·第二十七届中国科协年会学术论文·

深度卷积神经网络在面孔识别中的表现及与人类视觉系统的对比*

程羽慧**[1] 申天宇[1] 路子童**[2] 袁祥勇[3,4] 蒋 毅[3,4]
（[1]南京师范大学心理学院，南京，210097）（[2]麻省理工学院麦戈文脑研究所，剑桥，02139）
（[3]中国科学院心理研究所，认知科学与心理健康国家重点实验室，北京，100101）
（[4]中国科学院大学心理学系，北京，100049）

摘 要 面孔识别是人类社会交往中的核心认知能力。近年来，深度卷积神经网络（Deep Convolutional Neural Network，DCNN）在模拟和理解面孔加工中展现出强大的能力，为探究人类面孔识别的行为表现和神经机制提供了新的视角。因此，围绕识别能力、行为效应与神经机制三个方面，系统综述了DCNN与人类在面孔识别中的异同。1）首先，DCNN是否具备与人类相当的面孔识别能力？从面孔身份、性别、情绪等特征方面出发，评估DCNN在面孔识别任务中的表现；2）其次，尽管DCNN在识别准确性上表现优异，其加工策略是否与人类的行为机制一致？基于经典的面孔加工效应（如倒置效应、种族效应、熟悉性效应等）分析DCNN与人类加工策略上的相似性与差异性；3）进一步，DCNN的内部表征是否与人类面孔加工的神经机制相类似？从结构层级性和功能专门化两个方面，比较其表征方式与人类面孔识别系统的神经基础之间的对应关系。最后，当前模型在鲁棒性与泛化性、结果解释力、生物视觉系统模拟等方面仍存在一定局限性，未来研究也可进一步探索其与多模态网络及生成对抗网络的融合潜力。

关键词 面孔识别 卷积神经网络 梭状回面孔区 层级结构 功能分化

1 引言

面孔传递着丰富的社会信息（例如，年龄、性别、情绪等），对个体和种群的生存和发展具有重要的进化意义（O’Toole et al., 1998; Rhodes & Leopold, 2011）。面孔识别不仅是人类社会交往中的核心认知能力，也是视觉系统高度专门化的重要功能（Calder, 2011）。长期以来，心理学家、认知神经科学家和计算机视觉研究者一直致力于揭示面孔加工的认知神经机制（Behrmann & Avidan, 2022; O’Toole et al., 2018; Rossion, 2014）。已有研究表明，人类在处理面孔信息时与普通物体的方式不同，展现出高度的特异性，典型表现包括倒置效应（Yin, 1969）和面孔假想错觉等（Hadjikhani et al., 2009; Tauber et al., 2017）。这些现象表明，灵长类动物在长期进化过程中可能发展出高度特异的面孔加工神经机制（Tsao et al., 2008）。猕猴的电生理研究发现，其颞叶皮层，特别是梭状回区域，存在大量“面孔选择神经元”。这些神经元仅对面孔刺激产生特异性放电，且能区分不同个体面孔的身份和表

* 本研究得到国家自然科学基金青年项目（32400864）、南京师范大学引进人才科研启动项目（184080H201A45）和国家社会科学基金青年项目（23CYY048）的资助。

** 通讯作者：程羽慧，E-mail: chengyh@nnu.edu.cn；路子童，E-mail: zitonglu1996@gmail.com

DOI:10.16719/j.cnki.1671-6981.20250405

情（Desimone, 1991; Kadosh & Johnson, 2007），它们形成了一个紧密连接的特异性面孔识别系统；人类功能磁共振成像研究也揭示了一个核心的面孔加工网络，包括负责身份识别的梭状回面孔区（fusiform face area，FFA）、负责处理基本特征的枕叶面孔区（occipital face area，OFA）以及分析动态表情的后颞上沟（posterior superior temporal sulcus，pSTS）（Haxby et al., 2000; Kadosh & Johnson, 2007; Kanwisher et al., 1997）。

随着人工智能技术的快速发展，深度卷积神经网络（Deep Convolutional Neural Networks，DCNN）作为人工神经网络的一个重要分支，在视觉信息处理领域展现出显著的优势（LeCun et al., 2015）。DCNN由多层模拟神经元组成，这些神经元通过卷积和池化操作对输入（如面孔图像）进行特征提取，并将处理后的数据传递至网络后期的多个全连接层，最终实现面孔识别（见图一），代表性模型包括AlexNet、VGGNet和ResNet等（He et al., 2016; Krizhevsky et al., 2012; Simonyan & Zisserman, 2014b）。与之相比，传统的人脸识别方法如Eigenfaces（基于主成分分析）和Fisherfaces（基于线性判别分析）依赖线性投射与人工特征标注，它们在小规模数据集下具有一定效果，但在应对姿态、光照、表情等复杂变异时表现出鲁棒性不足和泛化能力有限（Belhumeur et al., 1997）。此外，受神经科学启发的模型（如HMAX和VisNet）则从模拟人脑视觉加工出发，强调层级结构与不变性学习。这类模型在一定程度上模拟了人类视觉系统的加工机制，但其识别精度、训练效率以及在大规模数据的扩展性方面，仍显著落后于当前神经网络模型（Riesenhuber & Poggio, 1999; Rolls & Milward, 2000）。因此，DCNN凭借其强大的非线性建模能力、对复杂视觉变异的高度鲁棒性以及对大规模训练数据的良好适应性，已成为当前研究面孔加工机制的主流计算工具。一方面，这些在大规模面孔数据上训练的DCNN模型，在识别精度上已达到甚至超过人类水平（Phillips et al., 2018; Taigman et al., 2014）；另一方面，DCNN的层级加工方式与人脑视觉通路具有一定的结构与功能相似性，在神经表征模式上也表现出与脑成像数据的相关性（Eickenberg et al., 2017; Grossman et al., 2019）。因此，深入比较DCNN与大脑在面孔加工机制上的异同，不仅有助于计算机视觉领域构建更类脑的智能模型，也为认知神经科学提供了一种强有力的建模工具，助力揭示大脑这一“黑箱”的运行机制。

近十年来，DCNN在面孔研究领域的文章数量逐年增多。作为一种前沿的计算模型，DCNN为理解人类面孔加工机制提供了新的研究手段。本文将系统综述DCNN与人类在面孔识别任务中的异同。首先，DCNN是否具备与人类相当的面孔识别能力？本文将从面孔身份、性别、情绪等多个特征维度出发，评估DCNN在识别任务中的表现；其次，尽管DCNN在识别准确性方面表现优异，其加工策略是否与人类的行为机制相一致？为此，本文将结合面孔识别中的经典行为效应（如倒置效应、种族效应、熟悉性效应等），探讨DCNN与人类在加工策略上的相似性与差异性；最后，DCNN的内部表征是否反映与人类神经系统类似的加工机制？本文将从结构层级性与功能专门化两个方面，比较DCNN的表征方式与人类面孔识别系统的神经基础之间的对应关系。在此基础上，本文将探讨DCNN在刻画人类面孔识别过程中的局限性，并提出未来研究的方向，旨在为认知科学与人工智能的交叉研究提供理论支持和方法启示。

