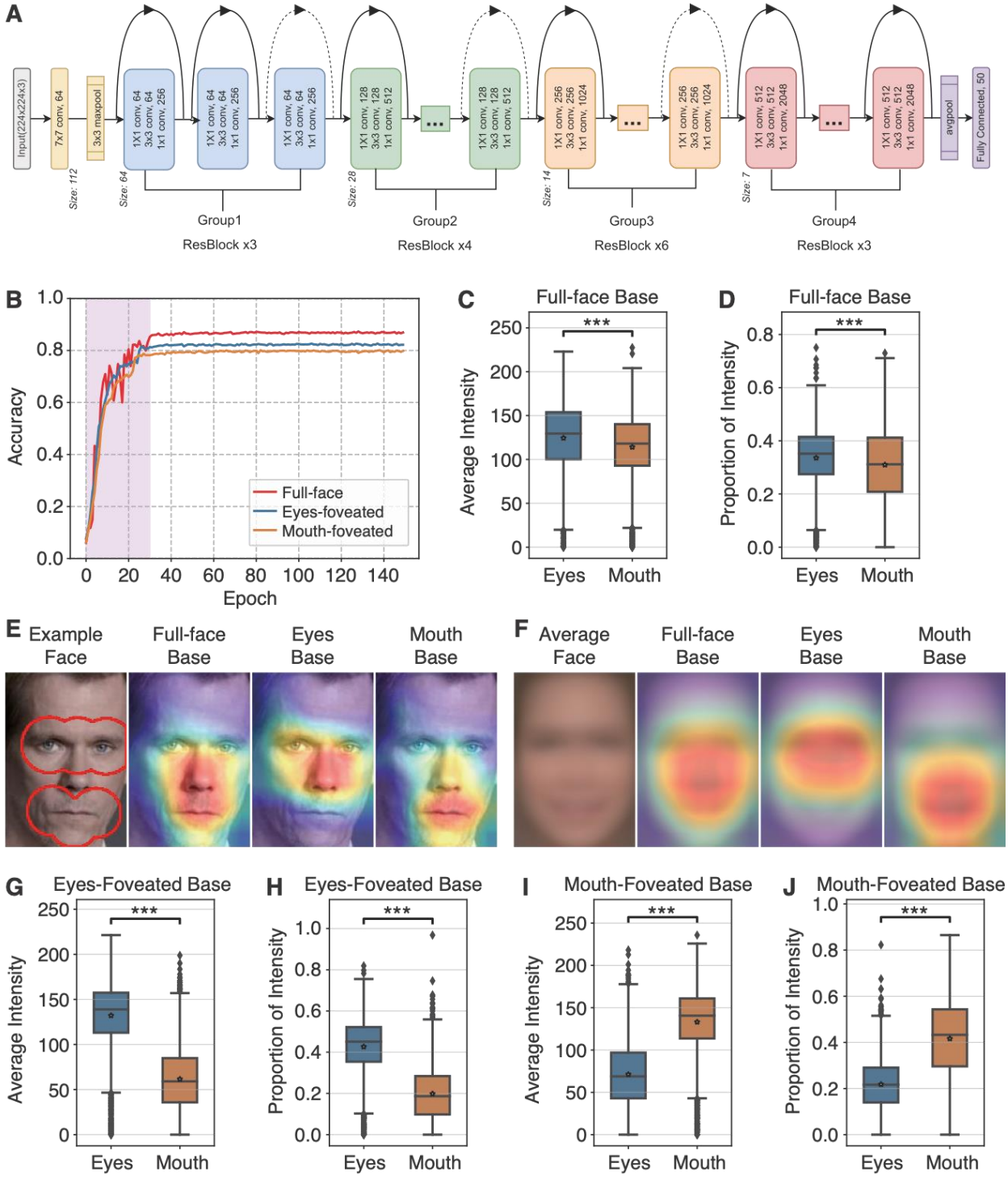


Figure 3.2 Face recognition models with different training stimuli.
(A) ResNet50 architecture. ResNet50 has 4 stages. It performs the initial convolution and max-pooling using 7×7 and 3×3 kernels, respectively. Subsequently, the inputs go through the 4 stages. All stages contain the basic residual blocks. For ResNet50, there are 3, 4, 6, and 3 residual blocks in stages 1, 2, 3, and 4, respectively. In each residual block, 3 convolution layers (1×1, 3×3, and 1×1) are stacked. The 1×1 convolution layers are responsible for reducing and then restoring the dimensions. The 3×3 layer is left as a bottleneck with smaller input/output dimensions. The curved arrows are skip connections or 'shortcuts'. Solid connections refer to the identity connection. The dashed connection denotes that the convolution operation in the residual block is performed with stride 2. As feature maps progress from one stage to another, the channel width is doubled and the size of the input is reduced to half. Finally, the network has an average pooling layer followed by a fully connected layer having 50 neurons (number of different identities). **(B)** Network learning curve. The validation accuracy of face identity recognition is plotted as a function of the model training epoch. The shaded area denotes the critical period. **(C, G, I)** Average Grad-CAM intensity for each region of interest (ROI). **(D, H, J)** The proportion of Grad-CAM intensity for each ROI. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, and the whiskers extend to the most extreme data points the algorithm considers to be not outliers. Asterisks indicate a significant difference using a two-tailed paired *t*-test. ***: $P < 0.001$. **(E)** The Grad-CAM intensity maps for an example face. **(F)** The Grad-CAM intensity maps for group average across faces. The intensity values indicate the contribution/importance of pixels for face recognition. The red contours in the example face delineate the eyes and mouth ROIs for this face. **(C, D)** Full-face model. **(G, H)** Eyes-foveated model. **(I, J)** Mouth-foveated model.

We next quantified information utilization in the images using Grad-CAM (see **Methods**). The heat maps reflected the regions in the face that contributed to the correct classification of face identities (**Figure 3.2E**). In the full-face model, we found that the network utilized information from both the eyes and mouth (see **Figure 3.2E** for an example; see **Figure 3.2F** for group average), although the network utilized more information from the eyes region than the mouth region (**Figure 3.2C**; average Grad-CAM intensity: eyes: 124.47±39.97, mouth: 114.10±37.51; two-tailed two-sample *t*-test: $t(7536) = 11.62$, $P < 10^{-30}$; **Figure 3.2D**; proportion of Grad-CAM intensity: eyes: 0.34±0.11, mouth: 0.31±0.14; $t(7536) = 8.79$, $P < 10^{-17}$).

To study the integration of facial information, we employed foveated imaging and created eyes-foveated and mouth-foveated images (**Figure 3.1B, C**; see **Methods**). As expected, with reduced information, the models reached a lower performance (**Figure 3.2B**). Interestingly, the eyes-foveated model had a better performance compared to the mouth-foveated model, indicating that the eyes contained more information than the mouth for face recognition. Notably, we observed a similar critical period (the first 30 epochs) compared to the full-face model, suggesting that the foveated models had a similar learning process.

We also quantified information utilization in the images in these models (**Figure 3.2E, F**). Indeed, the eyes-foveated model utilized more information in the eyes than the mouth (**Figure 3.2G**; average Grad-CAM intensity: eyes: 132.08±37.71, mouth: 61.59±35.62; $t(7536) = 83.42$, P

$< 10^{-30}$; **Figure 3.2H**; proportion of Grad-CAM intensity: eyes: 0.43±0.14, mouth: 0.20±0.13; $t(7536) = 73.61$, $P < 10^{-30}$), whereas the mouth-foveated model utilized more information in the mouth than the eyes (**Figure 3.2I**; average Grad-CAM intensity: eyes: 71.25±37.65, mouth: 133.14±40.25; $t(7536) = 68.95$, $P < 10^{-30}$; **Figure 3.2J**; proportion of Grad-CAM intensity: eyes: 0.22±0.11, mouth: 0.42±0.18; $t(7536) = 58.55$, $P < 10^{-30}$). This result confirmed that reducing certain visual inputs into training would lead to reduced utilization of the corresponding visual information. On the other hand, eyes-foveated model had a higher average Grad-CAM intensity in the eyes than the full-face model (**Figure 3.2G** vs. **Figure 3.2C**; two-tailed paired *t*-test: $t(3768) = 12.03$, $P < 10^{-30}$) and mouth-foveated model had a higher average Grad-CAM intensity in the mouth than the full-face model (**Figure 3.2I** vs. **Figure 3.2C**; $t(3768) = 30.39$, $P < 10^{-30}$), suggesting that the network could adjust to focus on available information.

Lastly, we showed that another popular DNN model for face recognition (i.e., VGG-Face) had a similar learning curve and critical period, and full-face models outperformed eyes-foveated models and mouth-foveated models (**Supplementary Figure 8A, B**). Therefore, we confirmed that our results were not idiosyncratic to the DNN model used in the present study.

3.2.2 Recovery with full-face images within vs. outside the critical period

Above, we have revealed a critical period during DNN training (i.e., learning face identities) and illustrated the information utilization during this process. We next investigated whether training with restricted stimuli (eyes-foveated faces or mouth-foveated faces) could be recovered with additional visual information.

We first used full-face images to recover impaired models. We found that providing full-face information to the network *within* the critical period led to a better performance, and this was the case for both eyes-foveated model (**Figure 3.3A**) and mouth-foveated model (**Figure 3.3B**). However, providing full-face information to the network *outside* the critical period did not improve the performance (**Figure 3.3A, B**), and it could even deteriorate the accuracy for the mouth-foveated model (**Figure 3.3B**). This result was confirmed with different starting points of recovery within or outside the critical period (**Supplementary Figure 7B, C**).

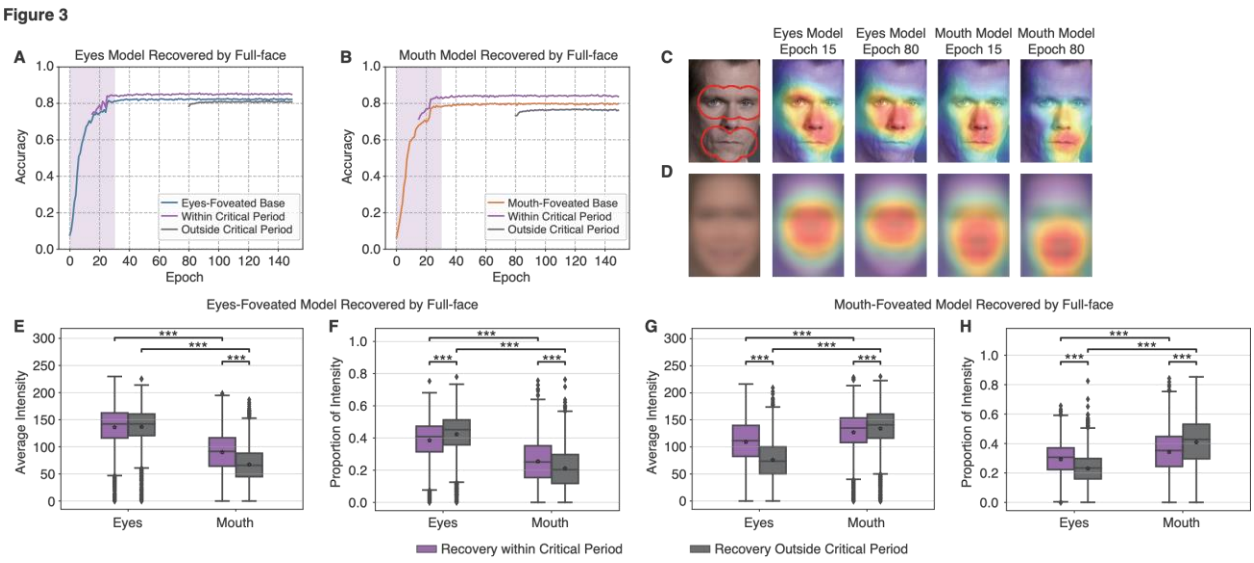


Figure 3.3 Recovery with full-face images.
(A) The learning curve for the eyes-foveated model. **(B)** The learning curve for the mouth-foveated model. **(C)** The Grad-CAM intensity maps for an example face. **(D)** The Grad-CAM intensity maps for group average across faces. **(E, G)** Average Grad-CAM intensity for each region of interest (ROI). **(F, H)** The proportion of Grad-CAM intensity for each ROI. **(E, F)** Eyes-foveated model. **(G, H)** Mouth-foveated model. Legend conventions as in **Figure 3.2**.

Importantly, the change in performance with recovery was associated with different utilization of facial information. For the eyes-foveated model, recovering within the critical period led to an increased utilization of mouth information compared to recovering outside the critical period (see **Figure 3.3C** for an example and **Figure 3.3D** for group average; **Figure 3.3E**; average Grad-CAM intensity: within: 90.03±37.21, outside: 67.27±32.00; $t(3768) = 53.68$, $P < 10^{-30}$; **Figure 3.3F**; proportion of Grad-CAM intensity: within: 0.25±0.13, outside: 0.21±0.12; $t(3768) = 45.65$, $P < 10^{-30}$). For both recovery conditions, the eyes still contributed more information than the mouth (**Figure 3.3E**; average Grad-CAM intensity: within: $t(7536) = 53.82$, $P < 10^{-30}$, outside: $t(7536) = 89.85$, $P < 10^{-30}$; **Figure 3.3F**; proportion of Grad-CAM intensity: within: $t(7536) = 43.84$, $P < 10^{-30}$, outside: $t(7536) = 72.29$, $P < 10^{-30}$).

Similarly, for the mouth-foveated model, recovering within the critical period led to an increased utilization of eyes information compared to recovering outside the critical period (see **Figure 3.3C** for an example and **Figure 3.3D** for group average; **Figure 3.3G**; average Grad-CAM intensity: within: 109.18±40.77, outside: 76.03±35.50; $t(3768) = 63.39$, $P < 10^{-30}$; **Figure 3.3H**; proportion of Grad-CAM intensity: within: 0.29±0.11, outside: 0.23±0.10; $t(3768) = 56.29$, $P < 10^{-30}$). For both recovery conditions, the mouth still contributed more information than the

eyes (**Figure 3.3G**; average Grad-CAM intensity: within: $t(7536) = 19.15$, $P < 10^{-30}$, outside: $t(7536) = 67.13$, $P < 10^{-30}$; **Figure 3.3H**; proportion of Grad-CAM intensity: within: $t(7536) = 16.34$, $P < 10^{-30}$, outside: $t(7536) = 54.58$, $P < 10^{-30}$).

