

2021 届研究生硕士学位论文

分类号: _____

学校代码: 10269

密 级: _____

学 号: 51183200130



华东师范大学

East China Normal University

硕士学位论文

MASTER'S DISSERTATION

论文题目: 人脑与深度卷积神经网络模型在面孔感知过程中的相似与不相似性表征

院 系: 心理与认知科学学院

专 业: 认知神经科学

研究方向: 感知觉和认知神经科学

指导教师: 库逸轩教授、王惠敏研究员

学位申请人: 路子童

2021 年 3 月

Dissertation for master degree in 2021

University code: 10269

Student ID: 51183200130

East China Normal University

Title: Similar and Dissimilar
Representations between the Human Brain
and the Deep Convolutional Neural
Network in Face Perception

Department: School of Psychology and Cognitive Science

Major: Cognitive Neuroscience

Research Direction: Sensory Perception and Cognitive Neuroscience

Supervisor: Prof. Yixuan Ku & Prof. Huimin Wang

Candidate: Zitong Lu

March 2021

华东师范大学学位论文原创性声明

郑重声明：本人呈交的学位论文《人脑与深度卷积神经网络模型在面孔感知过程中的相似与不相似性表征》，是在华东师范大学攻读硕士/博士（请勾选）学位期间，在导师的指导下进行的研究工作及取得的研究成果。除文中已经注明引用的内容外，本论文不包含其他个人已经发表或撰写过的研究成果。对本文的研究做出重要贡献的个人和集体，均已在文中作了明确说明并表示谢意。

作者签名：_____

日期： 年 月 日

华东师范大学学位论文著作权使用声明

《人脑与深度卷积神经网络模型在面孔感知过程中的相似与不相似性表征》系本人在华东师范大学攻读学位期间在导师指导下完成的硕士/博士（请勾选）学位论文，本论文的著作权归本人所有。本人同意华东师范大学根据相关规定保留和使用此学位论文，并向主管部门和学校指定的相关机构送交学位论文的印刷版和电子版；允许学位论文进入华东师范大学图书馆及数据库被查阅、借阅；同意学校将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于（请勾选）

- () 1. 经华东师范大学相关部门审查核定的“内部”或“涉密”学位论文*，于 年 月 日解密，解密后适用上述授权。
- () 2. 不保密，适用上述授权。

导师签名_____

本人签名_____

年 月 日

* “涉密”学位论文应是已经华东师范大学学位评定委员会办公室或保密委员会审定过的学位论文（需附获批的《华东师范大学研究生申请学位论文“涉密”审批表》方为有效），未经上述部门审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权）。

内容摘要

人脑能在极短时间内编码面孔所带来的各种面部信息，通过这些信息为各种其他认知行为与社会行为提供支持。而与生物脑不同，以计算模型形式存在的深度卷积神经网络却能够在很多任务中达到人类的表现水平。本研究突破传统心理学研究的瓶颈，尝试直接地解析人脑与深度卷积神经网络模型在面孔感知过程中对不同面孔信息的内在编码机制与表征异同。

本研究第一部分首先设计与实现了两个基于 Python 的可广泛用于前沿认知神经科学科研的工具包——NeuroRA 和 PyCTRSA。前者可进行跨模态神经数据表征分析，后者可进行基于全新的跨时域表征相似性分析的脑电或脑磁数据解码。

本研究第二部分基于脑电数据使用神经解码与表征相似性分析解析了人脑在面孔感知过程中对不同面孔信息的时序动态编码过程。结果发现，大脑会更早加工视觉相关信息，随后加工更复杂的面孔信息。本研究也首次直接分离了对熟悉面孔与不熟悉面孔的时序编码差异，大脑会更早进行前者的加工，而更晚加工后者。结合神经解码与表征相似性分析的结果，也发现面孔重复效应从 300ms 左右开始出现，且对于熟悉面孔效应更强。

本研究第三部分基于表征相似性分析探究了深度卷积神经网络模型在面孔感知过程中对不同面孔信息的分层表征方式。结果发现，经过面孔图片训练的 VGG-Face 模型与随机权重的 VGG-16 模型对面孔信息的表征存在巨大差异，在全连接层上，前者对三种类别的面孔（熟悉面孔、不熟悉面孔、乱相面孔）都存在一定类别内部表征相似性，后者却无。同时，深度卷积神经网络模型不特异性编码熟悉面孔。对其他面孔信息，VGG-Face 都表现出逐层表征增强，而未经过训练的模型则逐层表征减弱。

本研究第四部分基于假设构建了两个重复抑制模型对深度卷积神经网络模型进行激活修改，并使用分别经过两模型修改后的深度卷积神经网络模型与人脑进行跨模态表征相似性分析。结果发现，面孔识别过程中重复抑制效应更可能是由一种衰减机制造成，即是由对面孔刺激响应更强的神经元发生激活衰减造成，且深度卷积神经网络模型与人脑在面孔感知过程中存在一定时序上的分层表征相似性。

综上，本研究成果为之后认知神经科学家给予了很多方法上的便利与启示，而通过本研究里对面孔感知过程中人脑与深度卷积神经网络模型的内部表征分析的结果以及

两者之间的相似与不相似表征的结果，既对未来认知神经科学领域中探究面孔感知神经机制、也对类脑智能领域中优化面孔识别模型具有巨大意义。

关键词: 面孔感知，事件相关电位，深度卷积神经网络，神经解码，表征相似性分析，重复抑制

Abstract

The human brain has the ability to encode different kinds of facial information in a very short time, which can provide support for many other cognitive and social behaviors. Different from biological brain, the deep convolutional neural network (DCNN) exists in the form of a computational model. However, it can achieve the level of human performances in many tasks. Our study broke through the bottleneck of traditional psychological research, and attempted to directly analyze the internal coding mechanism in the human brain and DCNN model and similar and dissimilar representation between both.

Firstly, we designed and implemented two toolboxes, NeuroRA and PyCTRSA, which could be widely used in neuroscience domain based on Python. The former can be used to conduct representational analysis cross multi-modal neural data, while the latter can be used to conduct electroencephalography (EEG) and magnetoencephalography (MEG) decoding based on the novel cross-temporal representational similar analysis.

Secondly, we used neural decoding and representational similarity analysis (RSA) methods to analyze the temporal dynamic coding process of the human brain during face perception based on EEG. The results showed that the human brain processes visual information earlier, and then processes more complex facial information. Also, this study was the first study to directly separate the temporal coding differences between familiar faces and unfamiliar faces. Combined with the results of neural decoding and RSA, it was also found that face repetition effect began to appear from about 300ms, and this effect was stronger for familiar faces.

Thirdly, we explored the hierarchical representation of different facial information in the DCNN model during face perception based on RSA. The results showed that there were significant differences between pretrained VGG-Face model and nontrained VGG-16 model with random weights in the representation of facial information. With the full connected layer, VGG-Face model had certain within-type representational similarity for three types of faces (familiar faces, unfamiliar faces and scrambled faces), but randomly-weighted VGG-16 model had no similarity. Surprisingly, we found that the DCNN model didn't specifically

encode familiar faces. However, for other facial information, VGG-Face model showed layer-by-layer representation enhancement, while randomly-weighted VGG-16 model showed layer-by-layer representation weakening.

Finally, we constructed two repetition suppression (RS) models based on the hypothesis to modify the activations of the DCNN model and conducted cross-modal RSA between DCNN models modified by two RS models and the human brain. The results showed that the RS effect in face recognition was more likely caused by a fatigue mechanism, that was, the activation of neurons with stronger response to face stimulus was attenuated. In addition, the DCNN model had a certain hierarchical representation similarity with the human brain during face perception.

In summary, our research achievements have provided many convenience and enlightenment for cognitive neuroscientists. Through the results of internal representations in the human brain and the DCNN model in the process of face perception, as well as the results of similar and dissimilar representations between them in our study, it is not only helpful to explore the neural mechanism of face perception in cognitive neuroscience domain in the future, but also of great significance to optimize face recognition models in brain-inspired intelligence domain.

Keywords: face perception, ERP, DCNN, neural decoding, RSA, repetition suppression

目 录

1 绪论.....	1
1.1 面孔感知的神经机制.....	1
1.2 重复抑制及其相关研究.....	3
1.3 神经解码及其相关研究.....	5
1.4 表征相似性分析及其相关研究.....	6
1.5 比较人脑与深度卷积神经网络模型.....	8
1.6 研究目的与意义.....	10
2 用于神经数据表征分析的工具包.....	12
2.1 NEURORA：一个用于多模态神经数据表征分析的 PYTHON 工具包	12
2.1.1 NeuroRA 概述	12
2.1.2 模块与功能	14
2.1.3 使用示例	18
2.2 PYCTRSA：一个基于跨时域表征相似性分析的脑电/脑磁数据解码的 PYTHON 工具包	20
2.2.1 跨时域表征相似性分析	20
2.2.2 PyCTRSA 概述	21
2.2.3 模块与功能	21
2.2.4 使用示例	22
3 探究面孔信息在人脑中的动态表征.....	23
3.1 数据与实验.....	23
3.2 分析方法.....	24
3.2.1 脑电数据的预处理	24
3.2.2 基于脑电的分类解码	25
3.2.3 基于脑电的表征相似性分析	26
3.2.4 统计分析	29
3.2.5 基于脑电的面孔表征可视化	29
3.3 结果.....	29

3.4 讨论	36
4 探究面孔信息在深度卷积神经网络模型中的表征	39
4.1 深度卷积神经网络模型	39
4.2 分析方法	40
4.2.1 特征降维	40
4.2.2 基于深度卷积神经网络模型的表征相似性分析	41
4.2.3 基于深度卷积神经网络模型的面孔表征可视化	43
4.3 结果	43
4.4 讨论	48
5 探究人脑与深度卷积神经网络模型在面孔感知过程中的表征差异	51
5.1 分析方法	51
5.1.1 模型模拟	51
5.1.2 基于模型修改的表征不相似矩阵构建	53
5.1.3 跨模态的表征分析	53
5.2 结果	55
5.3 讨论	59
6 结语	61
6.1 总结与讨论	61
6.2 不足与展望	62
参考文献	65
附录	86
致谢	88

插图和附表目录

图 1-1 面孔感知的简易时序认知模型.....	2
图 1-2 重复抑制模型.....	4
图 1-3 MVPA 示意图	5
图 1-4 RSA 通过一个共同表征空间来连接不同模态的数据.....	7
图 1-5 分层卷积神经网络作为知觉皮层的模型.....	9
图 2-1 NeuroRA 总览	12
图 2-2 NeuroRA 中计算 RDM 的实现.....	15
图 2-3 两 RDMs 间计算的示意图.....	15
图 2-4 进行基于脑电或类似脑电数据的跨时间和导联的 RDMs 的相似性分析的示意图	16
图 2-5 使用 NeuroRA 进行核磁数据的表征分析的示意图	16
图 2-6 通过 NeuroRA 实现的可视化案例.....	18
图 2-7 示例结果.....	19
图 2-8 PyCTRSA 示例结果	22
图 3-1 实验流程示意图.....	24
图 3-2 实验条件（面孔类别与刺激状态）及对应简称示意图.....	24
图 3-3 神经 RDM 计算构建示意图	26
图 3-4 编码模型 RDMs (9×9)	27
图 3-5 神经 RDM 与编码模型 RDMs 间 GLM 计算示意图.....	28
图 3-6 基于脑电的面孔类别的分类解码结果.....	30
图 3-7 基于脑电的刺激状态的分类解码结果.....	32
图 3-8 神经 RDMs 及面孔表征可视化示例.....	34
图 3-9 人脑对不同面孔信息的动态表征.....	35
图 3-10 神经表征与编码模型的 GLM 结果.....	36
图 4-1 VGG-Face 模型结构	39
图 4-2 DCNN 模型 RDM 的构造过程示意	40
图 4-3 编码模型 RDMs (450×450)	42

图 4-4 神经 RDM 与编码模型 RDMs 间 GLM 计算示意图.....	43
图 4-5 DCNN 模型偶数层 RDMs.....	44
图 4-6 DCNNs 对三种类别面孔的内部表征相似性.....	45
图 4-7 DCNNs 对面孔信息的分层表征.....	46
图 4-8 基于 DCNNs 模型的面孔表征可视化结果.....	48
图 5-1 重复抑制的衰减模型与锐化模型示意图.....	52
图 5-2 跨模态相似性分析计算示意图.....	54
图 5-3 DCNN 模型经过重复抑制模型修改后第 16 层的表征	56
图 5-4 DCNN 模型经过重复抑制模型修改后第 16 层与人脑对面孔的有效表征相似性	57
图 5-5 人脑与基于模型修改的 DCNN 模型的时序分层表征相似性	58
附图 1 DCNN 模型经过两参数精细化重复抑制模型修改后第 16 层的表征.....	86
附图 2 DCNN 模型经过参数精细化模型修改后第 16 层与人脑对面孔的有效表征相似性	86
附图 3 人脑与基于参数精细化模型修改的 DCNN 模型的时序分层表征相似性.....	87

1 绪论

1.1 面孔感知的神经机制

人类对面孔识别的能力是不可思议的，通常，我们可以轻易地去识别一张脸、认出一个人、甚至通过他（她）的面部表情变化了解他（她）的内心情绪波动。通过面孔来识别与了解一个人——这一能力对于人类的各种社会功能起着至关重要的作用。认知神经科学领域也有大量关于面孔感知的研究(Behrman & Plaut, 2013; Duchaine & Yovel, 2015; Freiwald et al., 2016; Haxby et al., 2000; Kanwisher, 2000; Rossion, 2008)，深入地了解不同的精细的面孔信息是如何参与到对面孔的编码、学习与识别过程中是一个重要的挑战。

面孔处理被认为不同于非人脸客体处理，因为它更具有整体性，即面孔被表征为非分解的整体，而不是独立表征各组成部分（如眼睛、鼻子和嘴巴）及其之间关系的组合(Farah et al., 1998)。对面孔进行整体加工的证据来自许多行为范式，其中最常被引用的两种是部分-整体效应(Tanaka & Farah, 1993)和复合效应(Young et al., 1987)。在部分-整体效应中，被试在整张面孔的背景下识别两个面孔部分的能力要比单独识别时强。在复合效应中，被试在辨认与不一致的另半张面孔对齐的嵌合面孔的一半时，比辨认两个半张面孔不对齐时要慢(Young et al., 1987)。

由于 electroencephalography (EEG) and magnetoencephalography (MEG) 具有较高的时间分辨率，关于人脑面孔感知的时间加工过程的研究主要基于这两个技术。如图 1-1 所示，为一个基于大量研究调整后(Bruce & Young, 1986; James V. Haxby et al., 2000; Schweinberger & Burton, 2003)的简易时序上的面孔感知过程的认知模型(Schweinberger & Neumann, 2016)。假定人脸识别涉及多个功能阶段或者说是神经计算，在低级视觉分析（图片编码）时，将对视觉刺激及其图像元素进行详细分析。随后的阶段则是将基本信息（如相关轮廓、形状、颜色等等）整合到一个统一的整体的面孔表征中。尽管先前的模型通常没有确切指定哪种视觉信息对于人脸识别至关重要，但是单个二阶空间结构的编码（即特征之间按照度量标准布局排布的个体空间关系）已经被认为是个体面孔识别的关键信息。一些研究者认为二阶空间结构对学习新面孔可能非常重要，然而他们对于识别高度熟悉的面孔而言却没那么重要(Itz et al., 2014; Kaufmann et al., 2013)。一旦视觉刺激编码为一种合适的表征，它就需要和更持久性的熟悉面孔的表征——面孔识别单元 (face recognition units, FRUs) 进行比较。当传入的信息更类似于一个已经存储的表

征，则相应的 FRU 激活更强。一旦一个 FRU 被充分激活，该面孔可以被认为是熟悉的，并且一个“个体身份节点”（person identity node, PIN）被激活。PINs 被概念化为会聚节点，其可以通过面部、声音、身体或其他个性化提示来激活。然后 PIN 可以运行进一步提取其他语义或情节信息（如职业、居住地、最近遭遇等等），最终获取一个人的名字。

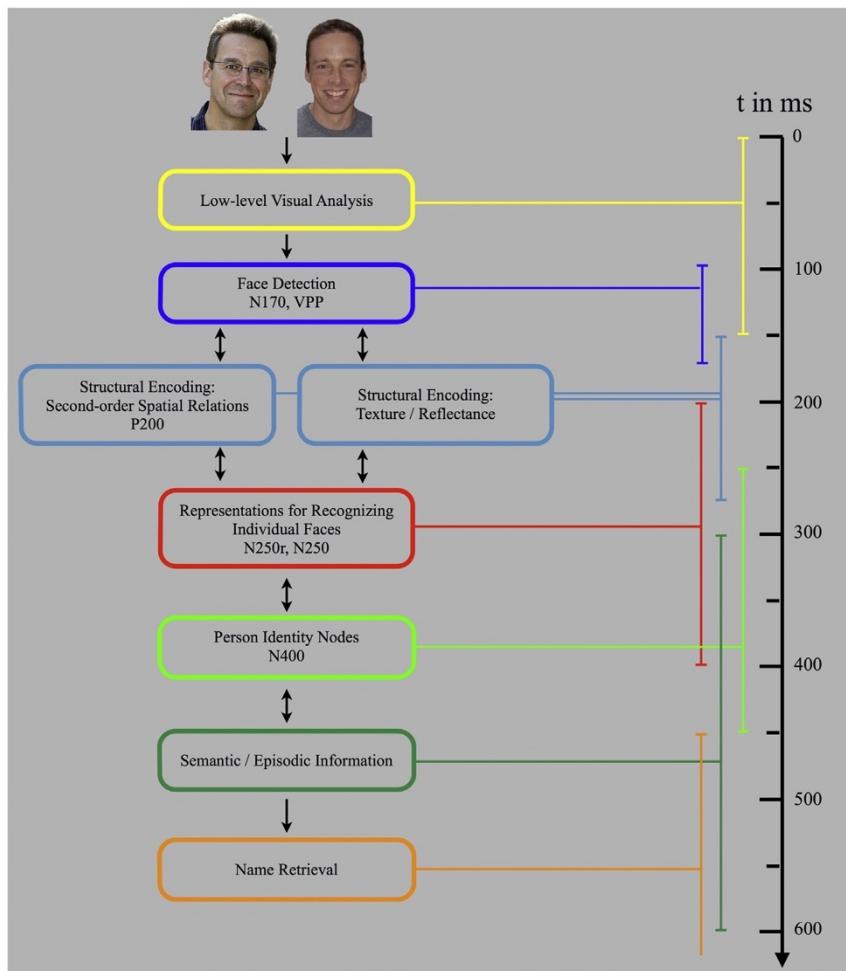


图 1-1 面孔感知的简易时序认知模型(Schweinberger & Neumann, 2016)

左侧不同颜色框之间的箭头：单向代表自顶向下，双向代表既有自底向上又有自顶向下；右侧为时间过程。

大量通过传统的事件相关电位(event-related potentials, ERPs)(Bentin & Deouell, 2000; Huddy et al., 2003; Itier & Taylor, 2002; Rossion & Caharel, 2011; Tanaka et al., 2006; Zheng et al., 2012)和时空模式(Cichy et al., 2014; H. Liu et al., 2009; Vida et al., 2017)的研究已经揭示了面孔加工的时间轮廓（包括低维视觉分析、面孔检测、结构编码、面孔再认的表征、自我识别、语义或情境信息编码和姓名提取等等）。但是相对而言，对于面孔信息的动态处理过程却所知甚少，包括对于面孔身份信息、面孔的熟悉度、面孔的完整性等等以及面孔重复效应下对面孔信息的动态处理过程的影响。

最近，很多研究开始使用模式分析方法来进行不同面孔类别与样本水平的区分(Carlson et al., 2013; Cauchoix et al., 2014; Cichy et al., 2014; Davidesco et al., 2014; Ghuman et al., 2014; Kaneshiro et al., 2015; Van de Nieuwenhuijzen et al., 2013)，包括使用EEG(Nemroodov et al., 2016)、MEG(Vida et al., 2017)和皮层脑电(electrocorticography, ECeG) (Ghuman et al., 2014)来进行表情不变的面孔身份区分，这些研究均发现了不同的时间窗对面孔信息敏感，而这些发现也与猴子神经电生理(W. A. Freiwald et al., 2009; Hung et al., 2005)和人类心理物理学研究(Crouzet et al., 2010; Lehky, 2000; Tanaka & Curran, 2001)的结果一致。而根据一些功能性磁共振成像(functional magnetic resonance imaging, fMRI)进行的皮层溯源研究发现面孔识别的相关信息主要来自于梭状回(fusiform gyrus, FG) (Anzellotti et al., 2014; Goesaert & Op de Beeck, 2013; Nestor et al., 2011)。

1.2 重复抑制及其相关研究

当重复感受相同刺激时，诱发的神经信号的强度总是会小于前一次观察到的神经信号，这一过程通常被称为重复抑制(repetition suppression, RS)效应。例如在很多猴子的研究中都可以观察到在刺激重复出现的条件下，其下颞皮层的视觉敏感的神经元响应会降低(Baylis & Rolls, 1987; Kaliukhovich & Vogels, 2011, 2012; Miller, Gochin, et al., 1991; Miller, Li, et al., 1991; Ringo, 1996; Sawamura et al., 2006; Sobotka & Ringo, 1994)。同样，在fMRI研究中，重复刺激可以降低血氧水平依赖性(blood oxygenation level-dependent, BOLD) (Henson & Rugg, 2003)。

在面孔重复表征的过程中，大量研究通过ERPs观察到了各种重复效应。在1995年两项研究中都发现连续呈现两面孔的过程中，若后一面孔与前一面孔一样时，相较后一面孔为新颖面孔的条件下，大脑腹部颞叶会观察到一个很强的负波，其时间范围大概在200-300ms，并且熟悉面孔相较不熟悉面孔其对应的重复效应更强(Begleiter et al., 1995; Schweinberger et al., 1995)。随后大量ERP的研究发现了不同的ERP成分可能与不同的面孔感知过程中的成分存在着联系。顶颞部的N170成分与面孔结构的检测有关，即N170这一成分对刺激“类别”敏感，当当前面孔的前一试次同样为一面孔刺激即使身份不同，N170的振幅都会发生抑制(Kloth et al., 2010; Kloth & Schweinberger, 2010; Kovács et al., 2006; Maurer et al., 2008; Mercure et al., 2011; Schweinberger et al., 2007; Walther et al., 2013)。随后大脑枕颞部的P200成分也发现是对面孔敏感的，其会受到面孔二阶空间结构的典型性的调节(Burkhardt et al., 2010; Kaufmann & Schweinberger, 2012; Latinus &

Taylor, 2006; Schulz et al., 2012; Zheng et al., 2012)。而来自腹侧颞叶的一个 ERP 负波成分——N250r (其中 r 代表重复, repetition), 在重复一个面孔时比不重复时更大, 且对面孔熟悉度高度敏感, 其重复效应在熟悉条件下比不熟悉条件下更大(Dörr et al., 2011; Herzmann et al., 2004; Pfütze et al., 2002; Schweinberger et al., 1995; Schweinberger & Burton, 2003; Wiese et al., 2013)。最后中顶叶区域与语义相关的 N400 成分则反映了对熟悉的人 (而非面孔) 的识别中的语义重复效应(Barrett & Rugg, 1989; Bentin et al., 1985; Rugg, 1985; Schweinberger, 1996; Stevenage et al., 2014)。

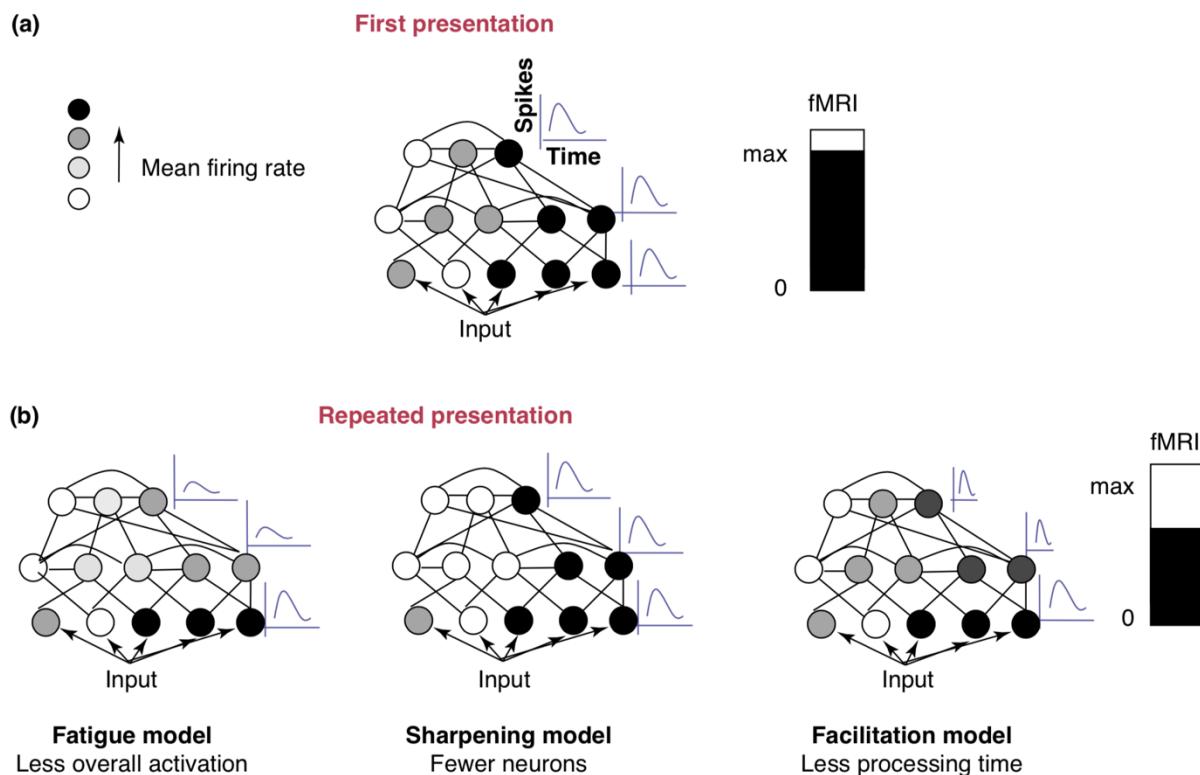


图 1-2 重复抑制模型(Grill-Spector et al., 2006)

(a) 初始条件下的表征; (b) 重复条件下基于衰减模型 (fatigue model)、锐化模型 (sharpening model) 和促进模型 (facilitation model) 的表征。

然而重复抑制效应的潜在神经机制仍然尚不清楚, Grill-Spector 等人(Grill-Spector et al., 2006)基于先前单细胞记录、fMRI 以及 EEG/MEG 的研究提出来三种可能的重复抑制模型 (基于 fMRI 的 BOLD 信号), 如图 1-2 所示。对于衰减 (fatigue) 模型, 其模型假设是对刺激产生最佳响应的神经元中发生重复抑制神经元的数量将比其他神经元中发生重复抑制神经元的数量多; 对于锐化 (sharpening) 模型, 其假设是编码与识别刺激无关的特征的神经元会出现重复抑制导致对刺激编码的稀疏表示; 对于促进 (facilitation) 模型, 其假设是重复会导致更快的刺激处理, 即更短的等待时间或更短的持续时间。一

些研究通过 fMRI 推断重复抑制的神经机制，例如通过单变量和多变量分析的方法进行建模比较不同神经机制得到的重复抑制效应并与实际神经反应比较(Alink et al., 2018; Weiner et al., 2010)。

1.3 神经解码及其相关研究

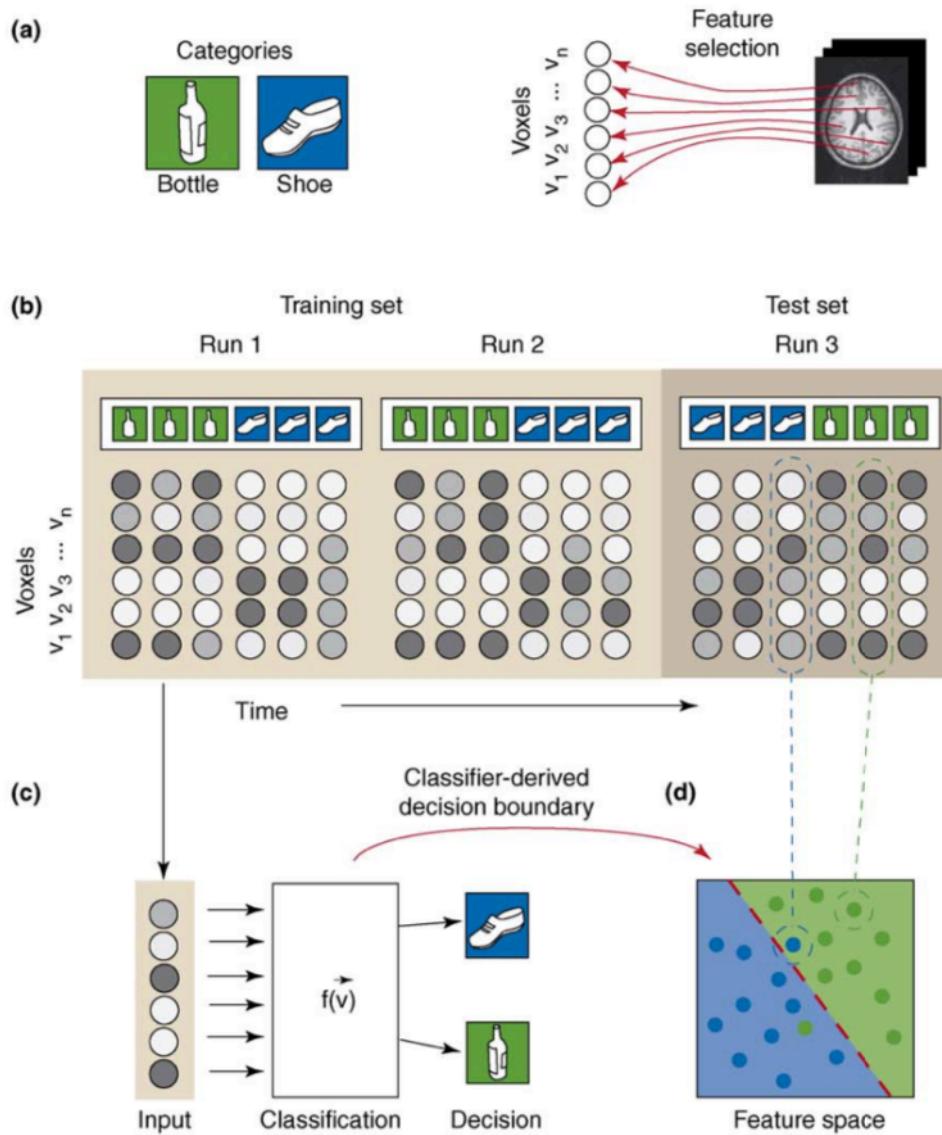


图 1-3 MVPA 示意图(Norman et al., 2006)

(a) 两图片对应的核磁激活反应；(b) 划分训练集与测试集；(c) 分类器分类。

在传统的神经科学的研究中往往限制于单变量分析，以核磁为例，对某一神经活动的相关研究常常是基于单体素激活值(Baddeley, 2003; Nee et al., 2013)，这样缺少对隐藏的神经模式进行有效评估(Norman et al., 2006)。多变量模式分析(multivariate pattern analysis, MVPA)则是一种用于深入探究大脑对不同信息隐藏的表征的更有效的方法(Norman et al., 2006)。如图 1-3 所示(Norman et al., 2006)，大脑对不同类别物体的编码模式不同，当

大脑对这两物体进行编码时，两物体对应的神经活动在高维空间中的表征模式存在差异，当大脑没有对这两物体进行编码时，对应的神经活动在高位空间上的分布是随机的。通过训练一个对两物体空间分布进行分类的分类器能有效探究两物体的表征模式是否存在差异，通常这里使用线性支持向量机（line support vector machine, Line-SVM）进行训练与测试，在二维的特征空间上对两种类型的表征进行区分。