2 DCNN是否具备与人类相当的面孔识别能力？

DCNN的发展源于研究者对面孔识别与分类任务的现实需求，尤其是在面孔识别准确性方面的挑战。研究者最关心的问题是，DCNN的识别准确性是否能够达到甚至超越人类水平。本文将从面孔身份、性别与情绪三个维度出发，探讨DCNN在这些任务中是否出现与人类相近的面孔加工能力。

2.1 身份信息

目前，大多数神经网络模型的训练数据主要基于面孔身份信息的特征提取，因此，身份识别的准确率成为衡量人类与DCNN面孔识别能力的重要指标。DCNN能否在面孔身份识别任务中达到或超越人类的表现水平？研究表明，计算模型在人脸识别任务中达到了人类的水平（O'Toole & Castillo, 2021;

Phillips & O'Toole, 2014; Taigman et al., 2014）。 其中，Taigman 等（2014）开发的 DeepFace 系统在大规模面孔数据库上取得了 97.35% 的识别准确率，达到了人类的判断水平（Taigman et al., 2014）。即使在高难度任务中，DCNN 也展现出优势。Phillips 等人（2018）的研究中分别要求人类被试和计算模型判断面孔在不同光照、角度和表情条件下是否属于同一身份。实验参与者包括法医检验人员（具备专业法庭面孔鉴定能力），超级识别者（天生具备超强面孔识别能力）和普通大学生。研究结果显示，DCNN 的识别准确率已达到了人类专家水平。此外，人机协作能显著提升面部识别的准确性 （Phillips et al., 2018）。这些发现表明，DCNN 不仅具备与人类相当的基础身份识别能力，在专业任务中也展现出独特的优势。

2.2 性别信息

人类不仅能识别面孔的身份信息，还能快速提取性别、情绪、年龄等其他特征信息（Dobs et al., 2019）。最新研究表明，DCNN 也能像人类一样准确判断面孔的性别（Dhar et al., 2020; Hill et al., 2019; Song et al., 2021）。例如，研究者采用反向相关分析方法，比较了经过性别分类迁移学习后的 DCNN（VGG-Face）与人类在性别分类任务中的表征差异（Song et al., 2021）。实验通过将中性面孔（基于男性和女性面孔的平均值生成）与随机噪声结合，分别测试人类和模型对性别判断的表现，并提取其各自对应的内部表征。结果显示，VGG-Face 的性别分类正确率可以达到 98.6%，其表征与人类的相关性高达 .73。进一步的研究发现模型主要依赖于低空间频率信息（如面部的整体形状）进行性别判断，表明两者在处理面部信息时采用了相似的加工方式。更重要的是，经过面孔识别任务训练的 VGG-Face 与人类表征的相似度较高，而经过物体训练网络的 AlexNet 与人类表征的相关性较低。这一结果表明，面孔识别中的性别判断能力高度依赖于特定的视觉经验（Song et al., 2021）。

2.3 情绪信息

研究者还发现，基于身份识别训练的 DCNN 能够自发地产生面孔情绪的表征（Colón et al., 2021; Zhou et al., 2022）。Zhou 及其同事首先从已经训练好的用于面孔身份信息识别的 DCNN（VGG-Face）中筛选出对六种基本表情（愤怒、厌恶、恐惧、快乐、悲伤和惊讶）具有选择性反应的“表情神经元”，发现它们可以正确分类面孔情绪。更有趣的是，通过对不同表情的渐变序列测试（morphed facial expressions）发现，这些神经元表现出与人类相似的表情混淆现象和类别知觉效应，例如模型与人类都容易将恐惧与惊讶、愤怒与厌恶混淆。研究还揭示，经过面孔身份训练的 DCNN（VGG-Face）相较于基于物体类别训练（VGG-16）或未经训练的 VGG-Face 网络（Zhou et al., 2022），不仅存在更多情绪选择性神经单元，也表现出显著更强的类似人类的情绪感知模式。这一结果说明人类的面部表情识别机制依赖于面孔的视觉经验，而非简单的通用视觉处理或神经网络的架构本身。

2.4 小结与讨论

综上所述，DCNN 具有接近甚至超越人类水平的面孔识别能力。虽然多数 DCNN 仅接受身份识别任务的训练，但它们仍能产生性别、情绪等社会性信息。近期有研究进一步表明，DCNN 也能够识别年龄和姿态角度等特征，且这些信息分布于网络的不同层级之中。具体而言，性别和姿态角度的表征在网络的浅层就已被编码；相比之下，年龄信息则随着网络层级的加深逐渐被表达（Dhar et al., 2020）。这些证据表明了 DCNN 在进行面孔识别过程中能自动表征各种面孔相关信息的能力。与这些结果类似，研究者采用漫画面孔发现，DCNN 形成了一个高度有组织的面孔相似性结构，其中身份、性别、光照和视角等信息呈现出嵌套的层级关系，即人脸身份嵌套于性别之下，而光照和视角信息则嵌套于身份之下（Hill et al., 2019）。因此，DCNN 可能像人脑一样存在“人脸空间结构”（face space），即大脑将每张人脸表征为多维特征空间中的一个点，每一个维度代表面孔的某种特征（如眼间距、鼻子宽度、皮肤颜色等），面孔之间的“距离”反映了它们在感知上的相似性或差异性（Nestor et al., 2016; O’Toole et al., 1993; O’Toole et al., 2018; Valentine et al., 2016）。研究发现，DCNN 可以通过多层非线性转换，将原始图像逐步映射至高度抽象的特征空间，其最终生成的面孔图像构成了一个高

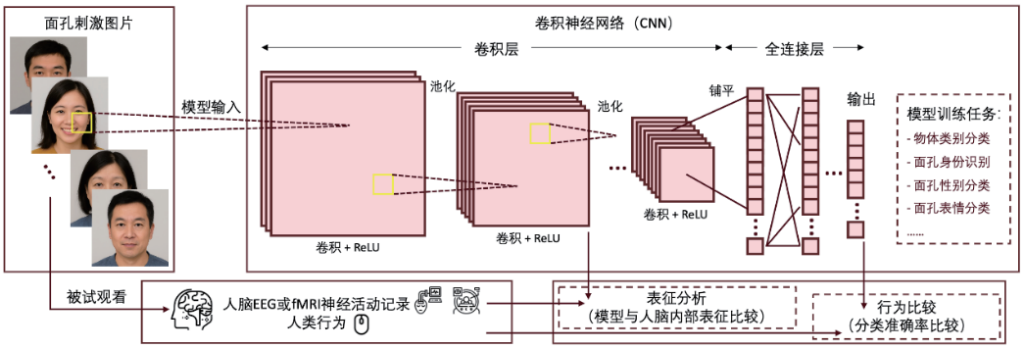

图 1　DCNN 与人类识别面孔的对比分析框架

注：面孔视觉刺激作为输入传入 DCNN 后，依次经过多个卷积层、激活函数、池化层和全连接层处理，最终得到模型的结果输出。DCNN 可以通过多种任务进行训练，例如物体分类、面孔身份识别、性别判断和情绪分类等。同样的刺激也可以让人类被试观看，研究者记录其行为反应或神经活动（包括 EEG 或 fMRI 信号），获得相应的人类数据。研究者通过将模型不同层级的内部表征与人脑在时间或空间维度上的神经表征进行对比，探究人类面孔识别过程中的表征机制。同时，研究者也可以通过操纵模型的训练数据、训练方式，进一步比较模型的分类结果与人类行为表现，以评估其行为一致性。