We next investigated the extent to which the foveated models recovered by comparing to the full-face model. Although the eyes-foveated model recovered with full-face images within the critical period improved performance, it did not fully reach the level of full-face model in model performance (85.67% vs. 87.32%; **Figure 3.3A** vs. **Figure 3.2B**) and utilization of facial information (**Figure 3.3E** vs. **Figure 3.2C**; average Grad-CAM intensity: eyes: $t(3768) = 22.72$, $P < 10^{-30}$, mouth: $t(3768) = 42.73$, $P < 10^{-30}$; **Figure 3.3F** vs. **Figure 3.2D**; proportion of Grad-CAM intensity: eyes: $t(3768) = 42.60$, $P < 10^{-30}$, mouth: $t(3768) = 45.06$, $P < 10^{-30}$), suggesting that the impaired eyes-foveated model could be partially recovered within the critical period. Similarly for the mouth-foveated model, when recovery with full-face images happened within the critical period, the resultant model improved model performance (84.66% vs. 87.32%; **Figure 3.3B** vs. **Figure 3.2B**) and utilization of facial information (**Figure 3.3G** vs. **Figure 3.2C**; average Grad-CAM intensity: eyes: $t(3768) = 26.30$, $P < 10^{-30}$, mouth: $t(3768) = 22.56$, $P < 10^{-30}$; **Figure 3.3H** vs. **Figure 3.2D**; proportion of Grad-CAM intensity: eyes: $t(3768) = 34.66$, $P < 10^{-30}$, mouth: $t(3768) = 34.40$, $P < 10^{-30}$) towards the full-face model, although the impaired mouth-foveated model was partially recovered.

Together, our results suggest that providing information in the critical period can recover model performance and information usage in the impaired models. However, providing information outside the critical period cannot recover impaired models anymore.

3.2.3 Recovery with complementary information within vs. outside the critical period

Above, we have shown recovery with full-face images, which contain full information of the faces. Can we recover impaired models with complementary information (i.e., providing eyes-foveated model with mouth-foveated images or providing mouth-foveated model with eyes-foveated images)?

To address this question, we input mouth-foveated images to the eyes-foveated model, within and outside the critical period. Recovering within the critical period using complementary stimuli led to a similar model performance (**Figure 3.4A**), but interestingly, recovering outside the

critical period even deteriorated the model performance (**Figure 3.4A**). As expected, recovering outside the critical period did not change information utilization: the eyes still contributed more information than the mouth (see **Figure 3.4C** for an example and **Figure 3.4D** for group summary; **Figure 3.4E**; average Grad-CAM intensity: eyes: 128.15±38.63, mouth: 77.75±37.45; $t(7536) = 57.51$, $P < 10^{-30}$; **Figure 3.4F**; proportion of Grad-CAM intensity: eyes: 0.38±0.13, mouth: 0.24±0.14; $t(7536) = 49.23$, $P < 10^{-57}$). However, notably, recovering within the critical led to an opposite pattern of information utilization: the mouth contributed more information than the eyes (see **Figure 3.4C** for an example and **Figure 3.4D** for group summary; **Figure 3.4E**; average Grad-CAM intensity: eyes: 97.56±41.72, mouth: 120.46±39.65; $t(7536) = 24.42$, $P < 10^{-30}$; **Figure 3.4F**; proportion of Grad-CAM intensity: eyes: 0.28±0.12, mouth: 0.36±0.16; $t(7536) = 21.95$, $P < 10^{-30}$), a pattern of results that was more similar to the mouth-foveated model. This result suggests that new complementary information provided during the critical period overrode the original information utilization. In other words, it indicates that the network mainly takes information provided later. It is worth noting that the model performance seemed to switch as well (**Figure 3.4A, B**): the eyes-foveated model turned into the mouth-foveated model.

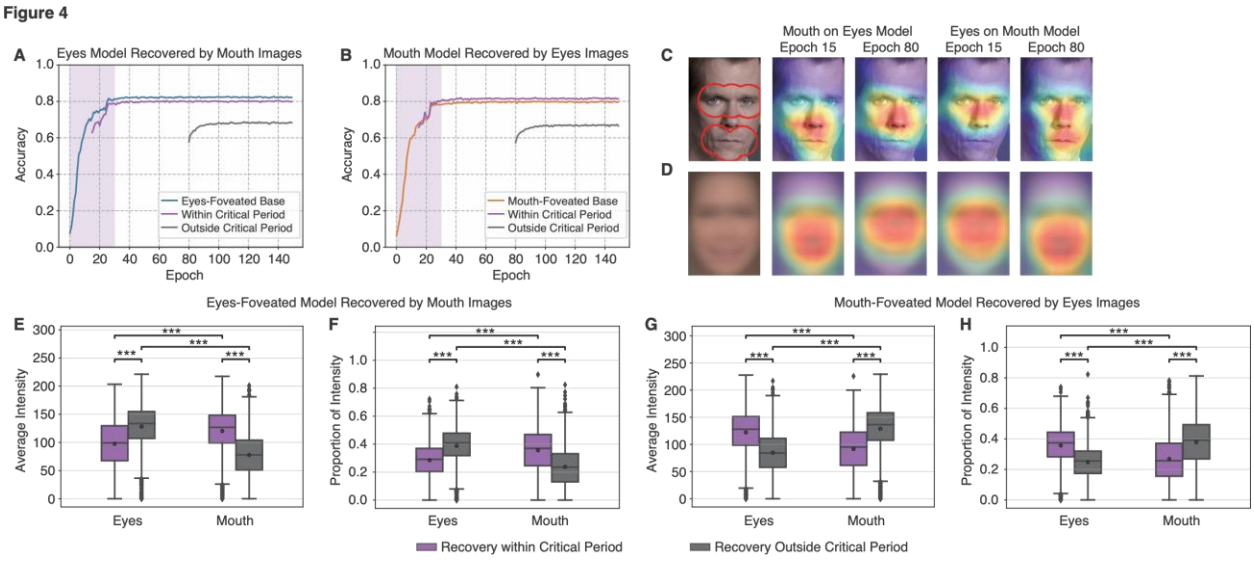


Figure 3.4 Recovery with images with complementary information.
(A) The learning curve for the eyes-foveated model. **(B)** The learning curve for the mouth-foveated model. **(C)** The Grad-CAM intensity maps for an example face. **(D)** The Grad-CAM intensity maps for group average across faces. **(E, G)** Average Grad-CAM intensity for each region of interest (ROI). **(F, H)** The proportion of Grad-CAM intensity for each ROI. **(E, F)** Eyes-foveated model. **(G, H)** Mouth-foveated model. Legend conventions as in **Figure 3.2**.

Similarly, when we input eyes-foveated images to the mouth-foveated model, we found that recovering outside the critical period even deteriorated the model performance (**Figure 3.4B**), and the critical period did not change information utilization: the mouth still contributed more information than the eyes (see **Figure 3.4C** for an example and **Figure 3.4D** for group summary; **Figure 3.4G**; average Grad-CAM intensity: eyes: 84.84±38.36, mouth: 128.98±41.65; $t(7536) = 47.86$, $P < 10^{-30}$; **Figure 3.4H**; proportion of Grad-CAM intensity: eyes: 0.24±0.11, mouth: 0.38±0.17; $t(7536) = 40.42$, $P < 10^{-30}$). However, again, recovering within the critical led to an opposite pattern of information utilization: the eyes contributed more information than the mouth (see **Figure 3.4C** for an example and **Figure 3.4D** for group summary; **Figure 3.4G**; average Grad-CAM intensity: eyes: 122.19±10.42, mouth: 91.48±40.97; $t(7536) = 32.76$, $P < 10^{-30}$; **Figure 3.4H**; proportion of Grad-CAM intensity: eyes: 0.36±0.13, mouth: 0.27±0.15; $t(7536) = 27.48$, $P < 10^{-30}$), a pattern of results that was more similar to the eyes-foveated model. This result again suggests that new complementary information provided during the critical period overrode the original information utilization, and the mouth-foveated model turned into the eyes-foveated model. In addition, the model performance seemed to switch as well (**Figure 3.4A, B**; i.e., the mouth-foveated model turned into the eyes-foveated model); and this result could be further replicated by a different DNN (**Supplementary Figure 8C, D**).

Together, using complementary information for recovery, we not only confirm that providing new information outside the critical period cannot alter the model anymore, but also show that providing new information within the critical period will override the original model.

3.2.4 Computational mechanism underlying recovery from the critical period

We next investigated why there was a critical period and why information provided in the critical period could override previous information. We hypothesize that the decrease in learning rate could explain the above results. This hypothesis is in line with the developmental trajectory of the primate visual system where learning is decreased as a function of age [110]. It is worth noting that during model training the learning rate evolved based on an adaptive rule (see **Methods**) that is consistent with neurodevelopment, and we did not preset the epoch-by-epoch learning rates (see also **Supplementary Figure 7A** for validation with different initial learning rates).

Indeed, we showed that the learning rate monotonically dropped as a function of training epoch (**Figure 3.5A**), suggesting that the learning became more local in the later stage, which is

consistent with the idea of critical period during development. However, notably, when we restored a larger learning rate in later epochs, the learning process was recovered (**Figure 3.5B**; here illustrated using the mouth-foveated model). We found that a larger learning rate could lead to a better recovery, and the same learning rate as the initial training phase (0.01; the learning rate in the critical period) could best recover the model (**Figure 3.5B**). Therefore, the inability to recover outside the critical period could be explained by the reduced learning rate: the network could not get out of the local minima to restore the learning for new information. This also explained why the network overrode the previously learned information for recovery within the critical period: the network converged to another local minimum with the new information for learning.

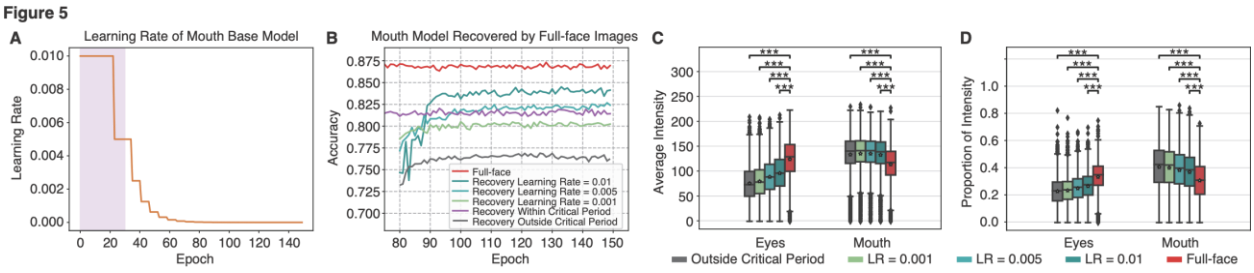


Figure 3.5 Improving model recovery by adjusting the learning rate.
(A) Learning rate as a function of training epoch in the mouth-foveated model. **(B)** The learning curve for each learning rate. Model recovery ability varied as a function of learning rate. The learning curves for the full-face model and the mouth-foveated model recovered by full-face images within and outside the critical period are shown as a reference. **(C)** Average Grad-CAM intensity for each learning rate. **(D)** The proportion of Grad-CAM intensity for each learning rate. LR: learning rate. Legend conventions as in **Figure 3.2**.

Furthermore, we showed that consistent with improved model performance, utilization of eyes information increased as a function of learning rate towards the full-face model (**Figure 3.5C**; average Grad-CAM intensity: recovery outside the critical period: 76.03±35.50, learning rate = 0.001: 80.11±35.13, learning rate = 0.005: 89.02±36.87, learning rate = 0.01: 96.50±37.92, full-face model: 124.47±39.97; one-way repeated-measure ANOVA across learning rates: $F(2,7536) = 1220$, $P < 10^{-20}$; **Figure 3.5D**; proportion of Grad-CAM intensity: recovery outside the critical period: 0.23±0.10, learning rate = 0.001: 0.24±0.10, learning rate = 0.005: 0.25±0.10, learning rate = 0.01: 0.27±0.11, full-face model: 0.34±0.11; $F(2,7536) = 1086$, $P < 10^{-20}$). Moreover, we derived similar results with the eyes-foveated model.

Together, our results suggest that learning rate is a determining factor for critical period and can explain the network performance and information utilization with respect to model recovery.

3.2.5 Knowledge distillation and attention transfer for model recovery

Can we achieve the same improvement in learning without modifying the learning rate but by applying knowledge distillation and attention transfer? Specifically, we used the full-face model as the teacher model and applied attention transfer to improve the mouth-foveated model outside the critical period (i.e., the student model, which is the same as the recovery outside the critical period; **Figure 3.6A**; see **Methods**). Indeed, attention transfer improved model performance outside the critical period (**Figure 3.6B**) and increased information utilization in the eyes (see **Figure 3.6C** for an example and **Figure 3.6D** for group average; **Figure 3.6E**; average Grad-CAM intensity: no attention transfer [i.e., recovery outside the critical period]: 76.03±35.50, attention transferred: 89.49±37.27; $t(3768) = 45.52$, $P < 10^{-30}$; **Figure 3.6F**; proportion of Grad-CAM intensity: no attention transfer: 0.23±0.10, attention transferred: 0.25±0.10; $t(3768) = 33.82$, $P < 10^{-30}$). However, the model after attention transferred still did not reach the same level of performance (**Figure 3.6B**; 80.76% vs. 87.32%; the performance of the attention transferred model was similar to recovery with a learning rate of 0.001 [80.53%]; **Figure 3.5B**) and utilization of eyes information (**Figure 3.6E**; average Grad-CAM intensity: $t(3768) = 47.91$, $P < 10^{-30}$; **Figure 3.6F**; proportion of Grad-CAM intensity: $t(3768) = 39.01$, $P < 10^{-30}$) as the full-face model (teacher model). Furthermore, similar results were found for the eyes-foveated model.