通过这种基于分类的神经解码方法应用到 fMRI，不仅可以探究大脑对不同类别的面孔和客体的编码情况还能探究大脑对朝向、位置、颜色和情绪等等特征的编码(Albers et al., 2013; Ban et al., 2012; Bannert & Bartels, 2013; Bo et al., 2021; Johnson & Johnson, 2014; Koch et al., 2020; Koenig-Robert & Pearson, 2019; Lescroart & Gallant, 2019; S. Liu et al., 2019; Reddy et al., 2010; Schlegel et al., 2013; Stokes et al., 2009; Vetter et al., 2014)。fMRI 具有较高的空间分辨率，能够通过神经解码的方法细化到不同脑区对不同信息的编码。而 EEG 和 MEG 技术具有较高的时间分辨率，能够利用 EEG 和 MEG 技术结合神经解码方法在时序上探究大脑对不同信息的动态编码情况(Bae & Luck, 2019a; Fahrenfort et al., 2017; Grootswagers et al., 2019; Hogendoorn & Burkitt, 2018; Hong et al., 2020; Long & Kuhl, 2019; Mares et al., 2020; Nemrodov et al., 2018; Noah et al., 2020; Robinson et al., 2019; Shatek et al., 2019; Smith & Smith, 2019; Teichmann et al., 2020; Xie et al., 2020)。在时序上逐时间点或逐时间窗进行分类器的训练与测试得到对应时间的解码正确率，即可得到在整个认知过程中的解码正确率曲线，通过与随机情况的统计分析可以追踪到编码对应信息的时间段。这种通过神经解码的方法不仅适用于感知觉研究，也逐渐开始应用于对记忆内容的探究(Bae & Luck, 2018, 2019b; Bocincova & Johnson, 2019; Cai et al., 2019; Christophe et al., 2012; Ester et al., 2013, 2015; Gosseries et al., 2018; Harrison & Tong, 2009; Rose et al., 2016; Serences et al., 2009; Sprague et al., 2016; Wolff et al., 2015, 2017; Xing et al., 2013)。

1.4 表征相似性分析及其相关研究

表征相似性分析（representational similarity analysis, RSA）是一种有效地通过比较不同条件之间表征差异来构建跨模态、跨物种之间表征模式相似性的方法(Kriegeskorte, Mur, & Bandettini, 2008)，其被迅速用于探究多种认知功能，包括知觉、记忆、语言和决策等等。尽管在脑科学领域各种神经记录的方法与技术快速发展，如无创的神经活动记录方法：EEG、MEG、fMRI 和功能性近红外成像（functional near-infrared spectroscopy,,

fNIRS) 以及一些有创的记录病人、猴子以及其他物种神经活动的方法: ECoG、立体脑电 (stereo-electro-encephalography, sEEG) 和一些其他电生理方法。然而, 对于跨模态、甚至跨物种结果的理解是十分困难的。RSA 方法则使用表征不相似性矩阵 (representational dissimilarity matrix, RDM) 提供了一种共同的表征空间以建立了跨模态、跨物种比较的桥梁。一个 RDM 反应的是同一模态下, 不同任务条件之间的表征(不)相似性。如图 1-4 所示(Popal et al., 2019), 不同模态以及不同物种的数据都可以投射到一个共同的表征空间上构建对应的 RDMs。之后, 再进行基于 RDMs 的比较来得到跨模态或是跨物种的表征相似性结果。例如, 一些研究尝试结合 fMRI 结果与电生理结果 (Kriegeskorte, et al., 2008)、结合 MEG 结果与电生理结果(Cichy et al., 2014)、结合 MEG 结果与 fMRI 结果(Cichy, Pantazis, et al., 2016)、结合 EEG 结果与 fMRI 结果(Muukkonen et al., 2020)或是结合行为学结果与 fMRI 结果(Bainbridge et al., 2017; Bainbridge & Rissman, 2018; Wang et al., 2018)。

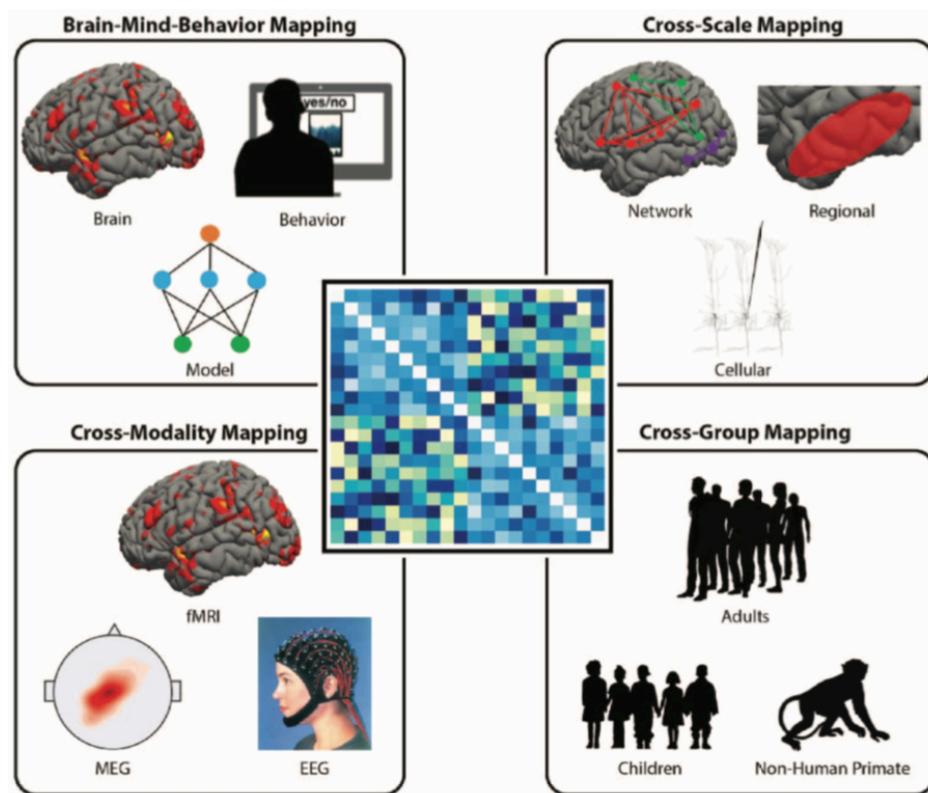


图 1-4 RSA 通过一个共同表征空间来连接不同模态的数据(Popal et al., 2019)

除了作为一种比较不同模态间表征差异的工具, RSA 也成为一个有效的方式去追踪多种任务间的多维表征。一方面, 研究者可以基于一些理论假设构建一些编码 RDMs, 然后用基于理论的编码 RDMs 和实际测得的神经活动得到的神经 RDMs 计算两者之间

的相似性程度(Alfred et al., 2018; Avery et al., 2021; Bainbridge et al., 2017, 2021; Bainbridge & Rissman, 2018; Cichy et al., 2014; Cichy, Pantazis, et al., 2016; Etzel et al., 2020; Feng et al., 2018; Hall-McMaster et al., 2019; Muukkonen et al., 2020; Yokoi & Diedrichsen, 2019)。基于此方案，可以推断出大脑对不同信息的加工模式、加工脑区与加工时间等等信息。另一方面，随着人工智能（artificial intelligence, AI）快速发展，RSA 可以被用来比较人工神经网络（artificial neural networks, ANN）和大脑活动之间的表征(Cichy, Khosla, et al., 2016; Güçlü & van Gerven, 2015; Kuzovkin et al., 2018; Rahmani Del Bakhshayesh et al., 2018; Urgen et al., 2019; Xie et al., 2020; Yamins et al., 2014)。因此，RSA 是一个有用的工具来连接多模态的神经数据、来深入理解大脑对不同信息的编码模式与加工过程。同时，它帮助建立了一个研究大脑与 AI 的更清晰的方式。

1.5 比较人脑与深度卷积神经网络模型

最近，大量的工作开始关注于比较生物大脑与人工神经网络模型（artificial neural network, ANN）之间对信息内在的处理机制的差异，这类工作主要集中在客体识别领域。大脑通过大量神经元形成的大尺度网络来实现客体识别能力，而深度卷积神经网络（deep convolutional neural networks, DCNNs）则是基于类似的想法构建的人工模型来完成这类任务。深度卷积神经网络已经逐渐成为计算机视觉领域最重要的模型，包括对于客体识别的 DCNN 模型(He et al., 2016; Krizhevsky et al., 2012; LeCun et al., 1998; Simonyan & Zisserman, 2015)、对面孔识别的 DCNN 模型(Hu et al., 2016; Li et al., 2015; Ranjan et al., 2017; Schroff et al., 2015; Taigman et al., 2014)、对音色识别的 DCNN 模型(Kell et al., 2018)、对表情识别的 DCNN 模型(Matsugu et al., 2003; Yu & Zhang, 2015)等等，在很多领域 DCNN 都达到了人类表现甚至在准确率上高于人类的识别能力。人脑作为一个超级智能“计算机”，而 DCNN 做为工程上的 AI 代表，尤其是其具有的分层结构，让比较人脑与 DCNN 的表征差异具有巨大的意义。

一方面，通过比较 DCNN 与人脑的一致性表征能一定程度上了解 DCNN 对信息是如何进行分层加工的。由于 DCNN 是一个工程学上的分类或回归模型，其内部的表征方式仍是一个黑盒子。人脑与 DCNN 都能在很多任务上达到类似的能力，越来越多的认知神经科学家与计算机科学家开始探究这两者在对信息加工方式上的异同。尤其是在视觉客体识别领域，结合人脑与 DCNNs 的研究已经发现了视觉腹侧通路与 DCNNs 的分层结构存在着类似的对视觉信息的加工模式(Cichy, et al., 2016; Güçlü & van Gerven, 2015;

Kietzmann et al., 2019; Yamins & DiCarlo, 2016), 如图 1-5 所示。

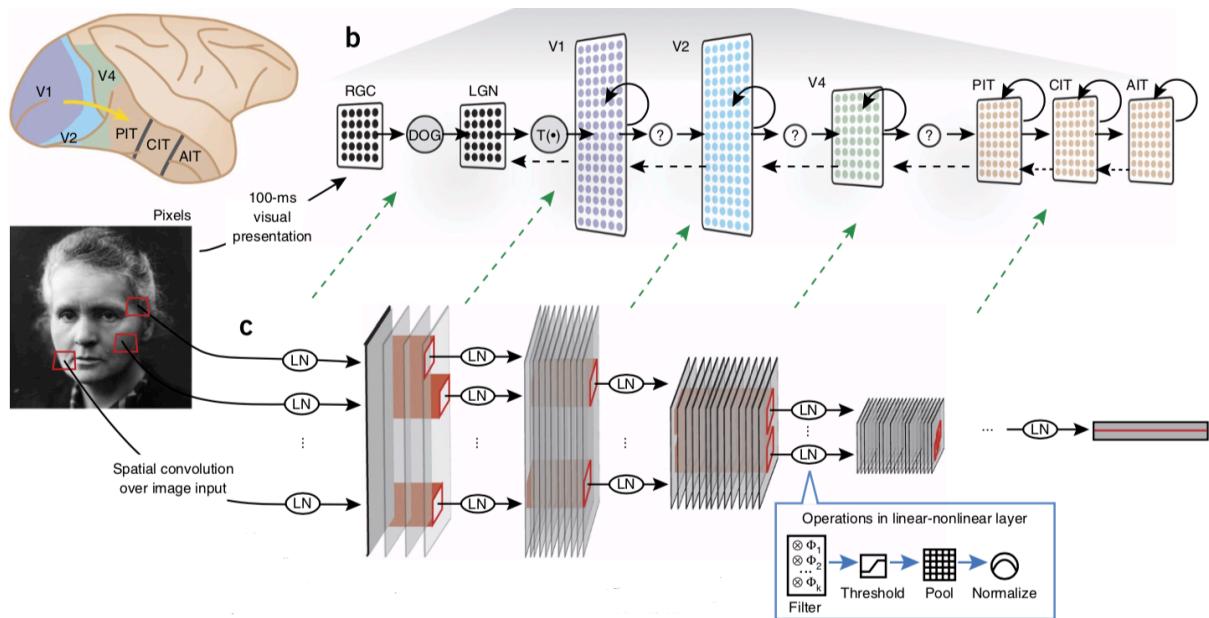


图 1-5 分层卷积神经网络作为知觉皮层的模型(Yamins & DiCarlo, 2016)

另一方面, DCNN 可以作为一个模型工具来探究哪种可能的机制与人脑的表征更一致。由于 DCNN 的早期层加工更低维的特征, 而晚期层加工更高维的特征, 可以借此来探究人脑在一个认知过程中对信息的编码情况是如何的。以及, 通过对 DCNN 模型来进行一些修改排除掉人类实验中难以排除的变量影响或是对 DCNN 的激活进行某些基于假设的修改, 进而比较 DCNN 的表征模式与神经活动的关系以探究人脑在某一认知过程中的神经机制(Jaegle et al., 2019; Xie et al., 2020; Xu et al., 2020)。

主流的用于比较大脑与 AI 的方法是建立编码模型和基于表征相似性分析两种, 前者通过构建深度卷积神经网络中激活值到神经活动信号的映射关系(Bashivan et al., 2019; Güçlü & van Gerven, 2015; Kubilius et al., 2016; Ponce et al., 2019), 后者通过表征相似性分析的方法构建两种模态下信息编码的表征不相似矩阵来比较相互间的表征差异(Bankson et al., 2018; Greene & Hansen, 2018; Kalfas et al., 2018; Kietzmann et al., 2019; Kuzovkin et al., 2018; Urgen et al., 2019; Xie et al., 2020)。在 AI 的面孔识别方面, 代表性的人脸识别系统包括 Deep Face(Taigman et al., 2014)、VGG face(Parkhi et al., 2015)、FaceNet(Schroff et al., 2015)和基于 ResNet101 的 L2 softmax 模型(Ranjan et al., 2017), 它们的识别能力已经达到了人类的水平(Phillips et al., 2018)。也有相关研究在进一步探究这些用于面孔识别的 DCNN 模型对面孔信息的空间表征(Hancock et al., 1996; Huang et al., 2012; Pande et al., 2017; Simonyan et al., 2013; Sun et al., 2014), 这些探究主要集中在

DCNN 模型对面孔身份信息的加工方式上。而在面孔识别或面孔感知领域，时序上比较人脑对面孔的加工过程与基于面孔识别的 DCNNs 之间表征差异的研究还很少(Dobs et al., 2019)，仍缺乏对其内部表征模式的深入探讨。

1.6 研究目的与意义

深入探究视觉面孔处理的动态过程有助于理解面孔感知中对不同面孔信息的加工模式及其对应神经机制，而传统的 ERP 或时频分析的研究往往只能给出一些时域上的指标却无法直接追踪不同面孔信息在大脑中时序上的编码过程，本研究希望使用神经解码以及表征相似性分析的方法来探究人脑在面孔感知过程中对面孔熟悉度、面孔完整性等等面孔信息的动态编码过程，以及在时序上追踪重复抑制效应的作用过程。

由于面孔感知是在生态和进化上是跨物种相关的(Charles & Sergent, 1992; W. Freiwald et al., 2016; Tsao & Livingstone, 2008; Weiner & Grill-Spector, 2015)，一个值得思考的问题是生物脑与从工程角度设计的非生物人工智能模型在对面孔的感觉加工过程中具有哪些表征差异？

一方面，需要先了解不同的面孔信息是如何在 DCNN 模型中进行表征的，包括正常面孔与乱相面孔、熟悉面孔与不熟悉面孔在 DCNN 模型中逐层的表征差异，以及这些面孔信息在经过预训练的模型和未经过预训练的模型之间的表征差异等等。

另一方面，对 DCNN 模型的激活进行基于重复抑制效应的模拟，进而比较人脑与 DCNN 模型之间的表征差异。由于 DCNN 模型不具有模拟时序信息编码的能力，本研究设定了衰减模型与锐化模型两种模型对 DCNNs 进行修改。通过比较真实神经活动与 DCNN 模型的激活之间的表征相似性来探究重复抑制背后可能存在的神经机制以及生物脑与 AI 脑之间在面孔感知过程中的相似表征模式。

本研究共分为四个部分：第一部分设计与实现了用于神经数据表征分析的工具包，为后续对神经活动与 DCNN 激活的表征模式的深入探究提供技术基础；第二部分着重基于 EEG 数据探究不同面孔信息在人脑中的动态加工过程；第三部分着重探究不同面孔信息在 DCNN 模型中是如何进行逐层编码的；第四部分基于两种重复抑制模型对 DCNN 模型的激活进行修改并与人脑神经活动进行比较，以深入探究重复抑制的神经机制以及人脑与 DCNN 模型之间跨模态的表征相似性。

本研究深入剖析了生物脑与 AI 脑之间在面孔感知过程中对面孔信息的编码异同，既探究了面孔感知过程中的脑机制，也为更清晰地理解了 DCNN 的内部加工方式，结合跨模态的比较，为之后神经科学领域与类脑智能领域的研究提供了新的思路。

2 用于神经数据表征分析的工具包

2.1 NeuroRA：一个用于多模态神经数据表征分析的 Python 工具包

2.1.1 NeuroRA 概述

NeuroRA 是一个十分易于使用的、用于多模态神经数据表征分析的 Python 工具包，用户可以使用 NeuroRA 来探究不同任务条件与不同模态下的神经表征及它们之间的表征差异。NeuroRA 的结构与功能总览如图 2-1 所示，其可以对几乎所有类型的神经数据（包括 EEG、MEG、fNIRS、fMRI 和其他电生理数据）和行为学数据进行分析。

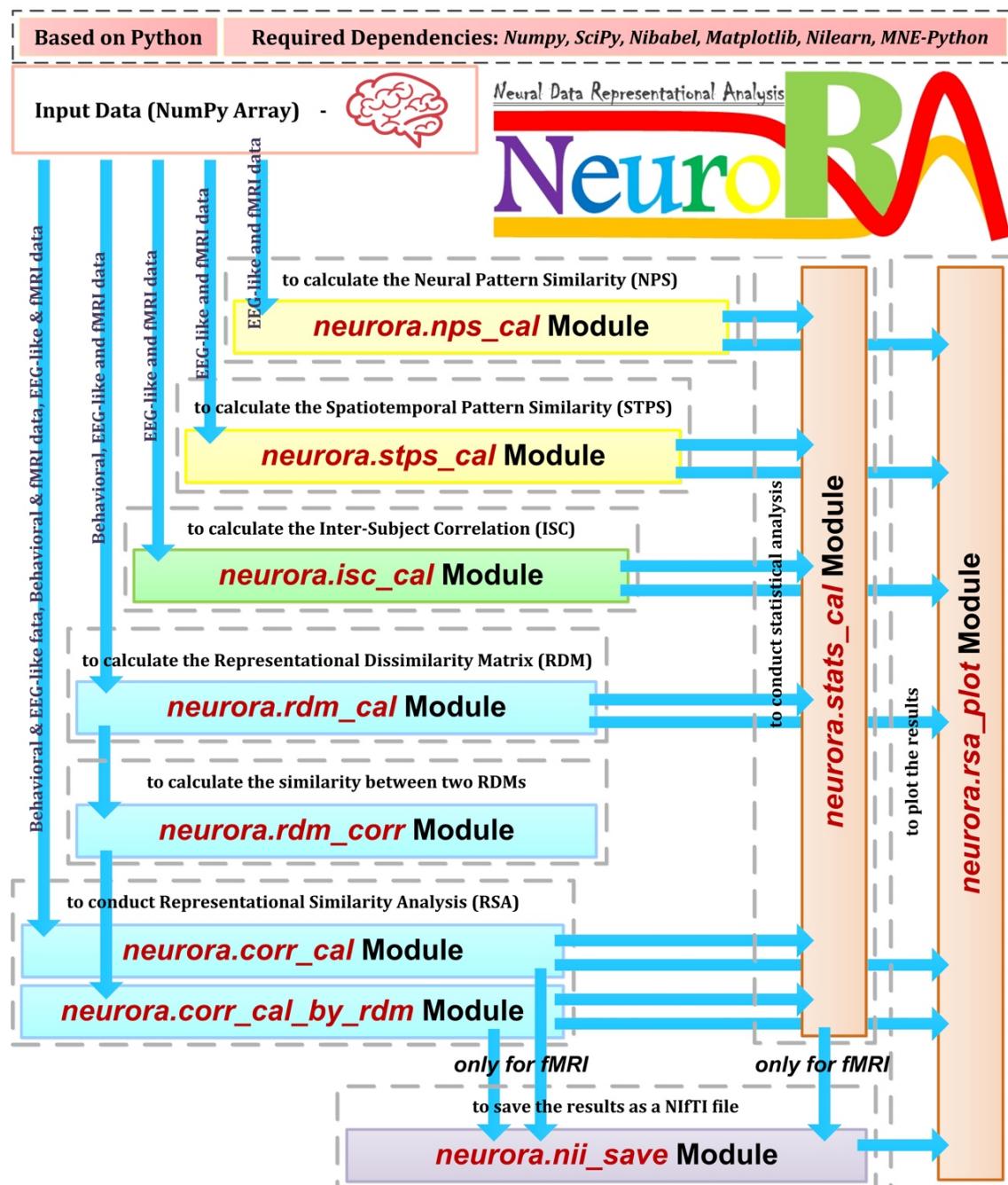


图 2-1 NeuroRA 总览

NeuroRA 是一个基于 Python 的工具包并且需要一些依赖包，包括 NumPy、SciPy、Matplotlib、Nilearn 和 MNE-Python。它包含一些主要部分：计算神经模式相似性（NPS）、时空模式相似性（STPS）、被试间相关（ISC）和表征不相似性矩阵（RDM），使用 RDMs 比较不同模态的表征差异，统计分析，将核磁数据的分析结果存为一个 NIfTI 文件，以及绘制结果。每一个计算部分对应 1-2 个模块。蓝色箭头说明可能的数据流向。

NeuroRA 提供了丰富的分析功能。（1）NPS（neural pattern similarity，神经模式相似性）模块用于计算两个不同条件下大脑活动的相关性(Cavanagh et al., 2018; J. V. Haxby et al., 2001)；（2）STPS（spatiotemporal pattern similarity，时空模式相似性）模块用于计算跨不同空间与不同时间的表征相似性(Lu et al., 2015; Xue et al., 2010)；（3）ISC（inter-subject correlation，被试间相关）模块用于计算同种条件下不同被试间大脑活动的一致性(Hasson et al., 2004)；（4）RDM 模块用于计算反应某一给定模态的神经数据在不同条件或刺激间表征（不）相似性的矩阵；（5）NeuroRA 提供了比较不同模态下 RDMs 的跨模态相关性分析方法，并且这一步骤提供了不同的参数来进行特定形式的计算，例如可以对每一个被试、每一个导联、每一个时间点单独进行计算或整合这些进行整体计算。

除了上述的计算功能，NeuroRA 也提供了一个统计模块来进行统计分析，以及提供了一个可视化模块来对计算结果进行绘图，例如绘制 RDMs、绘制时序的表征相似性结果和核磁数据的 RSA 结果等等。此外，NeuroRA 提供了一个独一无二的方法直接将基于核磁数据的 RSA 结果存为 NIfTI 格式的文件。

NeuroRA 所需要的依赖包包括 NumPy(Van Der Walt et al., 2011)、SciPy(Virtanen et al., 2020)、Matplotlib(Hunter, 2007)、Nibabel (<https://nipy.org/nibabel/>)、Nilearn (<https://nilearn.github.io/>) 和 MNE-Python(Gramfort et al., 2013, 2014)，用户在安装 NeuroRA 时会自动检测电脑环境中是否包含这些依赖包并进行自动下载安装。NumPy 帮助进行基于矩阵的计算，SciPy 帮助进行基本的统计分析，Matplotlib 和 Nilearn 被调用帮助实现一些画图功能，而 MNE-Python 被用来加载一些用户示例中的 MEG 数据。通过使用基于 Python 的高效计算工具包，NeuroRA 让用户可以直接且高效地进行神经数据的挖掘。用户只需要一行命令即可下载安装 NeuroRA：pip install neurora。工具包的官方网址是：<https://neurora.github.io/NeuroRA/>，其在线文档网址是：<https://neurora.github.io/documentation/>。此外，其源码可见 GitHub 的项目网址：<https://github.com/neurora/NeuroRA>。

2.1.2 模块与功能

NeuroRA 包含以下核心模块，并且将在未来补充更多必要的功能：

`nps_cal`: 一个基于神经数据计算神经模式相似性的模块；

`stps_cal`: 一个基于神经数据计算时空模式相似性的模块；

`isc_cal`: 一个基于神经数据计算被试间相关的模块；

`rdm_cal`: 一个基于多模态神经数据计算 RDMs 的模块；

`rdm_corr`: 一个基于不同算法（包括 Pearson 相关、Spearman 相关、Kendall's tau 相关、余弦相似度和欧式距离）计算两个 RDMs 相关系数的模块；

`corr_cal_by_rdm`: 一个计算不同模态下 RDMs 的表征相似性的模块；

`corr_cal`: 一个“一步”直接进行两不同模态下数据 RSA 的模块；

`niit_save`: 一个将核磁的表征分析结果存为一个.nii 文件的模块；

`stats_cal`: 一个计算统计结果的模块；

`rsa_plot`: 一个绘制表征分析结果的模块。它包含绘制 RDM、绘制基于脑电或类似脑电的时序数据（如脑磁数据）的表征分析的曲线图和热力图、绘制基于核磁数据的表征分析的结果（包括大脑切面图和大脑表面图）。

2.1.2.1 使用 NeuroRA 进行表征相似性分析

在 NeuroRA 中，首先可以通过输入同一模态的不同条件下的数据，计算任意两条件间的不相似性得到一个 RDM，不相似性的指标可以通过参数进行用户的自定义选择，包括计算欧式距离、马氏距离和 1-Pearson 相关系数。此外，在某些情况下研究者需要对每个被试、每个导联或每个时间点进行独立的 RDM 计算。因此，NeuroRA 提供给了用户一些可选参数来进行多 RDMs 的批量计算（如图 2-2 所示）。用户根据自身需求改变对应的计算参数，即可以对脑电或类似脑电的数据进行逐被试、逐导联或逐时间点的计算，以及可以对核磁数据进行全脑搜索式的计算或对特定的感兴趣脑区进行计算。

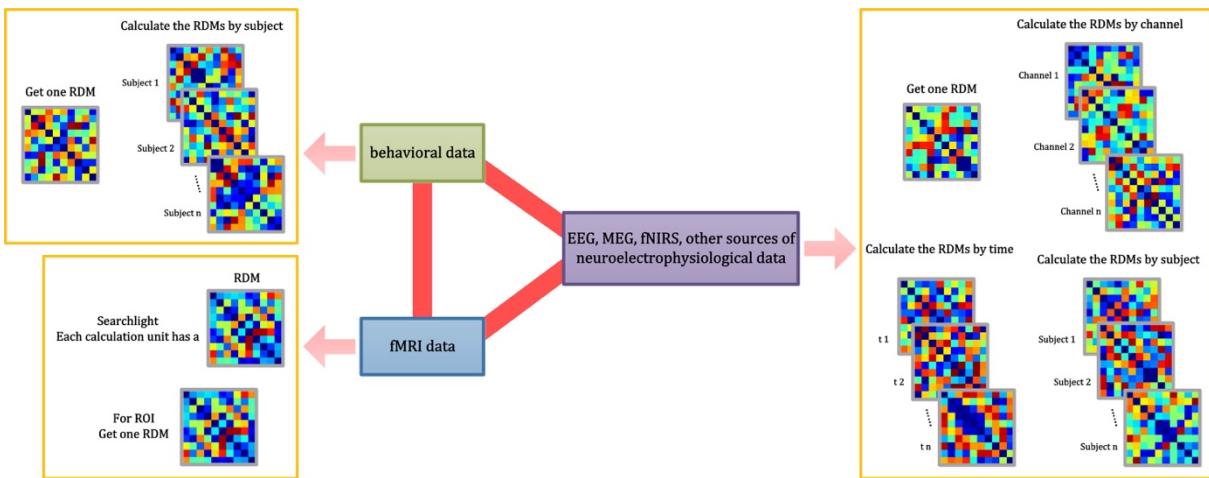


图 2-2 NeuroRA 中计算 RDM 的实现

NeuroRA 具有使用不同模态的数据计算 RDM(s)的能力，红线代表的是两种模态之间可进行跨模态计算，而粉色箭头对应的是该模态可以选择的计算方式。

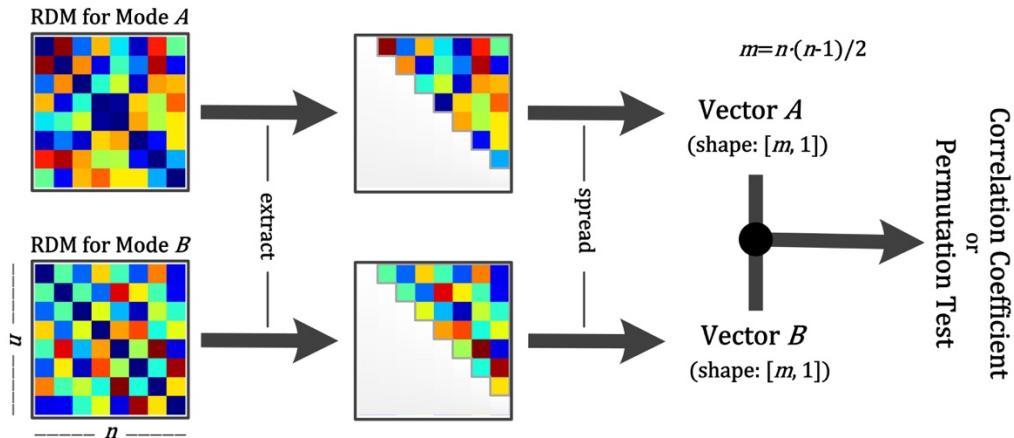


图 2-3 两 RDMs 间计算的示意图

进一步，对于 RDM 之间的比较也可以分为对两个 RDMs 进行相关性分析和对多个 RDMs 进行分析。在 NeuroRA 中，比较两个 RDMs 的相关性（相似性）时，由于 RDM 是一个对角对称的矩阵，需要先提取出矩阵中对角线上半部分（或下半部分）的值，进行平铺为向量，在进行两向量间的比较，包括计算相关系数（目前 NeuroRA 提供了计算 Pearson 相关系数、Spearman 相关系数和 Kendall's tau 相关系数的方法）、余弦相似度、欧式距离以及置换检验（如图 2-3 所示）。在此基础上，NeuroRA 提供了对于多个 RDMs 的计算方案，用户只需要输入两模态下的多个 RDMs，即可获得多个 RDMs 之间进行比较后的结果。更进一步，由于很多情况下用户不需要获得中间过程的 RDMs，NeuroRA 也提供了直接输入两种模态的数据、输出最终跨模态比较后的“一步到位”的方法。因此，用户可以调用 corr_cal 和 corr_cal_by_rdm 两种模块下的函数进行基于数据本身和基于 RDMs 的跨模态表征相似性分析。