维的人脸空间，在该空间中，同一身份的不同图像被紧密聚类，不同身份间则保持较大距离，实现了身份不变性的表征（Grossman et al., 2019; Hill et al., 2019; O’Toole et al., 2018）。正是由于 DCNN 展现出类人的行为特征与表征能力，使其成为研究人类面孔加工机制的一种新途径。然而，与人类相比，DCNN 在面孔识别上仍存在诸多局限性，详见后文讨论部分。

3　DCNN 是否能复现人类面孔识别的行为效应?

尽管 DCNN 在面孔识别任务中的准确性表现优越，但其加工策略是否与人类的行为机制一致仍然是一个值得探讨的问题。为此，研究者测试了 DCNN 是否能表现出典型的面孔加工的行为效应，例如倒置效应（Yin, 1969）、撒切尔效应（Dahl et al., 2010; Thompson, 1980）、种族效应（Bothwell et al., 1989）、熟悉性效应（Kramer et al., 2018）以及面孔假想错觉等（Hadjikhani et al., 2009）。这些现象被认为是面孔加工特异性的关键证据，表明与其他物体相比，人类识别面孔时具有独特的认知机制（Kanwisher, 2000; Tanaka & Sengco, 1997）。研究者进一步探究 DCNN 是否能展现出与人类相似的行为现象，并试图通过模型来揭示这些现象背后的深层原因。

3.1　倒置效应

面孔倒置效应（Face Inversion Effect）是指人类对正立面孔的识别能力优于倒置面孔，即当面孔被垂直旋转 180° 时，人们的识别能力会明显下降，而非面孔物体（如房屋、车、动物等）受倒置影响则较小（Yin, 1969）。研究者认为这反映了面孔识别依赖于整体加工（Holistic Processing），即人类在识别面孔时会同时整合各个面部特征（如眼睛、鼻子、嘴巴）及其空间关系。与此相关的还有撒切尔效应（Thatcher effect），即当面孔倒置时，即使其局部特征发生明显改变（如颠倒的眼睛和嘴巴），人们也难以察觉异常（Dahl et al., 2010; Thompson, 1980）。最近，研究者使用 DCNN 来探究倒置效应（Dobs et al, 2023; Tian et al, 2022）和撒切尔效应（Jacob et al, 2021）的内在机制。Tian 及其同事测试了 DCNN 对正倒立面孔和物体的分类准确率，结果发现 DCNN（VGG-Face）能自发产生面孔倒置效应（Tian et al., 2022）。Dobs 等人进一步通过身份匹配任务发现，只有基于面孔数据训练的网络才表现出倒置效应，而基于物体数据的网络则不具备这一特性（Dobs et al., 2023）。这表明模型出现倒置效应可以被视为专门为面孔识别优化的结果，而非一般视觉分类机制的副产物。

3.2　种族效应

种族效应（Own-Race Effect）是指个体在识别人脸时，对自身所属种族的面孔识别能力优于其他种族的面孔，这一现象通常被认为源于个体对本族面孔更丰富的接触经验（Bothwell et al., 1989）。为探究该效应是否也出现在 DCNN 中，研究者采用身

份匹配任务，评估 DCNN 对白人面孔与亚洲面孔的识别正确率。结果发现，接受白人面孔分类训练的 DCNN 在识别亚洲面孔时表现较差；反之，接受亚洲面孔分类训练的 DCNN 在识别白人面孔时表现较差。进一步，研究者训练了基于物体分类和白人（或亚洲）身份的面孔检测的模型，尽管在面部经验上进行了控制匹配，但该模型并未表现出种族效应。此外，无论是基于物体分类训练还是未经训练的 DCNN 均未表现出种族效应（Dobs et al., 2023）。这一发现与另一项研究结果相一致，当研究者使用包含更多亚洲面孔的数据集对 DCNN 进行重新训练时，模型对亚洲面孔而不是白人面孔的识别正确率更高。当数据集中亚洲面孔和白人面孔的数量达到平衡时，模型则不再表现出种族偏差（Tian et al, 2021）。这些结果表明，种族效应可能是面孔识别训练过程中自然形成的产物，而非源自种族面孔本身的固有差异。

3.3 熟悉效应

面孔熟悉效应（Familiarity Effect）是指熟悉面孔比陌生面孔更容易被识别和记忆。这一效应表明，面孔识别系统在加工熟悉面孔时更加高效。先前的理论认为，人们在感知陌生面孔身份时表现不佳，是因为大多数与身份无关的视觉变化在每个个体中具有特异性，因此每个面孔身份都必须“从零开始”进行学习（Kramer et al., 2018）。为验证这一观点，研究者采用 DCNN 考察了视觉经验对面孔识别的影响，比较了未经训练网络、物体专家网络以及面孔专家网络的表现（Blauch et al., 2021）。结果发现，只有经过面孔训练的网络，才能在新陌生身份的识别任务中实现显著的泛化能力。此外，随着先前学习的身份数量的增加，网络的泛化能力也随之提升。这表明，面孔图像中携带的身份信息具有一定的通用性，而非完全特异化。此外，模型也表现出类似人类的“熟悉化”现象：通过对原本不熟悉的面孔进行微调训练（fine-tuning），网络在识别任务中的表现显著提升。这些研究结果表明，增加面孔经验能够提升对陌生面孔的感知能力及新身份的学习效果。

3.4 面孔假想错觉

当人们看到模糊或抽象的图案（如插座、岩石、云朵或建筑纹理）时，常常会将其误认为面孔，这种现象被称为面孔假想错觉（Face Pareidolia），其成因可能与大脑的快速面孔检测机制有关（Hadjikhani et al., 2009; Taubert et al., 2017）。 研究采用脑磁图（magnetoencephalography, MEG）记录被试观看真实人脸、假想面孔和相匹配物体的神经活动。通过表征相似性分析（representational similarity analysis，RSA），比较 DCNN 内部表征与人脑神经活动的相似性（Gupta & Dobs, 2025）。结果发现包含物体分类任务训练的模型比未包含物体分类任务训练（仅基于面孔身份训练）的模型更符合人类表征假想面孔的 MEG 数据。进一步分析显示，在模型早期层，其对假想面孔的表征更接近真实面孔；而自第五层起，则逐渐趋近于普通物体的表征，这一动态变化与人类加工假想面孔时的时序激活模式相吻合（Wardle et al., 2020）。这些研究结果表明，面孔假想错觉现象可能并非源于专门进化的功能，而是大脑同时优化面孔识别与物体分类任务的附带产物（Gupta & Dobs, 2025）。