Figure 6

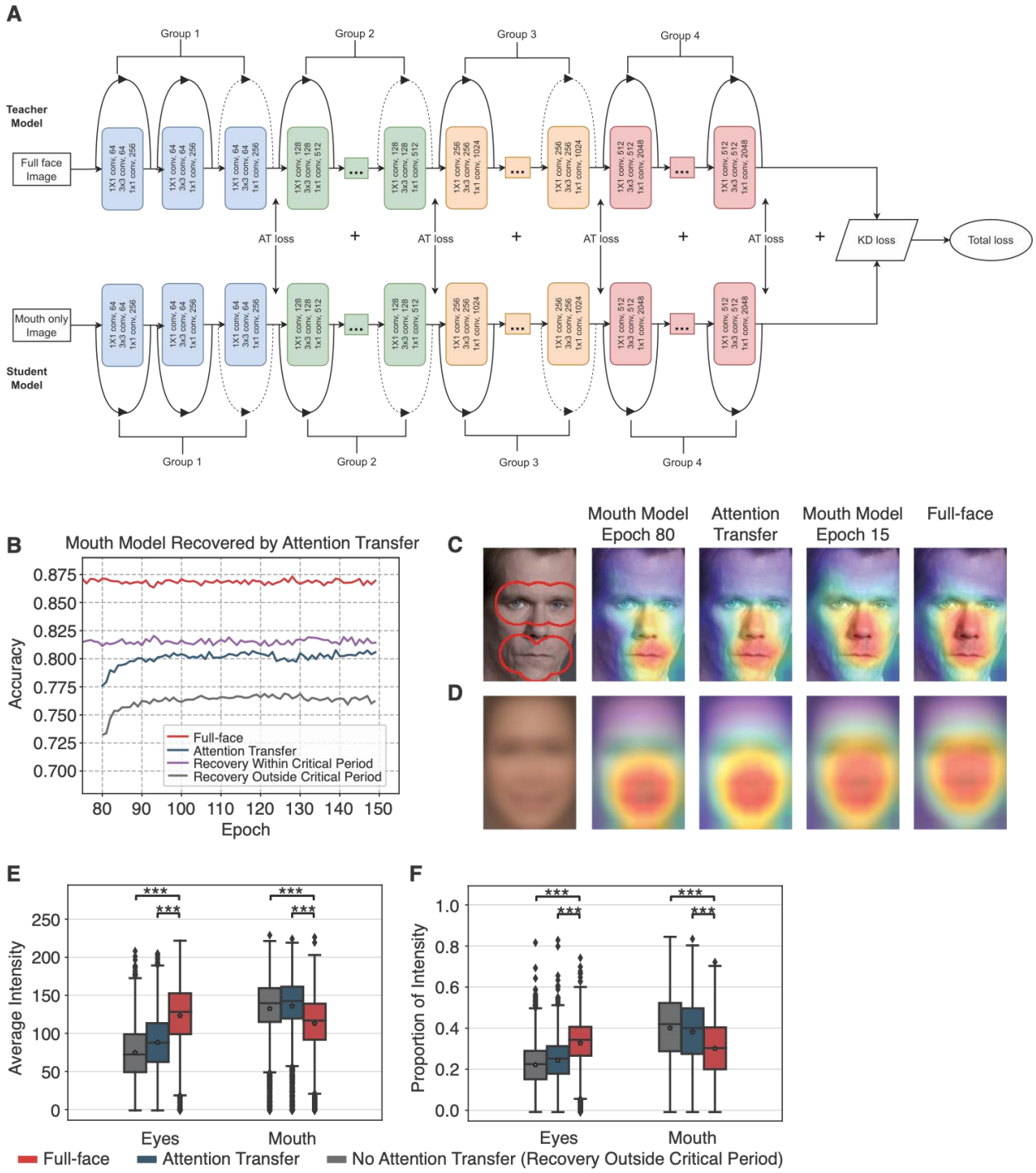


Figure 3.6 Improving model recovery by applying knowledge distillation and attention transfer. **(A)** A computational framework for knowledge distillation and attention transfer. We used the model trained by full-face images as the teacher model to improve the performance of the mouth-foveated model (student model) outside the critical period. Both the teacher model and the student model had the same ResNet50 architecture. **(B)** The learning curves for the mouth-foveated model with vs. without attention transfer. The learning curves for the full-face model and the mouth-foveated model recovered by full-face

images within and outside the critical period are shown as a reference. **(C)** The Grad-CAM intensity maps for an example face. **(D)** The Grad-CAM intensity maps for group average across faces. **(E)** Average Grad-CAM intensity for each region of interest (ROI). **(F)** The proportion of Grad-CAM intensity for each ROI. Legend conventions as in **Figure 3.2**.

Together, our results suggest that knowledge distillation and attention transfer can partially recover an impaired model outside the critical period (i.e., when learning rate is low), although the extent of recovery is limited compared to directly adjusting the learning rate.

3.2.6 Identity selectivity

We have shown before that identity-selective units (see **Methods**) are key building blocks of the DNN for face recognition [104]. We next investigated the change of identity selectivity during development of face recognition by summarizing the percentage of identity-selective units in each model. We focused on the top DNN layer (Conv4) where identity selectivity is most established and relevant [39, 104]. Indeed, we found that the full-face model (**Figure 3.7**; 87.5%) had a higher percentage of identity-selective units than the eyes-foveated model (78.9%; χ^2-test: $P < 10^{-10}$) and a control forehead-foveated model (72.8%; $P < 10^{-10}$; note that the eyes-foveated model also had a higher percentage of identity-selective units than the control forehead-foveated model: $P < 10^{-10}$). Importantly, for the eyes-foveated model, recovery with full-face images within the critical period increased the percentage of identity-selective units (**Figure 3.7**; 84.2%; $P < 10^{-10}$), but recovery outside the critical period did not increase the percentage of identity-selective units (79.1%; $P = 0.33$; within vs. outside: $P < 10^{-10}$). Similarly, for the forehead-foveated model, recovery with full-face images within the critical period increased the percentage of identity-selective units (**Figure 3.7**; 81.9%; $P < 10^{-10}$), although recovery outside the critical period also increased the percentage of identity-selective units (76.3%; $P < 10^{-10}$), albeit to a lesser extent (within vs. outside: $P < 10^{-10}$). Together, our results have further suggested a recovery mechanism using identity-selective units: restricted visual information impaired the formation of identity-selective units, and recovering within the critical period could increase and recover identity-selective units.

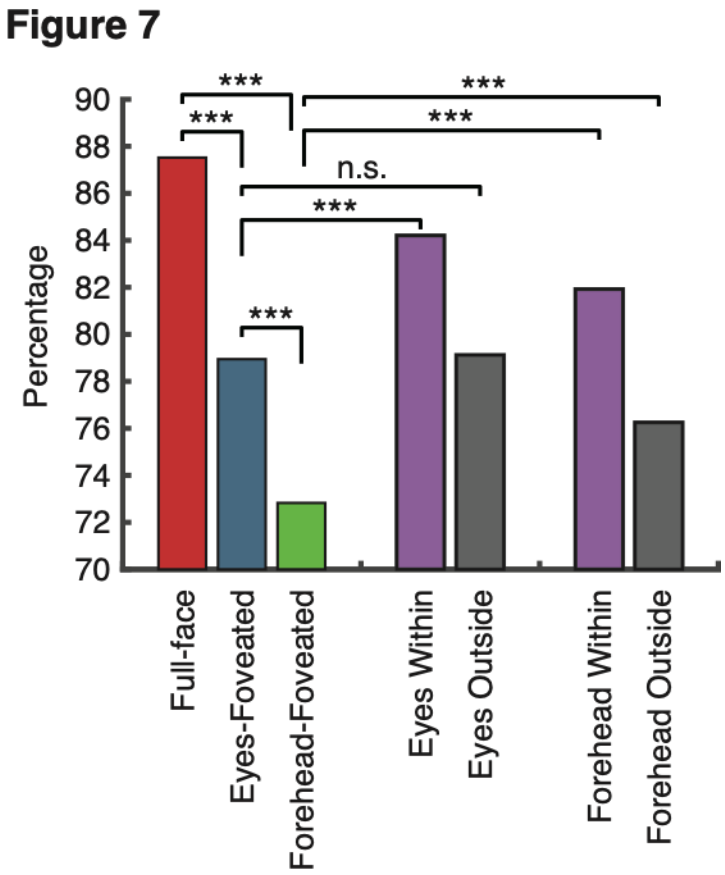


Figure 3.7 Percentage of identity-selective units for each model.
Eyes Within: eyes-foveated model recovered by full-face images within the critical period. Eyes Outside: eyes-foveated model recovered by full-face images outside the critical period. Forehead Within: forehead-foveated model recovered by full-face images within the critical period. Forehead Outside: forehead-foveated model recovered by full-face images outside the critical period. Asterisks indicate a significant difference between models using χ^2-test. ***: $P < 0.001$. n.s.: not significant.

3.2.7 Establishing the relationship between artificial DNN units and real primate neurons

The DNN performs face recognition tasks similarly as humans; and it has been suggested that DNNs share similarities with the primate visual system and can therefore help us better understand the neural mechanisms of face recognition [49, 50, 105]. We lastly investigated whether the development of face recognition in DNNs shared a similar biological basis.

First, we analyzed whether the ensemble of DNN units shared representational similarity with the ensemble of monkey inferotemporal (IT) neurons (**Figure 3.8A, C, D**). We used an independent set of stimuli from the CelebA dataset (500 natural face images of 50 celebrities) [111] to compare between artificial DNN units and real IT neurons. We recorded neuronal activity

using two Utah arrays in the anterior and central IT cortex (see **Methods**) while the monkey performed a passive viewing task (**Figure 3.8A**). We identified 53 multi-unit activity (MUA) channels that showed sufficient internal consistency and we focused on these channels for further analysis. We found that for all models the pairwise distance from the DNN significantly correlated with the neuronal pairwise distance from the monkey IT cortex (**Figure 3.8C**; see **Figure 3.8D** for temporal dynamics of each model); and for the full-face model, there was an increase of correlation towards the top DNN layer (**Figure 3.8C**). Notably, we found that the full-face model had a better correspondence with IT neurons than the eyes-foveated model and mouth-foveated model (**Figure 3.8C**). We also found that the eyes-foveated model recovered with full-face images within the critical period increased correspondence with IT neurons compared to recovery outside the critical period (Conv4: permutation $P < 0.001$). Therefore, the correspondence between DNN units and real IT neurons could reflect model performance and recovery.

Figure 8

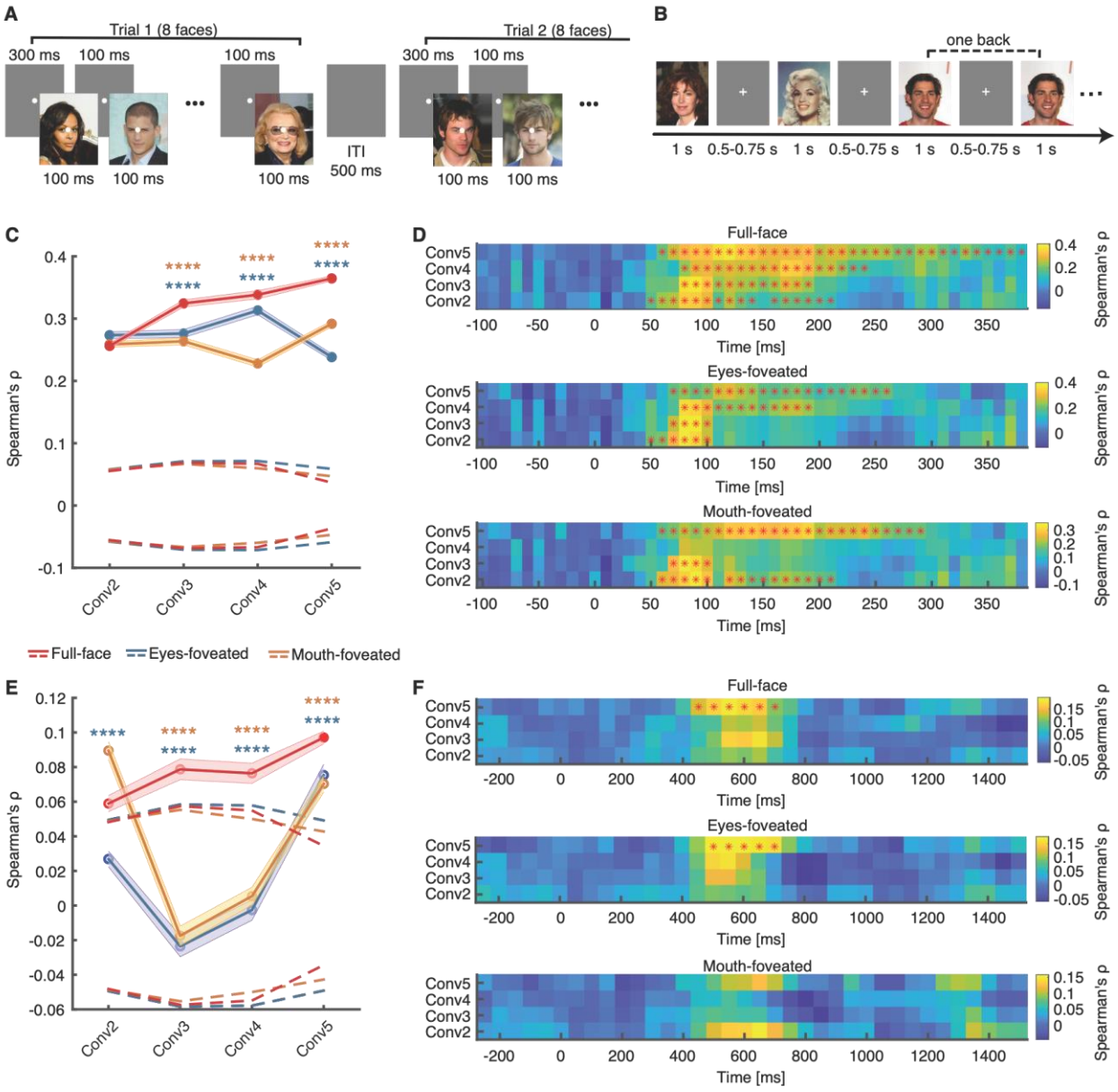


Figure 3.8 Match between the deep neural network (DNN) units and (real) primate neurons.