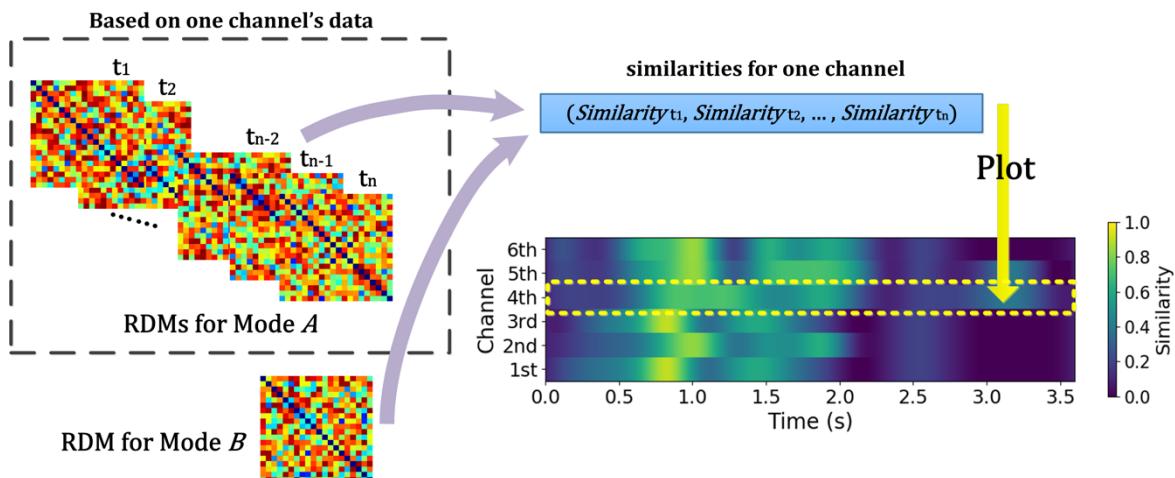


图 2-4 进行基于脑电或类似脑电数据的跨时间和导联的 RDMs 的相似性分析的示意图

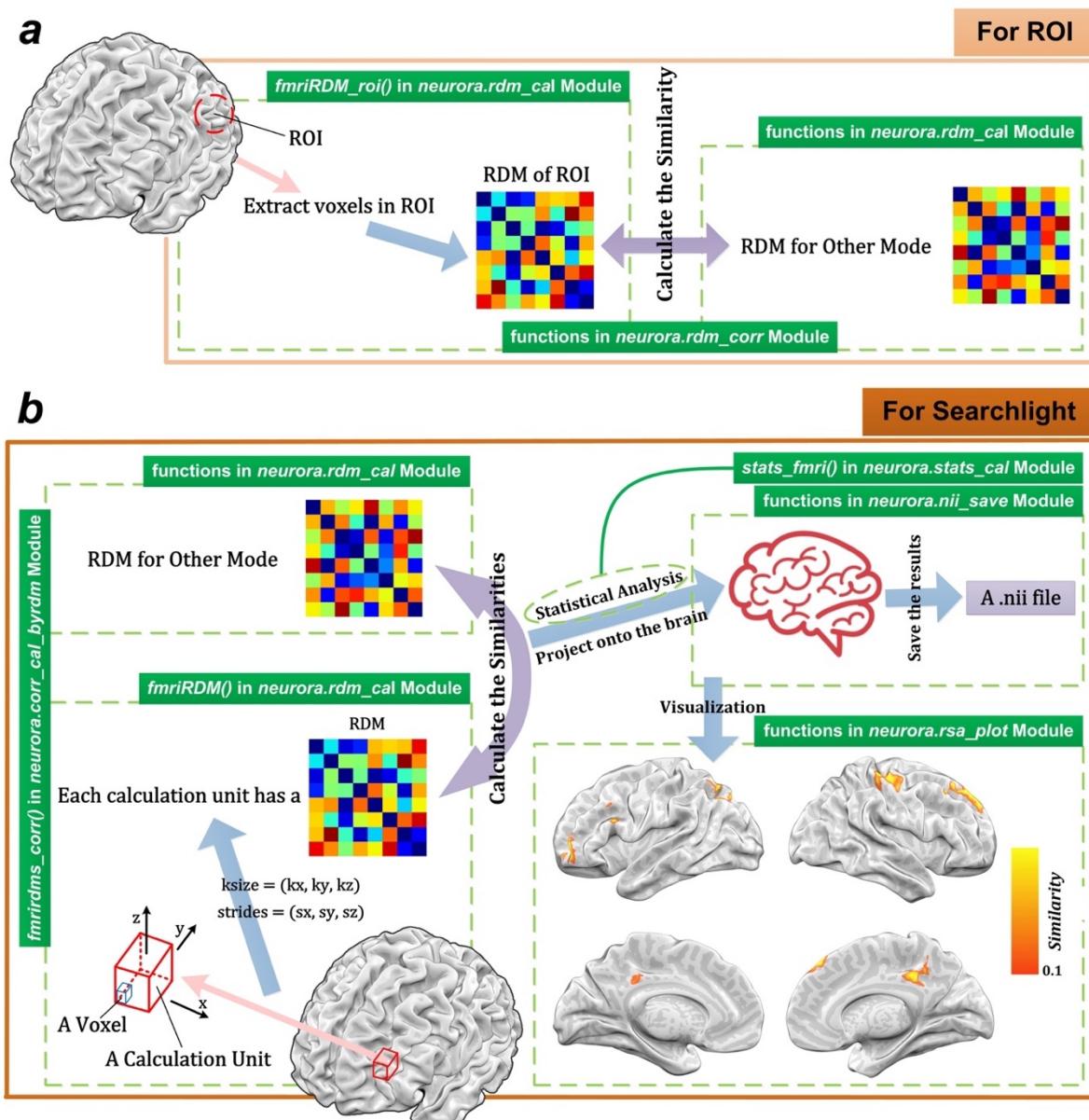


图 2-5 使用 NeuroRA 进行核磁数据的表征分析的示意图

(A) 基于感兴趣脑区的分析计算；对于每一个感兴趣的脑区，用户可以计算基于对应脑区的体素的 RDM 并计算该 RDM 与其他模态得到的 RDM 之间的相似性；(B) 全脑搜索式分析计算；对每一次搜索步骤，用户定义计算单元的大小与步长，每一次搜索对应的计算单元都会计算一个对应的 RDM，再与其他模态的 RDM 计算相似性计算，全脑搜索式计算后可以得到一个包含全脑相似性信息的 NIfTI 文件，右下的图片是对 NIfTI 文件的可视化结果（通过 NeuroELF 软件绘制，软件地址：<http://neuroelf.net>），有颜色的区域代表了两模态间的相似性显著，颜色代表相似性强度。图中绿色标签说明了对应的计算过程由 NeuroRA 中的哪一模块下的哪一函数实现。

以脑电数据为例，通过每一个时间点的某一导联的脑电数据可以构建一个对应的 RDM，同样通过行为学数据也可以构建一个 RDM。通过计算行为学 RDM 和脑电对应导联的时序 RDMs 之间的相似性可以得到一个相似性曲线，而对每一个导联的数据都进行同样的运算操作，则可以得到一个时间×导联的相似性热力图（如图 2-4 所示）。通过这种方式，就可以获得与行为表现相关的神经表征所对应的导联与时间段。而对于核磁数据而言，NeuroRA 提供了基于感兴趣脑区的计算（ROI-based computation）和全脑搜索式的计算(searchlight-based computation)，见图 2-5。

除了 RSA 外，用户也可以使用 NeuroRA 进行一些其他表征分析，如神经模式相似性、时空模式相似性和被试内相关，其具体用法和 RSA 部分类似。同时，NeuroRA 提供了独立的统计分析功能，支持对于上述各种表征分析方法的结果的统计分析。用户只需要输入每一个被试的相似性矩阵，就可以获得对应的统计结果（一个包含 t 值和 p 值的矩阵）。并且，NeuroRA 允许用户通过改变统计函数中的参数来选择是否进行置换检验。

更重要的是，NeuroRA 提供了一系列的对结果可视化的函数。一些典型的绘图功能见图 2-6 所示。在用于结果可视化的模块中，用户可以找到绘制 RDM、脑电或类似脑电的时序多导联数据的分析结果（包括曲线图、热力图等等）、核磁数据的分析结果（包括大脑二维切面图、三维表面图等等）。

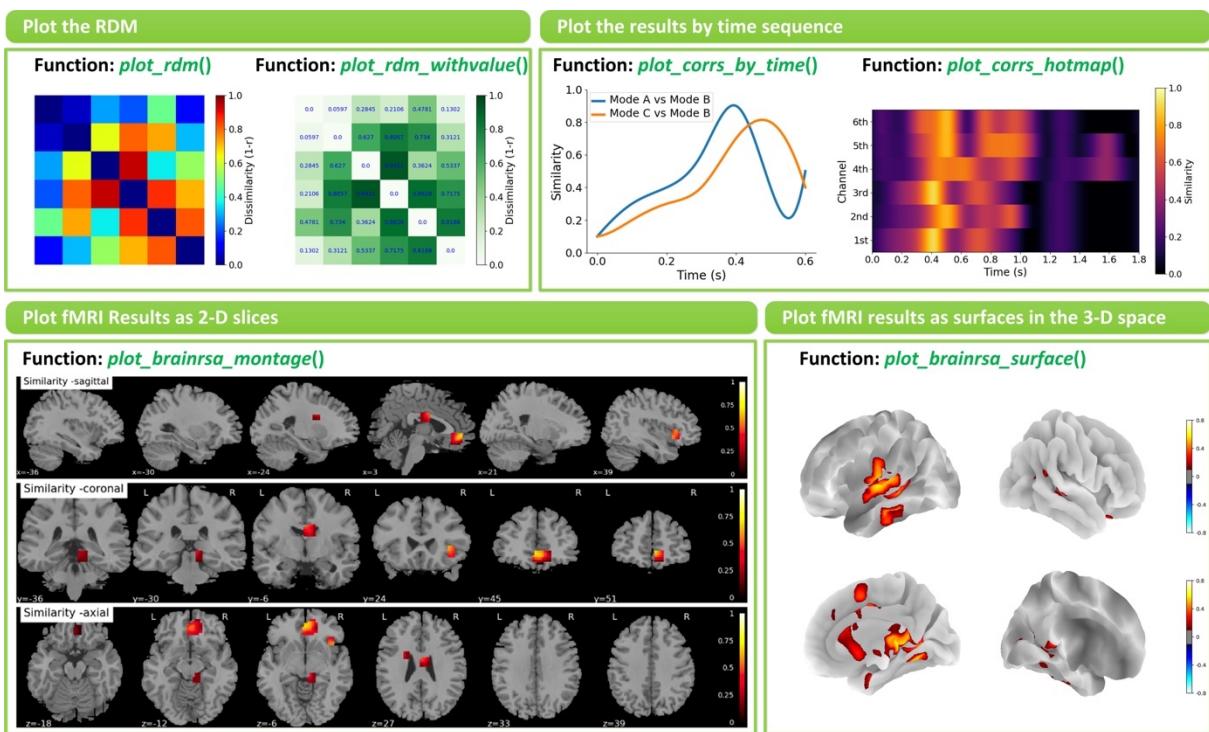


图 2-6 通过 NeuroRA 实现的可视化案例

左上：使用 `plot_rdm()` 和 `plot_rdm_withvalue()` 两函数分别绘制的同一 RDM；右上：通过 `plot_corrs_by_time()` 和 `plot_corrs_by_hotmap()` 两函数绘制的时序结果；左下：通过 `plot_brainrsa_montage()` 函数绘制的二维大脑切片结果；右下：通过 `plot_brainrsa_surface()` 函数绘制的三维空间的大脑表面结果。

2.1.3 使用示例

目前，NeuroRA 提供了几个基于公共数据集的代码示例来指导用户如何使用 NeuroRA 来处理脑电数据（和类似脑电数据）及核磁数据。

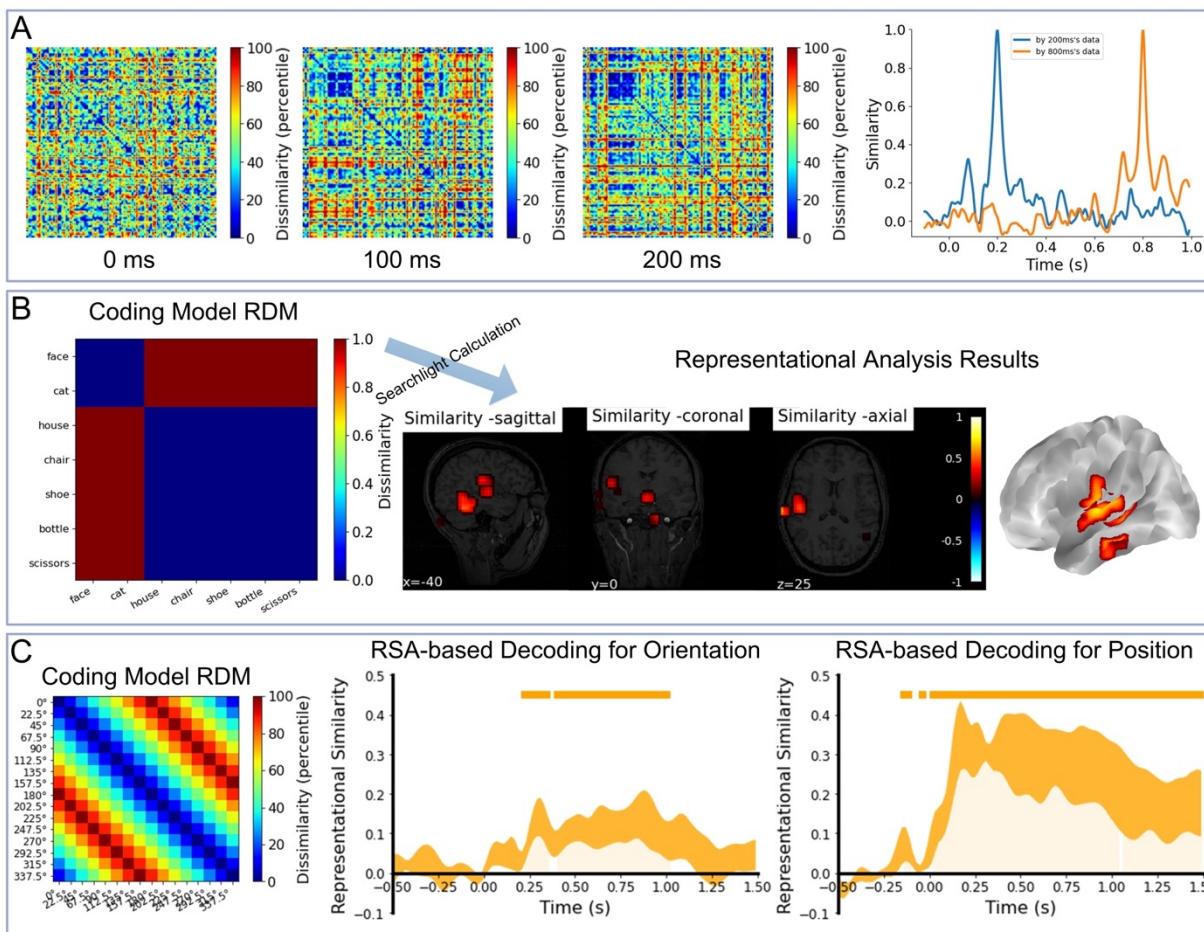


图 2-7 示例结果

(A)左侧: 基于前三个被试的全部 302 个被试的脑磁数据构建的 0ms、100ms 和 200ms 时的 RDMs; 右侧: 使用 200ms 和 800ms 的神经表征与所有时间点的神经表征计算相似性的结果。(B) 使用一个“生命体-非生命体”编码模型 RDM 与被试 2 的全脑 RDMs 进行全脑搜索式计算结果。在该编码模型 RDM 中, 假设生命客体与非生命客体内部表征一致。(C) 通过计算一个编码模型 RDM 与实验 2 中前五个被试的时序脑电 RDMs 之间的相关系数得到基于 RSA 的解码结果。在该编码模型 RDM 中, 假设两角度差异越大其对应神经活动的表征不相似性越大, 反之亦然。右侧两张图中, 小橙色的方块和浅橙色阴影代表显著时间段 ($p < 0.05$), 线的宽度反映的是加减一个标准误。

第一个例子基于视觉 92 分类任务的脑磁图数据(Cichy et al., 2014), 该示例中仅使用前三个被试的数据进行计算与演示。图 2-7a 显示了使用 NeuroRA 计算三个不同时间点下基于相关得到的 RDMs 以及时序上的神经表征分别和 200ms 和 800ms 的神经表征计算得到的时间上的相似性结果。这一示例的完整代码可见:

https://github.com/neurora/NeuroRA/blob/master/demo/NeuroRA_Demo1.ipynb。

第二个例子基于 Haxby 核磁数据(J. V. Haxby et al., 2001), 该示例中仅使用被试编号为 2 的被试数据进行计算与演示。图 2-7b 显示了使用一个“生命体-非生命体”编码模型

RDM 和基于核磁全脑搜索得到的 RDMs 进行 RSA 的结果，结果表明大脑颞叶皮层主要参与了生命体或非生命体信息的编码。这一示例的完整代码可见：

[https://github.com/neurora/NeuroRA/blob/master/demo/NeuroRA_Demo2.ipynb。](https://github.com/neurora/NeuroRA/blob/master/demo/NeuroRA_Demo2.ipynb)

第三个例子基于一个视觉工作记忆任务的脑电数据(Bae & Luck, 2018)，该示例中仅使用原文实验 2 中前五位被试的 ERP 数据。图 2-7c 显示了使用一个编码模型 RDM 和基于脑电数据得到的时序 RDMs 之间的相似性结果，结果表明在该视觉工作记忆任务中方向和位置信息都可以通过 ERP 数据解码成功出来。这一示例的完整代码可见：

[https://github.com/neurora/NeuroRA/blob/master/demo/NeuroRA_Demo3.ipynb。](https://github.com/neurora/NeuroRA/blob/master/demo/NeuroRA_Demo3.ipynb)

关于每一个模块、每一个函数的更详细的信息，包括数据输入类型、参数的选择以及数据输出类型，用户都可以参阅 NeuroRA 的教程文档或在线文档。

2.2 PyCTRSA：一个基于跨时域表征相似性分析的脑电/脑磁数据解码的 Python 工具包

2.2.1 跨时域表征相似性分析

在传统的 RSA 中，通常是使用一个基于假设的编码模型 RDM 或者一个由行为学数据得到的 RDM 来拟合由时间连续的神经数据得到的 RDMs。通过这一方法，可以在时序上解码出大脑对信息的编码情况或是大脑对行为学的影响。然而，传统的 RSA 方法无法进行跨时域的解码。在基于分类的解码方法中，跨时域的解码通过利用时间 i 的数据训练一个分类器来测试时间 j 的数据来实现。之所以要进行跨时域的分析是因为即使时间 i 和时间 j 都对同一信息进行了编码，但无法确定这两个时间点对该信息的编码模式是一样的。使用基于分类的解码方法，如果时间 i 的分类器能成功的分类时间 j 的数据则表明两时间对应的神经编码模式类似，若不能成功分类则表明虽然两时间都编码了该信息但大脑对其的编码模式是不同的。

基于跨时域表征相似性分析（cross-temporal representational similarity analysis, CTRSA）的解码方法是一种全新的基于 RSA 的对脑电/脑磁数据进行跨时域解码的算法。在很多神经解码过程中，往往并不能用分类的方法验证编码假设。因此 CTRSA 就是在传统 RSA 的基础上进行的一个跨时域的方法拓展。在 CTRSA 解码过程中，首先需要使用时间 i 和时间 j 的神经数据来构建一个跨时域表征不相似性矩阵（cross-temporal representational dissimilarity matrix, CTRDM）。沿着这个思路，可以计算得到 $N_{time} \times N_{time}$ 个 CTRDMs (N_{time} 代表分析用的神经数据的时间点数量)。然后类似传统 RSA 的思路，通过实验假设可以构建一个编码模型 RDM。最后用这个编码模型 RDM 和 $N_{time} \times N_{time}$ 个

CTRDMs 计算相似性从而得到跨时域解码结果。

2.2.2 PyCTRSA 概述

PyCTRSA 是一个全新的用于实现上述跨时域表征相似性分析的 Python 工具包。与 NeuroRA 一样，PyCTRSA 免费、开源、易于使用并且功能全面。心理学家和神经科学家可以使用 PyCTRSA 对脑电和脑磁数据进行基于 CTRSA 的解码分析。

PyCTRSA 所需要的依赖包包括 NumPy、SciPy、Matplotlib 和 NeuroRA，用户只需要一行命令即可下载安装：pip install pyctrsa。其源码可以参考 GitHub 的 PyCTRSA 项目网址：<https://github.com/ZitongLu1996/PyCTRSA>。通过使用 PyCTRSA，用户不仅可以基于这种全新的 CTRSA 方法计算跨时域的相似性来实现解码，也可以计算不同条件下神经数据的跨时域相似性来比较条件间的模式差异，同时也可以基于传统的时序 RDMs 来计算任意两时间点的两 RDMs 之间的表征模式的相似性进而得到跨时域相似性结果。

2.2.3 模块与功能

PyCTRSA 旨在提供易于理解的功能来通过很少的代码实现全新的基于 CTRSA 的跨时域解码方法，其包含以下四大主要的功能并对应四个独立的模块：

(1) 计算 CTRDM:

计算 CTRDM 的功能在 ctrdm 模块内，PyCTRSA 提供了对单个 CTRDM 计算的模块 ctrdm.single_cal 和对多个 CTRDM 进行批计算的模块 ctrdm.multi_cal。用户使用 PyCTRSA 可以对单个导联进行计算、对单个被试进行计算也可以对多导联多被试的数据计算得到 CTRDMs。

(2) 计算两 CTRDMs 之间的相似性:

计算两 CTRDMs 之间相似性的功能在 similarity 模块内，PyCTRSA 提供了类似 NeuroRA 中对两 RDMs 进行相似性计算的类似功能，用户可以选用 Pearson 相关、Spearman 相关、Kendall's tau 相关、余弦相似度和欧氏距离的方式来计算相似性。由于 CTRDMs 不像传统的 RDM 是一个对角对称的矩阵，在进行两 CTRDMs 比较时需要提取除对角线外的所有值平铺为向量再进行计算。

(3) 计算跨时域相似性:

计算跨时域相似性的功能在 ctsimilarity 模块内，PyCTRSA 提供了三种计算跨时域相似性的方案：使用 ctsimilarity.normaldatabased 模块可以计算两条件下神经数据的跨时域相似性；使用 ctsimilarity_normalrdmbased 模块可以基于传统的 RDMs 计算跨时域相

似性；使用 `fitctrdm` 模块可以计算一个编码模型 RDM 与 CTRDMs 之间的跨时域相似性。

(4) 绘制 CTRSA 结果：

绘制 CTRSA 结果的功能在 `plotting` 模块内，PyCTRSA 提供了对 CTRDM 的可视化模块 `plotting.ctrdm`、对时序相似性结果的可视化模块 `plotting.tbytsimilarities` 以及对跨时域相似性结果的可视化模块 `plotting.ctsimilarities`。

2.2.4 使用示例

目前，PyCTRSA 提供了一个基于公开的脑电数据(Bae & Luck, 2018)的示例，数据来源同 NeuroRA 的示例 3。该示例提供了完整的计算流程，包括下载示例数据、数据预处理、计算与绘制 CTRDMs 以及计算和绘制跨时域相似性结果共四个部分，来帮助用户学习与理解如何使用 PyCTRSA。图 2-8 显示了 PyCTRSA 这个示例中对视觉朝向和位置信息的跨时域解码结果。

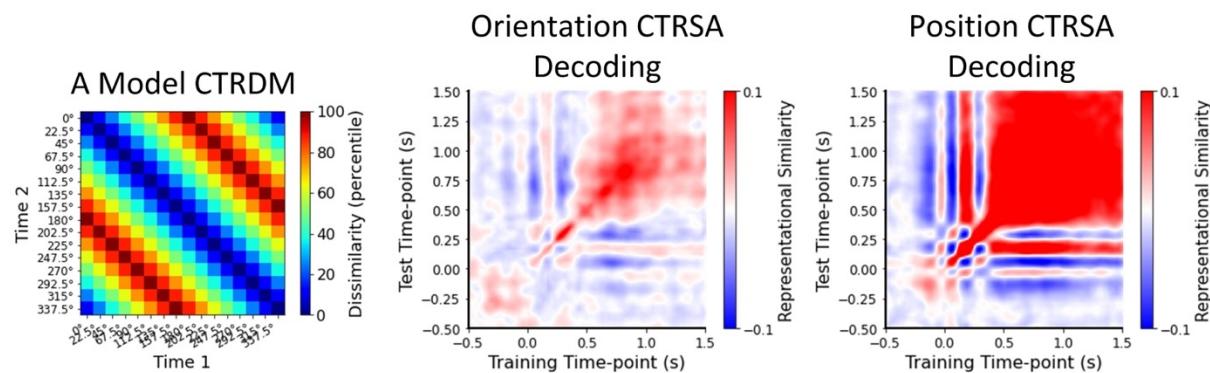


图 2-8 PyCTRSA 示例结果

左：一个基于假设的编码模型 CTRDM；中：基于 CTRSA 的朝向解码结果；右：基于 CTRSA 的位置解码结果。

PyCTRSA 提供了在线的教程与可直接演示的代码示例，同时基于同一批脑电数据，提供了基于分类的跨时域解码与基于 CTRSA 的跨时域解码的代码实现与结果。这些内容均可在以下地址获取：<https://github.com/ZitongLu1996/PyCTRSA/tree/master/docs>。

3 探究面孔信息在人脑中的动态表征

3.1 数据与实验

本研究所用数据来自 OpenNeuro 平台 (www.openneuro.org) 中一个面孔处理的 EEG 数据集，其属于一个多被试、多模态人类脑影像数据集项目，该项目记录了被试进行面孔感知任务的脑电、脑磁和核磁数据(Wakeman & Henson, 2015)。本研究仅选取其中 18 位被试的脑电数据进行分析，用于探究时序上对面孔信息的编码情况。

共 19 位被试参加了全部实验，年龄范围为 23 到 37 岁，其中 8 位女性，11 位男性。其中脑电数据集可能因为数据质量原因仅包含被试编号 002 至 019 的 18 位被试的脑电数据。

脑电部分的实验的面孔刺激共有 450 张灰度图像，其中熟悉面孔 (familiar faces)、不熟悉面孔 (unfamiliar faces) 和乱相面孔 (scrambled faces) 各 150 张。熟悉面孔来自知名人物，大多数英国成年人都能认出这些名人的面孔。而不熟悉面孔来自不知名人物（被试不认识的人），这些不熟悉面孔在性别和年龄上都会和熟悉面孔大致匹配。同时，所有照片都进行了缩放、裁剪只显示面部部分，这些照片覆盖了多种发型（长头发可能会被裁掉）、多种表情（主要是开心的或平静的表情）以及多种朝向（所有都是正中心到 3/4 视角）。而乱相面孔则是由相同序号的熟悉面孔和不熟悉面孔一起生成的。首先使用二维傅里叶变换打乱面孔的相位，然后再反变换回图像空间。

实验流程如图 3-1 所示。刺激呈现与被试正前方相距 1.3 米的屏幕正中央，水平与垂直视角分别大约 3.66° 和 5.38° 。照片背景为黑色，中间有一个固定的白色十字。每一个试次一开始先是出现一个随机的持续时间在 400 到 600ms 之间的固定交叉，然后刺激随机呈现 800 到 100 毫秒之间的随机持续时间。刺激开始之前的随机抖动是为了减少正在进行的神经震荡的叠加并避免任何刺激前的相位重置。在刺激的间期，屏幕显示的内容由一个中央白色圆圈组成，持续 1700ms。被试在实验中被要求全程集中注意力，从中心圆圈到中心十字的变化有助于参与者对每一个刺激做好准备。同时，被试还被要求在刺激呈现阶段尽量不要眨眼。

实验中，每一张图片呈现两次，第二次呈现要么是随着第一次呈现后接下来的试次立即重复出现，要么是间隔 5-15 个试次之后再延迟重复。三种类型的面孔均有 50% 的照片为第一种重复方式，另 50% 的照片为第二种重复方式。本研究中，将刺激第一次出

现的情况称为 New，第二次为立即重复出现的情况称为 Early，第三次为延迟重复出现的情况称为 Late（如图 3-2 所示）。为了确保被试对每个刺激的关注，被试需要持续用左手的食指按两个键中的一个，按键代表被试认为该刺激图片的对称性高于平均水平还是低于平均水平（平均水平由一个预先的练习实验得到，在练习实验中被试会观察 23 张正式实验中不会使用的面孔图片）。

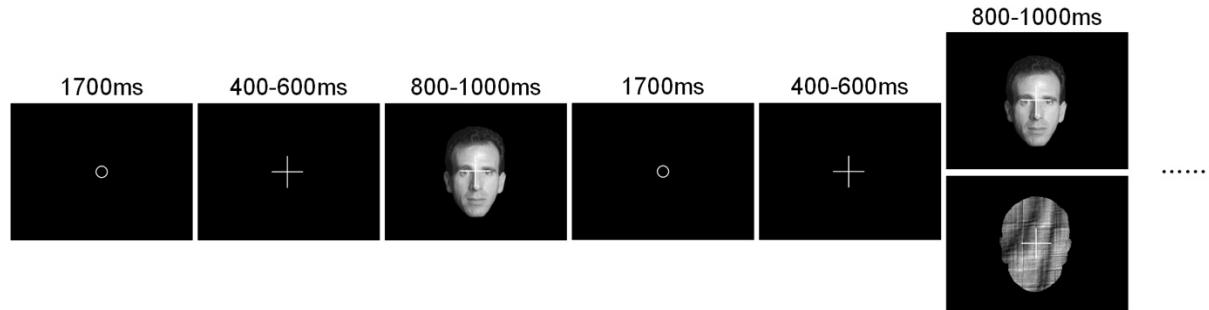


图 3-1 实验流程示意图



图 3-2 实验条件（面孔类别与刺激状态）及对应简称示意图

脑电数据使用 Elekta Neuromag Vectorview 306 系统在一个光磁屏蔽室内采集，被试需要头戴一个 70 导联的 Easycap 脑电帽来同步记录脑电信号，电极的布局符合拓展的 10-10% 系统。脑电参考电极放置在鼻子上、接地电极放置在左锁骨上，并额外一组双极电极来测量垂直(左眼)和水平眼动(VEOG 和 HEOG)，一组双极电极来测量心电(ECG) (左肋和右锁骨)。该数据集中的脑电数据已经过重采样，其采样率为 250Hz。

3.2 分析方法

3.2.1 脑电数据的预处理

在数据集已经经过简单预处理的基础上，使用 MATLAB 的 EEGLAB 工具包 (Delorme & Makeig, 2004) 对数据进行进一步的预处理。对数据进行 0.1Hz 的高通滤波和

30Hz 的低通滤波。然后，使用独立成分分析（independent component analysis, ICA）来识别和去除眨眼和眼动相关的成分(Drisdelle et al., 2017; Jung et al., 2000)。最后，选取每个试次的刺激呈现的前 500 毫秒到之后的 1500 毫秒进行数据分段进行后续分析。

3.2.2 基于脑电的分类解码

3.2.2.1 逐时间点解码

每一个试次的脑电数据都对应两种标签：面孔类别标签（熟悉面孔、不熟悉面孔和乱相面孔）和刺激状态标签（New、Early 和 Late），共进行了六次分类解码：熟悉面孔对不熟悉面孔、熟悉面孔对乱相面孔、不熟悉面孔对乱相面孔、New 对 Early、New 对 Late 和 Early 对 Late。在对面孔类别进行分类时不考虑刺激状态的影响，对刺激状态进行分类时不考虑面孔类别的影响。

使用 Linear-SVM 分别对每组分类条件进行二分类。首先对两个类别的试次标签进行二值化，分别将标签设为对应进行二分类的两个类别。对脑电数据进行降采样，每 20ms 作为一个时间窗（即 5 个时间采样点）进行数据的平均，因而原本 -500 到 1500ms 的 500 个时间点压缩为 100 个时间点。因此每个被试可以得到一个用于分类计算的标签向量与一个三维矩阵，包括时间维度、试次维度、导联维度三个维度。接下来对试次进行打乱，再每五个试次进行一次数据平均，再基于每一个时间点的数据进行分类器的训练与测试。这里进行十次迭代测试与训练，随机取 2/3 的数据进行训练，然后用剩余的 1/3 的数据进行测试以评估分类器的性能。整个试次随机打乱进行平均再进行分类训练与测试的过程再需要迭代十次。因此，经过了 100 次的重复，对所有迭代下的分类准确率进行平均，从而得到更可靠的解码精度。最终每个被试会得到对应分类条件下的时序解码准确率结果。

3.2.2.2 跨时域解码

在上述逐时间点的解码基础上，本研究进一步进行了基于脑电的跨时域的解码。跨时域的解码是为了构建信息解码的时间泛化矩阵，是对单一逐时间点解码的扩展。跨时域解码的本质是，输入一个时间点的数据样本到分类器中进行训练并用这个训练好的分类器去测试其他任意时间的数据来判断不同时间对信息的编码模式是否一致。因而在上述逐时间点解码的实现基础上，添加了对单个时间点数据进行分类器训练后的全时间范围内的测试计算，即对于任意时间点的数据进行分类器训练后，都需要用该分类来测试全部 100 个时间点的数据（包括该时间点本身）。同样，进行迭代计算，取 100 次迭代

后的平均准确率作为该被试该分类计算下的最终解码准确率。最终每个被试会得到对应分类条件下的跨时域解码准确率的时间泛化矩阵。上述全部脑电解码过程均基于 Python 的 scikit-learn 工具包(Pedregosa et al., 2011)实现。