3.5 小结与讨论

综上所述，DCNN 不仅在面孔识别能力上达到了人类水平，也能够一定程度上重现人类面孔识别相似的行为现象。更重要的是，DCNN 在传统研究的基础上提供了一个新的方法与思路让研究者借助 DCNN 尝试揭示产生这些面孔效应的深层原因。不论是探究倒置效应还是其他典型面孔效应，研究者尝试回答的一个根本问题是面孔识别是否依赖于专门的、特有的认知神经机制，即面孔加工是否具有领域特异性（domain specificity）还是通用性（domain-generality）（Kanwisher, 2000）。有学者提出了专家假说（expertise hypothesis），该理论认为，面孔识别中所涉及的认知神经机制其本质可能并非针对“面孔”这一类别本身的特异性加工，而是源于人类对高度专家化类别进行精细化区分的能力。因此，这一机制涉及所有需要精细辨别的视觉任务（如汽车、鸟类识别）。然而，行为和神经影像学研究在验证该假说时结果并不一致，因为在现实中，人类对非面孔类别（如汽车、鸟类）的经验程度难以与面孔经验相等同（McKone et al., 2007）。相比之下，使用 DCNN 进行模拟具有独特优势：研

究者可以严格控制网络的训练任务的数量和类型，从而克服传统研究中的经验偏差问题。研究发现当DCNN被训练用于识别汽车时，会在该模型中观察到“汽车倒置效应”（Dobs et al., 2023）；类似地，训练用于识别鸟类的DCNN也会表现出“鸟类倒置效应”（Yovel et al., 2023）。这些发现表明，倒置效应并非面孔独有，而是某类视觉任务优化后的泛化产物。然而，最新研究利用DCNN建模发现了相反的结论，即否定了“专家假说假说”，指出面孔加工机制并非一个通用于各类专家识别任务的通用模块，而更可能是一种具有领域特异性的加工系统（Kanwisher et al., 2023）。尽管当前关于该问题的结论尚存争议，研究者已开始尝试在保持模型结构不变的前提下操控任务类型（如训练数据与输出目标），分析模型内部单元对不同刺激的响应，或对模型内部的表征模型进行相似性比较等方式来阐明面孔加工的行为机制。综上，由于DCNN在模型结构、训练数据和方式上的高度可控性，研究者能够设计出严格的对照实验与条件比较，尝试开展一些在人类被试中难以实现的任务，从而推断潜在的面孔加工机制。然而，研究者也需要谨慎揭示与推断这些基于DCNN的实验对应的结果，其与人类面孔处理机制的对应关系依然不是完全明确的。

4 DCNN是否能模拟人脑面孔识别的神经机制?

近年来，研究者不仅关注DCNN在解释人类面孔加工行为表现的能力，也开始探讨其在模拟人类及非人灵长类（即猕猴）视觉加工机制中的潜力。在神经层面，当前研究者集中于探讨DCNN能否再现视觉系统的层级结构，以及面孔识别的功能分化特性。

4.1 结构层级性

研究发现，DCNN内部的计算层级与大脑视觉皮层之间存在高度对应关系。无论是人类还是非人类灵长类动物，其视觉皮层与DCNN中视觉特征的表征均呈现出从低级到高级的逐层递进。这种层级结构不仅体现在特征复杂度的逐步提升，还伴随着神经元感受野的扩大，以及对视觉不变性（如视角变化）响应能力的增强（Cadieu et al., 2014; Eickenberg et al., 2017; Güçlü & Van Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014）。具体而言，DCNN的早期层对应初级视觉皮层（如V1），主要编码低级视觉特征（如边缘、对比度、色彩、方向等）；网络的中间层对应于视觉通路的中间区域（如V2、V4），整合较为复杂的视觉特征（如形状、纹理）。对于面孔而言，该层主要编码特定视角信息(如头部姿态和光照方向)；而深层则对应较高级的视觉区域（如外侧枕叶区LO），主要编码具有视角不变性的抽象视觉表征，如个体身份、性别等（Eickenberg et al., 2017; J. Wang et al., 2022）。然而，DCNN中对于面孔信息的表征并不一定是随着深层加深就加工越强的。Grossman等人（2019）通过对33位癫痫患者的颅内脑电记录（iEEG）发现，将人脑面孔选择性神经元所形成的“人脸空间”与DCNN的中间层表征一致，而非浅层或深层（Grossman et al., 2019），这一结果也在核磁研究中被发现（Jiahui et al., 2023）。在灵长类动物研究中，Raman和Hosoya（2020）比较了DCNN与猕猴颞下皮层（inferotemporal cortex, IT）中较低层级的中部面孔区（middle face patch, ML）和较高层级的前部面孔区（anterior face patch, AM）。研究发现，DCNN中高层中可以准确复现AM区神经元对面孔身份与姿态变化的调谐特征，表现出较强的身份不变性（identity-invariant tuning）；但DCNN在模拟ML区神经元对低层级图像特征（如局部几何形状与结构）的调谐特性方面表现欠佳。这表明，尽管现有DCNN能够在一定程度上模拟高级面孔表征区域（如AM）的神经响应，但难以捕捉中层脑区(如ML)对面孔几何特征的早期加工机制(Raman & Hosoya, 2020）。这些研究表明现有DCNN模型虽能作为面孔表征的类脑模型，但其在模拟完整生物视觉系统中多层级面孔加工通路方面也存在局限性。

4.2 功能专门化

面孔识别在大脑中存在专门的计算表征。人类大脑中的梭状回面孔区（FFA）被认为是面孔识别的特异性脑区，并且猕猴的颞叶区域中同样存在大量面孔选择性神经元，这些跨物种的研究共同支持了面孔加工的功能专门化假设。DCNN中是否存在

对面孔的特异性加工呢？最近的研究发现 DCNN 能发展出类似人脑的面孔选择性神经元（Lu & Wang, 2025; Prince et al., 2024）。此外，研究者也尝试借助 DCNN 探究面孔识别依赖于大脑中通用的视觉识别系统，还是需要专门的神经加工过程，从而为理解大脑功能的专门化及其形成机制提供新的证据。例如，研究者探讨了 DCNN 同时执行人脸识别与物体分类任务时，是否会自发涌现类似于人脑的专门化功能结构。结果发现，单独训练于面孔或物体识别的网络在跨任务识别中表现较差，这表明面孔与物体识别依赖各自特异性的表征机制。而在双任务网络中，尽管模型未被预设任何偏置，它仍然通过内部滤波器的任务分工，自发形成了功能分化，类似于人脑中专门化的加工机制（Dobs et al., 2022）。研究者还发现在完全未经训练随机初始化的网络也会涌现出对面孔有选择性的单元，即没有经过训练的 AlexNet 其高层单元仍然可以对人脸图像表现出强烈的选择性。这些单元的选择性与猴子大脑中的人脸选择性神经元相匹配，并且可以用于进行面孔检测任务（Baek et al., 2021）。更关键的是，这些模型展现出功能专门化的特性，其表征模式与颅内脑电记录以及猕猴电生理实验中观察到的神经响应高度一致（Grossman et al., 2019; Raman & Hosoya, 2020; J. Wang et al., 2022）。更进一步，有研究者通过建立 DCNN 与神经活动之间的映射关系，成功预测了面孔特异性脑区在新面孔图片下的神经响应，并重构出最大化激活面孔特异性脑区的面孔图片（A. Luo et al., 2023; A. F. Luo et al., 2023; Ratan Murty et al., 2021）。然而，关于“面孔选择性神经元”是否真正反映了面孔特异性的编码机制，近期也出现了新的争议性证据。研究者通过分析灵长类视觉皮层中“面孔细胞”的神经响应模式，提出这些细胞的选择性并非源于对面孔类别的编码。研究发现通过非面孔图像也能有效激活这些“面孔细胞”，并利用模型重构的方法生成出高度激活这些细胞的非面孔图片，从而挑战了“面孔细胞”是面孔类别选择性编码单位的传统观点（Vinken et al., 2023）。这项研究提醒我们，所谓的“功能专门化”可能部分源于统计特征上的偏好，而非类别层面的先验模块化编码。这一结果为理解面孔加工中选择性表征的起源提供了新的思路，也提示研究者在使用 DCNN 解释神经功能专门化时需更谨慎地界定“选择性”与“专门性”的内涵。