(A, C, D) Monkey inferotemporal (IT) cortical neurons. **(B, E, F)** Human amygdala and hippocampal neurons. **(A)** Task used to acquire neural responses from a monkey. In each trial, 8 faces were presented for 100 ms each, followed by a fixed inter-stimulus-interval (ISI) of 100 ms. There was a central fixation point of 300 ms at the beginning of each trial and there was an inter-trial-interval (ITI) of at least 500 ms following each trial. The central fixation point persisted through the trial. **(B)** Task used to acquire single-neuron responses from humans. We employed a one-back task, in which patients responded whenever an identical famous face was repeated. Each face was presented for 1 s, followed by a jittered ISI of 0.5 to 0.75 s. **(C, E)** Correlation between pairwise distance in the primate neuronal face space and pairwise distance in the DNN face space. **(C)** Here we used the mean firing rate in a time window 70 ms to 180 ms after stimulus onset as the response to each face, and we averaged the responses to 10 faces for each face identity. **(E)** Here we used the mean firing rate in a time window 250 ms to 1000 ms after stimulus onset as the response to each face, and we averaged the responses to 10 faces for each face identity. Dashed lines

denote ±SD across permutation runs (n = 1000), and solid circles represent a significant correlation (permutation test: P < 0.05, Bonferroni correction across layers). Shaded area denotes ±SEM across bootstrap runs (n = 1000; each resample contained 35 identities), and asterisks indicate a significant difference between models using one-tailed two-sample t-test. ****: P < 0.001. **(D, F)** Temporal dynamics of correlation of pairwise distance between primate neurons and DNN units. **(D)** Monkey IT neurons (bin size = 40 ms, step size = 10 ms). **(F)** Human MTL neurons (bin size = 500 ms, step size = 50 ms). Color coding indicates the Spearman's correlation coefficient. Asterisks (*) indicate a significant correlation in that bin (permutation test: P < 0.05, false discovery rate [FDR] [51] corrected across time bins for each layer).

Second, we analyzed whether the ensemble of DNN units share representational similarity with the ensemble of human amygdala and hippocampal neurons (**Figure 3.8B, E, F**). We used the same stimuli (500 natural face images of 50 celebrities) as monkey recordings and recorded from 667 neurons in the human amygdala and hippocampus (340 neurons from the amygdala, 222 neurons form the anterior hippocampus, and 105 neurons from the posterior hippocampus; firing rate > 0.15 Hz) of 8 neurosurgical patients (23 sessions in total) [39]. Patients performed a one-back task (**Figure 3.8B**) and they could well recognize the faces [39]. The responses of 76/667 neurons (11.39%) differed between different face identities in a window 250-1250 ms following stimulus onset and these neurons were the real human identity-selective neurons. We grouped amygdala and hippocampal neurons as a single neuronal population (i.e., medial temporal lobe [MTL] neurons) for further analysis because they show very similar identity selectivity responses [39, 53]. We found that the pairwise distance from the top DNN layer significantly correlated with the neuronal pairwise distance from the human MTL, consistent with the processing stage along the ventral visual pathway (**Figure 3.8E**; see **Figure 3.8F** for temporal dynamics of the full-face model). We also found that the full-face model had a better correspondence with MTL neurons than the eyes-foveated model and mouth-foveated model (**Figure 3.8E, F**), and the mouth-foveated model recovered with full-face images within the critical period increased correspondence with MTL neurons compared to recovery outside the critical period (Conv2: permutation P = 0.001).

Together, by comparing between artificial units and real primate neurons, we not only revealed a systematic correspondence between the two face recognition systems but also showed that such correspondence was associated with DNN model performance and recovery.

3.3 Discussion

In this study, we systematically investigated face learning and facial information utilization during a critical period. Specifically, we revealed a critical period during development that has the following properties: (1) In the baseline condition, reduced facial information resulted in reduced model performance and subsequent inability to use information from the corresponding facial parts. (2) When full-face information was provided *within* the critical period, full recovery could be achieved; but recovery did not happen when full-face information was provided *outside* the critical period. (3) When complementary information was provided within the critical period, it could even override the original model and become a model like that trained with new information alone. We further provided a computational account with a learning rate that could explain the properties of critical periods, and we showed that by altering the learning rate, learning could be recovered. We also demonstrated an alternative approach, i.e., knowledge distillation and attention transfer, that can partially recover the model outside the critical period. Lastly, we showed that model performance and recovery were associated with identity-selective units as well as the correspondence with the primate visual systems. Together, our results not only highlight the importance of a critical period in face learning but also elucidate its underlying computational mechanism and restoration strategies.

Our present findings are consistent with the neurodevelopment concepts. First, brains have developmental critical periods. Classical studies have documented critical periods affecting a range of species and systems, from visual acuity in kittens [112, 113] to song learning in birds [114]. Uncorrected eye defects (e.g., strabismus, cataracts) during the critical period for visual development lead to amblyopia [88, 89]. Second, our results show that outside the critical period the learning process could not be restored, which is likely accounted for by the restricted learning rate. In analogy, brain rewiring is significantly reduced after the critical period [87], which may cause a reduced learning rate. On the other hand, pathways have been discovered in animal models through which critical periods may be re-opened in adults, making it possible to re-awaken a brain's youth-like plasticity and thus repair brain injury, recover sensory deficits, and treat neurodevelopmental disorders [115]. Third, network attention transfer could partially improve model performance, which is analogous to learning after development. Together, our present results highlight the importance of the critical period during development, which is highly consistent with the neurobiology of animal development.

Developmental prosopagnosia (DP) is an impairment in recognizing faces despite normal vision, intelligence, and socio-cognitive abilities and no history of brain damage [116]. Impaired development of face processing during the critical period may lead to prosopagnosia [117]. Here, we showed that information provided in the critical period determined subsequent information utilization, which is also consistent with the development of visual attention [118]. It has been recently argued that neural coding strategies during development may be highly dynamic [97]. More broadly, it has been shown that early emotional processing in young children differs from that observed in adolescents, who approach adults [119].

Our present results are relevant to neurodevelopmental disorders such as autism spectrum disorder (ASD). Many studies have documented abnormal face processing in people with ASD [120-128], and the such deficit has both a developmental [129] and genetic [130] root. In particular, people with ASD demonstrate impaired utilization of facial information. People with ASD have an increased tendency to saccade away from the eye region of faces when information is present in those regions [125], but instead have an increased preference to fixate on the location of the mouth [123]. During viewing naturalistic social videos, people with autism demonstrate abnormal patterns of social visual pursuit that are consistent with reduced saliency of eyes and increased saliency of mouths, bodies, and objects [122]. When viewing static faces, people with autism view non-feature areas of the faces significantly more often but core feature areas of the faces (e.g., eyes and mouth) significantly less often than controls [124] and they have piecemeal rather than configural strategies [131]. Similarly, some research suggests that people with ASD demonstrate active avoidance of fixating the eyes in faces, which in turn influences the recognition performance of emotions [121], whereas other research suggests that children with ASD demonstrate gaze indifference and passive insensitivity to the social signals in others' eyes at the time of initial diagnosis [132]. The atypical facial fixations are complemented by neuronal evidence of abnormal processing of information from the eye region of faces in blood-oxygen-level dependent (BOLD) fMRI [133] and single-neuron responses in the amygdala [134].

Our present study has provided a possible computational account for such deficit in ASD: reduced access to eye information during the critical period resulted in impaired utilization of eye information and thus gaze to the eyes after the development. Therefore, our results point to a potential way to recover from such face processing deficits by early training with guided fixation onto the eyes. On the other hand, although our results suggest that recovery outside the critical

period could not restore the normal function, the network attention transfer has provided an important alternative to recover the learning, which is also consistent with the behavioral training strategy currently being applied in ASD. It is worth noting that although our results highlight the importance of critical periods, a future study is needed to understand whether such deficits in ASD are the cause or consequence of a critical period.

Deep neural networks currently provide the most compelling quantitative models of the response patterns of neurons throughout the primate ventral visual stream. It has been shown that DNNs trained with unsupervised contrastive embedding can well simulate infant learning during development [105]. Interestingly, such unsupervised learning produces brain-like representations even when trained solely with real human child developmental data collected from head-mounted cameras. Furthermore, our current simulation approach with surrogate computational models can be generalized to other sensorimotor domains (e.g., auditory). It is worth noting that our findings were not restricted to the model or visualization method used, but could be replicated with other models and visualization methods. Moreover, such artificial neural network models for face recognition have a biological correspondence with both human and non-human primate neurons [104]. In particular, using human single-neuron recordings, it has been shown that neurons in the human amygdala and hippocampus encode facial features (e.g., the eyes and mouth) and eye movement to these facial features [135], which may in turn be related to abnormal facial feature representation in ASD [134, 136]. A future study is needed to directly investigate the neuronal mechanisms for face learning concerning critical periods.

3.4 Methods

3.4.1 Training and testing data

We used a subset of images from the CASIA_WebFace dataset as the training and testing data [137]. The CASIA-WebFace dataset has been used for various face verification and identification tasks. The dataset contains 494,414 face images of 10,575 real identities collected from the internet. In this study, we selected images from 50 identities that have more than 400 images in the dataset. The identities were diverse in race, sex, and age. Because the image quality varies, we manually removed the images that have a low resolution, major facial occlusion, or

extreme facial angles. As a result, our training and testing dataset contained 37,000 images from 50 different identities, with each identity having more than 300 image samples (**Figure 3.1A**).

3.4.2 Image processing

We applied the face and facial landmark detection tool, Multi-Task Cascaded Convolutional Neural Networks (MTCNN) [138], to crop the faces from the images and label the facial landmarks. MTCNN is a framework developed as a solution for both face detection and face alignment. It is one of the most popular and accurate face-detection tools. The process consists of three stages of convolutional networks (CNNs). It uses a shallow CNN as the first step to produce candidate windows quickly. Through a more intricate CNN, it improves the suggested candidate windows in the second step. To further refine the outcome and output face landmark positions, a third CNN that is more complicated than the others is used in the third step. After applying the MTCNN, we derived a tight bounding box outlining the face area as well as the coordinates for the centers of the eyes, nose tip, and two corners of the mouth.

We next applied foveated imaging to generate images that mimic human foveation/fixation (i.e., the spatial resolution is the highest at the point of the fovea and drops rapidly away from that point as a function of eccentricity, and thus the region around the point of fixation [or foveation point] is sampled with the highest intensity and perceived with the highest sensitivity; **Figure 3.1B, C**). Foveated imaging is a method of digital image processing where the level of detail or resolution varies across the image according to one or more fixation points. We utilized the open-source Python implementation of image retina transformation for foveated imaging (https://github.com/ouyangzhibo/Image_Foveation_Python) and produced two groups of foveated images (eyes-foveated and mouth-foveated) based on the facial landmarks identified by the MTCNN (**Figure 3.1B, C**). Specifically, in eyes-foveated images, only the eye region was clear and the rest of the image was blurry, whereas in mouth-foveated images, only the mouth region was clear and the rest of the image was blurry.

We lastly cropped the images based on the bounding box derived using the MTCNN (**Figure 3.1B, C**). All subsequent analyses were based on the cropped images.

3.4.3 Model training and testing

We used the well-known deep neural network (DNN) implementation based on the ResNet50 [139] convolutional neural network (CNN) architecture (see details in **Figure 3.2A**). Because the goal of the present study is to understand model performance during training, we trained the network from scratch.

We first trained three *base models*, using full-face images, eyes-foveated images, and mouth-foveated images. For each model, 80% of the images were used as the training set and the remaining 20% of the images were used as the testing/validation set. We used the SGD optimizer with an initial learning rate of 10^{-2} and all the models were trained for 150 epochs. An adaptive learning rate scheduler was applied, which halved the current learning rate when the loss of validation did not drop for 5 epochs. To update the weights, we computed the cross-entropy loss on random batches of 32 images (scaled to 224×224 pixels) for back-propagation. We derived similar results using different initial learning rates (**Supplementary Figure 7A**).

We next trained *recovery models* based on base models by providing base models with different information (i.e., same images with different foveation) during their training. We implemented the recovery at different stages of model training, and the model training only continued with the new set of images. It is worth noting that to facilitate a direct comparison between base models and recovery models, we used the same parameters for the recovery models as the corresponding base models, including the epoch-by-epoch learning rate. The recovery models were trained with the new set of images until they reached 150 epochs. For instance, if the base model was recovered at epoch 15, the recovery model would continue the training using the learning rate at epoch 15 and inherit all the subsequent learning rates from the base model. The recovery model would continue the training for another 135 epochs to have a total of 150 training epochs.

To compare between different models, we always used the same set of original images (not foveated) to test all models.

3.4.4 Critical period

The critical period is a time window of early post-natal development during which sensory deficits can lead to permanent skill impairment [85]. Similar to humans and other animals, deep artificial neural networks exhibit critical periods during which a temporary stimulus deficit can

impair the model performance [101]. In this study, we defined the critical period of the DNN as the epochs of the early fast learning phase (following the same definition as in [101]). Specifically, based on the learning curves of the base models, the first 30 epochs were defined as the critical period. We thus chose epoch 15 and epoch 80 to compare recovery models starting within vs. outside the critical period; and we derived similar results using other epochs (e.g., epoch 10 vs. epoch 90) to compare recovery models (**Supplementary Figure 7B, C**).

3.4.5 Model visualization and quantification

In our experiments, we adopted the Gradient-weighted Class Activation Mapping (Grad-CAM) [22] as our visualization tool. Grad-CAM is a popular technique for visualizing which regions in the original image contribute to the final output. It uses the gradients of the target category flowing into a certain convolution layer, usually the last one, to produce a coarse localization map highlighting the important pixels/regions in the image for predicting the category. This approach reveals the implicit attention of the model to make the real contributor of features in the input image distinguishable. Grad-CAM is an improvement from the previous approach, CAM [23], for both versatility and accuracy.

We further quantified Grad-CAM intensity in the eyes and mouth regions of interest (ROIs) (**Figure 3.2B**). We defined the eyes and mouth ROIs in the image based on the facial landmarks for each image. It is worth noting that the eyes ROI and the mouth ROI were of a similar size across images (eyes: 4894.76±1668.83 pixels, mouth: 4843.66±2110.94; two-tailed two-sample *t*-test: $t(7536) = 1.17$, $P = 0.24$). Because most of the nose region was covered by both the eyes and mouth ROIs, we did not separately analyze the Grad-CAM intensity for the nose region. In addition to the average Grad-CAM intensity in each ROI, we also calculated the proportion of total intensity in each ROI by dividing the total Grad-CAM intensity of an image.

3.4.6 Knowledge distillation and attention transfer

The basic idea behind knowledge distillation (KD) is to train a small, lightweight model using supervised information from a larger model with superior performance, in order to improve its own performance. It was first proposed by Hinton et al. in 2015 [140]. The large model is known as the teacher model, while the small model is characterized as the student model. The supervised information from the output of the teacher model is called “knowledge”, and the process of student

learning to migrate the supervised information from the teacher is called "distillation". Our recovery experiment, in contrast to the original KD concept, is built upon two identical architectures (**Figure 3.6A**). The gap in model performance was mainly reflected in the different stimuli. One model was trained by the full-face images, which were regarded as the teacher model (**Figure 3.6A** upper), whereas the other model was trained by the mouth-foveated images, which were regarded as the student model (**Figure 3.6A** lower). Our purpose here was to guide the mouth-foveated model to learn new features with information from the full-face model when the mouth-foveated model had already missed the critical period (i.e., recovery started at epoch 80, which is outside the critical period; note that this is the same model for recovery outside the critical period).