3.2.3 基于脑电的表征相似性分析

3.2.3.1 构建神经表征不相似性矩阵

首先构建基于脑电数据的 RDMs，由于有三种类型的面孔刺激、三种刺激状态，这里使用基于分类的脑电解码正确率作为不相似性指标构建 9×9 的神经 RDMs。并且，在传统的逐时间点构建 RDMs 的基础上，这里使用全新的方法来构建 CTRDMs 以进行 CTRSA。如图 3-3 所示，使用 t_A 时刻在熟悉面孔且延迟重复 (FL) 条件下的脑电数据训练一个 SVM 分类器，再使用这一训练好的分类器来测试 t_B 时刻在乱相面孔且初次看到 (SN) 条件下的脑电数据，将这一测试准确率作为该被试 $t_A \rightarrow t_B$ 跨时间的神经 RDM 中 t_A 的 FL 条件与 t_B 的 SN 条件之间的表征不相似性。由于同种条件之间无法进行分类解码，因此所有 RDMs 的对角线上表征不相似性值均为 0。此外， $t_A \rightarrow t_B$ 跨时间的神经 RDM 与 $t_B \rightarrow t_A$ 跨时间的神经 RDM 不同，前者为使用 t_A 时刻的数据训练 t_B 时刻的数据测试得到，后者为使用 t_B 时刻的数据训练 t_A 时刻的数据测试得到。因而，任意一对有向的时间组合均可以按上述方法构造一个 CTRDM。每一个被试即可根据脑电数据进行跨时域解码构建 100 (个时间点) $\times 100$ (个时间点) 个 CTRDMs。

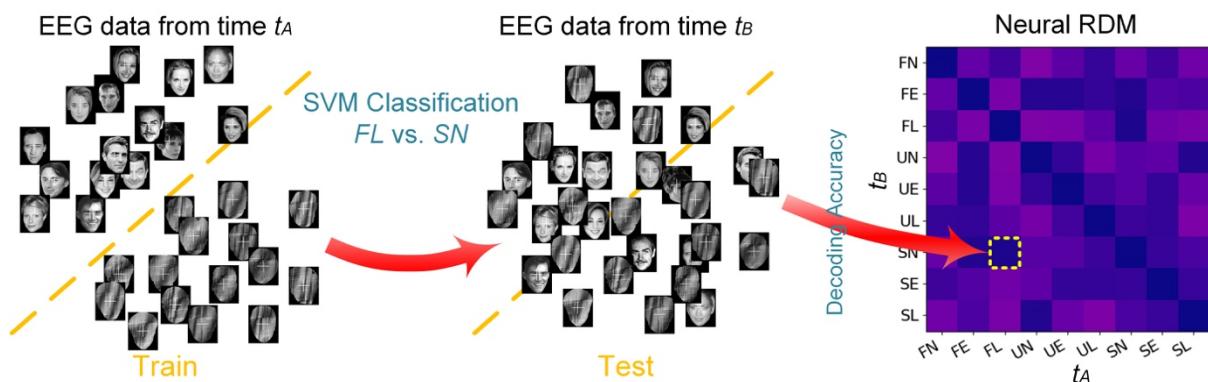


图 3-3 神经 RDM 计算构建示意图

3.2.3.2 构建编码模型表征不相似性矩阵

同时构建基于不同面孔信息的模型，这里共构建了 7 个不同的 9×9 的编码模型 RDMs (如图 3-4 所示)：

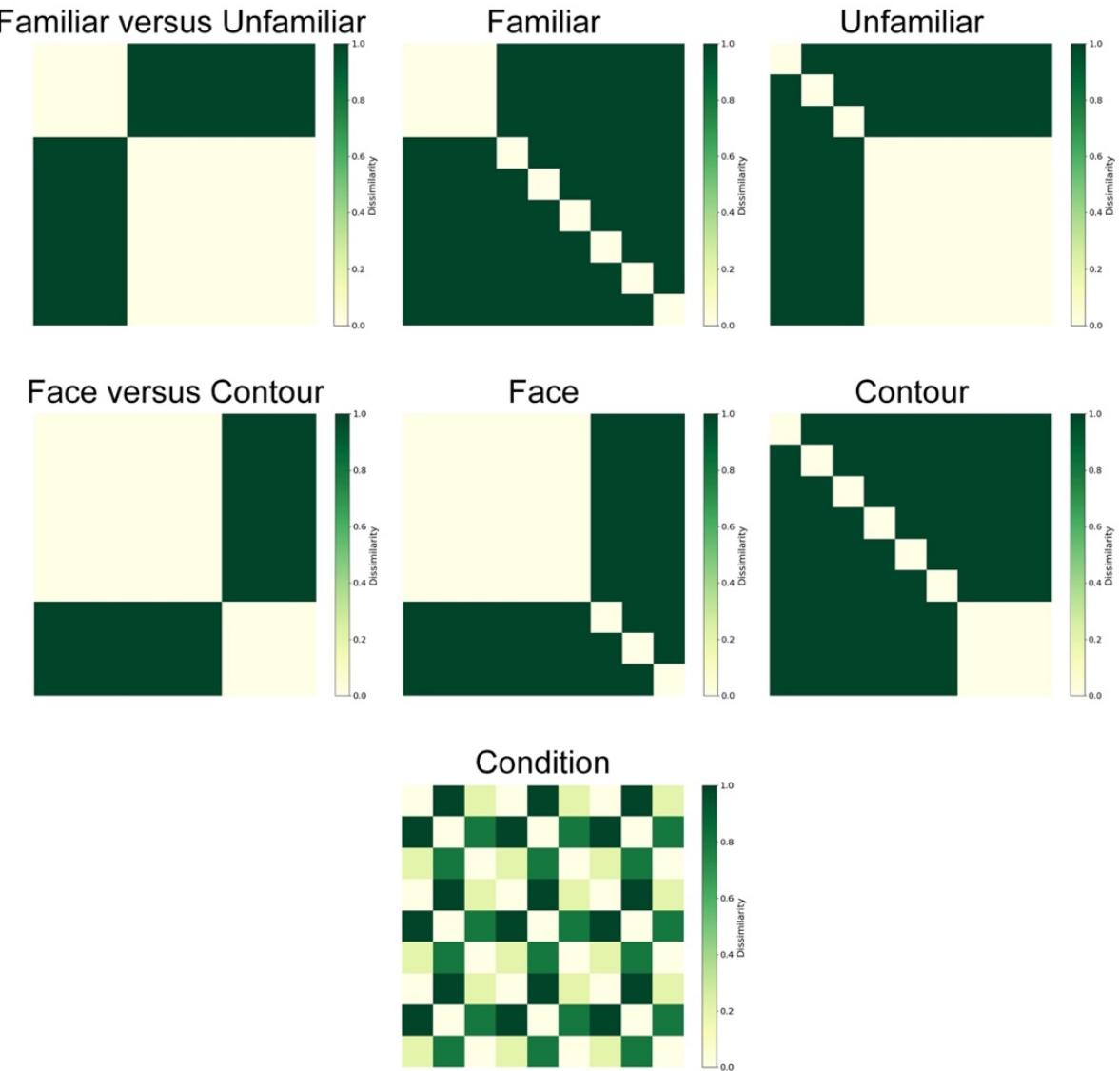


图 3-4 编码模型 RDMs (9×9)

第一行从左至右：熟悉度 (Familiar vs. Unfamiliar) 编码模型、熟悉面孔 (Familiar) 编码模型和不熟悉面孔 (Unfamiliar) 编码模型；第二行从左至右：面孔完整性 (Face vs. Contour) 编码模型、完整面孔 (Face) 编码模型和面孔轮廓 (Contour) 编码模型；第三行：刺激状态 (Condition) 编码模型。

- (1) 熟悉度 (Familiar vs. Unfamiliar) 编码模型：假设同为熟悉面孔或同为不熟悉面孔或乱相面孔（因为被试对乱相面孔也同样是不熟悉的）时的两不同条件间表征相似性高(不相似性设为0)，而熟悉面孔与不熟悉面孔之间的表征相似性低(不相似性为1)；
- (2) 熟悉面孔 (Familiar) 编码模型：假设仅同为熟悉面孔时的两不同条件间表征相似性高 (不相似性设为 0)，其他两不同条件间的表征相似性均很低 (不相似性设为 1)；
- (3) 不熟悉面孔 (Unfamiliar) 编码模型：假设同为不熟悉面孔或乱相面孔时的两不同条件间表征相似性高 (不相似性设为 0)，其他两不同条件间的表征相似性均很低 (不相

似性设为 1), 由于乱相面孔对被试来说也是不熟悉的, 因此这里将乱相面孔也加入了不熟悉面孔的条件; (4) 面孔完整性 (Face vs. Contour) 编码模型: 假设同为完整面孔或同为面孔轮廓时的两不同条件间表征相似性高 (不相似性设为 0), 其他两不同条件间的表征相似性均很低 (不相似性设为 1); (5) 完整面孔 (Face) 编码模型: 假设仅同为完整面孔时的两不同条件间表征相似性高 (不相似性设为 0), 其他两不同条件间的表征相似性均很低 (不相似性设为 1); (6) 面孔轮廓 (Contour) 编码模型: 假设仅同为乱相面孔时的两不同条件间表征相似性高 (不相似性设为 0), 其他两不同条件间的表征相似性均很低 (不相似性设为 1); (7) 刺激状态 (Condition) 编码模型: 假设同一刺激状态间表征相似性高 (不相似性设为 0), New 和 Early 间表征相似性最低 (不相似性设为 1), New 和 Later 间表征相似性较高 (不相似性设为 0.33), Early 和 Late 间表征相似性较低 (不相似性设为 0.67)。

3.2.3.3 人脑与模型间表征相似性分析

使用 Spearman 相关系数计算神经 RDMs 与编码模型 RDMs 之间的相似性。首先, 提取由同一时间点数据构成的 RDMs, 分别去和 7 个编码模型 RDMs 计算相关系数, 得到逐时间点的表征相似性结果。其次, 也用 CTRDMs 分别和 7 个编码模型 RDMs 计算相关系数, 得到跨时域的表征相似性结果。以上全部计算神经 RDMs、构建编码模型 RDMs 以及进行两者之间的相似性分析都基于 NeuroRA 和 PyCTRSA 实现。

除此之外, 本研究也额外使用了广义线性模型 (Generalized Linear Model, GLM) 的方法, 建立了神经 RDMs 与 7 个编码模型 RDMs 之间的联系。对于每一个时间点, 独立计算编码模型 RDMs 对该时间点的神经 RDM 的贡献。如图 3-5 所示, 即可得到时序上对应编码模型 i 的逐时间点的估算值 b_i 。

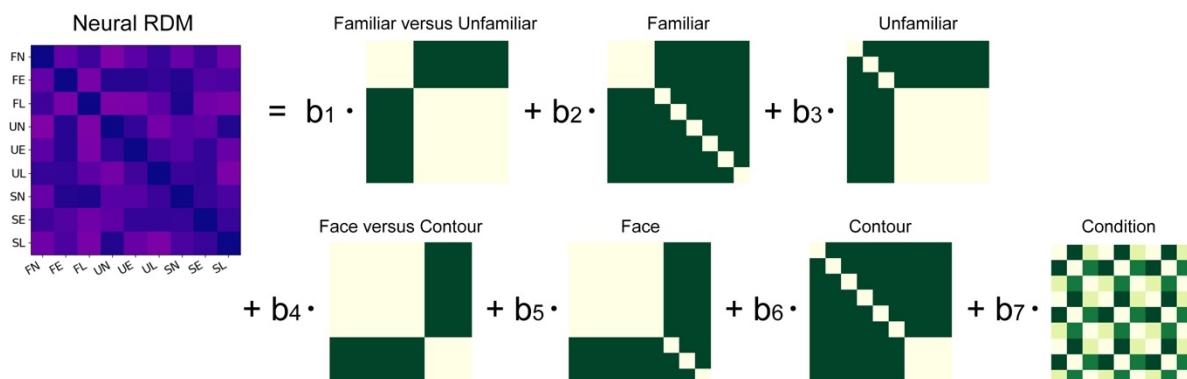


图 3-5 神经 RDM 与编码模型 RDMs 间 GLM 计算示意图

左侧为单个时间点的神经 RDM 示例, 右侧为 7 个模型编码 RDMs。

3.2.4 统计分析

对于基于分类的解码结果，如果某一时间内大脑的神经表征确实编码了某一信息，则该信息对应的两条件的神经表征在高维空间上的分布存在模式差异，进而认为对这一信息下两条件间的解码正确率应高于随机水平，即 50%。对于每一个解码时间点，都进行正确率与随机水平 0.5 之间的单样本均值检验得到每一时间点的 p 值，取 $p < 0.05$ 的时间点作为解码的显著性时段。

对于 RSA 结果，如果某一时间内大脑的神经表征符合假设的编码模型，则对应的编码模型 RDM 与神经 RDM 存在相关，进而比较时序 RDMs 与编码模型 RDM 的时序相似性（相关系数）是否显著大于 0。对于每一个时间点，都进行相似性与 0 之间的单样本均值检验得到每一时间点的 p 值，取 $p < 0.05$ 的时间点作为 RSA 的显著性时段。

对于跨时域解码和跨时域表征相似性分析的结果，在 t 检验的基础上进行了基于簇的置换检验（cluster-based permutation test）。首先提取跨时域结果的每一个显著性簇，然后计算每一个显著性簇内所有 t 值的和，作为该簇的 t 值。然后进行 1000 次置换来计算每一次迭代中拥有最大 t 值的簇的 t 值，从而得到一个最大簇 t 值的分布。最后，对每一个簇来比较其对应 t 值和随机最大簇 t 值的置换分布的显著性，取前者显著大于后者的簇 ($p < 0.05$) 作为最终显著的簇。

3.2.5 基于脑电的面孔表征可视化

使用多维尺度变换（multidimensional scaling, MDS）的方法可视化 9 种条件下神经表征的相似性与不相似性，因为 MDS 能很好的保留下原高维空间中不同条件之间的表征模式的距离远近，因而 MDS 提供了一种直观的方式感受不同条件之间的表征差异。由于 RDM 中的值即已经代表两条件之间的不相似性，这里直接选取一些时间点的 RDMs 作为 MDS 计算的输入，并在进行数据降维后的二维空间上进行投射。通过这种可视化面孔表征的方法，动态而直观地了解了时序上 9 种条件间的表征相似性变化。这一面孔表征可视化的计算与绘图基于 Python 的 scikit-learn 和 Matplotlib 工具包实现。

3.3 结果

基于脑电对三种类型的面孔分别在不同刺激状态下两两进行逐时间点分类解码的结果如图 3-6 第一行所示。

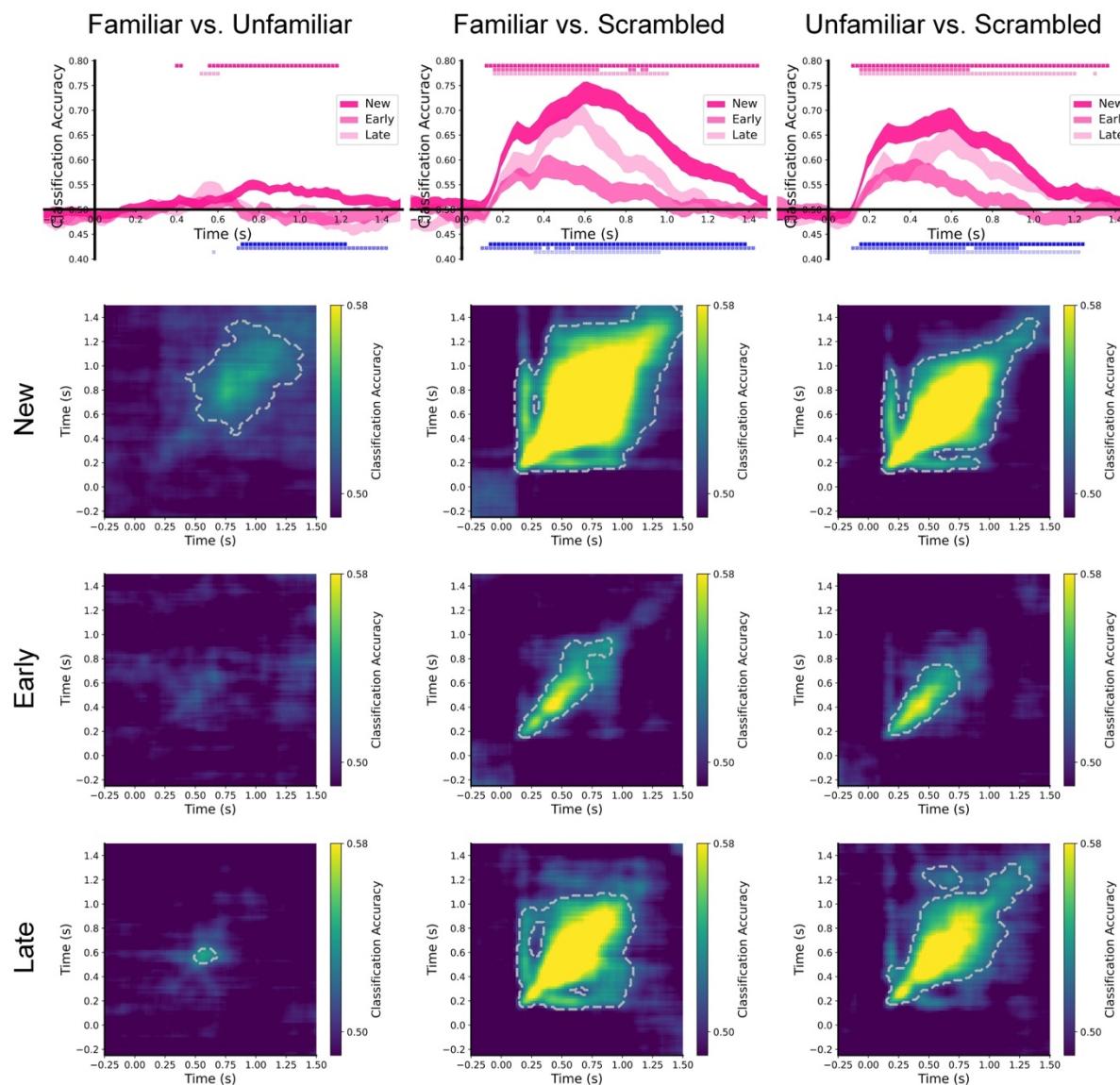


图 3-6 基于脑电的面孔类别的分类解码结果

从左到右三列依次是熟悉面孔和不熟悉面孔、熟悉面孔和乱相面孔以及不熟悉面孔和乱相面孔的两两解码结果；从上到下四行分别为逐时间解码结果、刺激为第一次看到的状态下的跨时域解码结果、刺激为立即重复状态下的跨时域解码结果以及刺激为延迟重复状态下的跨时域解码结果。对于逐时间点解码结果：曲线上方三种颜色的方块由深到浅分别代表第一次看到的状态、立即重复状态和延迟重复状态下解码正确率显著高于随机水平的时间，曲线下方三种颜色的方块由深到浅分别代表第一次看到的状态显著高于立即重复状态、第一次看到的状态显著高于延迟重复状态和立即重复状态显著高于延迟重复状态的时间，曲线宽度对应加减一个标准误。对于跨时域解码结果：灰色虚线勾勒的区域表示解码正确率显著高于随机水平。

熟悉面孔与不熟悉面孔的解码结果显示，当刺激为第一次看到的时候，两者的解码结果存在较长时间上（380-420ms 以及 540-1200ms）显著高于随机水平；当刺激为立即重复情况下看到时，两者的解码结果一直处于随机水平；当刺激为延迟重复情况下看到

时，两者的解码结果在短暂时上（500-600ms）显著高于随机水平。且在 720-1220ms 上刺激为第一次看到情况的解码正确率显著高于立即重复情况的解码正确率，在 400-1420ms 上刺激为第一次看到情况的解码正确率显著高于延迟重复情况的解码正确率。而熟悉面孔与乱相面孔的解码结果显示，当刺激为第一次看到时，两者的解码结果在 100-1440ms 上持续显著高于随机水平；当刺激为立即重复情况下看到时，两者的解码结果在 140-660ms、800-840ms 和 860-900ms 时显著高于随机水平；当刺激为延迟重复情况下看到时，两者的解码结果在 140-1020ms 时显著高于随机水平。且在 140-1380ms 上刺激为第一次看到情况的解码正确率显著高于立即重复情况的解码正确率，在 100-1420ms 上刺激为第一次看到情况的解码正确率显著高于延迟重复情况的解码正确率，在 360-960ms 上刺激为立即重复情况的解码正确率显著高于延迟重复情况的解码正确率。不熟悉面孔与乱相面孔的解码结果和熟悉面孔与乱相面孔的解码结果类似，当刺激为第一次看到时，两者的解码结果在 100-1380ms 上持续显著高于随机水平；当刺激为立即重复情况下看到时，两者的解码结果在 140-700ms 时显著高于随机水平；当刺激为延迟重复情况下看到时，两者的解码结果在 140-1200ms 时显著高于随机水平。且在 140-1240ms 上刺激为第一次看到情况的解码正确率显著高于立即重复情况的解码正确率，在 100-660ms 和 700-920ms 上刺激为第一次看到情况的解码正确率显著高于延迟重复情况的解码正确率，在 480-1220ms 上刺激为立即重复情况的解码正确率显著高于延迟重复情况的解码正确率。

进一步分析对面孔类别的跨时域分类解码结果，如图 3-6 的第二到第四行所示。跨时域分类解码的结果不仅说明了单个时间点自训练并进行分类测试的结果，也提供了不同时间之间对面孔类别差异的编码模式的异同。对比三种刺激状态下对熟悉面孔与不熟悉面孔的跨时域分类解码结果，当刺激为第一次看到时，解码正确率显著高于随机水平的区域在 500-1300ms 左右，而当刺激为延迟重复情况时，解码正确率显著高于随机水平的区域仅出现在 500-700ms 左右，但这两时间区域内的编码模式类似。同时，对于熟悉面孔与乱相面孔或不熟悉面孔与乱相面孔的跨时域分类解码结果，其在刺激为第一次看到的情况下解码正确率显著高于随机水平的区域最大，而刺激为立即重复的情况下解码正确率显著高于随机水平的区域最小，但三种刺激状态下的显著性的区域都是从较早阶段（100-140ms 左右）开始。

基于脑电对三种刺激状态分别在不同面孔类别下两两进行逐时间点分类解码的结

果如图 3-7 第一行所示。

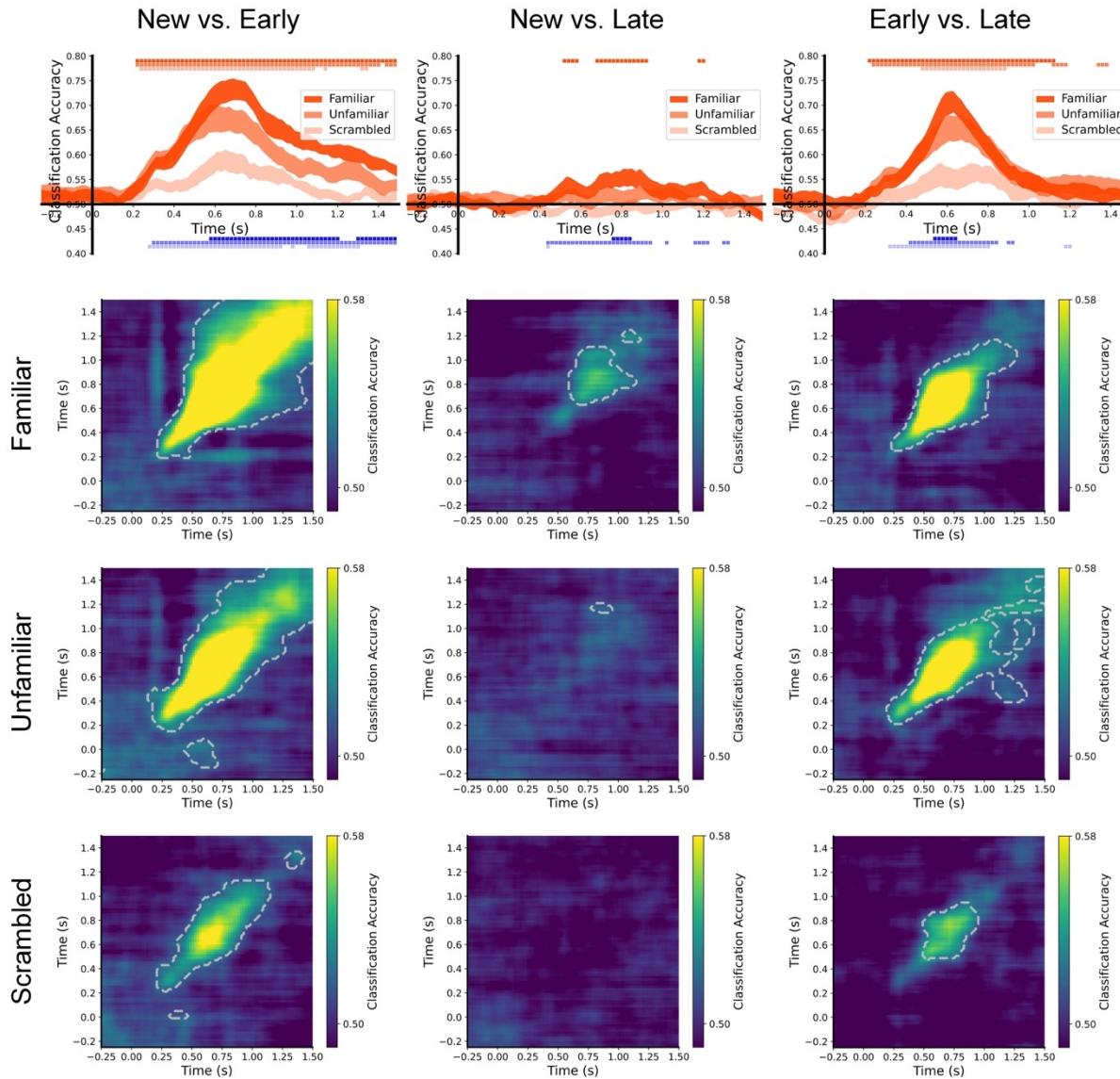


图 3-7 基于脑电的刺激状态的分类解码结果

从左到右三列依次是第一次看到的状态和立即重复状态、第一次看到的状态和延迟重复状态以及立即重复状态和延迟重复状态的两两解码结果；从上到下四行分别为逐时间解码结果、熟悉面孔的跨时域解码结果、不熟悉面孔的跨时域解码结果以及乱相面孔的跨时域解码结果。对于逐时间点解码结果：曲线上方三种颜色的方块由深到浅分别代表熟悉面孔、不熟悉面孔和乱相面孔情况下解码正确率显著高于随机水平的时间，曲线下方三种颜色的方块由深到浅分别代表熟悉面孔显著高于不熟悉面孔、熟悉面孔显著高于乱相面孔和不熟悉面孔显著高于乱相面孔的时间，曲线宽度对应加减一个标准误。对于跨时域解码结果：灰色虚线勾勒的区域表示解码正确率显著高于随机水平。

第一次看到的状态与立即重复状态的解码结果显示，当刺激为熟悉面孔和不熟悉面孔时，两者的解码结果均在较长时间上（200-1500ms）显著高于随机水平；当刺激为乱相面孔时，两者的解码结果在 220-1100ms 以及 1300-1340ms 显著高于随机水平。且在

580-1200ms 以及 1280-1500ms 上刺激为熟悉面孔的解码正确率显著高于不熟悉面孔的解码正确率，在 280-1500ms 上刺激为熟悉面孔的解码正确率显著高于乱相面孔的解码正确率，在 260-920ms 以及 1040-1300ms 上刺激为不熟悉面孔的解码正确率显著高于乱相面孔的解码正确率。而第一次看到的状态与延迟重复状态的解码结果显示，仅当刺激为熟悉面孔时存在显著，且两者的解码结果在 520-600ms、700-920ms 以及 1160-1200ms 上显著高于随机水平。且在 740-840ms 上刺激为熟悉面孔的解码正确率显著高于不熟悉面孔的解码正确率，在 420-940ms、1120-1220ms 以及 1280-1320ms 上刺激为熟悉面孔的解码正确率显著高于乱相面孔的解码正确率。立即重复状态与延迟重复状态的解码结果显示，当刺激为熟悉面孔时，两者的解码结果在 200-1120ms 上持续显著高于随机水平；当刺激为不熟悉面孔时，两者的解码结果在 220-1020ms、1100-1180ms 以及 1340-1400ms 时显著高于随机水平；当刺激为乱相面孔时，两者的解码结果在 460-880ms 时显著高于随机水平。且在 520-640ms 上刺激为熟悉面孔的解码正确率显著高于不熟悉面孔的解码正确率，在 400-840ms 和 880-920ms 上刺激为熟悉面孔的解码正确率显著高于乱相面孔的解码正确率，在 300-800ms 上刺激为不熟悉面孔的解码正确率显著高于乱相面孔的解码正确率。

进一步分析对刺激状态的跨时域分类解码结果，如图 3-7 的第二到第四行所示。这里跨时域分类解码的结果不仅说明了单个时间点进行自训练并进行分类测试的结果，也提供了不同时间之间对刺激状态差异的编码模式的异同。对比三种面孔类别下的跨时域分类解码结果，均体现出刺激为第一次看到的状态与立即重复状态间的解码正确率显著高于随机水平的区域最大，而刺激为第一次看到的状态与延迟重复状态间的解码正确率显著高于随机水平的区域很小，甚至没有。另一方面，三对分类解码的结果均体现出刺激为熟悉面孔时解码正确率显著高于随机水平的区域最大，而刺激为乱相面孔时解码正确率显著高于随机水平的区域最小。

基于脑电数据构建 9×9 的逐时间点 RDMs 和 CTRDMs，九个条件依次是：FN、FE、FL、UN、UE、UL、SN、SE 和 SL。图 3-8 为-250ms 时、0ms 时、250ms 时、500ms 时和 750ms 时得到的神经 RDMs 和对应的经过 MDS 后在二维空间上面孔表征可视化的结果。

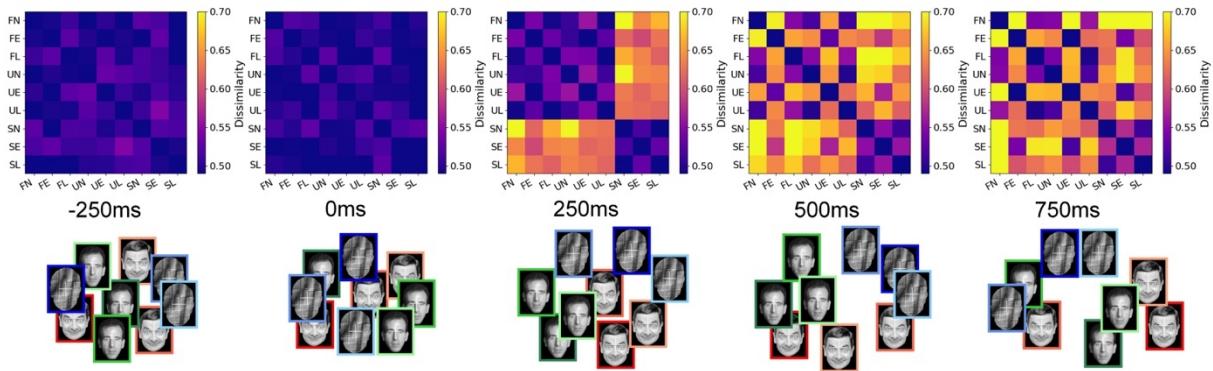


图 3-8 神经 RDMs 及面孔表征可视化示例

上一行五个时间点的 RDMs，颜色代表不相似性；下一行为五个时间点的 RDMs 经过 MDS 后的二维可视化结果，红框、蓝框和绿框分别代表熟悉面孔、不熟悉面孔和乱相面孔，框的色调由深到浅分别代表第一次看到的状态、立即重复状态和延迟重复状态。

刺激呈现之前的 RDMs 整体呈现出条件之间的低不相似性（高相似性），随着刺激出现，不同条件之间的表征差异出现，并且能看出在面孔感知阶段，完整面孔与面孔轮廓之间存在着表征模式的不同。

通过比较 7 个编码模型与神经表征之间的动态编码相似性，可以深入探究大脑对不同面孔信息的编码模式。图 3-9 显示了 7 个编码模型 RDMs 分别与脑电的时序 RDMs 和 CTRDMs 计算表征相似性后的结果。