4.3 小结与讨论

综上所述，得益于 DCNN 具有类似人脑视觉系统的层级结构以及 DCNN 内部激活的易分析性，越来越多的研究不仅把 DCNN 作为探究面孔加工行为表现的工具、也将其视为研究面孔加工神经机制的新手段。DCNN 通过其层级架构展现出对视觉信息的逐层编码特性，并在某些单元中自发涌现出类似人脑面孔选择性神经元的激活模式，进一步印证了其在模拟和解析人脑面孔识别机制方面的有效性和潜力。然而，当前 DCNN 与人脑尽管表现出一定程度的相似性，两者之间仍存在诸多本质性差异，例如，DCNN 仍然与人脑视觉系统在生物结构存在根本不同、缺乏其他模态数据对其的优化、其训练方式与人类视觉学习过程并不相同等等。因此，随着类脑智能领域的进一步发展，未来研究需要在现有 DCNN 基础上构建更接近人脑的视觉模型，从而研究者能更确信地使用模型来模拟并理解人脑面孔识别过程的神经机制。

5 问题与展望

近年来，人工智能的迅猛发展，尤其是深度神经网络的进步，为研究面孔识别与表征提供了前所未有的契机。一方面，DCNN 可作为工具，模拟人类面孔识别行为，帮助研究人员理解大脑的表征机制；另一方面，认知科学的实验范式亦可反向应用于揭示深度卷积神经网络内部的表征机制，从而逐步揭开其“黑箱”本质。然而，DCNNs 作为一种数据处理和预测工具，尚未能够完全解释人类视觉认知的复杂机制。特别是在面对复杂的视觉信息时，DCNN 仍然无法完美再现人类视觉感知的灵活性和复杂性。因此，DCNN 究竟能否作为模拟并解释人类视觉加工机制的有效模型，还是仅是一种研究工具，仍然值得深入思考（Wichmann & Geirhos, 2023）。以下，本文将从不同方面出发，进一步探讨模型的局限性。

首先，虽然 DCNNs 在面孔识别上表现出色，但它们对于不同角度、光照、图像失真等条件下的

鲁棒性远不如人类。例如，研究发现 DCNN 对视角变化非常敏感，准确率在“非典型视角”下会大幅下降，而人类能在不同 3D 视角下稳定识别物体（Dong et al., 2022）。此外，Geirhos 等人（2018）系统地研究了三个神经网络（ResNet-152、VGG-19 和 GoogLeNet）在面对 12 类图像失真曲下的表现。结果发现尽管 DCNN 在标准图像上可达到甚至超越人类性能，但其在面对未见过的图像失真时性能迅速下降，接近随机猜测水平。即使通过数据增强训练，DCNN 仍无法有效迁移至其他失真类型（Geirhos et al., 2018）。同时，DCNN 易受到对抗性攻击（adversarial attacks），即通过添加微小扰动就可能导致识别错误，说明模型的稳健性不如人类的视觉系统（Madry et al., 2017）。此外，人类能够在极少样本甚至一次接触的情况下学习和识别新面孔，体现出高度的学习效率（Lake et al., 2015）。相比之下，训练一个性能优越的人脸识别 DCNN 模型通常需要成千上万张图像样本，说明 DCNN 在有限样本学习能力方面仍显不足。这些研究结果共同说明，DCNN 的鲁棒性和泛化性仍显著弱于人类视觉系统。

其次，尽管 DCNN 在模拟人类面孔识别行为和神经机制方面取得了显著进展，但其解释性仍然存在一定的局限性。首先，尽管 DCNN 表现出和人一样的行为效应，但产生这些现象背后的原因存在多种可能的解释。不同的研究中进行测试的 DCNN 模型可能存在多个维度上的差异，包括：（1）不同结构的 DCNN 模型，如 AlexNet、VGG 以及 ResNet 等，它们可能包含了不同的模型层数、连接方式（Dobs et al., 2019; Song et al., 2021）；（2）不同的训练数据与训练目标，同样的模型结构可能基于客体分类、面孔身份分类以及面孔情绪分类等任务进行训练（Dobs et al., 2022; Dobs et al., 2023; Gupta & Dobs, 2025），且即便同样的训练任务使用的训练集也可能不同，造成模型内部的编码模式产生受任务和数据影响的偏差（M. Wang & Deng, 2020; Zhang et al., 2016）；（3）不同的训练方式，包括有监督、自监督训练或是无监督训练（Konkle & Alvarez, 2022; Zhuang et al., 2021）。这些差异都可能影响 DCNN 对面孔信息的编码以及面孔识别过程中的行为结果。因此，在使用 DCNN 进行面孔识别机制探究的时候，一方面需要严谨而全面地进行控制实验，包括控制模型结构、训练数据以及模型的训练方式等，另一方面需要谨慎地进行从模型到生物机制的类比推断、明确基于 DCNN 研究可能存在的局限性。

此外，DCNN 作为主流的“类脑视觉模型”尚未能完整体现生物视觉系统中的关键动态机制。目前主流的 DCNN 多采用前馈结构，而真实大脑视觉系统则展现出高度的时序依赖性和复杂的皮层间交互机制，例如注意力的动态调控、预测性编码以及前额叶脑区对早期视觉区域的反馈调节。此外，灵长类大脑的视觉系统中广泛存在横向和反馈连接（Kravitz et al., 2013），而这些连接恰恰是现有 DCNN 模型中缺乏的关键特征。已有研究表明，研究者比较前馈神经网络与递归神经网络在建模人类视觉系统表征观看不同自然物体类别（人脸、身体、自然物体等）的大脑反应。结果发现，在预测脑磁图时序数据中，递归网络显著优于同参数量的前馈模型（Kietzmann et al., 2019）。因此，为了增强模型的生物真实性与认知解释，未来研究亟需引入更具神经生理机制的架构特征，例如循环神经网络（Gu et al., 2017; Li, Zheng et al., 2018）、基于空间拓扑约束的神经网络（Blauch et al., 2022; Margalit et al., 2024）。此外，当前的大多数研究集中于静态图像下的面孔识别与表征，而忽略了更加生态、更加复杂的动态面孔加工过程（Jiahui et al., 2023; O'Toole et al., 2002）。近期发展出的双流卷积神经网络结构（two-stream CNN）以及循环回归神经网络为该领域提供了新的研究方向，未来研究可以同时关注空间信息和时间信息来进行动态面孔识别（Li et al., 2018; Simonyan & Zisserman, 2014a）。