DNN models can barely learn new features once they have passed the critical period, especially when the learning rate becomes extremely low. To reinforce the recovery effect, we used another technique called attention transfer (AT) [141], which can work together with KD. We used the average feature map of each group of convolutional layers as the attention and transferred the attention from the teacher model to the student model. It is worth noting that only the student model was updated during the process, while the teacher model acted as a supervisor, enabling the student model to learn from its own information, and as a result, all weights in the teacher model were frozen.

We made the learning rate of the student model identical to that of the foveated model to determine whether the KD-AT method could effectively aid in the recovery of the original foveated model under extremely low learning rates. We computed the attention transfer loss after each convolution group using the following loss function:

$$L_{AT} = L(W_{S,}x) + \sum_{j \in I} \left\| \frac{Q_S^j}{\left\| Q_S^j \right\|_2} - \frac{Q_T^j}{\left\| Q_T^j \right\|_2} \right\|_2$$

where $L(W_{S,}x)$ denotes the standard cross-entropy loss and I denotes the indices of all teacher-student activation layer pairs. $Q_S^j = vec\left(Avg\left(A_S^j\right)\right)$ and $Q_T^j = vec\left(Avg\left(A_T^j\right)\right)$ are the j-th pair of student and teacher attention maps in vectorized form, respectively. And the attention map is the cross-channel average of the activation tensor A.

Finally, we added the knowledge distillation loss between the output of the teacher model (y_t) and the student model (y_s) to the previous loss function. As a result, the final total loss was obtained as follows:

$$L_{AT} = L(W_{S,}x) + \sum_{j \in I} \left\| \frac{Q_S^j}{\left\| Q_S^j \right\|_2} - \frac{Q_T^j}{\left\| Q_T^j \right\|_2} \right\|_p + L_{KD}(y_t, y_s)$$

With this total loss function, we aimed to enable the student model not only to make correct predictions but also to learn similar feature representations as the teacher model.

3.4.7 Selection of identity-selective DNN units and primate neurons

To select identity-selective units [104], we used a one-way ANOVA to identify identity-selective units that had a significantly unequal response to different identities ($P < 0.01$). We further imposed an *additional criterion* to identify a *subset* of identity-selective units with selective identities: the response of an identity was 2 standard deviations (SD) above the mean of responses from all identities. These identified identities whose response stood out from the global mean were the encoded identities.

We followed the identical selection procedure for primate neurons [39, 104]. We used the mean firing rate in a time window 250-1000 ms after stimulus onset as the response to each face for primate neurons. Note that we also used this response to study the correlation between DNN units and primate neurons.

3.4.8 Neural recordings from a monkey

Detailed procedure has been described in our previous study [104]. Briefly, we recorded from the anterior and central inferotemporal (IT) cortex in one male rhesus macaque (Macaca mulatta) using two Utah arrays (Blackrock Microsystems) in (see [109, 142] for details). We detected the multi-unit spikes after the raw data were zero-phase band-pass filtered between 300-6000 Hz (Matlab ellip function, fourth order with 0.1 decibel pass-band ripple and 40 dB stop-band attenuation), and we used multi-unit activity (MUA) for analyses. To test with an independent dataset, the monkey passively viewed 500 images from the CelebA dataset [111]. In each trial, the monkey first viewed a white central fixation point (0.2 degrees of visual angle [DVA]) on a gray background for 300 ms to initiate a trial. Then, 8 faces were presented for 100 ms each, each followed by a blank (gray) screen for an inter-stimulus-interval (ISI) of 100 ms. The central

fixation point persisted through the trial, and fluid reward was given if the monkey successfully fixated through the entire trial. The inter-trial-interval (ITI) of blank gray screen was at least 500 ms. We recorded 4155 trials in total, and we rejected 666 trials where the monkey broke the fixation (±2 DVA). For each round of presentation, we generated a random sequence for the 500 faces; and we used different sequences for different rounds of presentation. All procedures conformed to local and U.S. National Institutes of Health guidelines, including the U.S. National Institutes of Health Guide for Care and Use of Laboratory Animals. All experiments were performed with the approval of the MIT Institutional Animal Care and Use Committee (IACUC).

3.4.9 Neural recordings from human neurosurgical patients

Detailed procedure has been described in our previous study [104, 143]. Briefly, we recorded from implanted depth electrodes in the amygdala and hippocampus from 8 neurosurgical patients (23 sessions in total) with pharmacologically intractable epilepsy. Bipolar wide-band recordings (0.1-9000 Hz), using one of the eight microwires as reference, were sampled at 32 kHz and stored continuously for off-line analysis with a Neuralynx system. The raw signal was filtered with a zero-phase lag 300-3000 Hz bandpass filter and spikes were sorted using a semi-automatic template matching algorithm as described previously [83]. Units were carefully isolated and recording and spike sorting quality were assessed quantitatively. Only units with an average firing rate of at least 0.15 Hz (entire task) were considered. Only single units were considered. Trials were aligned to stimulus onset and we used the mean firing rate in a time window 250 ms to 1000 ms after stimulus onset as the response to each face. We employed a 1-back task with the same 500 CelebA images as monkey recordings. In each trial, a single face was presented at the center of the screen for a fixed duration of 1 second, with uniformly jittered inter trial interval (ITI) of 0.5-0.75 seconds. Patients pressed a button if the present face image was identical to the immediately previous image. All participants provided written informed consent using procedures approved by the Institutional Review Board of West Virginia University (WVU).

3.4.10 Match between DNN units and primate neurons

We employed a pairwise distance metric [50] to compare neural coding of face identities between primate neurons and DNN units. For each pair of identities, we used the dissimilarity value (1−Pearson's r) [84] as a distance metric. The primate neuronal distance metric was

calculated between firing rates of all recorded neurons and the DNN distance metric was calculated between activation of all DNN units. We then correlated the primate neuronal distance metric and the DNN distance metric. To determine statistical significance, we used a non-parametric permutation test with 1000 runs. In each run, we randomly shuffled the face labels and calculated the correlation between the primate neuronal distance metric and the DNN distance metric. The distribution of correlation coefficients computed *with* shuffling (i.e., null distribution) was eventually compared to the one *without* shuffling (i.e., observed response). If the correlation coefficient of the observed response was greater than 95% of the correlation coefficients from the null distribution, it was considered *significant*. A significant correlation indicated that the DNN face space corresponded to the primate neuronal face space [50]. We computed the correlation for each DNN layer so that we could determine the specific layer that the neuronal population encoded. For each face identity, we averaged the response of all faces of that identity to get a single mean firing rate.

To get temporal dynamics, for human neurons, we used a moving window with a bin size of 500 ms and a step size of 50 ms (given the sparseness of human MTL neurons, this time window is commonly used [39, 62, 136]). The first bin started −300 ms relative to trial onset (bin center was thus 50 ms before trial onset), and we tested 19 consecutive bins (the last bin was thus from 600 ms to 1100 ms after trial onset). For monkey neurons, we used a moving window with a bin size of 40 ms and a step size of 10 ms. The first bin started −70 ms relative to stimulus onset (bin center was thus 50 ms before stimulus onset), and we tested 26 consecutive bins (the last bin was thus from 180 ms to 220 ms after stimulus onset). We used Bonferroni correction to correct for multiple comparisons across DNN layers and false discovery rate (FDR) [51] to correct for multiple comparisons across time bins.

We used a bootstrap with 1000 runs to compare between models (full-face vs. eyes-foveated and full-face vs. mouth-foveated). In each run, data from 70% of the identities (i.e., 35 identities) was randomly selected to calculate the correspondence between DNN units and primate neurons. We thus created a distribution of correspondence for each model.

We further used a permutation test with 1000 runs to statistically compare the correspondence for recovery within vs. outside the critical period. In each run, we shuffled the recovery labels (within vs. outside) and calculated the difference in correspondence between

recoveries. We then compared the observed difference in correspondence between recoveries with the permuted null distribution to derive statistical significance.

3.4.11 Data availability

All data and statistical analysis code are available on GitHub (https://github.com/JingeW/Critical-Period-Analysis).

Chapter 4: Contributions and future work

4.1 Summary

In this dissertation, we have conducted a comprehensive review of the complex history between artificial intelligence (AI) and neuroscience, examining how both fields have contributed to each other's development while overcoming their own challenges. This exploration has underscored the potential and significance of interdisciplinary research in advancing our understanding of both AI and neuroscience.

Our investigation delved into the computational mechanisms underlying face perception, identifying Deep Neural Networks (DNNs) as an appropriate framework for modeling this cognitive process. We introduced several approaches that can be tailored to different tasks, stimulus types, and training paradigms. Utilizing these adaptive techniques, we examined the neural foundations of face perception, gaining a deeper understanding of the processes involved in this complex task.

In our systematic study of DNN layers and units, we scrutinized their architectural principles, learning rules, and optimization methods. This in-depth analysis allowed us to explore factors that contribute to the effectiveness of DNNs in complex pattern recognition tasks, including face perception. This study not only makes our computational models more scientifically grounded and rational, but also provides an important reference for simulating the perceptual processes of the human brain.

To bridge the gap between AI and neuroscience, we correlated DNNs features with brain activity data, investigating how the computational processes of deep learning models parallel the neural processes in the human brain. This approach enabled us to identify similarities and differences, providing valuable insights into the potential of DNNs as models of neural computation.

Furthermore, we emphasize the importance of interpretability in computational models, arguing that making these models more transparent and understandable is critical for both scientific progress and practical applications. For our task, the interpretability of artificial neural networks can not only help us better understand the compatibility and rationality of the model but also provide strong support for understanding biological hypotheses. With good interpretability, researchers can build trust in AI systems and ensure their benefits are effectively exploited.

4.2 Contribution

Our investigation into the brain's encoding revealed a novel face code of neurons in the human medial temporal lobe (MTL). The research found identity neurons in the human MTL that encode face identities based on visual similarity rather than conceptual similarity, and feature neurons that exhibit region-based feature coding. This discovery has opened new avenues for understanding the neural mechanisms underlying face recognition and further highlights the importance of both biological and computational approaches to this field of study.

Our analysis of the response characteristics of the face recognition DNN identified a subset of DNN units selective for facial identity. These identity-selective units can extend their discriminative abilities to face identities displayed in various styles, as well as to those not involved in training. This finding demonstrates the robustness and flexibility of such models, suggesting their potential application in developing more effective and adaptable face recognition systems. Additionally, further correlation analysis with brain data reveals similarities between our computational model and primate brain processes during face recognition, creating a bridge between the sparse coding hypothesis in neuroscience and deep learning. This connection enhances our understanding of the underlying principles governing both artificial and biological systems.

Furthermore, our study of critical periods deepened our understanding of the learning and development process for face recognition. This analysis not only revealed the relationship between face recognition development and critical periods but also provided strategies for addressing face processing deficits. This insight into the significance of the critical period can inform interventions and treatments for individuals with face recognition difficulties, ultimately improving their quality of life.

All these neuroscience-related discussions have been supported by AI algorithms and artificial neural networks, which have proven to be valuable tools for simulating real neural networks. These silico experiments enable researchers to test numerous parameters and achieve better control over experimental conditions, which is often not feasible when working with human or animal subjects. Our experiments design presents a framework for utilizing computer science techniques to address neuroscience challenges.

In conclusion, our research makes a substantial contribution to the interdisciplinary field of computer science and neuroscience, particularly in the area of facial recognition. By investigating the neural and computational mechanisms of face perception in both biological

systems and artificial neural networks, we have deepened our understanding of these processes and pinpointed key factors that impact their development and efficacy. Capitalizing on the strengths of both fields enables us to further comprehend the human brain, propel the evolution of artificial intelligence systems, and ultimately unravel the intricate mechanisms governing perception, cognition, and action. Moreover, our findings not only offer valuable insights for the design of more effective AI systems but also hold the potential to aid in the development of novel therapies and interventions for treating facial perception disorders or other related neurological conditions.

4.3 Future work

Building upon our research, it would be intriguing to investigate how neurons encode facial social features during perception. Another topic worth investigating is the role of explicit attention mechanisms, such as the transformer, in face perception and determining whether these mechanisms can provide a plausible explanation for certain neuroscientific hypotheses. Furthermore, incorporating third-generation artificial neural networks, spiking neural networks (SNNs), as computational models could provide valuable insights. By leveraging its powerful bio-interpretability, we can delve deeper into the underlying mechanisms of face perception, thus continuing to advance the intersection of artificial intelligence and neuroscience. These explorations will not only enrich our understanding of the human brain but also contribute to the development of more sophisticated AI systems.