对于逐时间点的 RSA 结果，神经表征与面孔熟悉度编码模型在 400-780ms 以及 1080-1160ms 内显著相似，与熟悉面孔编码模型在 180-400ms 内显著相似，与不熟悉面孔编码模型在 500-980ms 以及 1040-1220ms 内显著相似，与面孔完整性编码模型在 140-1020ms 内显著相似，与完整面孔编码模型在 140-660ms 以及 760-920ms 内显著相似，与面孔轮廓编码模型在 140-1500ms 内显著相似，与刺激状态编码模型在 300-360ms 以及 440-1500ms 内显著相似。

对于 CTRSA 结果，神经表征与面孔熟悉度编码模型、熟悉面孔编码模型与不熟悉面孔编码模型的相似性整体更弱一些，其中大脑对熟悉面孔信息的编码在较早阶段（大约 140-500ms）出现，而对面孔熟悉度信息和不熟悉面孔信息的编码则在较晚阶段具有持续一致的编码（大约 500-1400ms）。而大脑对面孔完整性信息和完整面孔信息的编码都在大约 140-1000ms 内具有持续的表征一致性，而对面孔轮廓信息和刺激状态信息的编码持续性更强，并且对前者的编码开始时间要早于后者，大脑前者大约在 140-1500ms 内具有持续一致的表征，对后者大约在 450-1500ms 内具有持续一致的表征。

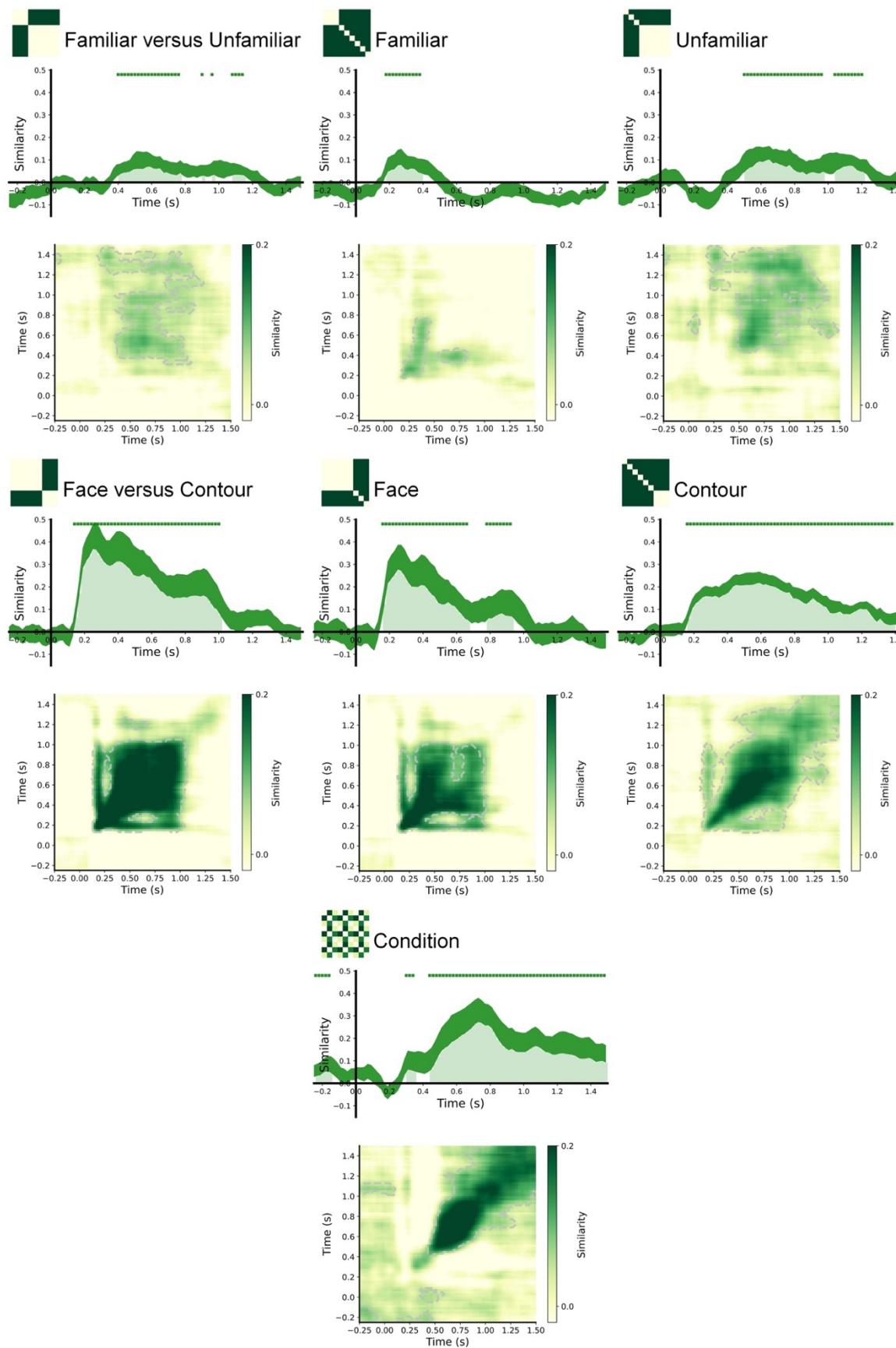


图 3-9 人脑对不同面孔信息的动态表征

七种模型分别对应了一张逐时间点 RSA 的结果图和一张 CTRSA 的结果图。在逐时间点 RSA 的结

果图中，曲线上方的绿点以及曲线与 x 轴之间的阴影表示显著性的时间段 ($p<0.05$)，曲线的宽度反映的是加减一个标准误。在 CTRSA 的结果图里，灰色曲线框出的区域为显著的区域（基于簇的置换检验， $p<0.05$ ）。

对于基于脑电的神经 RDMs 与基于假设的不同编码模型 RDMs 之间的相似性比较，也进行了基于 GLM 的计算，结果见图 3-10 所示，计算得到的 7 个模型的时序 Beta 估计值的显著性时间区间与图 3-9 一致。

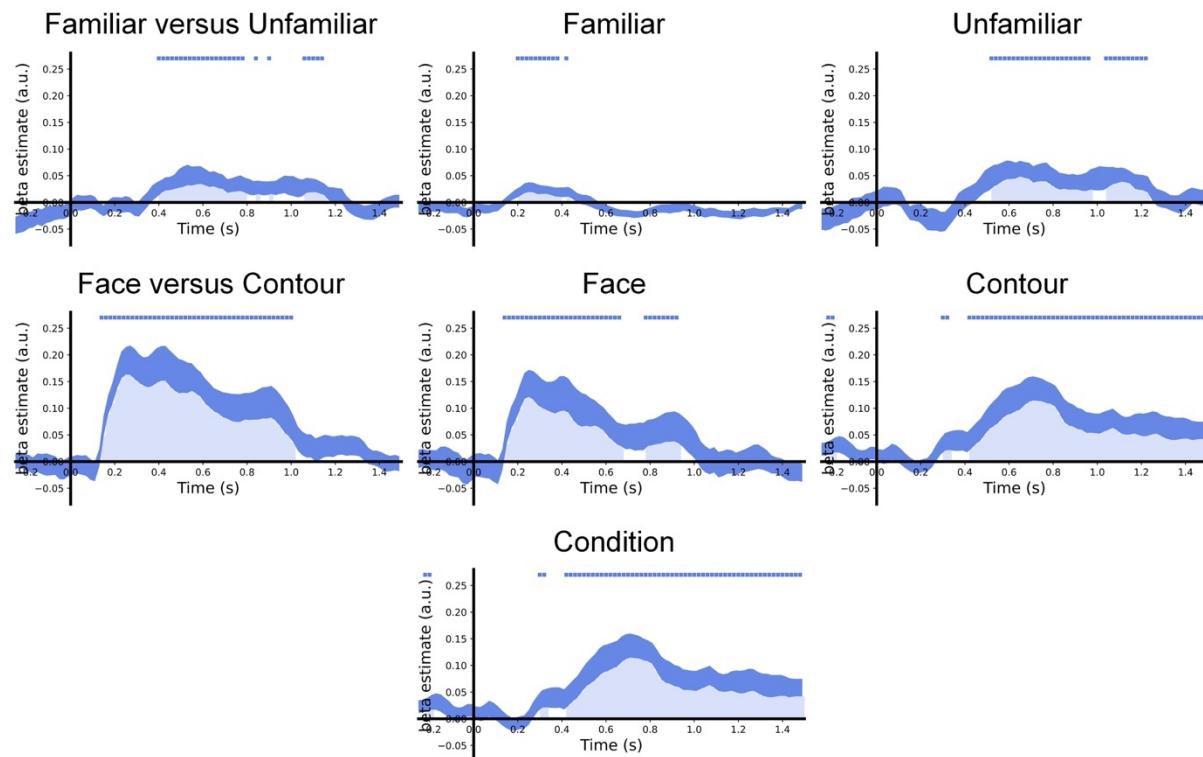


图 3-10 神经表征与编码模型的 GLM 结果

曲线上方的绿点与曲线与 x 轴之间的阴影表示显著性的时间段 ($p<0.05$)，曲线的宽度反映的是加减一个标准误。

即通过两种方式验证了这种基于 RDM 在时序上追踪大脑在面孔感知过程中对不同面孔信息编码情况的方案是可行的，其结果也是可靠的。

3.4 讨论

首先通过基于脑电对面孔类别的分类解码结果说明，人脑在面孔感知过程中当面孔为第一次看到的状态，其对不同面孔类别的加工差异更大，即体现在时序与跨时域的解码正确率上要在很长一段时间上显著高于面孔为两种重复状态时的结果。同时，人脑对于完整的面孔与乱相面孔之间的表征差异在三种刺激状态下都存在，即对于熟悉面孔与乱相面孔的解码和不熟悉面孔与乱相面孔的解码都在三种刺激状态下存在显著。这一结

果也在时序上印证了重复抑制效应的存在，当三种刺激状态下对两面孔类别解码正确率存在差异，第一次看到的状态要强于延迟重复状态强于立即重复状态，则说明立即重复状态的重复抑制效应更强，延迟重复状态的重复抑制效应强。对应出现显著的开始时间也是当刺激为第一次看到的状态时更早（100ms 左右），这也说明大脑不仅在第一次看到面孔时更强地加工面孔类别信息，也会更早地开始进行加工。而人脑对于熟悉面孔与不熟悉面孔的表征差异则要弱于完整面孔与乱相面孔之间的表征差异，因为仅在刺激为第一次看到的状态和延迟重复状态下存在一定时间段上的显著，并且前者的显著时长（380-420ms 以及 540-1200ms）要长于后者（500-600ms）。与此同时，其发生显著的时间（400ms 左右）也要晚于完整面孔与乱相面孔之间解码出现显著的时间（100-140ms 左右），这也说明大脑对熟悉度信息的加工要晚于对于面孔完整性的加工。

而通过基于脑电对刺激状态的分类解码结果直接追踪了状态之间的时序表征差异，当两种刺激状态下的解码正确率存在显著差异，则说明这两种状态间的表征差异大。无论是对熟悉面孔、不熟悉面孔还是乱相面孔，都可以在刺激第一次看到的状态与早期重复状态之间以及立即重复状态与延迟重复状态之间存在长时间的显著解码结果，而第一次看到的状态与延迟重复状态之间的解码结果很弱。这也说明了重复抑制效应的存在，且立即重复状态下的重复效应更前，而延迟重复状态下的表征情况更接近于第一次看到该刺激的状态。同时，对于刺激状态的两两解码正确率而言，熟悉面孔下的解码正确率要高于不熟悉面孔高于乱相面孔。这说明重复抑制效应在熟悉面孔上更强，而在乱相面孔上这一效应最弱。

基于条件中存在的不同面孔信息构建不同的模型，再进行跨模态的人脑与模型之间的表征相似性分析能在时序与跨时域上有效地追踪到大脑对不同信息的动态编码情况。首先发现大脑对面孔的视觉信息的编码要早于对更高级面孔信息的编码，即对人脑表征与面孔完整性编码模型的发生显著相似性的时间（140ms 左右）要早于与面孔熟悉度编码模型的发生显著相似性的时间（400ms 左右）。其次，人脑对面孔轮廓的加工时长要长于对完整面孔的加工时长，即更持久地保持了对轮廓信息的编码，而对轮廓内详细的面孔视觉特征的编码则仅维持到看到面孔刺激后的 1000ms 左右。另一方面，分别构建熟悉面孔的编码模型与不熟悉面孔的编码模型，尝试分离大脑对熟悉与不熟悉面孔信息的加工过程。结果也显示大脑会在更早进行对熟悉面孔信息的加工（180-400ms），在较晚进行对不熟悉面孔信息的加工（500-1220ms）。利用这种直接分离熟悉与不熟悉信息的

手段成功发现了大脑时序上对两种类型面孔的动态编码过程，对熟悉面孔的加工甚至在早期知觉加工阶段就出现，这可能是由于熟悉面孔在早期激活了面孔识别单元造成了大脑更早地对它进行响应，在早期就增强了对面孔信息的表征。而不熟悉面孔则更晚地被加工，但由于其新颖性也造成了大脑对不熟悉面孔信息的晚期更久的编码。最后，人脑与刺激状态模型之间的表征也找到了较长时序上（从 300ms 开始）的持续一致性，不同于先前关于重复抑制的研究大多找到的是一些 ERP 的成分作为指标，本研究发现了在面孔感知过程中重复抑制效应的持续存在。

4 探究面孔信息在深度卷积神经网络模型中的表征

4.1 深度卷积神经网络模型

本研究选用用于面孔识别领域的一个 DCNN 模型——VGG-Face 模型，该模型使用 2622 个人的每人 1000 张面孔图片的数据集进行训练，其在 IFW 数据集和 YouTube Faces 数据集上的测试准确率分别达到了 97.27% 和 92.8%。VGG-Face 本质上是一个 VGG-16 模型结构，其包含 13 个卷积层和 3 个全连接层，具体结构见图 4-1 所示。本研究把 VGG-Face 模型作为一个习得面孔特征的 DCNN 模型。作为对照，额外选用一个随机权重的 VGG-16 模型，其初始权重为随机赋值未经过任何数据输入训练。将随机权重的 VGG-16 模型作为一个未学习面孔特征的 DCNN 模型。

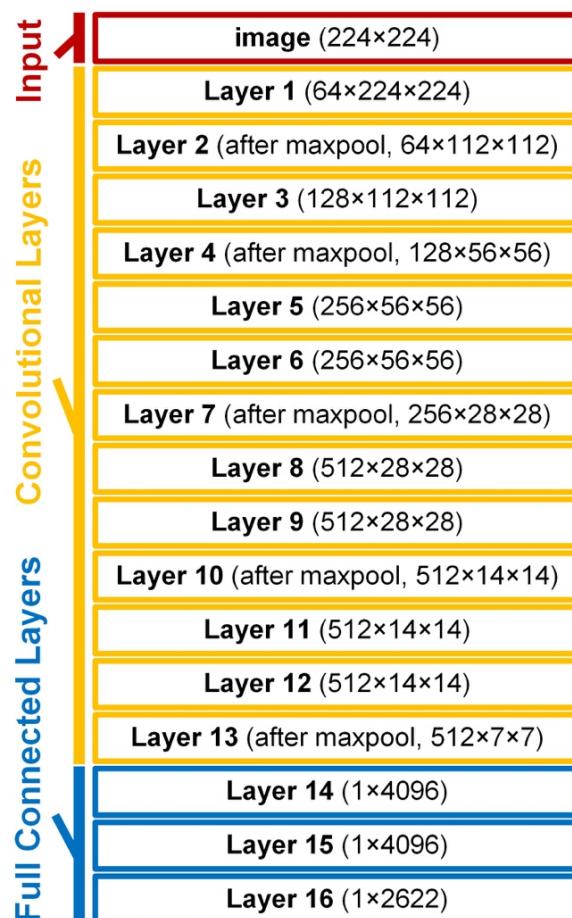


图 4-1 VGG-Face 模型结构

红色代表图片输入，橙色为卷积层，蓝色为全连接层，括号内为对应层的特征输出维度，当某一卷积层包含一最大池化层，则将经过最大池化层之后的结果作为这一层的输出。

4.2 分析方法

4.2.1 特征降维

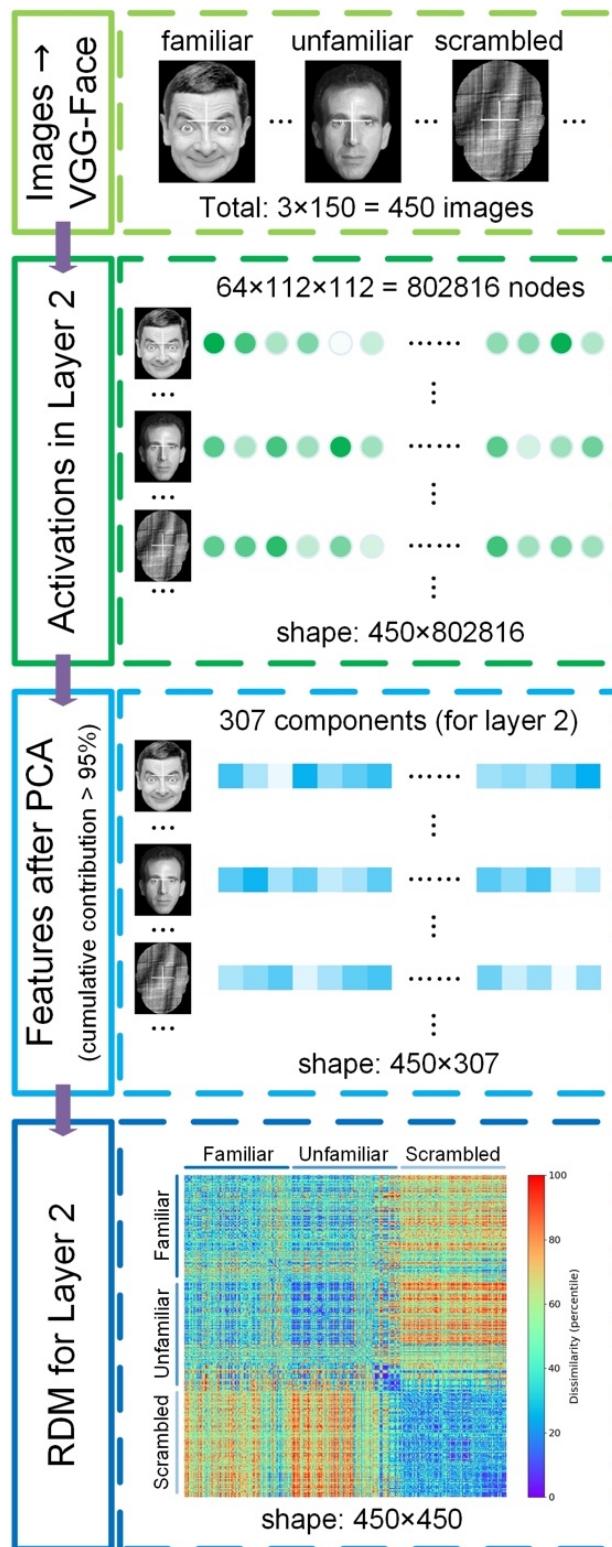


图 4-2 DCNN 模型 RDM 的构造过程示意

该示意图以 VGG-Face 模型的第二层为例，由上至下：图片输入->计算第二层的激活向量->PCA 降维获得特征向量->得到第二层的 RDM。

由于 VGG-16 模型每一层都有大量的节点，以第 2 层为例，包含 64 个 112×112 的特征图，即 802816 个节点。即每一张图片输入到 VGG-16 模型中，第 2 层的激活都可以输出为一个 1×802816 的向量。这里，首先使用主成分分析法（principal component analysis, PCA）将第 2 层的 802816 个特征维度进行降维，将经过 PCA 计算后的主成分按其贡献率从大到小排列，选取总贡献率大于 95% 的主要贡献成分，作为降维后的特征维度。这里如图 4-2 所示，第 2 层的特征维度降维后为 307 个特征维度。依照此方式对每一层的特征维度都进行降维（取总贡献率超过 95% 的成分），从而每一张图片在任意一层上的激活都对应了一个降维后的激活向量。上述 PCA 降维过程分别对 VGG-Face 模型和随机权重的 VGG-16 模型的每一层进行计算，这一降维部分全部基于 Python 的 scikit-learn 工具包实现。

4.2.2 基于深度卷积神经网络模型的表征相似性分析

4.2.2.1 构建 VGG-16 表征不相似性矩阵

每一张图片输入到 VGG-16 模型中都可以得到每一层特征经过降维后的激活向量，450 张面孔图片在 VGG-16 的每一层则都可以对应 450 个激活向量。由于计算量巨大，所有计算都只基于偶数层，即只对 VGG-16 的第 2 层、第 4 层、第 6 层……第 16 层计算。对于偶数层中的任意一层，计算任意两图片对应激活向量之间的 Pearson 相关系数 r ，用 $1-r$ 作为不相似性指标，如图 4-2 所示，按 150 张熟悉面孔、150 张不熟悉面孔、150 张乱相面孔的顺序构建这一层的 450×450 的 RDM。对 VGG-Face 模型和随机权重的 VGG-16 模型的每一层都计算其对应的 RDM，计算 DCNN 的 RDMs 过程基于 NeuroRA 工具包实现。

4.2.2.2 构建编码模型表征不相似性矩阵

类似 3.2.3.2 部分，构建基于不同面孔信息的模型。对于上述 DCNN 的 RDMs 构建，没有模拟同神经 RDMs 类似的刺激状态的变化，因而在构建编码模型 RDMs 时也无法构建对应的 Condition 编码模型。由于 DCNN 的 RDMs 尺寸为 450×450 ，这里共构建了 6 个不同的 450×450 的编码模型 RDMs（如图 4-3 所示）。(1) Familiar vs. Unfamiliar 编码模型：假设同为熟悉面孔或同为不熟悉面孔或乱相面孔（因为被试对乱相面孔也同样是不熟悉的）时的两不同条件间表征相似性高（不相似性设为 0），而熟悉面孔与不熟悉面孔之间的表征相似性低（不相似性为 1）；(2) Familiar 编码模型：假设仅同为熟悉面孔时的两不同条件间表征相似性高（不相似性设为 0），其他两不同条件间的表征相似性

均很低（不相似性设为 1）；（3）Unfamiliar 编码模型：假设同为不熟悉面孔或乱相面孔时的两不同条件间表征相似性高（不相似性设为 0），其他两不同条件间的表征相似性均很低（不相似性设为 1），由于乱相面孔对被试来说也是不熟悉的，因此这里将乱相面孔也加入了不熟悉面孔的条件；（4）Face vs. Contour 编码模型：假设同为完整面孔或同为面孔轮廓时的两不同条件间表征相似性高（不相似性设为 0），其他两不同条件间的表征相似性均很低（不相似性设为 1）；（5）Face 编码模型：假设仅同为完整面孔时的两不同条件间表征相似性高（不相似性设为 0），其他两不同条件间的表征相似性均很低（不相似性设为 1）；（6）Contour 编码模型：假设仅同为乱相面孔时的两不同条件间表征相似性高（不相似性设为 0），其他两不同条件间的表征相似性均很低（不相似性设为 1）。

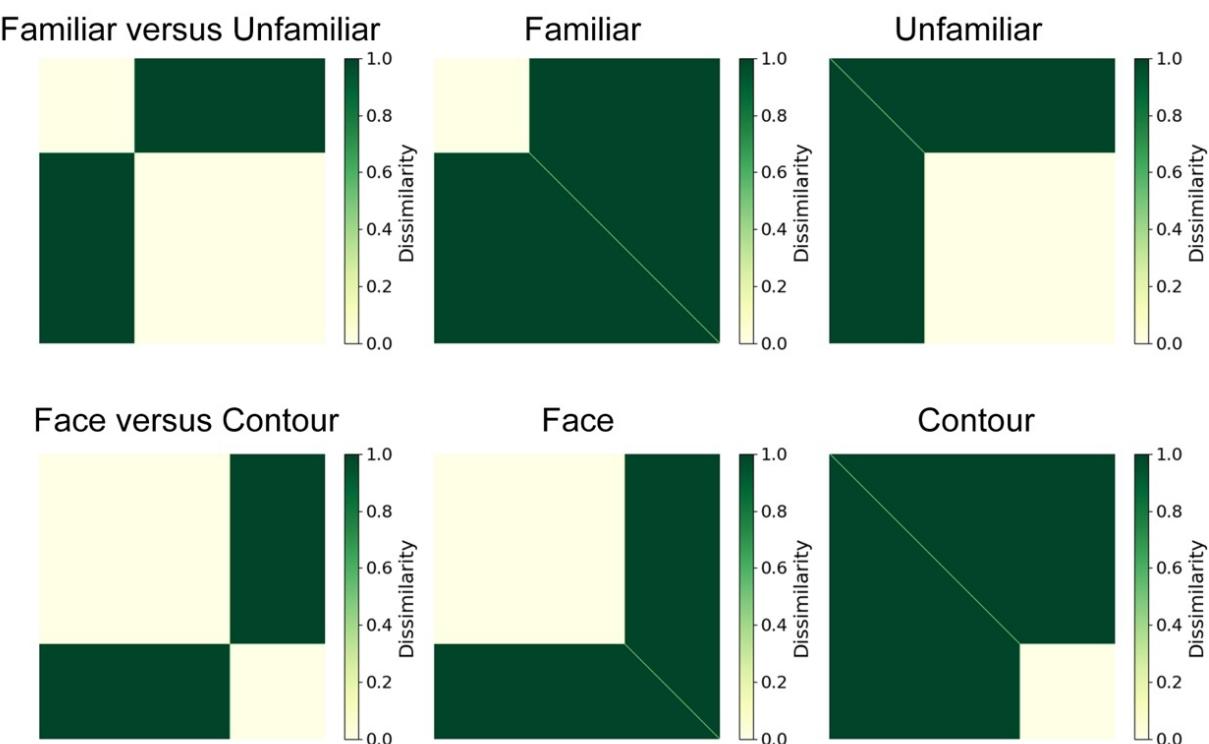


图 4-3 编码模型 RDMs (450×450)

4.2.2.3 VGG-16 类别内部表征相似性

为了探究 DCNN 分别对三种类别的面孔的一致性表征，逐层计算其类别内部表征相似性。这里，计算 VGG-Face 模型和随机权重的 VGG-16 模型对三种类别面孔的各自内部任意两不同图片之间的 Pearson 相关系数，逐层分别计算每一类别面孔的平均相关系数作为对应的类别内部表征相似性指标。即可以得到两个 VGG-16 模型其类别内部表征相似性的逐层变化。

4.2.2.4 VGG-16 与模型间表征相似性分析

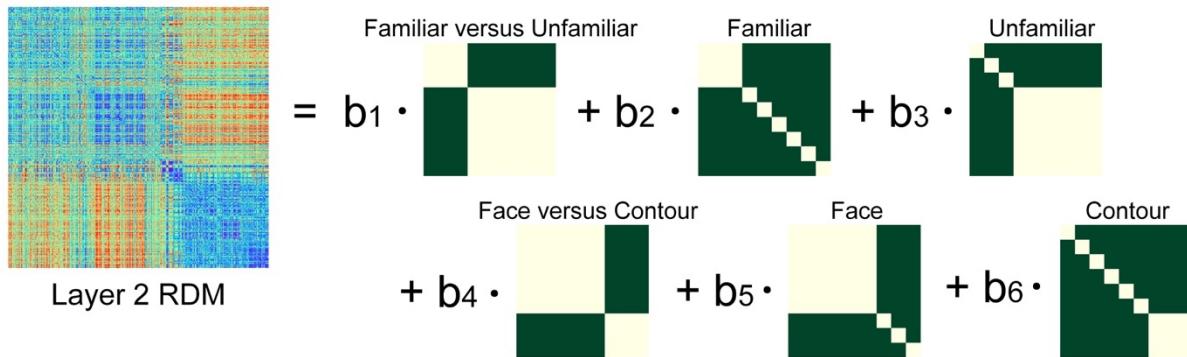


图 4-4 神经 RDM 与编码模型 RDMs 间 GLM 计算示意图

对于 VGG-16 与模型间的表征分析,这里使用 GLM 的方法,建立了 DCNN 的 RDMs 与 6 个编码模型 RDMs 之间的联系。对于 VGG-Face 模型和随机权重的 VGG-16 模型的每一个偶数层的 RDM,计算编码模型 RDMs 对该 DCNN 模型的该层的 RDM 的贡献。如图 4-4 所示,即可得到 VGG-Face 模型和随机权重的 VGG-16 模型每一个偶数层的表征对应编码模型 i 的估算值 b_i 。

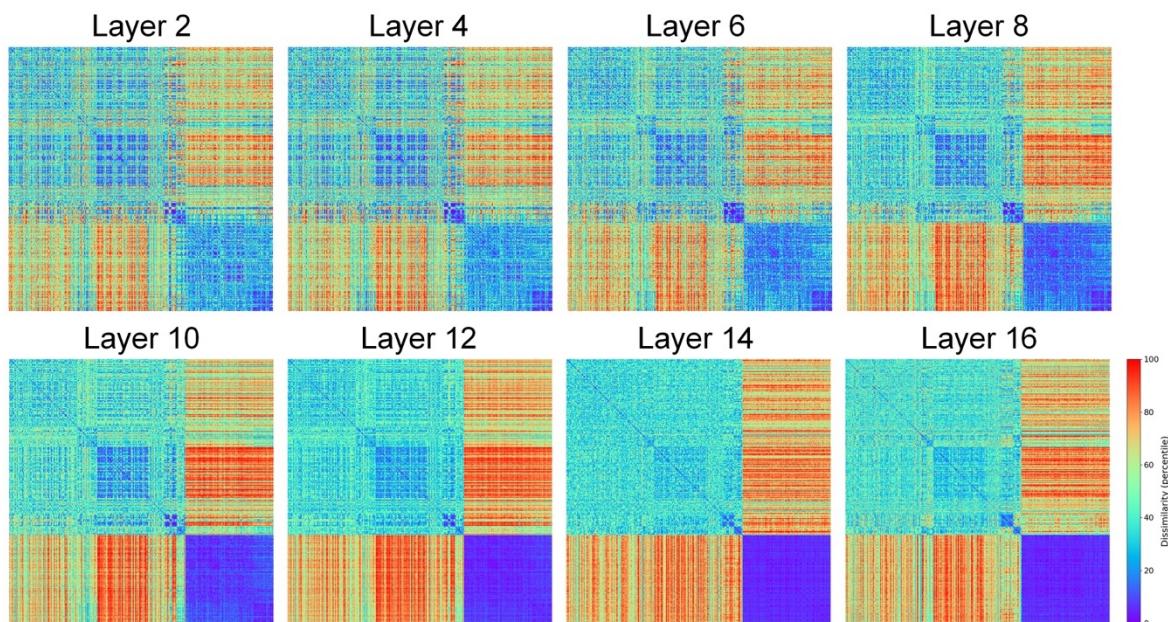
4.2.3 基于深度卷积神经网络模型的面孔表征可视化

使用 t 分布的随机邻居嵌入 (t-distributed stochastic neighbor embedding, t-SNE) 算法将 DCNN 中对图片表征的高维特征投射到二维平面空间上, 经过 t-SNE 算法降维可视化能有效地保留不同图片之间在对应特征空间上的表征相似性。若在二维空间上两图片间隔距离越近, 则说明两图片在 DCNN 的这一层上的表征模式更相似。这里分别对 VGG-Face 模型和随机权重的 VGG-16 模型的早期层 (第 2 层) 和晚期层 (第 16 层) 进行 t-SNE 降维并绘图, 这一过程基于 Python 的 scikit-learn 工具包实现。