再者，现有神经网络模型主要聚焦于面孔知觉层面的表征机制，但人类对面孔的加工远不止于感知阶段，还涉及更高层次的语义加工。例如，当我们看到一位名人的面孔时，不仅能够识别其外貌特征，还会联想到其身份、职业、社会地位，甚至激发与之相关的记忆与情感联结。最近的研究比较了视觉模型（即 VGG, 专注于图像特征提取）、语义模型（即 sentence generative pretrained transformer，SGPT，处理来自 Wikipedia 的文字描述）和视觉 - 语义融合模型（即 contrastive image-language pre-

training，CLIP， 将图像与语义共同嵌入的多模态模型）的表现，发现人类在知觉任务中更依赖视觉表征，而在记忆任务中则更多依赖语义信息；CLIP 模型能够有效整合视觉与语义信息，更接近人类加工机制（Shoham et al., 2024）。这提示未来面孔识别模型的发展需进一步融合语言、记忆与社会情境等高阶认知要素，以更全面地模拟人类的面孔的表征机制（Lu & Wang, 2025）。

最后，除 DCNN 以外，计算机视觉中的生成式对抗网络（Generative Adversarial Network，GAN）也受到了研究者的关注，它作为一种基于深度学习的生成框架模型，可以生成高度真实且符合语义的面孔。这种高质量、可控的人脸图像不仅为面孔知觉的研究提供了丰富的新型实验材料，也在建模上与 DCNN 形成了互补关系。DCNN 擅长模拟人类在面孔知觉阶段的特征提取过程，而 GAN 则提供了一种逆向映射路径，能够直观地“可视化”DCNN 内部的表征内容。具体而言，研究者可以将 CNN 提取的嵌入向量映射回GAN的生成器，从而构建“心理等效面孔”，并揭示模型对身份、表情等抽象属性的编码方式。例如， GAN 生成的面孔在被试者主观评估中，甚至比真实面孔更具自然感（Lago et al., 2021）。Shoura 等人（2025）利用 StyleGAN2 探讨了种族效应的表征机制，发现 GAN 生成面孔的潜在空间结构可有效映射人类对同族与异族面孔的感知差异（Shoura et al., 2025）。可见，GAN 不仅是生成高质量刺激的技术手段，GAN 与 DCNN 结合也增强了面孔表征机制的理解。

参考文献

Baek, S., Song, M., Jang, J., Kim, G., & Paik, S.-B. (2021). Face detection in untrained deep neural networks. *Nature communications, 12*(1), 7328.

Behrmann, M., & Avidan, G. (2022). Face perception: computational insights from phylogeny. *Trends in cognitive sciences, 26*(4), 350–363.

Belhumeur, P. N., Hespanha, J. P., & Kriegman, D. J. (1997). Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on pattern analysis and machine intelligence, 19*(7), 711–720.

Blauch, N. M., Behrmann, M., & Plaut, D. C. (2021). Computational insights into human perceptual expertise for familiar and unfamiliar face recognition. *Cognition, 208,* 104341.

Blauch, N. M., Behrmann, M., & Plaut, D. C. (2022). A connectivity-constrained computational account of topographic organization in primate high-level visual cortex. *Proceedings of the National Academy of Sciences, 119*(3), e2112566119.

Bothwell, R. K., Brigham, J. C., & Malpass, R. S. (1989). Cross-racial identification. *Personality and Social Psychology Bulletin, 15*(1), 19–25.

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., ...DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLOS Computational Biology, 10*(12), e1003963.

Calder, A. J. (2011). Oxford handbook of face perception: Oxford University Press.

Col ó n, Y. I., Castillo, C. D., & O’Toole, A. J. (2021). Facial expression is retained in deep networks trained for face identification. *Journal of Vision, 21*(4), 4–4.

Dahl, C. D., Logothetis, N. K., B ü lthoff, H. H., & Wallraven, C. (2010). The Thatcher illusion in humans and monkeys. *Proceedings of the Royal Society B: Biological Sciences, 277*(1696), 2973–2981.

Desimone, R. (1991). Face-selective cells in the temporal cortex of monkeys. *Journal of cognitive neuroscience, 3*(1), 1–8.

Dhar, P., Bansal, A., Castillo, C. D., Gleason, J., Phillips, P. J., & Chellappa, R. (2020). How are attributes expressed in face DCNNs? *Paper presented at the 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition* (FG 2020).

Dobs, K., Isik, L., Pantazis, D., & Kanwisher, N. (2019). How face perception unfolds over time. *Nature communications, 10*(1), 1258.

Dobs, K., Martinez, J., Kell, A. J. E., & Kanwisher, N. (2022). Brain-like functional specialization emerges spontaneously in deep neural networks. *Science advances, 8*(11), eabl8913.

Dobs, K., Yuan, J., Martinez, J., & Kanwisher, N. (2023). Behavioral signatures of face perception emerge in deep neural networks optimized for face recognition. *Proceedings of the National Academy of Sciences, 120*(32), e2220642120.

Dong, Y., Ruan, S., Su, H., Kang, C., Wei, X., & Zhu, J. (2022). Viewfool: Evaluating the robustness of visual recognition to adversarial viewpoints. *Advances in neural information processing systems, 35,* 36789–36803.

Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017). Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage, 152,* 184–194.

Geirhos, R., Temme, C. R. M., Rauber, J., Sch ü tt, H. H., Bethge, M., & Wichmann, F. A. (2018). *Generalisation in humans and deep neural networks. Advances in neural information processing systems, 31.*

Grossman, S., Gaziv, G., Yeagle, E. M., Harel, M., M é gevand, P., Groppe, D. M., ...Mehta, A. D. (2019). Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks. *Nature communications, 10*(1), 4934.

Gu, J., Yang, X., De Mello, S., & Kautz, J. (2017). Dynamic facial analysis: From bayesian filtering to recurrent neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1548–1557).

Güçlü, U., & Van Gerven, M. A. J. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of neuroscience, 35*(27), 10005–10014.

Gupta, P., & Dobs, K. (2025). Human-like face pareidolia emerges in deep neural networks optimized for face and object recognition. *PLOS Computational Biology, 21*(1), e1012751.

Hadjikhani, N., Kveraga, K., Naik, P., & Ahlfors, S. P. (2009). Early (M170) activation of face-specific cortex by face-like objects. *Neuroreport, 20*(4), 403–407.

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in cognitive sciences, 4*(6), 223–233.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hill, M. Q., Parde, C. J., Castillo, C. D., Colon, Y. I., Ranjan, R., Chen, J.-C., ...O'Toole, A. J. (2019). Deep convolutional neural networks in the face of caricature. *Nature Machine Intelligence, 1*(11), 522–529.