Bibliography

1. Fan, J., et al., *From Brain Science to Artificial Intelligence.* Engineering, 2020. **6**(3): p. 248-252.
2. Hebb, D.O., 2005.
3. Rosenblatt, F., *The perceptron: A probabilistic model for information storage and organization in the brain.* Psychological Review, 1958. **65**(6): p. 386-408.
4. Hubel, D.H. and T.N. Wiesel, *Receptive fields of single neurones in the cat's striate cortex.* The Journal of Physiology, 1959. **148**(3): p. 574-591.
5. LeCun, Y., et al., *Backpropagation Applied to Handwritten Zip Code Recognition.* Neural Computation, 1989. **1**(4): p. 541-551.
6. Krizhevsky, A., I. Sutskever, and G.E. Hinton, *ImageNet classification with deep convolutional neural networks.* Communications of the ACM, 2017. **60**(6): p. 84-90.
7. LeCun, Y., Y. Bengio, and G. Hinton, *Deep learning.* Nature, 2015. **521**(7553): p. 436-444.
8. Yamins, D.L.K., et al., *Performance-optimized hierarchical models predict neural responses in higher visual cortex.* Proceedings of the National Academy of Sciences, 2014. **111**(23): p. 8619.
9. Bao, P., et al., *A map of object space in primate inferotemporal cortex.* Nature, 2020.
10. Richards, B.A., et al., *A deep learning framework for neuroscience.* Nature Neuroscience, 2019. **22**(11): p. 1761-1770.
11. Miller, G.A., *The cognitive revolution: a historical perspective.* Trends in Cognitive Sciences, 2003. **7**(3): p. 141-144.
12. Turing, A.M., *Computing Machinery and Intelligence.* 2009: p. 23-65.
13. Minsky, M. and S.A. Papert, 2017.
14. Brauwers, G. and F. Frasincar, *A General Survey on Attention Mechanisms in Deep Learning.* IEEE Transactions on Knowledge and Data Engineering, 2021: p. 1-1.
15. McCarthy, J., *Defending AI research: a collection of essays and reviews.* 1996.
16. Bengio, Y., P. Frasconi, and P. Simard. *The problem of learning long-term dependencies in recurrent networks*. in *IEEE international conference on neural networks*. 1993. IEEE.
17. Lillicrap, T.P., et al., *Backpropagation and the brain.* Nature Reviews Neuroscience, 2020. **21**(6): p. 335-346.
18. Maass, W., *Networks of spiking neurons: the third generation of neural network models.* Neural networks, 1997. **10**(9): p. 1659-1671.
19. Chrol-Cannon, J. and Y. Jin, *Computational modeling of neural plasticity for self-organization of neural networks.* Biosystems, 2014. **125**: p. 43-54.
20. Maass, W., *Liquid State Machines: Motivation, Theory, and Applications.* 2011: p. 275-296.
21. Maass, W., T. Natschläger, and H. Markram, *Real-Time Computing Without Stable States: A New Framework for Neural Computation Based on Perturbations.* Neural Computation, 2002. **14**(11): p. 2531-2560.
22. Selvaraju, R.R., et al. *Grad-cam: Visual explanations from deep networks via gradient-based localization*. in *Proceedings of the IEEE international conference on computer vision*. 2017.
23. Zhou, B., et al. *Learning deep features for discriminative localization*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
24. Freeman, W.J., *Mass action in the nervous system*. Vol. 2004. 1975: Citeseer.
25. LeCun, Y., J. Denker, and S. Solla, *Optimal brain damage.* Advances in neural information processing systems, 1989. **2**.

26. Hinton, G.E., et al., *Improving neural networks by preventing co-adaptation of feature detectors.* arXiv preprint arXiv:1207.0580, 2012.
27. Hensch, T.K., *Critical period regulation.* Annual Review of Neuroscience, 2004. **27**(1): p. 549-579.
28. Taylor, D., *Critical period for deprivation amblyopia in children.* Transactions of the ophthalmological societies of the United Kingdom, 1979. **99**(3): p. 432-439.
29. Mitchell, D.E., *The extent of visual recovery from early monocular or binocular visual deprivation in kittens.* The Journal of physiology, 1988. **395**(1): p. 639-660.
30. Tsao, D.Y., et al., *A Cortical Region Consisting Entirely of Face-Selective Cells.* Science, 2006. **311**(5761): p. 670-674.
31. Rolls, E., *Neurons in the cortex of the temporal lobe and in the amygdala of the monkey with responses selective for faces.* Hum Neurobiol., 1984. **3**(4): p. 209-22.
32. Kreiman, G., C. Koch, and I. Fried, *Category-specific visual responses of single neurons in the human medial temporal lobe.* Nat Neurosci, 2000. **3**(9): p. 946-953.
33. Turk, M.A. and A.P. Pentland. *Face recognition using eigenfaces*. in *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 1991. IEEE.
34. Freiwald, W.A., D.Y. Tsao, and M.S. Livingstone, *A face feature space in the macaque temporal lobe.* Nat Neurosci, 2009. **12**(9): p. 1187-1196.
35. Han, T., et al., *Replicating neuroscience observations on ML/MF and AM face patches by deep generative model.* Neural Computation, 2019. **31**(12): p. 2348-2367.
36. O'Toole, A.J., et al., *Face Space Representations in Deep Convolutional Neural Networks.* Trends in Cognitive Sciences, 2018. **22**(9): p. 794-809.
37. VanRullen, R. and L. Reddy, *Reconstructing faces from fMRI patterns using deep generative neural networks.* Communications Biology, 2019. **2**(1): p. 193.
38. Grossman, S., et al., *Deep Convolutional modeling of human face selective columns reveals their role in pictorial face representation.* bioRxiv, 2018: p. 444323.
39. Cao, R., et al., *Feature-based encoding of face identity by single neurons in the human medial temporal lobe.* bioRxiv, 2020: p. 2020.09.01.278283.
40. Tran, L., X. Yin, and X. Liu. *Disentangled representation learning gan for pose-invariant face recognition*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
41. Parde, C.J., et al., *Single Unit Status in Deep Convolutional Neural Network Codes for Face Identification: Sparseness Redefined.* arXiv preprint arXiv:2002.06274, 2020.
42. Srivastava, N., et al., *Dropout: a simple way to prevent neural networks from overfitting.* The journal of machine learning research, 2014. **15**(1): p. 1929-1958.
43. Zoph, B. and Q.V. Le, *Neural architecture search with reinforcement learning.* arXiv preprint arXiv:1611.01578, 2016.
44. Frankle, J. and M. Carbin, *The lottery ticket hypothesis: Finding sparse, trainable neural networks.* arXiv preprint arXiv:1803.03635, 2018.
45. Simoncelli, E.P. and B.A. Olshausen, *Natural Image Statistics and Neural Representation.* Annual Review of Neuroscience, 2001. **24**(1): p. 1193-1216.
46. Yin, R.K., *Looking at upside-down faces.* Journal of experimental psychology, 1969. **81**(1): p. 141.
47. Simonyan, K. and A. Zisserman, *Very deep convolutional networks for large-scale image recognition.* arXiv preprint arXiv:1409.1556, 2014.
48. Paszke, A., et al., *Pytorch: An imperative style, high-performance deep learning library.* Advances in neural information processing systems, 2019. **32**.
49. Yamins, D.L.K. and J.J. DiCarlo, *Using goal-driven deep learning models to understand sensory cortex.* Nature Neuroscience, 2016. **19**: p. 356.

50. Grossman, S., et al., *Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks.* Nature Communications, 2019. **10**(1): p. 4934.
51. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society. Series B (Methodological), 1995. **57**(1): p. 289-300.
52. Mormann, F., et al., *Latency and Selectivity of Single Neurons Indicate Hierarchical Processing in the Human Medial Temporal Lobe.* The Journal of Neuroscience, 2008. **28**(36): p. 8865-8872.
53. Quian Quiroga, R., et al., *Invariant visual representation by single neurons in the human brain.* Nature, 2005. **435**(7045): p. 1102-1107.
54. Chang, L., et al., *Explaining face representation in the primate brain using different computational models.* Current Biology, 2021. **31**(13): p. 2785-2795.e4.
55. Baek, S., et al., *Face detection in untrained deep neural networks.* Nature Communications, 2021. **12**(1): p. 7328.
56. Allen-Zhu, Z. and Y. Li, *Backward feature correction: How deep learning performs deep learning.* arXiv preprint arXiv:2001.04413, 2020.
57. Quian Quiroga, R., *Concept cells: the building blocks of declarative memory functions.* Nature Reviews Neuroscience, 2012. **13**: p. 587.
58. Cadieu, C.F., et al., *Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition.* PLOS Computational Biology, 2014. **10**(12): p. e1003963.
59. Gross, C., et al., *Inferior temporal cortex as a pattern recognition device*, in *Computational Learning and Cognition*, E. Baum, Editor. 1993, Society for Industrial and Applied Mathematics: Philadelphia. p. 44-73.
60. DiCarlo, James J., D. Zoccolan, and Nicole C. Rust, *How Does the Brain Solve Visual Object Recognition?* Neuron, 2012. **73**(3): p. 415-434.
61. Quian Quiroga, R., *No Pattern Separation in the Human Hippocampus.* Trends in Cognitive Sciences, 2020. **24**(12): p. 994-1007.
62. Wang, S., et al., *The human amygdala parametrically encodes the intensity of specific facial emotions and their categorical ambiguity.* Nature Communications, 2017. **8**: p. 14821.
63. Wang, S., et al., *Neurons in the human amygdala selective for perceived emotion.* Proceedings of the National Academy of Sciences, 2014. **111**(30): p. E3110-E3119.
64. Wang, S., et al., *Abstract goal representation in visual search by neurons in the human pre-supplementary motor area.* Brain, 2019. **142**(11): p. 3530-3549.
65. Wang, S., et al., *Encoding of Target Detection during Visual Search by Single Neurons in the Human Brain.* Current Biology, 2018. **28**(13): p. 2058-2069.e4.
66. Cao, R., et al., *Encoding of facial features by single neurons in the human amygdala and hippocampus.* bioRxiv, 2020.
67. Cao, R., et al., *A neuronal social trait space of faces in the human amygdala and hippocampus.* bioRxiv, 2021.
68. Parkhi, O.M., A. Vedaldi, and A. Zisserman, *Deep face recognition.* 2015.
69. Taigman, Y., et al. *Deepface: Closing the gap to human-level performance in face verification*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.
70. Shen, G., et al., *Deep image reconstruction from human brain activity.* PLoS Comput Biol, 2019. **15**(1): p. e1006633.
71. Shen, G., et al., *End-to-End Deep Image Reconstruction From Human Brain Activity.* Frontiers in Computational Neuroscience, 2019. **13**(21).

72. Yildirim, I., et al., *Efficient inverse graphics in biological face processing.* Science Advances, 2020. **6**(10): p. eaax5979.
73. Dobs, K., et al., *Brain-like functional specialization emerges spontaneously in deep neural networks.* bioRxiv, 2021: p. 2021.07.05.451192.
74. Liu, Z., et al. *Deep Learning Face Attributes in the Wild*. in *Proceedings of International Conference on Computer Vision (ICCV)*. 2015.
75. Chen, Y., Y.-K. Lai, and Y.-J. Liu. *Cartoongan: Generative adversarial networks for photo cartoonization*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.
76. Mishra, A., et al. *IIIT-CFW: A benchmark database of cartoon faces in the wild*. in *European Conference on Computer Vision*. 2016. Springer.
77. Russakovsky, O., et al., *Imagenet large scale visual recognition challenge.* International journal of computer vision, 2015. **115**(3): p. 211-252.
78. Hinton, G.E. and S.T. Roweis, *Stochastic neighbor embedding.* Advances in neural information processing systems, 2003: p. 857-864.
79. van der Maaten, L. and G. Hinton, *Visualizing Data using t-SNE.* Journal of Machine Learning Research, 2008. **9**: p. 2579-2605.
80. Pedregosa, F., et al., *Scikit-learn: Machine learning in Python.* the Journal of machine Learning research, 2011. **12**: p. 2825-2830.
81. Kar, K., et al., *Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior.* Nat Neurosci, 2019. **22**(6): p. 974-983.
82. Kar, K. and J.J. DiCarlo, *Fast Recurrent Processing via Ventrolateral Prefrontal Cortex Is Needed by the Primate Ventral Stream for Robust Core Visual Object Recognition.* Neuron, 2021. **109**(1): p. 164-176 e5.
83. Rutishauser, U., E.M. Schuman, and A.N. Mamelak, *Online detection and sorting of extracellularly recorded action potentials in human medial temporal lobe recordings, in vivo.* Journal of Neuroscience Methods, 2006. **154**(1-2): p. 204-224.
84. Kriegeskorte, N., M. Mur, and P. Bandettini, *Representational similarity analysis - connecting the branches of systems neuroscience.* Frontiers in Systems Neuroscience, 2008. **2**(4).
85. Berardi, N., T. Pizzorusso, and L. Maffei, *Critical periods during sensory development.* Current Opinion in Neurobiology, 2000. **10**(1): p. 138-145.
86. Birdsong, D., *Second language acquisition and the critical period hypothesis*. 1999: Routledge.
87. Hensch, T.K., *Critical period plasticity in local cortical circuits.* Nature Reviews Neuroscience, 2005. **6**(11): p. 877-888.
88. Daw, N.W., *Critical Periods and Amblyopia.* Archives of Ophthalmology, 1998. **116**(4): p. 502-505.
89. Hensch, T.K. and E.M. Quinlan, *Critical periods in amblyopia.* Visual Neuroscience, 2018. **35**: p. E014.
90. Geldart, S., et al., *The effect of early visual deprivation on the development of face processing.* Developmental Science, 2002. **5**(4): p. 490-501.
91. McKone, E., et al., *A critical period for faces: Other-race face recognition is improved by childhood but not adult social contact.* Scientific Reports, 2019. **9**(1): p. 12820.
92. Parr, L.A., *The evolution of face processing in primates.* Philosophical Transactions of the Royal Society B: Biological Sciences, 2011. **366**(1571): p. 1764-1777.
93. Pascalis, O., M. Fort, and P.C. Quinn, *Development of face processing: are there critical or sensitive periods?* Current Opinion in Behavioral Sciences, 2020. **36**: p. 7-12.
94. Pierce, K., et al., *Face processing occurs outside the fusiform `face area' in autism: evidence from functional MRI.* Brain, 2001. **124**(10): p. 2059-2073.