4.3 结果

将 450 张面孔图片按照熟悉面孔、不熟悉面孔和乱相面孔分别输入到 VGG-Face 模型和权重随机的 VGG-16 模型中得到两个 DCNN 模型的偶数层的 RDMs 如图 4-5 所示, 这里每一个 RDM 的从左到右 (或从上到下) 450 个条件依次是 150 张熟悉面孔、150 张不熟悉面孔和 150 张乱相面孔。

a Pretrained VGG-Face



b Random VGG-16

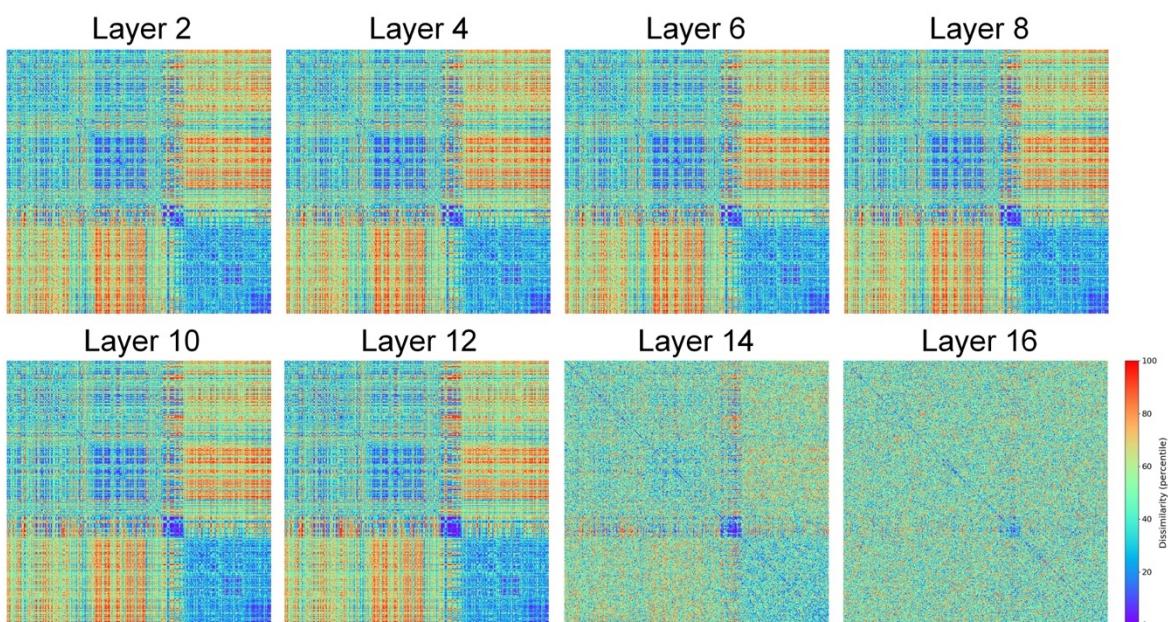


图 4-5 DCNN 模型偶数层 RDMs

(a) VGG-Face 模型的偶数层 RDMs; (b) 随机权重 VGG-16 模型的偶数层 RDMs。

通过图 4-5a 能直观的看出,对于 VGG-Face 模型,对乱相面孔的内部表征不相似性随着层数的增加逐渐减弱,而对熟悉的面孔的内部表征的不相似性比对不熟悉面孔的内部表征的不相似性高,且随着层数的增加,不熟悉面孔的内部表征不相似性逐渐增强。同时,完整面孔(熟悉面孔加不熟悉面孔)与面孔轮廓(乱相面孔)之间的表征存在差

异，并且完整面孔与面孔轮廓之间的表征差异在卷积层上逐层增强。

对于权重随机的 VGG-16 模型，如图 4-5b 所示。也能在卷积层观察到完整面孔与面孔轮廓之间的表征存在差异，但这种差异要弱于在 VGG-Face 模型中出现的差异。同时，完整面孔与面孔轮廓之间的表征差异在全连接层几乎不再存在。类似 VGG-Face，在随机权重 VGG-16 模型中也观察到了对熟悉的面孔的内部表征的不相似性比对不熟悉面孔的内部表征的不相似性高，但是没有明显的层级变化。

对两个 DCNN 模型逐偶数层 RDMs 分别计算熟悉面孔、不熟悉面孔和乱相面孔类别内部的表征相似性，结果如图 4-6 所示。对于 VGG-Face 模型，其对熟悉面孔和不熟悉面孔的类别内部表征相似性一直较低但高于 0，且前者要低于后者，而对于乱相面孔的类别内部表征相似性在前 6 层逐渐降低但随后逐渐升高，并在全连接层上达到接近 0.8 的水平。对于随机权重的 VGG-16 模型，其对熟悉面孔和不熟悉面孔的类别内部表征相似性在卷积层要高于 VGG-Face 模型的结果，同时，其对乱相面孔的类别内部表征相似性也在第 10 层之前都高于 VGG-Face 模型的结果。但在全连接层，其对三种面孔的类别内部表征相似性全部降为 0 左右的水平。

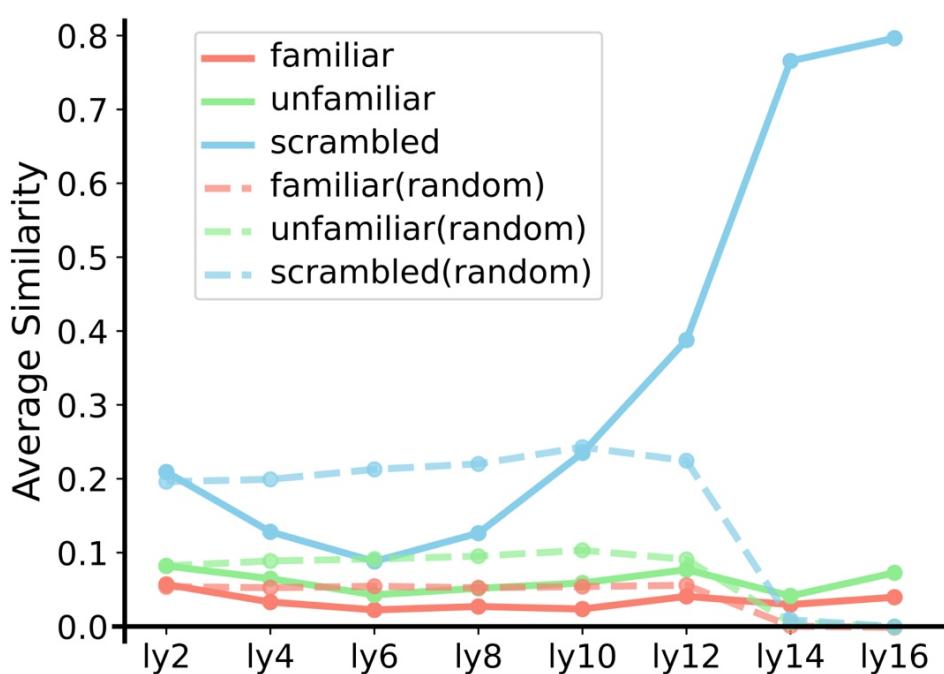


图 4-6 DCNNs 对三种类别面孔的内部表征相似性

红、绿、蓝线分别对应熟悉面孔、不熟悉面孔和乱相面孔，其中粗线代表 VGG-Face 模型的类别内部表现相似性结果，虚线代表随机权重的 VGG-16 模型的类别内部表征相似性结果。

深蓝色的点和线代表 VGG-Face 模型的计算结果，蓝绿色的点和线代表随机权重的 VGG-16 模型的计算结果。

更进一步，使用 GLM 计算两 DCNNs 逐偶数层对面孔信息的编码情况，如图 4-7 所示。

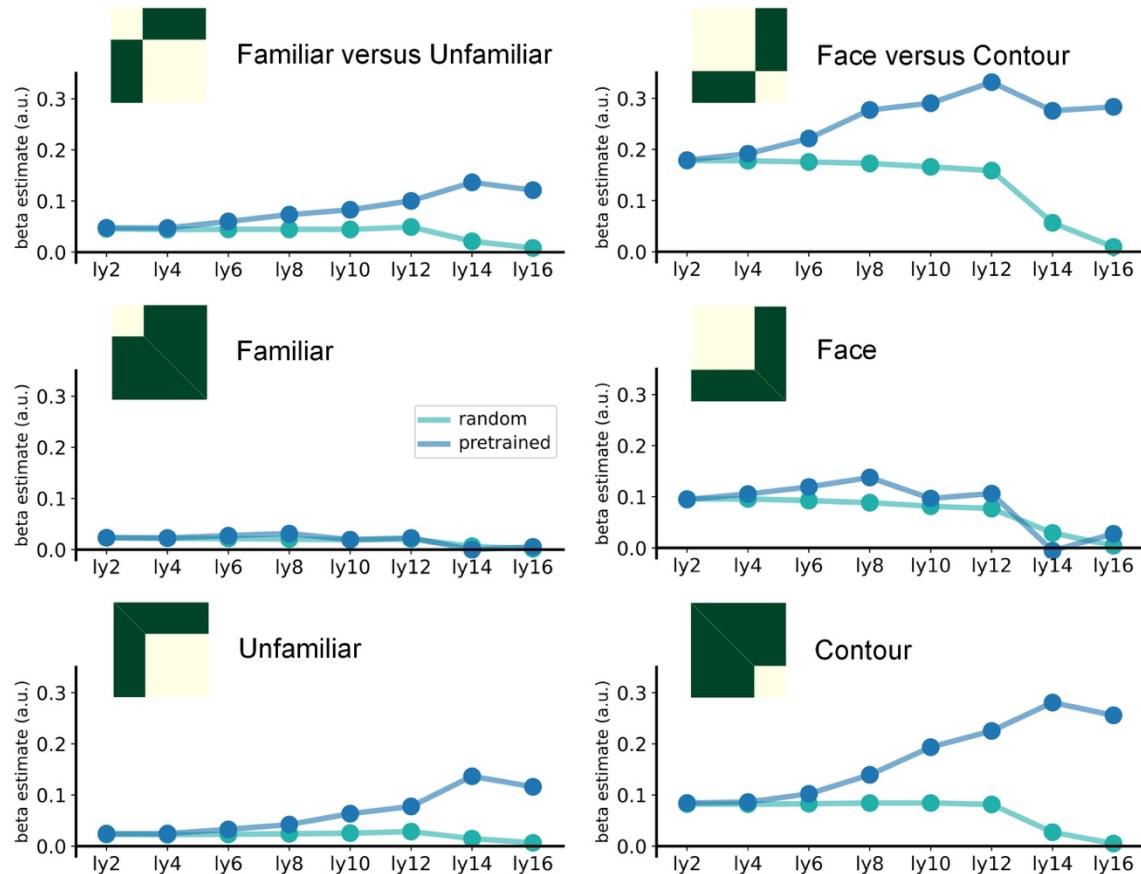


图 4-7 DCNNs 对面孔信息的分层表征

深蓝色代表 VGG-Face 模型的结果，蓝绿色代表随机权重的 VGG-16 模型的结果。

对于熟悉面孔（Familiar）编码模型，两 DCNNs 模型在所有层上的 Beta 估计值都趋近于 0，即两 DCNNs 模型在卷积层和全连接层上均没有对熟悉面孔进行可以编码；对于不熟悉面孔（Unfamiliar）编码模型、面孔熟悉度（Familiar vs. Unfamiliar）编码模型和面孔轮廓（Contour）编码模型，两 DCNNs 模型在第 2 层和第 4 层上的 Beta 估计值接近且都很低，随着层数加深，VGG-Face 模型的 Beta 估计值升高，而随机权重的 VGG-16 模型的 Beta 估计值降低，甚至在全连接最后一层（第 16 层）降为 0，即 VGG-Face 模型对不熟悉面孔、面孔熟悉度和面孔轮廓的编码随着网络层数增高而增强，而随机权重的 VGG-16 模型则对这三种面孔信息的编码随着网络层数增高而减弱。对于完整面孔（Face）编码模型，两 DCNNs 模型在层级上的 Beta 估计值变化比较一致，在卷积

层上的 Beta 估计值都在 0.1 左右的水平,而在全连接层上均降到接近 0 的水平,即 VGG-Face 模型和随机权重的 VGG-16 模型都在卷积层上存在对完整面孔的编码,但在全连接层上不再对这一信息进行刻意编码了。对于面孔完整性 (Face vs. Contour) 编码模型,在 VGG-Face 模型上,其 Beta 估计值始终处于相对较高水平,且在卷积层随着层数增高 Beta 估计值增大,而在全连接层略减小。在随机权重的 VGG-16 模型上,其 Beta 估计值在卷积层相对稳定,略微会随着层数增高而减小,而在全连接层减小明显,并在第 16 层降到接近 0 的水平。即 VGG-Face 模型对面孔完整性的编码随着卷积层的增高而增强,并在全连接层维持着一个稳定较高的水平,而随机权重的 VGG-16 模型在卷积层还存在对面孔完整性的编码,但在全连接层其编码逐渐减弱直至最后一层时不再进行该信息的编码。

最后使用 t-SNE 的方法直观地将 450 张面孔图片投射到基于两 DCNNs 模型早期层和晚期层特征空间的二维空间上,如图 4-8 所示。这里选择第 2 层作为早期层与第 16 层作为晚期层以作比较。

在第 2 层的降维可视化结果中 (如图 4-8a 所示),无论是 VGG-Face 模型的结果还是随机权重的 VGG-16 模型的结果,都能够明显看到蓝色框的乱相面孔的聚集,且与红色框的熟悉面孔和绿色框的不熟悉面孔存在分离,而熟悉面孔与不熟悉面孔则混杂在一起。

而对于第 16 层的降维可视化, VGG-Face 模型的结果 (如图 4-8b 左侧所示) 和随机权重的 VGG-16 模型的结果 (如图 4-8b 右侧所示) 明显存在空间表征模型的差异。在随机权重的 VGG-16 模型的结果中,三种类别的面孔全部混杂在了一起,在二维表征空间上完全无法分离开来。在 VGG-Face 模型的结果中,依旧能够看到乱相面孔的聚集以及与另外两种面孔间的分离。虽然熟悉面孔与不熟悉面孔依然在表征空间上有重叠,但绿框的不熟悉面孔在二维表征空间的左侧存在两块紧密的聚集。并且,仔细观察二维空间上这两块不熟悉面孔的聚簇和周围的熟悉面孔,可以发现偏上的聚簇里的不熟悉面孔和周围的熟悉面孔全为男性面孔,靠下的聚簇里的不熟悉面孔全为女性面孔。

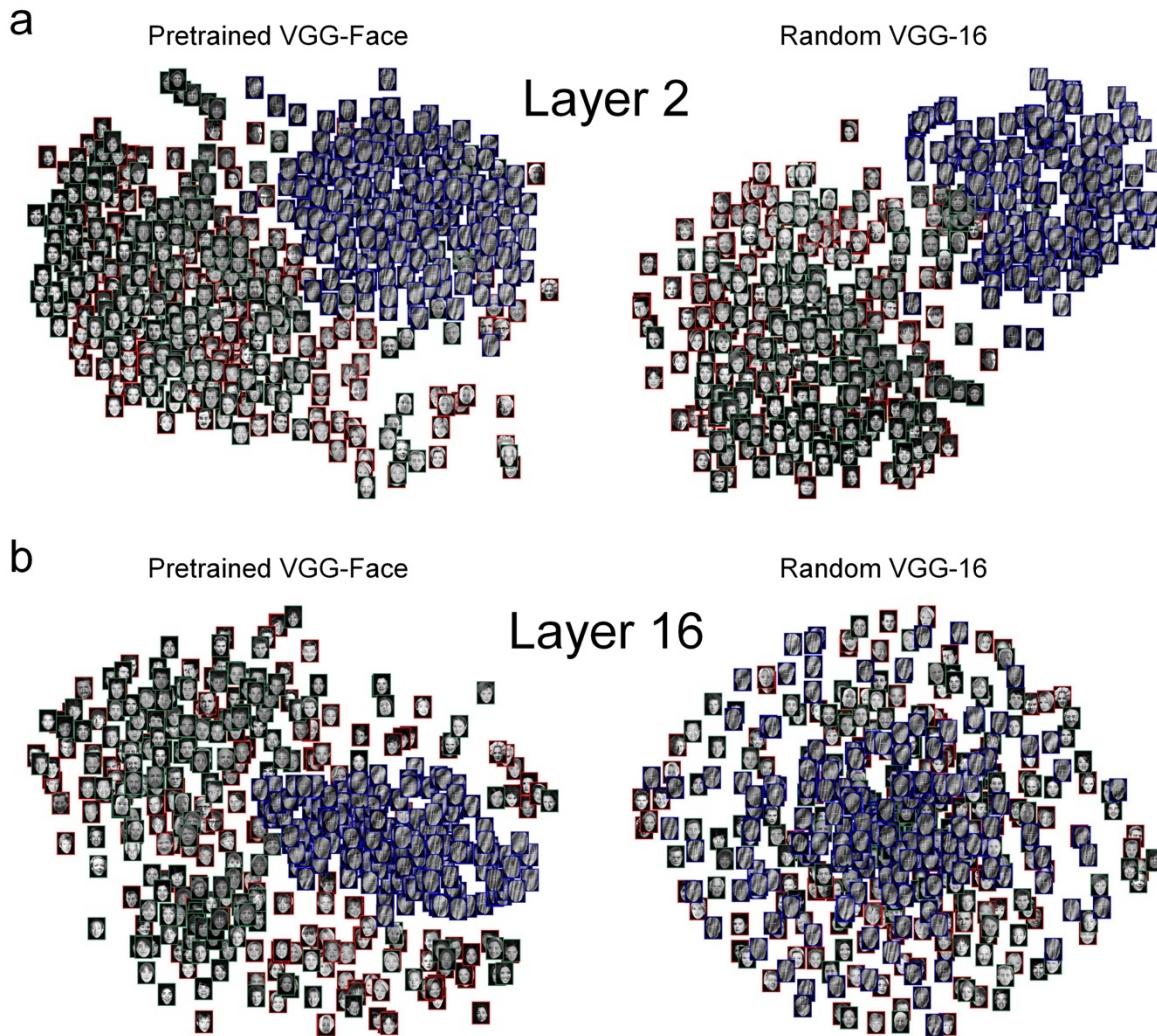


图 4-8 基于 DCNNs 模型的面孔表征可视化结果

(a) 450 张面孔分别在两 DCNNs 模型的第 2 层特征空间上的二维表征可视化结果；(b) 450 张面孔分别在两 DCNNs 模型的第 16 层特征空间上的二维表征可视化结果。左侧为基于 VGG-Face 模型的结果，右侧为基于随机权重的 VGG-16 模型的结果。红色框的面孔为熟悉面孔，绿色框的面孔为不熟悉面孔，蓝色框的面孔为乱相面孔。

4.4 讨论

VGG-Face 模型作为一个面孔识别能力达到人类水平的习得了面孔信息的模型，权重随机的 VGG-16 模型作为一个仅保留网络结构但未经过分类学习的模型。通过输入 450 张面孔图片构建基于两 DCNNs 的 RDMs 能从信息表征的角度来研究 DCNNs 学到了哪些面孔信息、以及这些面孔信息是如何在 DCNN 模型中分层加工的。

对于 DCNNs 对不同面孔类别的内部表征，两 DCNNs 模型都表现出对乱相面孔的

类别内部表征相似性高于不熟悉面孔高于熟悉面孔。乱相面孔的类别内部表征相似性高可能是由于乱相面孔仅包含面孔的轮廓信息，更容易造成内部相互之间的高相关。不熟悉面孔的类别内部表征相似性高于熟悉面孔的类别内部表征相似性，一方面可能说明 DCNNs 对不熟悉面孔有更一致性的表征模式，另一方面也可能说明 DCNNs 对熟悉面孔的加工具有更强的面孔特异性。然而，随机权重的 VGG-16 模型在全连接层对三种类别的面孔的类别内部表征相似性直接降到 0，但 VGG-Face 模型则在全连接层保持了类别内部的表征相似性，并且对乱相面孔的类别内部表征相似性达到了很高的水平。全连接层加工更高维的信息，且直接与面孔识别相关，这说明真正经过面孔数据训练过后的模型在具有面孔识别的表征空间上学习到了对不同类别信息编码的内部一致性模式，且在全连接层的高维信息空间上对乱相面孔之间的表征更一致。而未经过训练的 VGG-16 模型在低维特征信息经过全连接层的变化之后，既没有具备面孔识别的能力、也没有学到对三种面孔类别的一致性加工模式。

逐层比较两 DCNNs 对面孔的表征与编码模型的表征差异，无论是 VGG-Face 模型还是随机权重的 VGG-16 模型与熟悉面孔编码模型的 Beta 估计值都几乎为 0。这一结果说明了 DCNNs 对面孔感知的加工机制与人脑不同，人脑会分别对熟悉面孔信息和不熟悉面孔信息加工，而 DCNNs 没有特异地编码熟悉面孔信息。对于完整面孔信息的编码，两 DCNNs 模型都在卷积层有编码，但在全连接层都不再编码。而对于面孔熟悉度信息、不熟悉面孔信息、面孔完整性信息和面孔轮廓信息，两模型在第 2 层和第 4 层的结果接近，但是在随后的层数上都出现了表征差异。随机权重的 VGG-16 模型都在卷积层上存在一定程度的编码，并在全连接层对这些信息的编码减弱，而 VGG-Face 模型则在卷积层上对这些信息的编码逐层增强并在全连接层维持一个较高水平，这些结果体现了 DCNNs 模型在面孔识别过程中对不同面孔信息的分层编码，且用于面孔识别的 VGG-Face 模型会对面孔熟悉度信息、不熟悉面孔信息、面孔完整性信息和面孔轮廓信息进行特异性的、逐层增加的加工。

通过对 450 张图片在两 DCNNs 模型第 2 层和第 16 层的二维面孔空间表征可视化，能直观地根据面孔在二维空间的聚集看到 VGG-Face 模型对面孔完整性信息和面孔轮廓信息的编码，并且在第 16 层观察到对不熟悉面孔信息的编码，而随机权重的 VGG-16 模型仅在卷积层存在乱相面孔的聚集。意外的是，对 VGG-Face 模型第 16 层的二维可视化投射结果可以发现表征空间上对男性面孔与女性面孔的分离，虽然在数据分析过程

中没有对信息进行刻意加标签以区分，但这也有力地说明在 VGG-Face 的晚期层数也存在对性别信息的编码。

5 探究人脑与深度卷积神经网络模型在面孔感知过程中的表征差异

5.1 分析方法

5.1.1 模型模拟

为了能比较人脑与 DCNN 模型对面孔信息的加工，首先需要对 DCNN 的面孔激活进行修改来模拟人脑对面孔的重复抑制状态下的神经表征。这里建立两种重复抑制模型，一种为衰减模型（Fatigue model），另一种为锐化模型（Sharpening model）。通过修改模型参数来模拟人脑对面孔刺激的立即重复状态和延迟重复状态的表征。

设定面孔图片 p 在 DCNN 的第 1 层的激活向量（进行 PCA 降维前）为 $A = (a_1, a_2, \dots, a_n)$ ，其中激活值由大到小排列，即 $a_1 > a_2 > a_3 > \dots > a_{n-1} > a_n > a_n = a_{n+1} = \dots = a_{n+m} = 0$ ， n 表示 DCNN 的第 1 层有 n 个激活值不为 0 的节点、 m 个激活值为 0（不激活）的节点。

对于衰减模型，设定在重复状态下对刺激激活响应高的节点的激活进行减弱，对刺激激活响应低的节点的激活不变，且激活响应强度越高的节点其衰减也越多。若重复观看图片 p ，则基于衰减模型得到的激活向量为 $A' = (a'_1, a'_2, \dots, a'_n)$ 观看该面孔图片，设定基于衰减模型其激活向量变为 $A' = (a'_1, a'_2, \dots, a'_n)$ ，其内部激活则为：

$$a'_i = \begin{cases} \left(1 - \alpha + \frac{(i-1) \cdot \alpha}{\beta \cdot n}\right) \cdot a_i & (1 \leq i \leq \beta \cdot n) \\ a_i & (\beta \cdot n < i \leq n) \end{cases}$$

其中， α 表示最大衰减系数， β 表示进行衰减的激活值的比例。即前 $\beta \cdot n$ 个激活值在重复观看对应图片时会发生衰减，且激活最强的节点到第 $\beta \cdot n$ 个节点的衰减从 α 到 $\alpha / (\beta \cdot n)$ 等比减弱。

对于锐化模型，设定对刺激激活响应低的节点在重复状态下不再激活，对刺激激活响应高的节点的激活不变。设定重复观看该面孔时基于锐化模型的激活向量变为 $A' = (a'_1, a'_2, \dots, a'_n)$ ，其内部激活则为：

$$a'_i = \begin{cases} a_i & (1 \leq i < (1 - \theta) \cdot n) \\ 0 & ((1 - \theta) \cdot n \leq i \leq n) \end{cases}$$

其中， θ 表示进行锐化的激活值的比例。即后 $\theta \cdot n$ 个激活值在重复观看对应图片时激活值变为 0。

图 5-1 为一示意图，假设仅有 30 个存在激活的节点（激活值不为 0），某张图片直

接获得的激活值按 30 个激活节点从大到小的顺序排列为图 5-1a，而基于衰减模型修改 ($\alpha=0.5$, $\beta=0.9$) 后的重复抑制状态下的激活情况如图 5-1b 所示，基于锐化模型修改 ($\theta=0.3$) 后的重复抑制状态下的激活情况如图 5-2 所示。对于 450 张面孔图片，均分别输入到 VGG-Face 模型和权重随机的 VGG-16 模型中，即可类似地计算所有偶数层的基于两种模型的重复抑制状态下的激活向量。

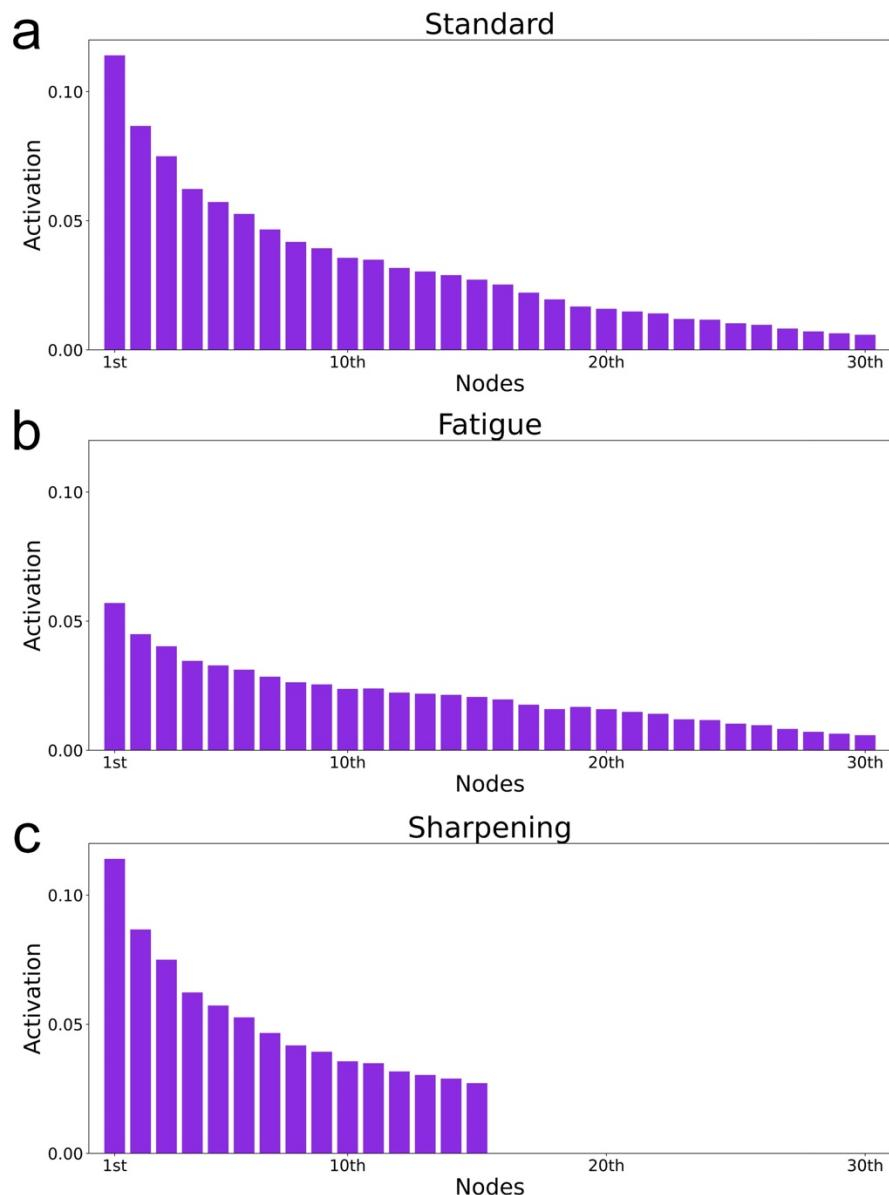


图 5-1 重复抑制的衰减模型与锐化模型示意图

以 20 个激活节点的简化情况作为示例：(a) 标准的激活情况；(b) 基于衰减模型修改后的激活情况，此时衰减模型的参数为： $\alpha=0.5$, $\beta=0.5$ ；(c) 基于锐化模型修改后的激活情况，此时锐化模型的参数为： $\theta=0.5$ 。

5.1.2 基于模型修改的表征不相似矩阵构建

首先基于上述两种重复抑制模型，这里给定三种面孔刺激对应的模型参数：对于衰减模型，Early 和 Late 状态下三种面孔类别条件下的 β 值均设为 0.9，即将按激活强度由高到低排列的前 90% 的激活不为 0 的节点设为激活响应强度高的节点，这 90% 的激活响应强度高的节点发生衰减。并设置 Early 状态下的最大衰减系数 α 为 0.5，Late 状态下的最大衰减系数为 0.05。对于锐化模型，将 Early 状态下的 θ 值设为 0.5，Late 状态下的 θ 值设为 0.05。

因此，每张图片在 DCNN 中对应每一层的输出都可以计算得到三个激活向量，分别对应人面孔感知过程中的 New、Early 和 Late 状态。再将这些向量分别输入到 PCA 降维空间中去，得到一一对应的降维后的特征向量。对于 DCNN 的每一层，450 张面孔图片即可获得 1350 个特征向量。对于同一层得到的特征向量，计算任意两向量之间的不相似性（1-Pearson 相关系数），按 FN、FE、FL、UN、UE、UL、SN、SE 和 SL 的顺序构建 VGG-Face 模型和随机权重的 VGG-16 模型的每一个偶数层的 1350×1350 的 RDM。

此外，根据基于脑电的分类解码部分的结果，New 和 Early 的解码表现在熟悉面孔上最高、在乱相面孔上最低，而 New 和 Late 的解码表现仅在熟悉面孔上高于随机水平。本研究也进一步对不同面孔类别条件修改了两模型的参数，构建了两更精细的模型。对于衰减模型，这里将熟悉面孔、不熟悉面孔和乱相面孔 Early 状态下的 α 分别设为 0.5、0.45 和 0.4，而 Late 状态下的 α 分别设为 0.05、0.01、0.01， β 始终为 0.5。对于锐化模型，Early 状态下熟悉面孔、不熟悉面孔和乱相面孔的 θ 分别 0.5、0.49 和 0.48，而 Late 状态下的 θ 分别设为 0.05、0.01、0.01。对应的分析结果如附图 1 至附图 3 所示，两精细模型得到的结果与上述未对三种面孔类别进行系数区分的模型得到的结果类似。

5.1.3 跨模态的表征分析

5.1.3.1 压缩 VGG-16 的表征不相似性矩阵

对于 VGG-Face 模型和随机权重的 VGG-16 模型，分别可以得到 8 个偶数层的 1350×1350 的 RDMs。对于脑电数据，每个被试可以得到时序上 100 个时间点的 9×9 的 RDMs 及跨时域上的 100×100 个时间点的 9×9 的 CTRDMs。为了建立 DCNN 与人脑之间的联系，首先对 DCNN 的所有 1350×1350 的 RDMs 进行平均操作，对 9 种条件 (FN、FE、FL、UN、UE、UL、SN、SE 和 SL) 下两两匹配计算不相似性的值进行平均，使所有 DCNN 的 RDMs 压缩为 9×9 (如图 5-2 右侧所示)。

5.1.3.2 脑电与 VGG-16 间表征相似性分析

使用 Spearman 相关系数计算神经 RDMs 与深度卷积神经网络模型 RDMs 之间的相似性。如图 5-2 所示，首先，提取由同一时间点数据构成的神经 RDMs，分别去和两个 DCNN 模型的每一个偶数层对应的 RDM 计算相关系数，得到逐时间点的表征相似性结果，即可以得到 2 (VGG-Face 模型和权重随机的 VGG-16 模型) 个 100 (个时间点) \times 8 (个偶数层) 的相似性矩阵。其次，也用神经 CTRDMs 分别和 DCNN 模型的 RDMs 计算相关系数，得到跨时域的表征相似性结果。以上跨模态的 RSA 和 CTRSA 计算都基于 NeuroRA 和 PyCTRSA 实现。