Jacob, G., Pramod, R. T., Katti, H., & Arun, S. P. (2021). Qualitative similarities and differences in visual object representations between brains and deep networks. *Nature communications, 12*(1), 1872.

Jiahui, G., Feilong, M., Visconti di Oleggio Castello, M., Nastase, S. A., Haxby, J. V., & Gobbini, M. I. (2023). Modeling naturalistic face processing in humans with deep convolutional neural networks. *Proceedings of the National Academy of Sciences, 120*(43), e2304085120.

Kadosh, K. C., & Johnson, M. H. (2007). Developing a cortex specialized for face perception. *Trends in cognitive sciences, 11*(9), 367–369.

Kanwisher, N. (2000). Domain specificity in face perception. *Nature neuroscience,* 3(8), 759–763.

Kanwisher, N., Gupta, P., & Dobs, K. (2023). CNNs reveal the computational implausibility of the expertise hypothesis. *Iscience, 26*(2).

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience, 17*(11), 4302–4311.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLOS Computational Biology, 10*(11), e1003915.

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K. A., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences, 116*(43), 21854–21863.

Konkle, T., & Alvarez, G. A. (2022). A self-supervised domain-general learning framework for human ventral stream representation. Nature communications, 13(1), 491.

Kramer, R. S. S., Young, A. W., & Burton, A. M. (2018). Understanding face familiarity. Cognition, 172, 46–58.

Kravitz, D. J., Saleem, K. S., Baker, C. I., Ungerleider, L. G., & Mishkin, M. (2013). The ventral visual pathway: an expanded neural framework for the processing of object quality. *Trends in cognitive sciences, 17*(1), 26–49.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems, 25.*

Lago, F., Pasquini, C., Böhme, R., Dumont, H., Goffaux, V., & Boato, G. (2021). More real than real: A study on human visual perception of synthetic faces [applications corner]. *IEEE Signal Processing Magazine, 39*(1), 109–116.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science, 350*(6266), 1332–1338.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature, 521*(7553), 436–444.

Li, Y., Zheng, W., Cui, Z., & Zhang, T. (2018). Face recognition based on recurrent regression neural network. *Neurocomputing, 297*, 50–58.

Lu, Z., & Wang, Y. (2025). *Category-Selective Neurons in Deep Networks: Comparing Purely Visual and Visual-Language Models.* arXiv preprint arXiv:2502.16456.

Luo, A., Henderson, M., Wehbe, L., & Tarr, M. (2023). Brain diffusion for visual exploration: Cortical discovery using large scale generative models. Advances in neural information processing systems, 36, 75740–75781.

Luo, A. F., Henderson, M. M., Tarr, M. J., & Wehbe, L. (2023). *Brainscuba: Fine-grained natural language captions of visual cortex selectivity.* arXiv preprint arXiv:2310.04420.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). *Towards deep learning models resistant to adversarial attacks.* arXiv preprint arXiv:1706.06083.

Margalit, E., Lee, H., Finzi, D., DiCarlo, J. J., Grill-Spector, K., & Yamins, D. L. K. (2024). A unifying framework for functional organization in early and higher ventral visual cortex. *Neuron, 112*(14), 2435–2451.

McKone, E., Kanwisher, N., & Duchaine, B. C. (2007). Can generic expertise explain special processing for faces? *Trends in cognitive sciences, 11*(1), 8–15.

Nestor, A., Plaut, D. C., & Behrmann, M. (2016). Feature-based face representations and image reconstruction from behavioral and neural data. *Proceedings of the National Academy of Sciences, 113*(2), 416–421.

O'Toole, A. J., & Castillo, C. D. (2021). Face recognition by humans and machines: three fundamental advances from deep learning. *Annual Review of Vision Science, 7*(1), 543–570.

O'Toole, A. J., Roark, D. A., & Abdi, H. (2002). Recognizing moving faces: A psychological and neural synthesis. *Trends in cognitive sciences, 6*(6), 261–266.

O'Toole, A. J., Abdi, H., Deffenbacher, K. A., & Valentin, D. (1993). Low-dimensional representation of faces in higher dimensions of the face space. *Journal of the Optical Society of America A, 10*(3), 405–411.

O'Toole, A. J., Castillo, C. D., Parde, C. J., Hill, M. Q., & Chellappa, R. (2018). Face space representations in deep convolutional neural networks. *Trends in cognitive sciences, 22*(9), 794–809.

O'Toole, A. J., Deffenbacher, K. A., Valentin, D., McKee, K., Huff, D., & Abdi, H. (1998). The perception of face gender: The role of stimulus structure in recognition and classification. *Memory & cognition, 26*, 146–160.

Phillips, P. J., & O'Toole, A. J. (2014). Comparison of human and computer performance across face recognition experiments. *Image and Vision Computing, 32*(1), 74–85.

Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K., ...Sankaranarayanan, S. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences, 115*(24), 6171–6176.

Prince, J. S., Alvarez, G. A., & Konkle, T. (2024). Contrastive learning explains the emergence and function of visual category-selective regions. *Science advances, 10*(39), eadl1776.

Raman, R., & Hosoya, H. (2020). Convolutional neural networks explain tuning properties of anterior, but not middle, face-processing areas in macaque inferotemporal cortex. *Communications biology, 3*(1), 221.

Ratan Murty, N. A., Bashivan, P., Abate, A., DiCarlo, J. J., & Kanwisher, N. (2021). Computational models of category-selective brain regions enable high-throughput tests of selectivity. *Nature communications, 12*(1), 5540.

Rhodes, G., & Leopold, D. A. (2011). Adaptive norm-based coding of face identity. *The Oxford handbook of face perception,* 263–286.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature neuroscience, 2*(11), 1019–1025.

Rolls, E. T., & Milward, T. (2000). A model of invariant object recognition in the visual system: learning rules, activation functions, lateral inhibition, and information-based performance measures. *Neural computation, 12*(11), 2547–2572.

Rossion, B. (2014). Understanding face perception by means of human electrophysiology. *Trends in cognitive sciences, 18*(6), 310–318.

Shoham, A., Grosbard, I. D., Patashnik, O., Cohen-Or, D., & Yovel, G. (2024). Using deep neural networks to disentangle visual and semantic information in human perception and memory. *Nature Human Behaviour, 8*(4), 702–717.

Shoura, M., Walther, D. B., & Nestor, A. (2025). Unraveling other-race face perception with GAN-based image reconstruction. *Behavior Research Methods, 57*(4), 1–14.

Simonyan, K., & Zisserman, A. (2014a). Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems, 27*.

Simonyan, K., & Zisserman, A. (2014b). *Very deep convolutional networks for large-scale image recognition.* arXiv preprint arXiv:1409.1556.

Song, Y., Qu, Y., Xu, S., & Liu, J. (2021). Implementation-independent representation for deep convolutional neural networks and humans in processing faces. *Frontiers in computational neuroscience, 14*, 601314.