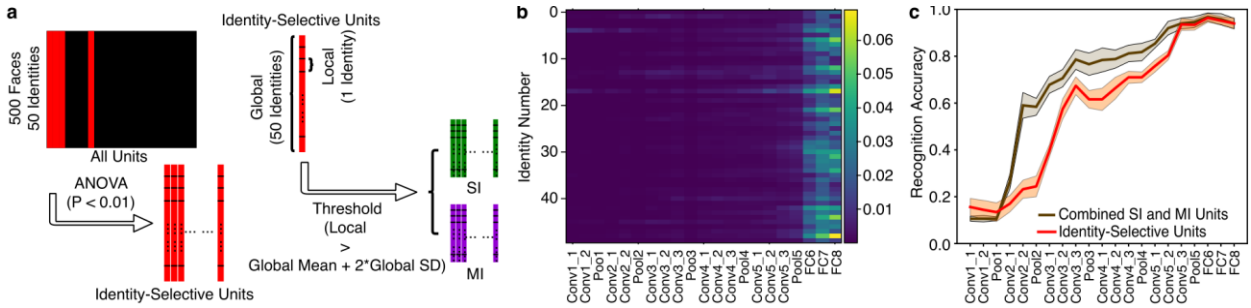
95. Röder, B., et al., *Sensitive periods for the functional specialization of the neural system for human face processing.* Proceedings of the National Academy of Sciences, 2013. **110**(42): p. 16760-16765.
96. Sugita, Y., *Innate face processing.* Current Opinion in Neurobiology, 2009. **19**(1): p. 39-44.
97. Avitan, L. and G.J. Goodhill, *Code Under Construction: Neural Coding Over Development.* Trends in Neurosciences, 2018. **41**(9): p. 599-609.
98. Golarai, G., et al., *Differential development of high-level visual cortex correlates with category-specific recognition memory.* Nature Neuroscience, 2007. **10**(4): p. 512-522.
99. Gomez, J., et al., *Microstructural proliferation in human cortex is coupled with the development of face processing.* Science, 2017. **355**(6320): p. 68.
100. Golarai, G., K. Grill-Spector, and A.L. Reiss, *Autism and the development of face processing.* Clinical Neuroscience Research, 2006. **6**(3): p. 145-160.
101. Achille, A., M. Rovere, and S. Soatto, *Critical learning periods in deep neural networks.* ICLR, 2019.
102. Parkhi, O., A. Vedaldi, and A. Zisserman, *Deep face recognition*. 2019, British Machine Vision Association. p. 1-12.
103. Schroff, F., D. Kalenichenko, and J. Philbin. *Facenet: A unified embedding for face recognition and clustering*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.
104. Wang, J., et al., *Face identity coding in the deep neural network and primate brain.* Communications Biology, 2022. **5**(1): p. 611.
105. Zhuang, C., et al., *Unsupervised neural network models of the ventral visual stream.* Proceedings of the National Academy of Sciences, 2021. **118**(3): p. e2014196118.
106. Grill-Spector, K., et al., *The functional neuroanatomy of face perception: from brain measurements to deep neural networks.* Interface Focus, 2018. **8**(4): p. 20180013.
107. Hill, M.Q., et al., *Deep convolutional neural networks in the face of caricature.* Nature Machine Intelligence, 2019. **1**(11): p. 522-529.
108. Dobs, K., et al., *Brain-like functional specialization emerges spontaneously in deep neural networks.* Science Advances, 2022. **8**(11): p. eabl8913.
109. Kar, K., et al., *Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior.* Nature Neuroscience, 2019. **22**(6): p. 974-983.
110. Deng, W., J.B. Aimone, and F.H. Gage, *New neurons and new memories: how does adult hippocampal neurogenesis affect learning and memory?* Nature Reviews Neuroscience, 2010. **11**(5): p. 339-350.
111. Liu, Z., et al. *Deep Learning Face Attributes in the Wild*. in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2015.
112. Wiesel, T.N. and D.H. Hubel, *Effects of visual deprivation on morphology and physiology of cells in the cat's lateral geniculate body.* Journal of neurophysiology, 1963. **26**(6): p. 978–993.
113. Wiesel, T.N., *Postnatal development of the visual cortex and the influence of environment.* Nature, 1982. **299**(5884): p. 583.
114. Konishi, M., *Birdsong: from behavior to neuron.* Annual review of neuroscience, 1985. **8**(1): p. 125–170.
115. Hensch, T.K. and P.M. Bilimoria. *Re-opening windows: manipulating critical periods for brain development*. in *Cerebrum: the Dana forum on brain science*. 2012. Dana Foundation.
116. Grill-Spector, K., et al., *The Functional Neuroanatomy of Human Face Perception.* Annual Review of Vision Science, 2017. **3**(1): p. 167-196.

117. Rivolta, D., *Cognitive and Neural Aspects of Face Processing*, in *Prosopagnosia: When all faces look the same*. 2014, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 19-40.
118. Amso, D. and G. Scerif, *The attentive brain: insights from developmental cognitive neuroscience.* Nature Reviews Neuroscience, 2015. **16**(10): p. 606-619.
119. Batty, M. and M.J. Taylor, *The development of emotional face processing during childhood.* Developmental Science, 2006. **9**(2): p. 207-220.
120. Adolphs, R., L. Sears, and J. Piven, *Abnormal Processing of Social Information from Faces in Autism.* Journal of Cognitive Neuroscience, 2001. **13**(2): p. 232-240.
121. Kliemann, D., et al., *Atypical Reflexive Gaze Patterns on Emotional Faces in Autism Spectrum Disorders.* The Journal of Neuroscience, 2010. **30**(37): p. 12281-12287.
122. Klin, A., et al., *Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism.* Arch Gen Psychiatry, 2002. **59**: p. 809-816.
123. Neumann, D., et al., *Looking you in the mouth: abnormal gaze in autism resulting from impaired top-down modulation of visual attention.* Social Cognitive and Affective Neuroscience, 2006. **1**(3): p. 194-202.
124. Pelphrey, K., et al., *Visual Scanning of Faces in Autism.* Journal of Autism and Developmental Disorders, 2002. **32**(4): p. 249-261.
125. Spezio, M.L., et al., *Analysis of face gaze in autism using "Bubbles".* Neuropsychologia, 2007. **45**(1): p. 144-151.
126. Spezio, M.L., et al., *Abnormal Use of Facial Information in High-Functioning Autism.* Journal of Autism and Developmental Disorders, 2007. **37**(5): p. 929-939.
127. Wang, S. and R. Adolphs, *Reduced specificity in emotion judgment in people with autism spectrum disorder.* Neuropsychologia, 2017. **99**: p. 286-295.
128. Ruan, M., et al., *Deep Neural Network Reveals the World of Autism From a First‐Person Perspective.* Autism Research, 2021. **14**(2): p. 333-342.
129. Jones, W. and A. Klin, *Attention to eyes is present but in decline in 2-6-month-old infants later diagnosed with autism.* Nature, 2013. **504**(7480): p. 427-431.
130. Constantino, J.N., et al., *Infant viewing of social scenes is under genetic control and is atypical in autism.* Nature, 2017. **547**(7663): p. 340-344.
131. Dawson, G., S.J. Webb, and J. McPartland, *Understanding the Nature of Face Processing Impairment in Autism: Insights From Behavioral and Electrophysiological Studies.* Developmental Neuropsychology, 2005. **27**(3): p. 403-424.
132. Moriuchi, J.M., A. Klin, and W. Jones, *Mechanisms of Diminished Attention to Eyes in Autism.* American Journal of Psychiatry, 2017. **174**(1): p. 26-35.
133. Kliemann, D., et al., *The Role of the Amygdala in Atypical Gaze on Emotional Faces in Autism Spectrum Disorders.* The Journal of Neuroscience, 2012. **32**(28): p. 9469-9476.
134. Rutishauser, U., et al., *Single-Neuron Correlates of Atypical Face Processing in Autism.* Neuron, 2013. **80**(4): p. 887-899.
135. Cao, R., et al., *Encoding of facial features by single neurons in the human amygdala and hippocampus.* Communications Biology, 2021. **4**(1): p. 1394.
136. Cao, R., et al., *A neuronal social trait space for first impressions in the human amygdala and hippocampus.* Molecular Psychiatry, 2022. **27**(8): p. 3501-3509.
137. Yi, D., et al., *Learning face representation from scratch.* arXiv preprint arXiv:1411.7923, 2014.
138. Zhang, K., et al., *Joint face detection and alignment using multitask cascaded convolutional networks.* IEEE signal processing letters, 2016. **23**(10): p. 1499-1503.
139. He, K., et al. *Deep residual learning for image recognition*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
140. Hinton, G., O. Vinyals, and J. Dean, *Distilling the knowledge in a neural network.* arXiv preprint arXiv:1503.02531, 2015. **2**(7).

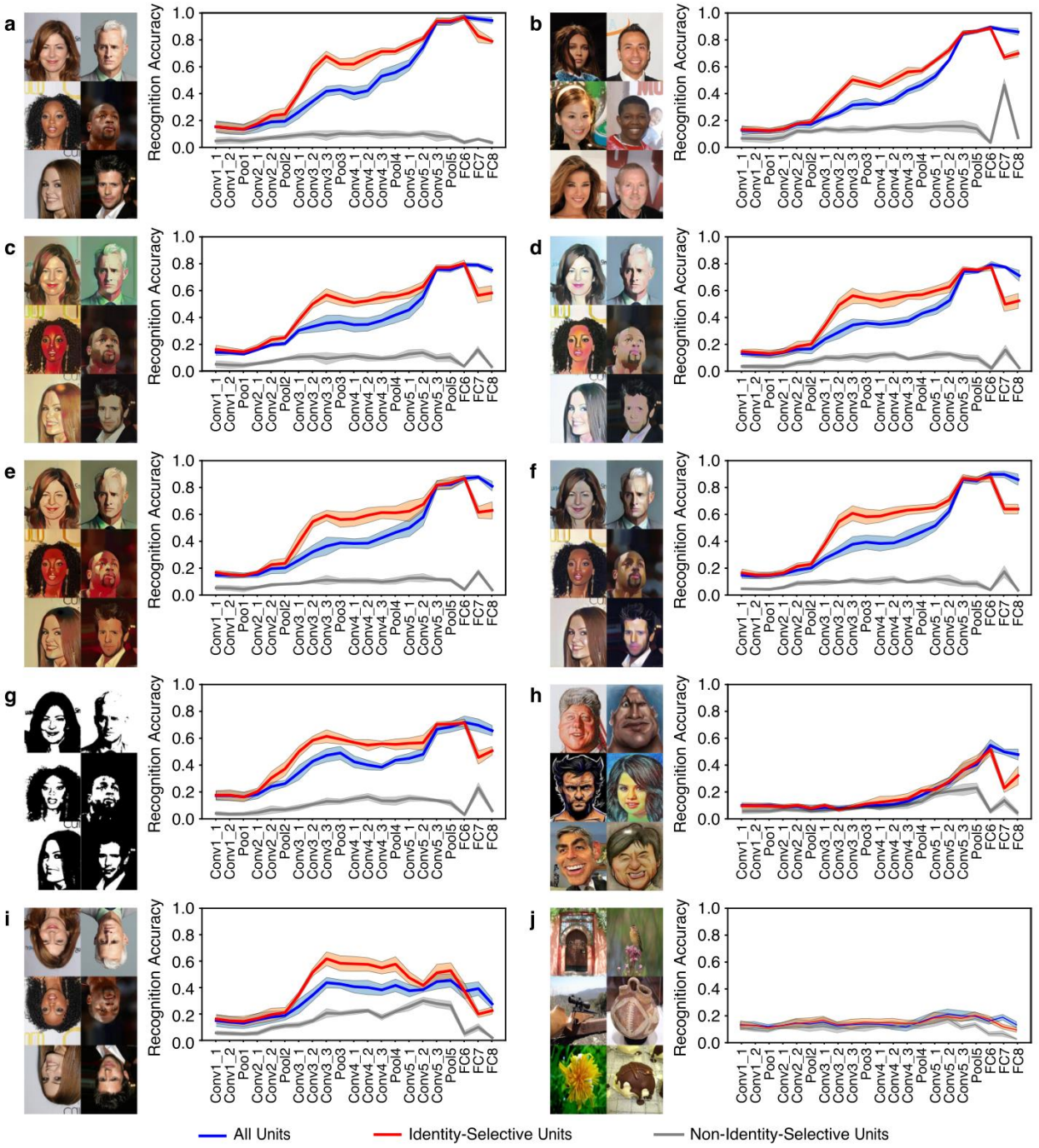
141. Zagoruyko, S. and N. Komodakis, *Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer.* ICLR, 2017.
142. Kar, K. and J.J. DiCarlo, *Fast Recurrent Processing via Ventrolateral Prefrontal Cortex Is Needed by the Primate Ventral Stream for Robust Core Visual Object Recognition.* Neuron, 2021. **109**(1): p. 164-176.e5.
143. Cao, R., et al., *A human single-neuron dataset for face perception.* Scientific Data, 2022. **9**(1): p. 365.

Appendix A – Supplementary Figures of Chapter 2

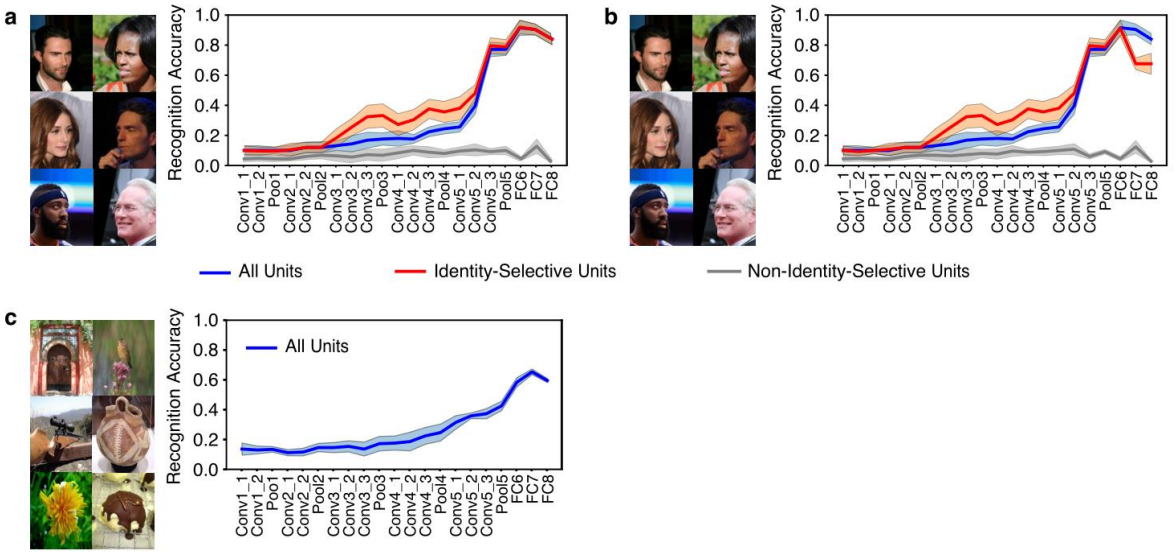
Supplementary Fig. 1. Sample stimuli. Faces from 50 celebrity identities were used for neural recordings. The identities were diverse in race, gender, and age, with a variety of facial expressions. All images had the same resolution, and the faces had a similar size and position in the images.