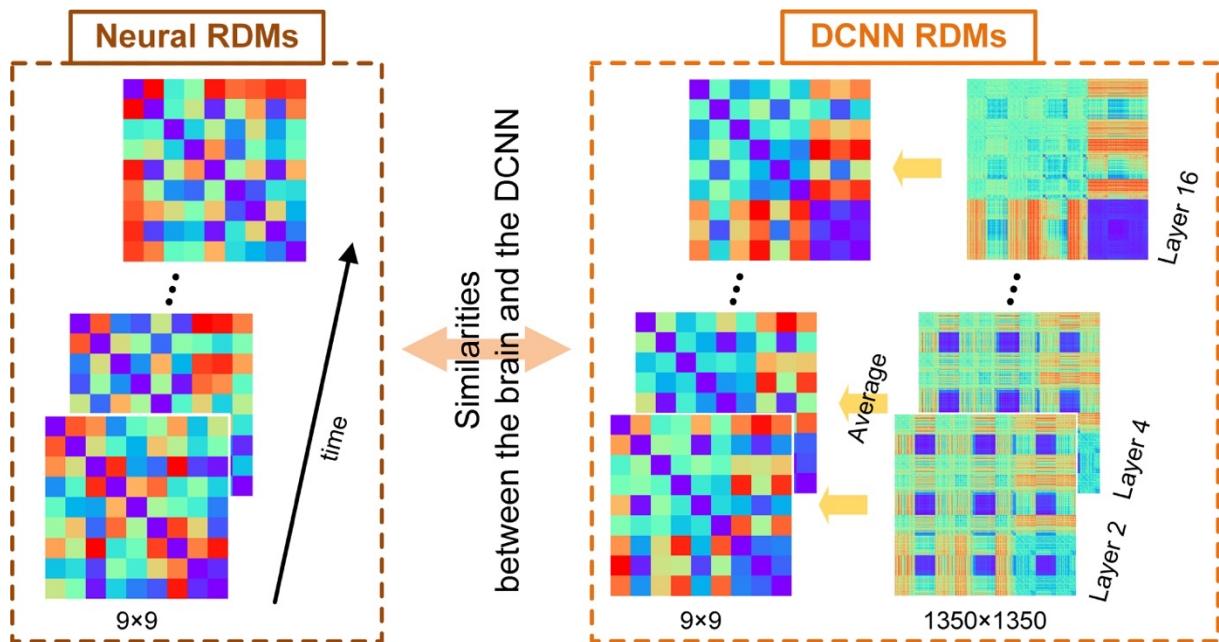


图 5-2 跨模态相似性分析计算示意图

5.1.3.3 统计分析

对于 RSA 结果，如果某一时间内大脑的神经表征与 DCNN 某一层的表征模型相似，则对应的 DCNN 模型的 RDM 与神经 RDM 存在相关，进而比较时序 RDMs 与 DCNN 模型的 RDM 的时序相似性（相关系数）是否显著大于 0。对于每一个时间点，都进行相似性与 0 之间的单样本均值检验得到每一时间点的 p 值，取 $p < 0.05$ 的时间点作为 RSA 的显著性时段。

对于 CTRSA 结果，在 t 检验的基础上进行了基于簇的置换检验。首先提取跨时域结果的每一个显著性簇，然后计算每一个显著性簇内所有 t 值的和，作为该簇的 t 值。然后进行 1000 次置换来计算每一次迭代中拥有最大 t 值的簇的 t 值，从而得到一个最大簇 t 值的分布。最后，对每一个簇来比较其对应 t 值和随机最大簇 t 值的置换分布的显

著性，取前者显著大于后者的簇 ($p<0.05$) 作为最终显著的簇。

5.2 结果

经过上述两重复抑制模型对图片在 DCNN 模型中激活情况的修改，分别计算了 450 张图片输入到 VGG-Face 模型与随机权重的 VGG-16 模型中所有偶数层激活得到的 1350×1350 的 RDMs。图 5-3a 展示了基于衰减模型与基于锐化模型对 VGG-Face 模型和权重随机的 VGG-16 模型的第 16 层激活构建的 RDMs，由于经过大量图片训练的 VGG-Face 其目的是进行面孔识别，而第 16 层网络作为全连接层的最后一层与面孔识别直接相关，并且通过 4.3 节中编码模型与 DCNN 模型的表征比较可知 VGG-Face 第 16 层中包含了对多种面孔信息的编码。这里，首先对 DCNNs 第 16 层在面孔感知过程中的激活进行单独分析。

分别计算两 DCNNs 模型经过两种重复抑制模型修改后第 16 层得到的 RDMs 与基于脑电得到的人脑在面孔感知过程中时序的神经 RDMs 之间的表征相似性，如图 5-3b 所示。结果发现四个 DCNN 模型的 RDMs 都在很长时间范围内与神经 RDMs 存在显著相似，基于衰减模型修改的 VGG-Face 模型其第 16 层的表征在 120-1380ms 与人脑表征模式显著相似，基于衰减模型修改的随机权重的 VGG-16 模型其第 16 层的表征在 120-1360ms 与人脑表征模式显著相似，基于锐化模型修改的 VGG-Face 模型其第 16 层的表征在 120-1060ms 以及 1160-1300ms 与人脑表征模式显著相似，基于锐化模型修改的随机权重的 VGG-16 模型其第 16 层的表征在 120-1260ms 与人脑表征模式显著相似。而对于衰减模型而言，经过面孔数据集训练的 DCNN 模型 (VGG-Face 模型) 其在 140-860ms 的时间范围内与人脑的表征相似性要显著高于未经过训练的 DCNN 模型 (随机权重的 VGG-16 模型)。对于锐化模型而言，经过训练的 DCNN 模型则在 140-360ms 的时间范围内与人脑的表征相似性要显著高于未经过训练的 DCNN 模型。

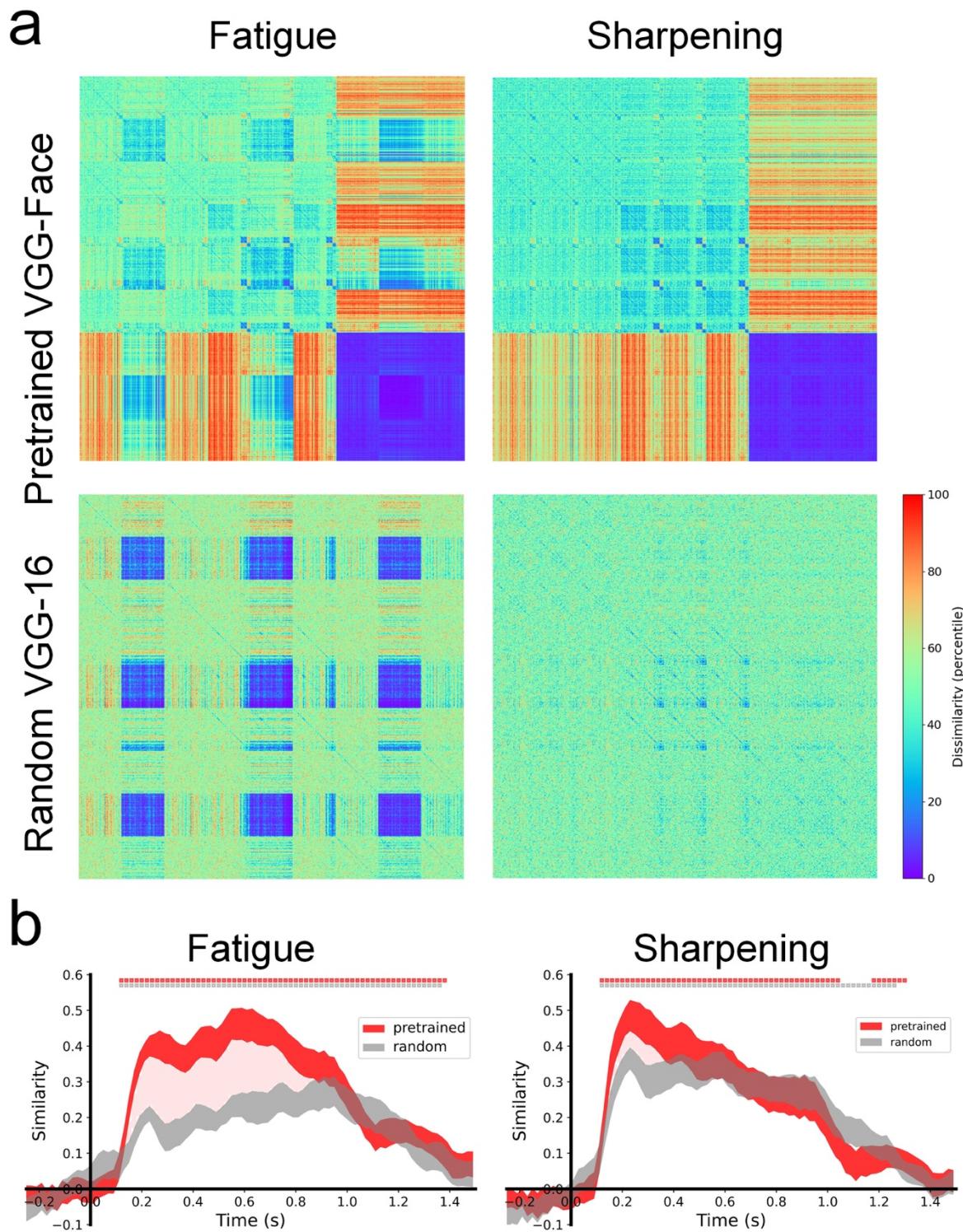


图 5-3 DCNN 模型经过重复抑制模型修改后第 16 层的表征

(a) 左上：基于衰减模型修改的 VGG-Face 模型的第 16 层的 RDM；右上：基于锐化模型修改的 VGG-Face 模型的第 16 层的 RDM；左下：基于衰减模型修改的随机权重的 VGG-16 模型的第 16 层的 RDM；右下：基于锐化模型修改的随机权重的 VGG-16 模型的第 16 层的 RDM。(b) 左侧：基于衰减模型修改的两 DCNNs 第 16 层 RDMs 与神经 RDMs 之间的相似性；右侧：基于锐化模型修改的

两 DCNNs 第 16 层 RDMs 与神经 RDMs 之间的相似性。衰减模型的参数为： $\beta=0.9$ ，Early 状态下 $\alpha=0.5$ ，Late 状态下 $\alpha=0.05$ ；锐化模型的参数为：Early 状态下 $\theta=0.5$ ，Late 状态下 $\theta=0.05$ 。曲线上的红色方块与灰色方块分别代表 VGG-Face 模型与随机权重的 VGG-16 模型与人脑表征间存在显著相似的时间，浅红色阴影区域对应 VGG-Face 模型与人脑表征的相似性显著高于随机权重的 VGG-16 模型与人脑表征的相似性的时间，曲线的宽度反映的是加减一个标准误。

为了获取由于数据训练造成模型学到了用于面孔识别的相应面孔信息的 DCNN 模型与人脑的表征相似性部分，这里将 VGG-Face 模型与人脑间的表征相似性减去权重随机的 VGG-16 模型与人脑间的表征相似性作为真正的有效表征相似性，即：

$$S_{valid} = S_{Pretrained_VGG-Face} - S_{Random_VGG-16}$$

上式中 S_{valid} 为用于面孔识别的 DCNN 模型与人脑的有效表征相似性， $S_{Pretrained_VGG-Face}$ 为经过数据训练的 VGG-Face 模型与人脑的原始表征相似性， S_{Random_VGG-16} 为未经过数据训练的随机权重的 VGG-16 模型与人脑的原始表征相似性。

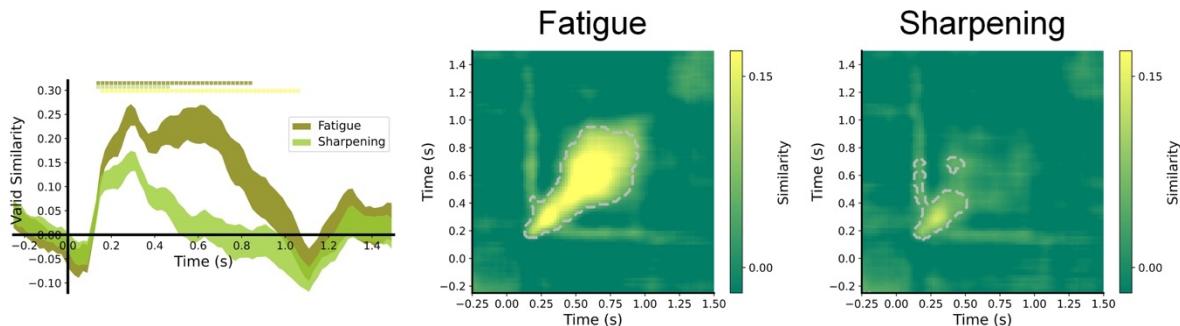


图 5-4 DCNN 模型经过重复抑制模型修改后第 16 层与人脑对面孔的有效表征相似性

左：时序上基于两重复抑制模型得到的 DCNN 与人脑对面孔表征的有效相似性结果；中：跨时域上基于衰减模型得到的 DCNN 与人脑对面孔表征的有效相似性结果；右：跨时域上基于锐化模型得到的 DCNN 与人脑对面孔表征的有效相似性结果。逐时间点结果中：曲线上的橄榄绿色方块、青绿色方块和黄绿色方块分别代表基于衰减模型计算得到有效相似性的显著时间、基于锐化模型计算得到有效相似性的显著时间和前者显著高于后者的时间，曲线的宽度反映的是加减一个标准误。跨时域结果中：灰色虚线勾勒的区域表示与人脑的表征具有显著的有效相似性。

对 DCNN 模型第 16 层的激活计算逐时间的有效表征相似性和跨时间的有效表征相似性，结果如图 5-4 所示。经过衰减模型修改后的 DCNN 在第 16 层的表征与人脑具有更长时间的显著的有效相似性（140-860ms），且这一过程的持续相似性结果存在两个峰值（300ms 左右和 620ms 左右），并在 400ms 左右存在一个谷值，结合跨时域的结果，谷值前后的相似性结果不具有时间持续性，可能可以划分为 140-400ms 和 400-860ms 两

个 DCNN 模型与人脑表征相似的阶段。而经过锐化模型修改后的 DCNN 则在第 16 层的表征与人脑的有效相似性显著时间较短,仅在 140-360ms 存在显著。并且,在 160-1080ms 内衰减模型得到的有效表征相似性结果都要显著高于锐化模型得到的有效表征相似性结果。

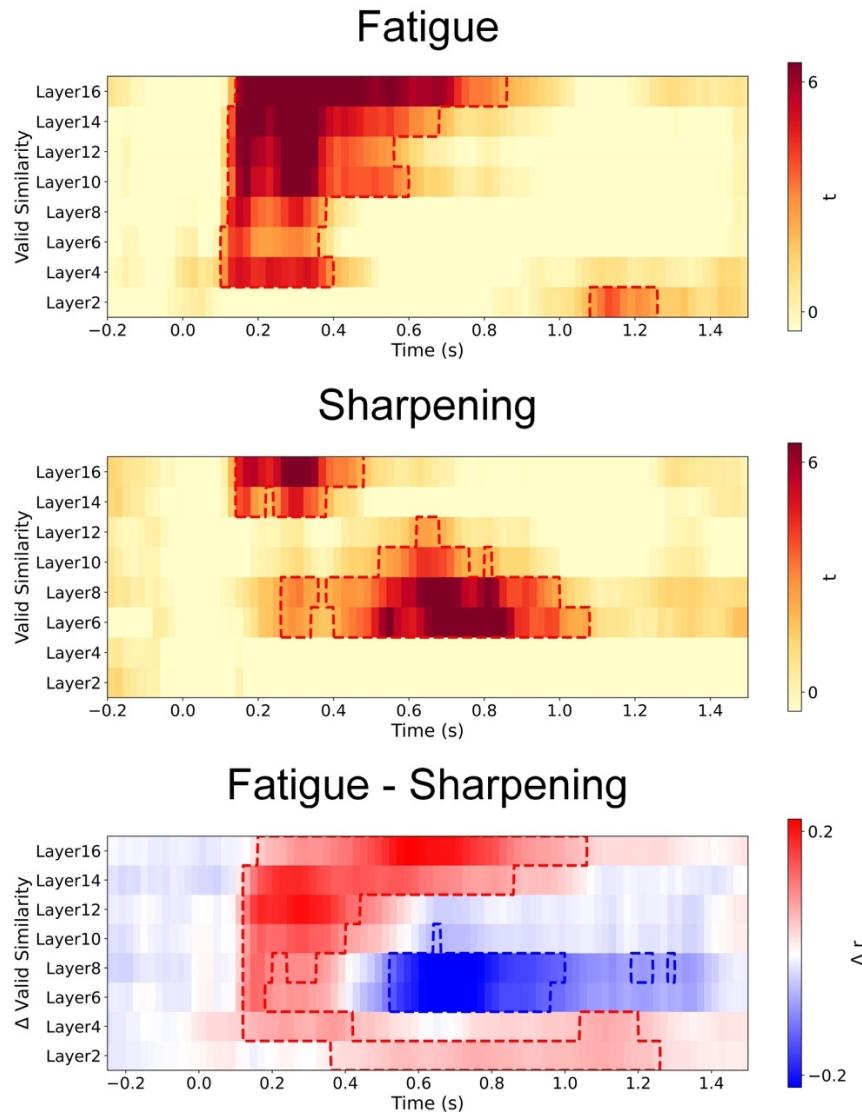


图 5-5 人脑与基于模型修改的 DCNN 模型的时序分层表征相似性

从上到下依次是基于衰减模型修改后的 DCNN 模型与人脑表征相似性结果、基于锐化模型修改后的 DCNN 模型与人脑表征相似性结果以及基于两重复抑制模型修改后跨模态表征的相似性差异的结果。

对应的逐偶数层的 DCNN 模型与人脑的有效表征相似性结果如图 5-5 所示。基于衰减模型改进的 DCNN 在全部偶数层都存在与人脑的显著的表征相似性,从第 2 层到第 16 层的偶数层分别与人脑的面孔表征在 1080-1260ms、100-400ms、100-360ms、120-380ms、120-600ms、120-560ms、120-680ms 和 140-860ms 显著相似。而基于锐化模型改进的

DCNN 在第 6 层到第 16 层的所有参与计算的偶数层都与人脑存在显著表征相似性，第 6 层对应 260-340ms 以及 400-1080ms、第 8 层对应 260-360ms 以及 380-1000ms、第 10 层对应 520-760ms、第 12 对应层 620-680ms、第 14 层对应 140-220ms 以及 240-380ms、第 16 层对应 140-480ms。将经过两重复抑制模型修改对应的表征相似性结果做差，基于衰减模型修改的 DCNN 模型相较基于锐化模型修改的 DCNN 在第 2 层到第 16 层逐偶数层分别与人脑面孔感知过程的 380-1260ms、120-420ms 以及 1020-1200ms、120-180ms、120-200ms 以及 240-320ms、120-400ms、120-440ms、120-860ms、160-1060ms 时段的神经表征之间的相似性更高，反之基于锐化模型修改的 DCNN 在第 6 层和第 8 层分别与人脑在 520-960ms、520-1000ms 以及 1180-1240ms 存在更高的表征相似性。

5.3 讨论

本部分的研究通过对 DCNN 模型的激活进行修改模拟出对图片立即重复状态与延迟重复状态的激活，以构建 1350×1350 的 RDMs，再对表征矩阵进行降维的方式来建立 DCNN 模型与人脑在面孔感知过程中的联系。

以与面孔识别直接关联的 DCNN 模型第 16 层的表征为例，通过直接比较其与人脑在面孔感知过程得到的时序相似性，发现即使是未经过训练的随机权重的 VGG-16 模型也与人脑的表征在整个时序过程中存在显著相关，这一相关很可能是由于经过重复抑制模型对激活向量进行修改后建立的 RDMs 成功拟合出来抑制效应。而真正模型经过面孔训练集学到的信息与人脑直接的表征成分，这里通过用 VGG-Face 模型得到的结果减去随机权重的 VGG-16 模型得到的结果作为模型训练后的 DCNN 与人脑之间的有效表征相似性。基于此有效表征相似性的结果，在第 16 层，经过衰减模型修改的 DCNN 模型的表征与人脑的表征在更长时间上显著相似，跨时域的结果也显示其与人脑的表征模式在更广范围内显著相似。这说明面孔感知过程中的重复抑制效应更有可能是一种衰减机制造成的，而且 DCNN 模型与人脑在 140-860ms 上都存在类似的对面孔信息的编码模式。

对所有偶数层进行基于两重复抑制模型修改后的跨模态有效表征相似性计算，得到 DCNN 模型分层表征与人脑时序表征之间的联系。结果显示基于衰减模型修改的 DCNN 模型与人脑的表征具有一定的层次性，随着 DCNN 的层数加深，其与人脑对面孔感知的显著表征相似性时长也越长。使用大量面孔图片训练的 DCNN 模型随着层数加深，加工的视觉信息也从更低维的视觉特征到更高维的面孔特征，而人脑对面孔的编码也是早期

加工更低维的视觉特征、晚期加工更高维的面孔特征。这一跨模态分层加工结果的发现也与客体识别领域中人脑与 DCNN 模型进行比较得到的分层加工结果类似。

同时，这一结果也再次说明面孔感知过程中，更有可能是一种衰减机制的存在造成了对面孔的重复抑制效应。不过，结果也发现在一些层上基于锐化模型修改的 DCNN 模型与人脑的表征存在一定相似性，这一结果主要集中在卷积层的中间几层，时间范围大约是 500-1000ms 左右。这些结果表明，处理面孔信息的神经元在重复抑制过程中大多发生的是原本激活更强的神经元其激活发生更多衰减，且随着处理的信息维度更高，更多神经元在重复抑制过程中发生这种衰减，而还存在少部分、且主要处理低维面孔信息的神经元，在重复抑制过程中是原本激活更弱的神经元不再激活。

6 结语

6.1 总结与讨论

面孔感知对正常的社会认知功能至关重要，例如面孔提供了一些关键的面部信息，人们用它们来区分一个人、来了解一个人。神经解码与表征分析的方法能突破 ERP 成分作为时序指标的限制，深入了解时序上大脑是如何逐步加工不同的面孔信息有助于深入理解大脑对面孔信息的加工模式。DCNNs 作为工程学上达到与人类类似表现水平的 AI 模型，了解 VGG-Face 是如何在这一分层网络结构中加工面孔信息有利于进一步揭开 AI 的表征本质。而进一步比较人脑与 DCNNs 模型在面孔感知过程中的表征差异，既能将 DCNNs 模型作为一个工具来窥探人脑中可能存在神经机制，也能通过找到生物脑与 AI 脑之间的异同给予未来创造更优的类脑智能模型一些启发。

得益于 Python 强大的科学计算性能与相关数据分析包的迅速发展，本研究开发与设计了两个基于 Python 的可以广泛用于前沿认知神经科学领域研究的工具包——NeuroRA 和 PyCTRSA。前者囊括了几乎各种类型的行为学与神经数据，包括 EEG、MEG、fMRI、fNIRS 和其他电生理数据等等，用户可以设定各种相关参数来仅通过很少的代码实现各种表征分析，包括 NPS、STPS、ISC 和 RSA 等等，以及通过一到两行代码实现对核磁结果的存储、统计分析以及绘图功能。后者则是在传统 RSA 的基础上进行了全新跨时域功能的拓展，用户通过少量代表就可以实现基于 CTRSA 的 EEG/MEG 解码并对结果进行绘制。本研究中随后所有的表征分析部分均基于这两工具包实现。

本研究基于 EEG 数据使用神经解码与 RSA 的方法有效地探究了面孔感知过程中人脑对不同面孔信息的时序加工过程。人脑对面孔熟悉度、熟悉面孔、不熟悉面孔、面孔完整性、完整面孔、面孔轮廓以及刺激状态都有响应的表征，只是存在时序上的动态差异。大脑会在更早（140ms 左右开始）进行视觉相关信息的加工，随后开始加工更复杂的面孔信息，如面孔熟悉度（400ms 左右开始）。本研究也首次直接分离了对熟悉面孔与不熟悉面孔的时序编码差异，对于熟悉面孔，大脑会更早进行加工（180-400ms 左右），而对于不熟悉面孔，大脑会更晚进行加工（500-1220ms 左右）。同时，通过对不同刺激状态的解码以及基于刺激状态模型的 RSA 比较，发现重复效应从 300ms 开始出现，并且对于熟悉面孔的重复效应最强，对于乱相面孔的重复效应最弱。

而对于 DCNN 模型对不同面孔信息的编码情况的探究，发现了许多有趣的结果。经过大量面孔图片训练的 VGG-Face 模型与随机权重的 VGG-16 模型在面孔感知过程中，

其前两层的表征模式相似，随着层数加深，表征差异变大。随机权重的 VGG-16 模型对熟悉面孔、不熟悉面孔和乱相面孔的类别内部表征相似性在全连接层都降为 0，而 VGG-Face 模型对这三种类别的面孔都具有各自一定的相似表征，并且在全连接层对于乱相面孔具有极高的类别内部表征相似性。令人感觉惊讶的是，DCNN 没有特异性编码熟悉面孔。对于其他面孔信息，VGG-Face 和随机权重的 VGG-16 都逐层减弱了对完整面孔信息的编码，而对于面孔熟悉度、不熟悉面孔、面孔完整性和面孔轮廓信息的编码，前者逐层增强，后者逐层减弱。经过学习的 DCNN 模型更多地加工面孔内部的信息，并能特异性区分面孔与非完整面孔。通过对面孔表征的空间可视化，甚至发现 VGG-Face 在基于全连接层的二维表征空间上成功分离了男性面孔与女性面孔，这意味着其编码了面孔的性别信息。

最后，本研究通过构建基于重复抑制效应的模型对 DCNN 模型进行激活修改，以建立人脑与 DCNN 模型之间的表征比较。结果发现 DCNN 模型与人脑在面孔感知过程中从 140-860ms 都存在高度相似性的表征，且经过衰减模型修改的 DCNN 模型与人脑的表征显著性结果要多于经过锐化模型修改的 DCNN 模型与人脑的表征显著性结果。面孔识别过程中，重复抑制效应更有可能是一种衰减机制造成，且经过衰减模型修改的 DCNN 与人脑的表征具有一定的层次性，随着 DCNN 层数增加，其与人脑也在更长时间上具有表征相似性，这种从早期层对应低维视觉特征到晚期层对应更高维面孔特征的结果也符合人脑对面孔刺激早期加工更低维视觉特征、晚期加工更高维面孔特征的发现。这一分层表征的结果与客体识别领域中人脑与 DCNN 进行比较的结果类似(Cichy, et al., 2016; Güçlü & van Gerven, 2015; Kietzmann et al., 2019; Yamins & DiCarlo, 2016)。结合两重复抑制模型修改条件下的结果，人脑与 DCNN 模型之间的相似性表征结果可能表明，更广泛的神经元在重复抑制过程中是原本对面孔刺激产生特异性激活更强的神经元的激活衰减更多，而还存在少部分的神经元中是在重复抑制过程中原本对面孔刺激产生较弱激活的神经元不再激活，且衰减机制对编码更高维信息的神经元群效应更强，锐化机制更可能存在于处理低维信息的神经元群中。

6.2 不足与展望

对于本研究中工具包部分的工作，后期仍然有大量的功能需要添加、改进与优化。比如：(1) 缺少直接读取采集到的原始数据并进行自动化地数据转换的模块，后期考虑添加；(2) 提高 NeuroRA 中迭代计算中的计算效率（尤其是对 fMRI 数据进行分析）；

(3) 增加图形用户界面方便用户使用拖拽式与点击式的功能进行相关计算操作；(4) 扩展绘图模块的功能，包括添加 MDS、t-SNE 的可视化结果以及增加动态表征结果的呈现功能；(5) 增加更完善的用户示例，包括使用更多公共数据集来展示更多函数的使用示例、添加更详细的使用说明；(6) 增加更细致的代码测试模块，当前工具包仅使用了单元测试进行测试；(7) 增加基于分类解码的正确率作为不相似性指标的功能。

对于本研究中人脑表征分析部分的工作，由于受到公共数据集的原实验限制，仍然有很多值得探究的面孔信息没有进行细致深入地分析，例如面孔的一些低维特征，如发型、肤色、面孔呈现的角度、正立或倒置等等，以及面孔的一些高维更复杂的特征，如性别（gender）、身份（identity）、种族（race）、面部表征（facial expression）和可记忆性（memorability）等等。因此，还需要加入类别更多的实验刺激来进一步探究更为丰富而复杂的面孔感知过程。同时，本部分的研究全部基于 EEG 数据，其目的是探究不同面孔信息在时序上的动态表征情况。而今后还可以基于 fMRI 以及 MEG 数据，来进一步探究具体不同面孔信息在大脑中分别由哪些脑区参与了这些信息的加工，并且利用跨模态 RSA 的方法来探究不同脑区对面孔信息的表征在时序上对应的不同编码阶段。

对于本研究中 DCNN 模型表征分析部分的工作，只选取了具有代表性的 VGG 网络结构进行分析，之后的工作还需要进一步测试在 VGG-Face 模型与随机权重的 VGG-16 模型上的结果能否在其他 DCNN 模型中重现。此外，由于 VGG-Face 模型是使用大量面孔数据以面孔身份识别为目的训练得到的，模型在学习过程中没有接受过非面孔刺激的训练，这也是与实际人脑对面孔信息处理上的一些差异。但是由于发现 DCNN 模型不特异性编码熟悉面孔信息，这可能是其与人脑对面孔信息加工方式之间的一个巨大差异。而之后的研究也应该深入探究这一不编码熟悉面孔机制的内在原因，并尝试给予模型对熟悉面孔进行特异性编码以优化与开发进一步的类脑智能模型。

对于本研究中人脑与 DCNN 模型比较部分的工作，首先，由于 DCNN 模型不具有时序处理能力，本研究仅测试了两种重复抑制模型：衰减模型与锐化模型。此外，类似前两部分的不足，使用 fMRI 数据能找到人脑与 DCNN 模型逐层相似性表征对应的脑区变化，而更换不同的网络结构或权重能找到这一跨模态表征相似性对于 AI 模型的差异所在。此外除了 DCNN 模型，循环神经网络（recurrent neural network, RNN）模型与脉冲神经网络（spiking neural network, SNN）模型也被广泛的用于计算机视觉与类脑智能领域，未来还可以进一步比较这些 AI 模型与人脑之间在不同认知过程中对不同信息的

表征差异。由于 AI 模型更多地是基于工程学角度构建出来的，对人脑信息表征的深入探究与跨模态的表征差异的比较有利于给予 AI 模型更多神经科学层面的启发，这将对未来类脑智能领域的发展提供至关重要的启示。

参考文献

- Albers, A. M., Kok, P., Toni, I., Dijkerman, H. C., & De Lange, F. P. (2013). Shared representations for working memory and mental imagery in early visual cortex. *Current Biology*, 23(15), 1427–1431. <https://doi.org/10.1016/j.cub.2013.05.065>
- Alfred, K. L., Connolly, A. C., & Kraemer, D. J. M. (2018). Putting the pieces together: Generating a novel representational space through deductive reasoning. *NeuroImage*, 183, 99–111. <https://doi.org/10.1016/j.neuroimage.2018.07.062>
- Alink, A., Abdulrahman, H., & Henson, R. N. (2018). Forward models demonstrate that repetition suppression is best modelled by local neural scaling. *Nature Communications*, 9(1), 1–10. <https://doi.org/10.1038/s41467-018-05957-0>
- Anzellotti, S., Fairhall, S. L., & Caramazza, A. (2014). Decoding Representations of Face Identity That are Tolerant to Rotation. *Cerebral Cortex*, 24(8), 1988–1995. <https://doi.org/10.1093/cercor/bht046>
- Avery, J. A., Liu, A. G., Ingeholm, J. E., Gotts, S. J., & Martin, A. (2021). Viewing images of foods evokes taste quality-specific activity in gustatory insular cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 118(2). <https://doi.org/10.1073/pnas.2010932118>
- Baddeley, A. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience*, 4, 829–839. <https://doi.org/10.1038/nrn1201>
- Bae, G. Y., & Luck, S. J. (2018). Dissociable decoding of spatial attention and working memory from EEG oscillations and sustained potentials. *Journal of Neuroscience*, 38(2), 409–422. <https://doi.org/10.1523/JNEUROSCI.2860-17.2017>
- Bae, G. Y., & Luck, S. J. (2019a). Decoding motion direction using the topography of sustained ERPs and alpha oscillations. *NeuroImage*, 184, 242–255. <https://doi.org/10.1016/j.neuroimage.2018.09.029>
- Bae, G. Y., & Luck, S. J. (2019b). Reactivation of Previous Experiences in a Working Memory Task. *Psychological Science*, 30(4), 587–595. <https://doi.org/10.1177/0956797619830398>
- Bainbridge, W. A., Dilks, D. D., & Oliva, A. (2017). Memorability: A stimulus-driven