Taigman, Y., Yang, M., Ranzato, M. A., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1701–1708).

Tanaka, J. W., & Sengco, J. A. (1997). Features and their configuration in face recognition. *Memory & cognition, 25*(5), 583–592.

Taubert, J., Wardle, S. G., Flessert, M., Leopold, D. A., & Ungerleider, L. G. (2017). Face pareidolia in the rhesus monkey. *Current Biology, 27*(16), 2505–2509.

Thompson, P. (1980). Margaret Thatcher: A new illusion. *Perception, 9*(4), 483–484.

Tian, F., Xie, H., Song, Y., Hu, S., & Liu, J. (2022). The face inversion effect in deep convolutional neural networks. *Frontiers in computational neuroscience, 16*, 854218.

Tian, J., Xie, H., Hu, S., & Liu, J. (2021). Multidimensional face representation in a deep convolutional neural network reveals the mechanism underlying AI racism. *Frontiers in computational neuroscience, 15*, 620281.

Tsao, D. Y., Moeller, S., & Freiwald, W. A. (2008). Comparing face patch systems in macaques and humans. *Proceedings of the National Academy of Sciences, 105*(49), 19514–19519.

Valentine, T., Lewis, M. B., & Hills, P. J. (2016). Face-space: A unifying concept in face recognition research. *Quarterly journal of experimental psychology, 69*(10), 1996–2019.

Vinken, K., Prince, J. S., Konkle, T., & Livingstone, M. S. (2023). The neural code for "face cells" is not face-specific. *Science advances, 9*(35), eadg1736.

Wang, J., Cao, R., Brandmeir, N. J., Li, X., & Wang, S. (2022). Face identity coding in the deep neural network and primate brain. *Communications biology, 5*(1), 611.

Wang, M., & Deng, W. (2020). Mitigating bias in face recognition using skewness-aware reinforcement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9322–9331).

Wardle, S. G., Taubert, J., Teichmann, L., & Baker, C. I. (2020). Rapid and dynamic processing of face pareidolia in the human brain. *Nature communications, 11*(1), 4518.

Wichmann, F. A., & Geirhos, R. (2023). Are deep neural networks adequate behavioral models of human visual perception? *Annual Review of Vision Science, 9*(1), 501–524.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences, 111*(23), 8619–8624.

Yin, R. K. (1969). Looking at upside-down faces. *Journal of experimental psychology, 81*(1), 141.

Yovel, G., Grosbard, I., & Abudarham, N. (2023). Deep learning models challenge the prevailing assumption that face-like effects for objects of expertise support domain-general mechanisms. *Proceedings of the Royal Society B, 290*(1998), 20230093.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2016). *Understanding deep learning requires rethinking generalization.* arXiv preprint arXiv:1611.03530.

Zhou, L., Yang, A., Meng, M., & Zhou, K. (2022). Emerged human-like facial expression representation in a deep convolutional neural network. *Science advances, 8*(12), eabj4383.

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., & Yamins, D. L. K. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences, 118*(3), e2014196118.

The Performance of Deep Convolutional Neural Networks in Face Recognition and the Comparison with the Human Visual System

Cheng Yuhui[1], *Shen Tianyu*[1], *Lu Zitong*[2], *Yuan Xiangyong*[3,4], *Jiang Yi*[3,4]

([1] School of Psychology, Nanjing Normal University, Nanjing, 210097)

([2] Massachusetts Institute of Technology McGovern Institute for Brain Research, Cambridge, MA, 02139)

([3]State Key Laboratory of Cognitive Science and Mental Health, Institute of Psychology, Chinese Academy of Sciences, 16 Lincui Road, Beijing, 100101)

([4] Department of Psychology, University of Chinese Academy of Sciences, 19A Yuquan Road, Beijing, 100049)

Abstract Face recognition is a fundamental cognitive function that plays a crucial role in human social interaction, as the human brain exhibits a remarkable sensitivity to facial stimuli. For decades, psychologists, cognitive neuroscientists, and computer vision researchers have been dedicated to uncovering the behavioral and neural mechanisms underlying face processing. Existing studies have demonstrated that humans process facial information differently from other objects, supporting the existence of highly specialized mechanisms for face perception. In particular, the fusiform face area (FFA) in the human brain has been identified as a specialized region for face recognition, and numerous face-selective neurons have been observed in the temporal lobe of macaques. In recent years, Deep Convolutional Neural Networks (DCNNs) have demonstrated remarkable performance in modeling and understanding face processing, providing new computational perspectives for exploring the neural mechanisms underlying face recognition. DCNNs are a class of artificial neural networks that have achieved impressive performance in visual recognition tasks, including face recognition. These models typically begin by applying a series of convolutional and pooling operations to extract increasingly abstract features, which are then passed through one or more fully connected layers to perform classification tasks. Consequently, there has been a growing interest in investigating the applications of DCNNs in face recognition.

First, this review examines the performance of DCNNs in identifying key facial attributes. Although most DCNNs are trained only for face identity tasks, they can still infer social information such as gender and expression. In addition, this review also discusses the similarities and differences between DCNNs and humans in well-known face processing phenomena, such as the inversion, own-race, and familiarity effects. Evidence suggests that DCNNs can produce face-specific cognitive effects similar to those observed in humans. To better understand the computational validity of DCNNs, this review compares their internal representations with the neural mechanisms involved in human face recognition. On the one hand, this paper analyzes the hierarchical processing architecture that emerges in trained DCNNs and evaluates its correspondence with the hierarchical structure of the human visual system, spanning from early visual areas (e.g., V1–V4) to higher-level face-selective regions such as the FFA. On the other hand, this review further discusses evidence for brain-like functional specialization within DCNNs, examining whether units selective to different facial attributes can be mapped onto the functionally specialized cortical areas observed in neuroimaging and electrophysiological studies.

Lastly, this paper highlights several limitations of current models and outlines promising directions for future research. First, although DCNNs excel at face recognition, they remain far less robust than humans when faced with challenges such as viewpoint shifts, image distortions, adversarial perturbations, and limited training data. Second, although DCNNs exhibit behavioral effects like those observed in humans, there are multiple possible explanations for the underlying mechanisms responsible for these phenomena. The DCNN models examined in different studies often vary in terms of architecture, task objectives, and training datasets, which may affect the comparability of their results. Third, the extent to which current models can capture essential features of the biological visual system remains unclear. Specifically, many DCNNs operate as feedforward architectures and lack critical elements such as recurrent processing, top-down feedback, and dynamic attentional modulation, all of which are fundamental characteristics of the human visual system. Fourth, current neural network models primarily focus on the perceptual stage underlying face recognition. Future research should aim to incorporate semantic-level processing to more fully capture the complexity of human face perception. Fifth, generative Adversarial Networks (GANs) have recently attracted significant attention, which are powerful tools for generating diverse facial stimuli, enabling more controlled and flexible investigations of face perception. Integrating GANs with DCNNs has also enhanced our understanding of the mechanisms underlying facial representation, making it a promising direction for future research..

Key words face recognition, Convolutional Neural Network, fusiform face area, hierarchical structure, functional specialization