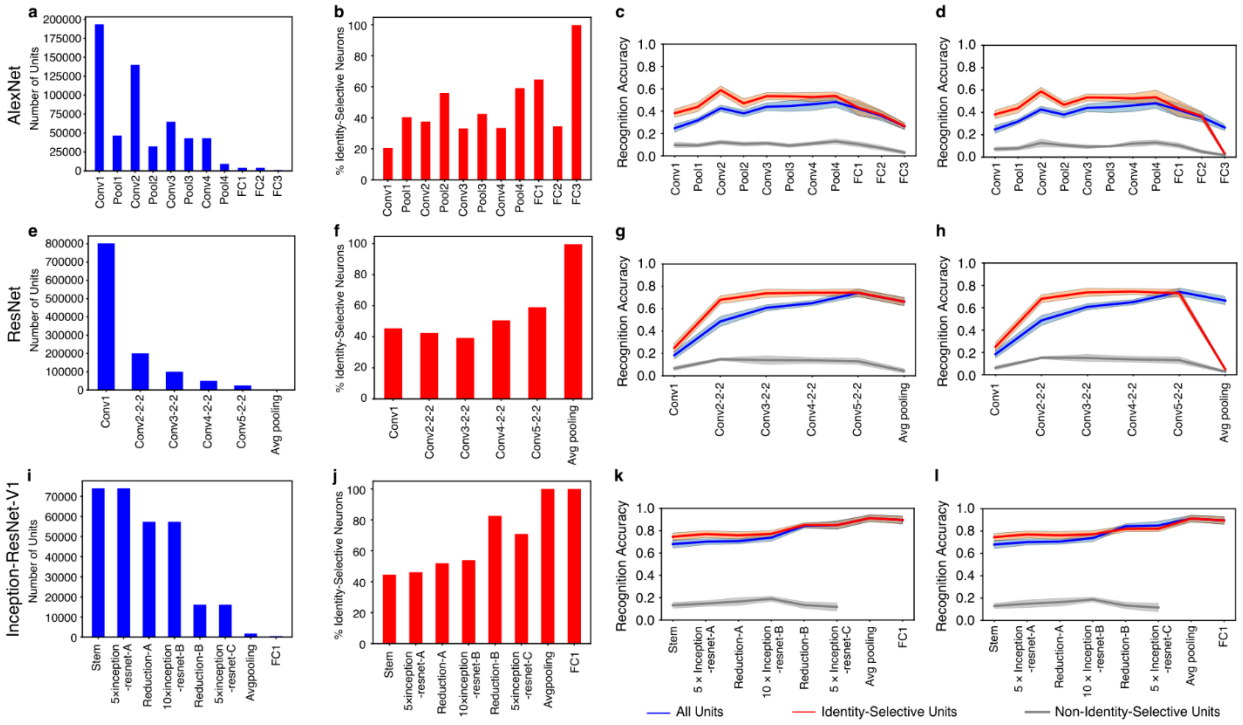
Supplementary Fig. 2. Additional results for identity-selective DNN units. (**a**) Procedure for selecting identity-selective DNN units. We used a one-way ANOVA to identify identity-selective DNN units that had a significantly unequal response to different identities ($P < 0.01$). We further imposed an additional criterion to identify a subset of identity-selective DNN units with selective identities: the response of an identity was 2 standard deviations (SD) above the mean of responses from all identities. These identified identities whose response stood out from the global mean were the encoded identities. We refer to the DNN units that encoded a single identity as single-identity (SI) units and we refer to the DNN units that encoded multiple identities as multiple-identity (MI) units. (**b**) The encoding frequency of each face identity in each deep neural network (DNN) layer. Color coding indicates the encoding frequency of each DNN layer (i.e., the number of DNN units encoding an identity divided by the total number of DNN units of a layer). (**c**) Combined SI and MI units demonstrated even better discriminability of face identities than identity-selective units. Identity recognition accuracy is shown for each DNN layer. Error shade denotes one standard deviation across 5-fold cross validation. Brown: combined SI and MI units. Red: identity-selective units.



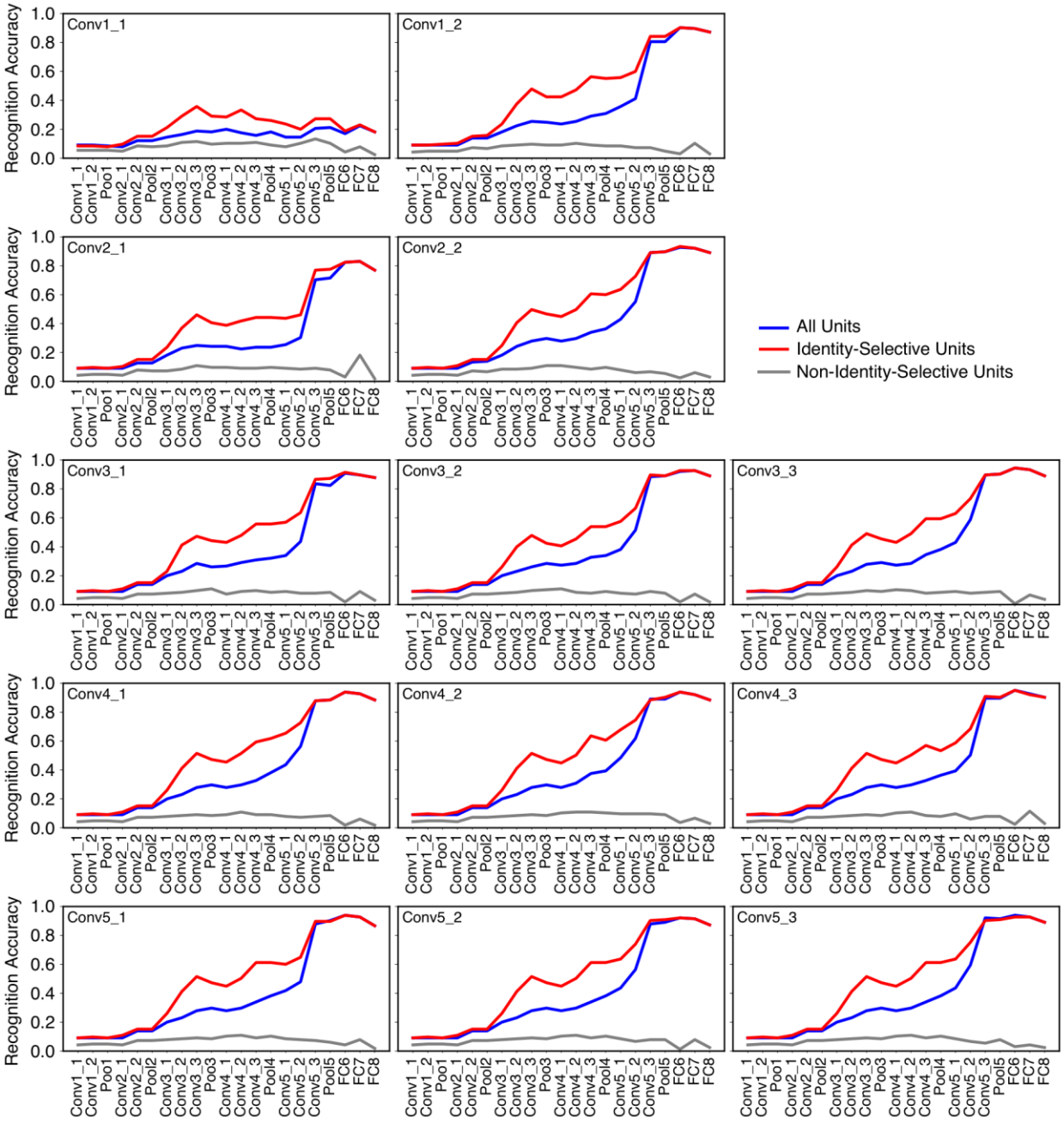
Supplementary Fig. 3. Recognition accuracy when we equated the number of identity-selective units and non-identity-selective units per layer. (**a**) Original faces used to identify identity- selective units. (**b**) Faces from a different set of 50 identities randomly selected from the CelebA database. (**c**) Original faces in the cartoon style Hayao. (**d**) Original faces in the cartoon style Hosoda. (**e**) Original faces in the cartoon style Paprika. (**f**) Original faces in the cartoon style Shinkai. (**g**) Original faces in the Mooney style. (h) A different set of celebrity caricature faces. (**i**) Original faces in inversion. (**j**) A set of non-face objects selected from the ImageNet stimuli. Identity recognition accuracy is shown for each deep neural network (DNN) layer. Error shade denotes one standard deviation across 5-fold cross validation. Blue: all units from each DNN layer. Red: identity-selective units. Gray: non-identity-selective units.



Supplementary Fig. 4. Recognition accuracy for profile faces and non-face objects. (**a, b**) Profile faces. For each face identity in Fig. 2a, we selected 5 profile faces from the CelebA dataset for analysis. (**a**) Analysis with all identity-selective and non-identity-selective units. (**b**) Analysis with equal number of identity-selective units and non-identity-selective units per layer. (**c**) Non-face objects. Legend conventions as in Supplementary Fig. 3.



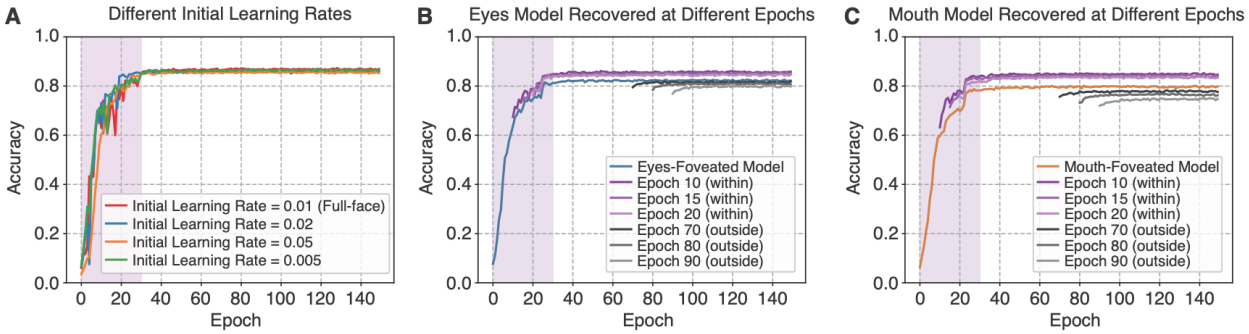
Supplementary Fig. 5. Recognition accuracy for different DNNs. (**a-d**) AlexNet pre-trained by images from ImageNet. (**e-h**) ResNet pre-trained by images from MS-Celeb-1M. (**i-l**) Inception-ResNet-V pre-trained by images from CASIA-WebFace. No fine-tuning was applied to these DNNs. (**a, e, i**) Number of neurons per layer. (**b, f, j**) Percentage of identity- selective neurons per layer. Note that all units in the last two layers of Inception-ResNet-V1 were identity-selective units. (**c, g, k**) Recognition accuracy. (**d, h, l**) Recognition accuracy with equal number of identity-selective units and non-identity-selective units per layer. Legend conventions as in Supplementary Fig. 3.



Supplementary Fig. 6. Recognition accuracy following a random dropout of a single layer of DNN units. 30% of the DNN units were dropped. The label in each plot indicates the layer that dropout was performed. Legend conventions as in Supplementary Fig. 3.

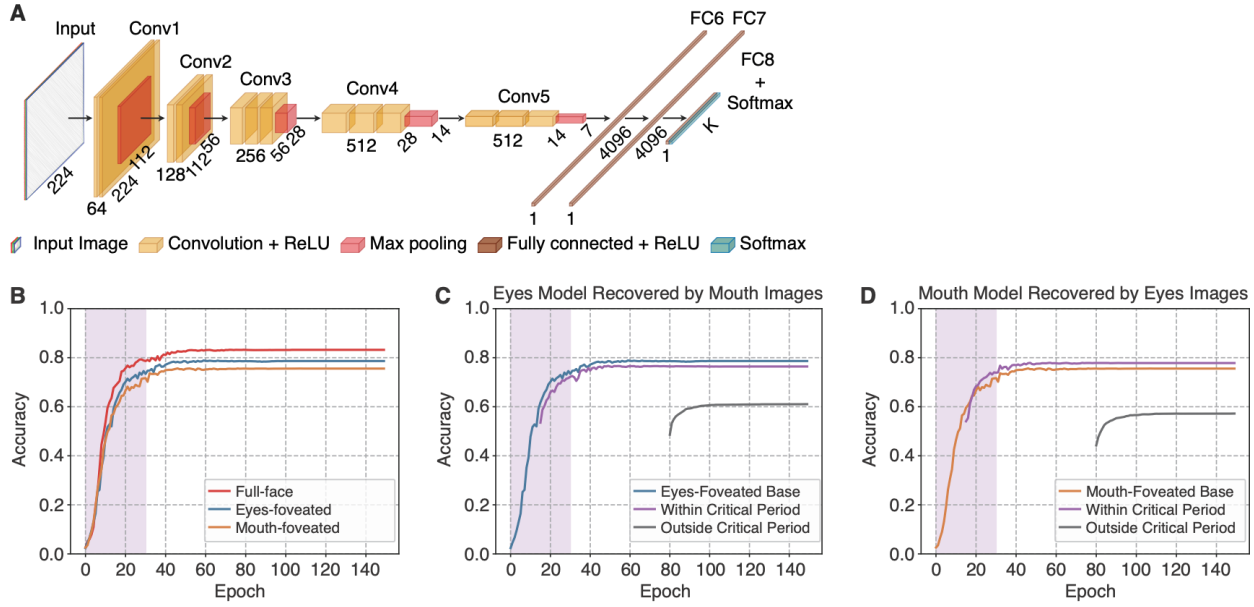
Appendix B – Supplementary Figures of Chapter 3

Figure S1



Supplementary Fig. 7. Model control analyses. **(A)** The learning curves for different initial learning rates. **(B, C)** Recovery by full-face images starting from different epochs. Shown are learning curves within or outside the critical period. **(B)** Eyes-foveated model. **(C)** Mouth-foveated model.

Figure S2



Supplementary Fig. 8. Model learning in VGG-16. **(A)** VGG-16 architecture. The convolutional neural network (CNN) consisted of a feature extraction section (13 convolutional layers) and a classification section (3 fully connected [FC] layers). The feature extraction section was consistent with the typical architecture of a CNN. A 3×3 filter with 1-pixel padding and 1-pixel stride was applied to each convolutional layer, which was followed by Rectified Linear Unit (ReLU) operation. Every convolutional block was followed by a max-pooling operation with a stride of 2 pixels. There were 3 FC layers in each classification section: the first two had 4096 channels each, and the third performed a *K*-way classification. Each FC layer was followed by a ReLU and 50% dropout to avoid overfitting. A nonlinear Softmax operation was applied to the final output of the VGG-16 network to make the classification prediction identities. **(B-D)** Network learning curve. The validation accuracy of face identity recognition is plotted as a function of the model training epoch. The shaded area denotes the critical period. **(B)** Full-face, eyes-foveated, and mouth-foveated base models. **(C)** Eyes-foveated model recovered by mouth information (i.e., mouth-foveated images). **(D)** Mouth-foveated model recovered by eyes information (i.e., eyes-foveated images). Legend conventions as in **Fig. 2**.