- perceptual neural signature distinctive from memory. *NeuroImage*, 149, 141–152.
<https://doi.org/10.1016/j.neuroimage.2017.01.063>
- Bainbridge, W. A., Hall, E. H., & Baker, C. I. (2021). Distinct Representational Structure and Localization for Visual Encoding and Recall during Visual Imagery. *Cerebral Cortex*, 31(4), 1898–1913. <https://doi.org/10.1093/cercor/bhaa329>
- Bainbridge, W. A., & Rissman, J. (2018). Dissociating neural markers of stimulus memorability and subjective recognition during episodic retrieval. *Scientific Reports*, 8(1), 1–11. <https://doi.org/10.1038/s41598-018-26467-5>
- Ban, H., Preston, T. J., Meeson, A., & Welchman, A. E. (2012). The integration of motion and disparity cues to depth in dorsal visual cortex. *Nature Neuroscience*, 15, 636–643. <https://doi.org/10.1038/nn.3046>
- Bankson, B. B., Hebart, M. N., Groen, I. I. A., & Baker, C. I. (2018). The temporal evolution of conceptual object representations revealed through models of behavior, semantics and deep neural networks. *NeuroImage*, 178, 172–182.
<https://doi.org/10.1016/j.neuroimage.2018.05.037>
- Bannert, M. M., & Bartels, A. (2013). Decoding the yellow of a gray banana. *Current Biology*, 23(22), 2268–2272. <https://doi.org/10.1016/j.cub.2013.09.016>
- Barrett, S. E., & Rugg, M. D. (1989). Event-related potentials and the semantic matching of faces. *Neuropsychologia*, 27(7), 913–922. [https://doi.org/10.1016/0028-3932\(89\)90067-5](https://doi.org/10.1016/0028-3932(89)90067-5)
- Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science*, 364(6439). <https://doi.org/10.1126/science.aav9436>
- Baylis, G. C., & Rolls, E. T. (1987). Responses of neurons in the inferior temporal cortex in short term and serial recognition memory tasks. *Experimental Brain Research*, 65(3), 614–622. <https://doi.org/10.1007/BF00235984>
- Begleiter, H., Porjesz, B., & Wang, W. (1995). Event-related brain potentials differentiate priming and recognition to familiar and unfamiliar faces. *Electroencephalography and Clinical Neurophysiology*, 94(1), 41–49. [https://doi.org/10.1016/0013-4694\(94\)00240-L](https://doi.org/10.1016/0013-4694(94)00240-L)
- Bentin, S., & Deouell, L. Y. (2000). Structural encoding and identification in face processing:

- ERP evidence for separate mechanisms. *Cognitive Neuropsychology*, 17(1–3), 35–55.
<https://doi.org/10.1080/026432900380472>
- Bentin, S., McCarthy, G., & Wood, C. C. (1985). Event-related potentials, lexical decision and semantic priming. *Electroencephalography and Clinical Neurophysiology*, 60(4), 343–355. [https://doi.org/10.1016/0013-4694\(85\)90008-2](https://doi.org/10.1016/0013-4694(85)90008-2)
- Bo, K., Yin, S., Liu, Y., Hu, Z., Meyyapan, S., Andreas, K., & Ding, M. (2021). Decoding Multivoxel Representations of Affective Scenes in Retinotopic Visual Cortex. *Cerebral Cortex*, 31(4), 1898–1913. <https://doi.org/10.1093/cercor/bhaa329>
- Bocincova, A., & Johnson, J. S. (2019). The time course of encoding and maintenance of task-relevant versus irrelevant object features in working memory. *Cortex*, 111, 196–209. <https://doi.org/10.1016/j.cortex.2018.10.013>
- Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, 77(3), 305–327. <https://doi.org/10.1111/j.2044-8295.1986.tb02199.x>
- Burkhardt, A., Blaha, L. M., Jurs, B. S., Rhodes, G., Jeffery, L., Wyatte, D., Delong, J., & Busey, T. (2010). Adaptation modulates the electrophysiological substrates of perceived facial distortion: Support for opponent coding. *Neuropsychologia*, 48(13), 3743–3756. <https://doi.org/10.1016/j.neuropsychologia.2010.08.016>
- Cai, Y., Sheldon, A. D., Yu, Q., & Postle, B. R. (2019). Overlapping and distinct contributions of stimulus location and of spatial context to nonspatial visual short-term memory. *Journal of Neurophysiology*, 121(4), 1222–1231.
<https://doi.org/10.1152/jn.00062.2019>
- Carlson, J. M., Cha, J., & Mujica-Parodi, L. R. (2013). Functional and structural amygdala - Anterior cingulate connectivity correlates with attentional bias to masked fearful faces. *Cortex*, 49(9), 2595–2600. <https://doi.org/10.1016/j.cortex.2013.07.008>
- Cauchoix, M., Barragan-Jason, G., Serre, T., & Barbeau, E. J. (2014). The neural dynamics of face detection in the wild revealed by MVPA. *Journal of Neuroscience*, 34(3), 846–854.
<https://doi.org/10.1523/JNEUROSCI.3030-13.2014>
- Cavanagh, S. E., Towers, J. P., Wallis, J. D., Hunt, L. T., & Kennerley, S. W. (2018). Reconciling persistent and dynamic hypotheses of working memory coding in prefrontal

- cortex. *Nature Communications*, 9(1), 1–16. <https://doi.org/10.1038/s41467-018-05873-3>
- Charles, C. G., & Sergent, J. (1992). Face recognition. *Current Opinion in Neurobiology*, 2(2), 156–161. [https://doi.org/10.1016/0959-4388\(92\)90004-5](https://doi.org/10.1016/0959-4388(92)90004-5)
- Christophe, T. B., Hebart, M. N., & Haynes, J. D. (2012). Decoding the contents of visual short-term memory from human visual and parietal cortex. *Journal of Neuroscience*, 32(38), 12983–12989. <https://doi.org/10.1523/JNEUROSCI.0184-12.2012>
- Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6(1), 1–13.
<https://doi.org/10.1038/srep27755>
- Cichy, R. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature Neuroscience*, 17(3), 455–462. <https://doi.org/10.1038/nn.3635>
- Cichy, R. M., Pantazis, D., & Oliva, A. (2016). Similarity-Based Fusion of MEG and fMRI Reveals Spatio-Temporal Dynamics in Human Cortex During Visual Object Recognition. *Cerebral Cortex*, 26(8), 3563–3579. <https://doi.org/10.1093/cercor/bhw135>
- Crouzet, S. M., Kirchner, H., & Thorpe, S. J. (2010). Fast saccades toward faces: Face detection in just 100 ms. *Journal of Vision*, 10(4), 1–17. <https://doi.org/10.1167/10.4.16>
- Davidescu, I., Zion-Golumbic, E., Bickel, S., Harel, M., Groppe, D. M., Keller, C. J., Schevon, C. A., McKhann, G. M., Goodman, R. R., Goelman, G., Schroeder, C. E., Mehta, A. D., & Malach, R. (2014). Exemplar Selectivity Reflects Perceptual Similarities in the Human Fusiform Cortex. *Cerebral Cortex*, 24(7), 1879–1893.
<https://doi.org/10.1093/cercor/bht038>
- Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, 134(1), 9–21. <https://doi.org/10.1016/j.jneumeth.2003.10.009>
- Dobs, K., Isik, L., Pantazis, D., & Kanwisher, N. (2019). How face perception unfolds over time. *Nature Communications*, 10(1), 1–10. <https://doi.org/10.1038/s41467-019-09239-1>
- Dörr, P., Herzmann, G., & Sommer, W. (2011). Multiple contributions to priming effects for

- familiar faces: Analyses with backward masking and event-related potentials. *British Journal of Psychology*, 102(4), 765–782. <https://doi.org/10.1111/j.2044-8295.2011.02028.x>
- Drisdelle, B. L., Aubin, S., & Jolicoeur, P. (2017). Dealing with ocular artifacts on lateralized ERPs in studies of visual-spatial attention and memory: ICA correction versus epoch rejection. *Psychophysiology*, 54(1), 83–99. <https://doi.org/10.1111/psyp.12675>
- Ester, E. F., Anderson, D. E., Serences, J. T., & Awh, E. (2013). A neural measure of precision in visual working memory. *Journal of Cognitive Neuroscience*, 25(5), 754–761. https://doi.org/10.1162/jocn_a_00357
- Ester, E. F., Sprague, T. C., & Serences, J. T. (2015). Parietal and Frontal Cortex Encode Stimulus-Specific Mnemonic Representations during Visual Working Memory. *Neuron*, 87(4), 893–905. <https://doi.org/10.1016/j.neuron.2015.07.013>
- Etzel, J. A., Courtney, Y., Carey, C. E., Gehred, M. Z., Agrawal, A., & Braver, T. S. (2020). Pattern Similarity Analyses of FrontoParietal Task Coding: Individual Variation and Genetic Influences. *Cerebral Cortex*, 30(5), 3167–3183.
<https://doi.org/10.1093/cercor/bhz301>
- Fahrenfort, J. J., Grubert, A., Olivers, C. N. L., & Eimer, M. (2017). Multivariate EEG analyses support high-resolution tracking of feature-based attentional selection. *Scientific Reports*, 7, 1886. <https://doi.org/10.1038/s41598-017-01911-0>
- Farah, M. J., Wilson, K. D., Drain, M., & Tanaka, J. N. (1998). What is “Special” about Face Perception? *Psychological Review*, 105(3), 482–498. <https://doi.org/10.1037/0033-295X.105.3.482>
- Feng, C., Yan, X., Huang, W., Han, S., & Ma, Y. (2018). Neural representations of the multidimensional self in the cortical midline structures. *NeuroImage*, 183, 291–299.
<https://doi.org/10.1016/j.neuroimage.2018.08.018>
- Freiwald, W. A., Tsao, D. Y., & Livingstone, M. S. (2009). A face feature space in the macaque temporal lobe. *Nature Neuroscience*, 12(9), 1187–1196.
<https://doi.org/10.1038/nn.2363>
- Freiwald, W., Duchaine, B., & Yovel, G. (2016). Face Processing Systems: From Neurons to

- Real-World Social Perception. *Annual Review of Neuroscience*, 39(1), 325–346.
<https://doi.org/10.1146/annurev-neuro-070815-013934>
- Ghuman, A. S., Brunet, N. M., Li, Y., Konecky, R. O., Pyles, J. A., Walls, S. A., Destefino, V., Wang, W., & Richardson, R. M. (2014). Dynamic encoding of face information in the human fusiform gyrus. *Nature Communications*, 5(1), 1–10.
<https://doi.org/10.1038/ncomms6672>
- Goesaert, E., & Op de Beeck, H. P. (2013). Representations of facial identity information in the ventral visual stream investigated with multivoxel pattern analyses. *Journal of Neuroscience*, 33(19), 8549–8558. <https://doi.org/10.1523/JNEUROSCI.1829-12.2013>
- Gosseries, O., Yu, Q., Larocque, J. J., Starrett, M. J., Rose, N. S., Cowan, N., & Postle, B. R. (2018). Parietal-occipital interactions underlying control-and representation-related processes in working memory for nonspatial visual features. *Journal of Neuroscience*, 38(18), 4357–4366. <https://doi.org/10.1523/JNEUROSCI.2747-17.2018>
- Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., & Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, 7(7 DEC), 267.
<https://doi.org/10.3389/fnins.2013.00267>
- Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Parkkonen, L., & Hämäläinen, M. S. (2014). MNE software for processing MEG and EEG data. *NeuroImage*, 86, 446–460. <https://doi.org/10.1016/j.neuroimage.2013.10.027>
- Greene, M. R., & Hansen, B. C. (2018). Shared spatiotemporal category representations in biological and artificial deep neural networks. *PLOS Computational Biology*, 14(7), e1006327. <https://doi.org/10.1371/journal.pcbi.1006327>
- Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: Neural models of stimulus-specific effects. In Trends in Cognitive Sciences (Vol. 10, Issue 1, pp. 14–23). Elsevier Current Trends. <https://doi.org/10.1016/j.tics.2005.11.006>
- Grootswagers, T., Robinson, A. K., Shatek, S. M., & Carlson, T. A. (2019). Untangling featural and conceptual object representations. *NeuroImage*, 202, 116083.
<https://doi.org/10.1016/j.neuroimage.2019.116083>

- 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. <https://doi.org/10.1523/JNEUROSCI.5023-14.2015>
- Hall-McMaster, S., Muhle-Karbe, P. S., Myers, N. E., & Stokes, M. G. (2019). Reward Boosts Neural Coding of Task Rules to Optimize Cognitive Flexibility. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, 39(43), 8549–8561. <https://doi.org/10.1523/JNEUROSCI.0631-19.2019>
- Hancock, P. J. B., Burton, A. M., & Bruce, V. (1996). Face processing: Human perception and principal components analysis. *Memory and Cognition*, 24(1), 26–40. <https://doi.org/10.3758/BF03197270>
- Harrison, S. A., & Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature*, 458, 632–635. <https://doi.org/10.1038/nature07832>
- Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject Synchronization of Cortical Activity during Natural Vision. *Science*, 303(5664), 1634–1640. <https://doi.org/10.1126/science.1089506>
- Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539), 2425–2430. <https://doi.org/10.1126/science.1063736>
- Haxby, James V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. In Trends in Cognitive Sciences (Vol. 4, Issue 6, pp. 223–233). Elsevier Current Trends. [https://doi.org/10.1016/S1364-6613\(00\)01482-0](https://doi.org/10.1016/S1364-6613(00)01482-0)
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-December, 770–778. <https://doi.org/10.1109/CVPR.2016.90>
- Henson, R. N. A., & Rugg, M. D. (2003). Neural response suppression, haemodynamic repetition effects, and behavioural priming. *Neuropsychologia*, 41(3), 263–270. [https://doi.org/10.1016/S0028-3932\(02\)00159-8](https://doi.org/10.1016/S0028-3932(02)00159-8)
- Herzmann, G., Schweinberger, S. R., Sommer, W., & Jentzsch, I. (2004). What's special about personally familiar faces? A multimodal approach. *Psychophysiology*, 41(5), 688–

701. <https://doi.org/10.1111/j.1469-8986.2004.00196.x>
- Hogendoorn, H., & Burkitt, A. N. (2018). Predictive coding of visual object position ahead of moving objects revealed by time-resolved EEG decoding. *NeuroImage*, 171, 55–61.
<https://doi.org/10.1016/j.neuroimage.2017.12.063>
- Hong, X., Bo, K., Meyyappan, S., Tong, S., & Ding, M. (2020). Decoding attention control and selection in visual spatial attention. *Human Brain Mapping*, 41(14), 3900–3921.
<https://doi.org/10.1002/hbm.25094>
- Hu, G., Yang, Y., Yi, D., Kittler, J., Christmas, W., Li, S. Z., & Hospedales, T. (2016). When Face Recognition Meets with Deep Learning: An Evaluation of Convolutional Neural Networks for Face Recognition. *Proceedings of the IEEE International Conference on Computer Vision*, 2016-Febru, 384–392. <https://doi.org/10.1109/ICCVW.2015.58>
- Huang, G. B., Lee, H., & Learned-Miller, E. (2012). Learning hierarchical representations for face verification with convolutional deep belief networks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2518–2525.
<https://doi.org/10.1109/CVPR.2012.6247968>
- Huddy, V., Schweinberger, S. R., Jentzsch, I., & Burton, A. M. (2003). Matching faces for semantic information and names: An event-related brain potentials study. *Cognitive Brain Research*, 17(2), 314–326. [https://doi.org/10.1016/S0926-6410\(03\)00131-9](https://doi.org/10.1016/S0926-6410(03)00131-9)
- Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science*, 310(5749), 863–866.
<https://doi.org/10.1126/science.1117593>
- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science and Engineering*, 9(3), 90–95. <https://doi.org/10.1109/MCSE.2007.55>
- Itier, R. J., & Taylor, M. J. (2002). Inversion and contrast polarity reversal affect both encoding and recognition processes of unfamiliar faces: A repetition study using ERPs. *NeuroImage*, 15(2), 353–372. <https://doi.org/10.1006/nimg.2001.0982>
- Itz, M. L., Schweinberger, S. R., Schulz, C., & Kaufmann, J. M. (2014). Neural correlates of facilitations in face learning by selective caricaturing of facial shape or reflectance. *NeuroImage*, 102(P2), 736–747. <https://doi.org/10.1016/j.neuroimage.2014.08.042>

- Jaegle, A., Mehrpour, V., Mohsenzadeh, Y., Meyer, T., Oliva, A., & Rust, N. (2019). Population response magnitude variation in inferotemporal cortex predicts image memorability. *ELife*, 8. <https://doi.org/10.7554/eLife.47596>
- Johnson, M. R., & Johnson, M. K. (2014). Decoding individual natural scene representations during perception and imagery. *Frontiers in Human Neuroscience*, 8(1 FEB), 59. <https://doi.org/10.3389/fnhum.2014.00059>
- Jung, T. P., Makeig, S., Westerfield, M., Townsend, J., Courchesne, E., & Sejnowski, T. J. (2000). Removal of eye activity artifacts from visual event-related potentials in normal and clinical subjects. *Clinical Neurophysiology*, 111(10), 1745–1758. [https://doi.org/10.1016/S1388-2457\(00\)00386-2](https://doi.org/10.1016/S1388-2457(00)00386-2)
- Kalfas, I., Vinken, K., & Vogels, R. (2018). Representations of regular and irregular shapes by deep Convolutional Neural Networks, monkey inferotemporal neurons and human judgments. *PLOS Computational Biology*, 14(10), e1006557. <https://doi.org/10.1371/journal.pcbi.1006557>
- Kaliukhovich, D. A., & Vogels, R. (2011). Stimulus Repetition Probability Does Not Affect Repetition Suppression in Macaque Inferior Temporal Cortex. *Cerebral Cortex*, 21(7), 1547–1558. <https://doi.org/10.1093/cercor/bhq207>
- Kaliukhovich, D. A., & Vogels, R. (2012). Stimulus repetition affects both strength and synchrony of macaque inferior temporal cortical activity. *Journal of Neurophysiology*, 107(12), 3509–3527. <https://doi.org/10.1152/jn.00059.2012>
- Kaneshiro, B., Perreau Guimaraes, M., Kim, H.-S., Norcia, A. M., & Suppes, P. (2015). A Representational Similarity Analysis of the Dynamics of Object Processing Using Single-Trial EEG Classification. *PLOS ONE*, 10(8), e0135697. <https://doi.org/10.1371/journal.pone.0135697>
- Kaufmann, J. M., Schulz, C., & Schweinberger, S. R. (2013). High and low performers differ in the use of shape information for face recognition. *Neuropsychologia*, 51(7), 1310–1319. <https://doi.org/10.1016/j.neuropsychologia.2013.03.015>
- Kaufmann, J. M., & Schweinberger, S. R. (2012). The faces you remember: Caricaturing shape facilitates brain processes reflecting the acquisition of new face representations.

- Biological Psychology, 89(1), 21–33. <https://doi.org/10.1016/j.biopsycho.2011.08.011>
- Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, 98(3), 630-644.e16. <https://doi.org/10.1016/j.neuron.2018.03.044>
- Kietzmann, T. C., Spoerer, C. J., Sørensen, 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 of the United States of America*, 116(43), 21854–21863. <https://doi.org/10.1073/pnas.1905544116>
- Kloth, N., & Schweinberger, S. R. (2010). Electrophysiological correlates of eye gaze adaptation. *Journal of Vision*, 10(12), 1–13. <https://doi.org/10.1167/10.12.17>
- Kloth, N., Schweinberger, S. R., & Kovács, G. (2010). Neural correlates of generic versus gender-specific face adaptation. *Journal of Cognitive Neuroscience*, 22(10), 2345–2356. <https://doi.org/10.1162/jocn.2009.21329>
- Koch, G. E., Paulus, J. P., & Coutanche, M. N. (2020). Neural Patterns are More Similar across Individuals during Successful Memory Encoding than during Failed Memory Encoding. *Cerebral Cortex* (New York, N.Y. : 1991), 30(7), 3872–3883. <https://doi.org/10.1093/cercor/bhaa003>
- Koenig-Robert, R., & Pearson, J. (2019). Decoding the contents and strength of imagery before volitional engagement. *Scientific Reports*, 9, 3504. <https://doi.org/10.1038/s41598-019-39813-y>
- Kovács, G., Zimmer, M., Bankó, É., Harza, I., Antal, A., & Vidnyánszky, Z. (2006). Electrophysiological Correlates of Visual Adaptation to Faces and Body Parts in Humans. *Cerebral Cortex*, 16(5), 742–753. <https://doi.org/10.1093/cercor/bhj020>
- Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2, 4. <https://doi.org/10.3389/NEURO.06.004.2008>
- Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., & Bandettini, P. A. (2008). Matching Categorical Object Representations in Inferior

- Temporal Cortex of Man and Monkey. *Neuron*, 60(6), 1126–1141.
<https://doi.org/10.1016/j.neuron.2008.10.043>
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 25.
- Kubilius, J., Bracci, S., & Op de Beeck, H. P. (2016). Deep Neural Networks as a Computational Model for Human Shape Sensitivity. *PLoS Computational Biology*, 12(4), e1004896. <https://doi.org/10.1371/journal.pcbi.1004896>
- Kuzovkin, I., Vicente, R., Petton, M., Lachaux, J. P., Baciu, M., Kahane, P., Rheims, S., Vidal, J. R., & Aru, J. (2018). Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex. *Communications Biology*, 1(1), 1–12. <https://doi.org/10.1038/s42003-018-0110-y>
- Latinus, M., & Taylor, M. J. (2006). Face processing stages: Impact of difficulty and the separation of effects. *Brain Research*, 1123(1), 179–187.
<https://doi.org/10.1016/j.brainres.2006.09.031>
- LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2323.
<https://doi.org/10.1109/5.726791>
- Lehky, S. R. (2000). Fine discrimination of faces can be performed rapidly. *Journal of Cognitive Neuroscience*, 12(5), 848–855. <https://doi.org/10.1162/089892900562453>
- Lescroart, M. D., & Gallant, J. L. (2019). Human Scene-Selective Areas Represent 3D Configurations of Surfaces. *Neuron*, 101(1), 178-192.e7.
<https://doi.org/10.1016/j.neuron.2018.11.004>
- Li, H., Lin, Z., Shen, X., Brandt, J., & Hua, G. (2015). A convolutional neural network cascade for face detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June, 5325–5334.
<https://doi.org/10.1109/CVPR.2015.7299170>
- Liu, H., Agam, Y., Madsen, J. R., & Kreiman, G. (2009). Timing, Timing, Timing: Fast Decoding of Object Information from Intracranial Field Potentials in Human Visual

- Cortex. Neuron, 62(2), 281–290. <https://doi.org/10.1016/j.neuron.2009.02.025>
- Liu, S., Yu, Q., Tse, P. U., & Cavanagh, P. (2019). Neural Correlates of the Conscious Perception of Visual Location Lie Outside Visual Cortex. Current Biology, 29(23), 4036-4044.e4. <https://doi.org/10.1016/j.cub.2019.10.033>
- Long, N. M., & Kuhl, B. A. (2019). Decoding the tradeoff between encoding and retrieval to predict memory for overlapping events. NeuroImage, 201, 116001.
<https://doi.org/10.1016/j.neuroimage.2019.07.014>
- Lu, Y., Wang, C., Chen, C., & Xue, G. (2015). Spatiotemporal neural pattern similarity supports episodic memory. Current Biology, 25(6), 780–785.
<https://doi.org/10.1016/j.cub.2015.01.055>
- Mares, I., Ewing, L., Farran, E. K., Smith, F. W., & Smith, M. L. (2020). Developmental changes in the processing of faces as revealed by EEG decoding. NeuroImage, 211, 116660. <https://doi.org/10.1016/j.neuroimage.2020.116660>
- Matsugu, M., Mori, K., Mitari, Y., & Kaneda, Y. (2003). Subject independent facial expression recognition with robust face detection using a convolutional neural network. Neural Networks, 16(5–6), 555–559. [https://doi.org/10.1016/S0893-6080\(03\)00115-1](https://doi.org/10.1016/S0893-6080(03)00115-1)
- Maurer, U., Rossion, B., & McCandliss, B. D. (2008). Category specificity in early perception: face and word N170 responses differ in both lateralization and habituation properties. Frontiers in Human Neuroscience, 2(DEC), 18.
<https://doi.org/10.3389/neuro.09.018.2008>
- Mercure, E., Kadosh, K. C., & Johnson, M. H. (2011). The N170 Shows Differential Repetition Effects for Faces, Objects, and Orthographic Stimuli. Frontiers in Human Neuroscience, 5(JANUARY), 1–10. <https://doi.org/10.3389/fnhum.2011.00006>
- Miller, E. K., Gochin, P. M., & Gross, C. G. (1991). Habituation-like decrease in the responses of neurons in inferior temporal cortex of the macaque. Visual Neuroscience, 7(4), 357–362. <https://doi.org/10.1017/S0952523800004843>
- Miller, E. K., Li, L., & Desimone, R. (1991). A neural mechanism for working and recognition memory in inferior temporal cortex. Science, 254(5036), 1377–1379.
<https://doi.org/10.1126/science.1962197>

- Muukkonen, I., Ölander, K., Numminen, J., & Salmela, V. R. (2020). Spatio-temporal dynamics of face perception. *NeuroImage*, 209, 116531.
<https://doi.org/10.1016/j.neuroimage.2020.116531>
- Nee, D. E., Brown, J. W., Askren, M. K., Berman, M. G., Demiralp, E., Krawitz, A., & Jonides, J. (2013). A meta-Analysis of executive components of working memory. *Cerebral Cortex*, 23(2), 264–282. <https://doi.org/10.1093/cercor/bhs007>
- Nemrodov, D., Niemeier, M., Mok, J. N. Y., & Nestor, A. (2016). The time course of individual face recognition: A pattern analysis of ERP signals. *NeuroImage*, 132, 469–476. <https://doi.org/10.1016/j.neuroimage.2016.03.006>
- Nemrodov, D., Niemeier, M., Patel, A., & Nestor, A. (2018). The neural dynamics of facial identity processing: Insights from EEG-based pattern analysis and image reconstruction. *ENeuro*, 5(1). <https://doi.org/10.1523/ENEURO.0358-17.2018>
- Nestor, A., Plaut, D. C., & Behrmann, M. (2011). Unraveling the distributed neural code of facial identity through spatiotemporal pattern analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 108(24), 9998–10003.
<https://doi.org/10.1073/pnas.1102433108>
- Noah, S., Powell, T., Khodayari, N., Olivan, D., Ding, M., & Mangun, G. R. (2020). Neural Mechanisms of Attentional Control for Objects: Decoding EEG Alpha When Anticipating Faces, Scenes, and Tools. *Journal of Neuroscience*, 40(25), 4913–4924.
<https://doi.org/10.1523/JNEUROSCI.2685-19.2020>
- Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. In *Trends in Cognitive Sciences* (Vol. 10, Issue 9, pp. 424–430). Elsevier Current Trends.
<https://doi.org/10.1016/j.tics.2006.07.005>
- Parde, C. J., Castillo, C., Hill, M. Q., Colon, Y. I., Sankaranarayanan, S., Chen, J.-C., & O'Toole, A. J. (2017). Face and image representation in deep cnn features. 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition.
<https://doi.org/10.1109/FG.2017.85>
- Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep Face Recognition. *Proceedings of*

- the British Machine Vision Conference 2015, 41.1-41.12. <https://doi.org/10.5244/c.29.41>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M.,
Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Cournapeau, D., Passos, A.,
Brucher, M., Perrot Andéouardand'andéouard Duchesnay, M., & Perrot, M. (2011).
Scikit-learn: Machine Learning in Python. In Machine Learning in Python. Journal of
Machine Learning Research (Vol. 12). Micromote Publishing. <https://hal.inria.fr/hal-00650905v2>
- Pfütze, E. M., Sommer, W., & Schweinberger, S. R. (2002). Age-related slowing in face and
name recognition: Evidence from event-related brain potentials. *Psychology and Aging*,
17(1), 140–160. <https://doi.org/10.1037/0882-7974.17.1.140>
- Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K., Cavazos, J. G.,
Jeckeln, G., Ranjan, R., Sankaranarayanan, S., Chen, J. C., Castillo, C. D., Chellappa, R.,
White, D., & O'Toole, A. J. (2018). Face recognition accuracy of forensic examiners,
superrecognizers, and face recognition algorithms. *Proceedings of the National Academy
of Sciences of the United States of America*, 115(24), 6171–6176.
<https://doi.org/10.1073/pnas.1721355115>
- Ponce, C. R., Xiao, W., Schade, P. F., Hartmann, T. S., Kreiman, G., & Livingstone, M. S.
(2019). Evolving Images for Visual Neurons Using a Deep Generative Network Reveals
Coding Principles and Neuronal Preferences. *Cell*, 177(4), 999-1009.e10.
<https://doi.org/10.1016/j.cell.2019.04.005>
- Popal, H., Wang, Y., & Olson, I. R. (2019). A Guide to Representational Similarity Analysis
for Social Neuroscience. *Social Cognitive and Affective Neuroscience*, 14(11), 1243–
1253. <https://doi.org/10.1093/scan/nsz099>
- Rahmani Del Bakhshayesh, A., Annabi, N., Khalilov, R., Akbarzadeh, A., Samiei, M.,
Alizadeh, E., Alizadeh-Ghodsi, M., Davaran, S., & Montaseri, A. (2018). Recent
advances on biomedical applications of scaffolds in wound healing and dermal tissue
engineering. In *Artificial Cells, Nanomedicine and Biotechnology* (Vol. 46, Issue 4, pp.
691–705). Taylor and Francis Ltd. <https://doi.org/10.1080/21691401.2017.1349778>
- Ranjan, R., Sankaranarayanan, S., Castillo, C. D., & Chellappa, R. (2017). An All-In-One

- Convolutional Neural Network for Face Analysis. Proceedings - 12th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2017 - 1st International Workshop on Adaptive Shot Learning for Gesture Understanding and Production, ASL4GUP 2017, Biometrics in the Wild, Bwild 2017, Heteroge, 17–24.
<https://doi.org/10.1109/FG.2017.137>
- Reddy, L., Tsuchiya, N., & Serre, T. (2010). Reading the mind's eye: Decoding category information during mental imagery. *NeuroImage*, 50(2), 818–825.
<https://doi.org/10.1016/j.neuroimage.2009.11.084>
- Ringo, J. L. (1996). Stimulus specific adaptation in inferior temporal and medial temporal cortex of the monkey. *Behavioural Brain Research*, 76(1–2), 191–197.
[https://doi.org/10.1016/0166-4328\(95\)00197-2](https://doi.org/10.1016/0166-4328(95)00197-2)
- Robinson, A. K., Grootswagers, T., & Carlson, T. A. (2019). The influence of image masking on object representations during rapid serial visual presentation. *NeuroImage*, 197, 224–231. <https://doi.org/10.1016/j.neuroimage.2019.04.050>
- Rose, N. S., LaRocque, J. J., Riggall, A. C., Gosseries, O., Starrett, M. J., Meyering, E. E., & Postle, B. R. (2016). Reactivation of latent working memories with transcranial magnetic stimulation. *Science*, 354(6316), 1136–1139. <https://doi.org/10.1126/science.aah7011>
- Rossoni, B., & Caharel, S. (2011). ERP evidence for the speed of face categorization in the human brain: Disentangling the contribution of low-level visual cues from face perception. *Vision Research*, 51(12), 1297–1311.
<https://doi.org/10.1016/j.visres.2011.04.003>
- Rugg, M. D. (1985). The Effects of Semantic Priming and Word Repetition on Event-Related Potentials. *Psychophysiology*, 22(6), 642–647. <https://doi.org/10.1111/j.1469-8986.1985.tb01661.x>
- Sawamura, H., Orban, G. A., & Vogels, R. (2006). Selectivity of neuronal adaptation does not match response selectivity: A single-cell study of the fMRI adaptation paradigm. *Neuron*, 49(2), 307–318. <https://doi.org/10.1016/j.neuron.2005.11.028>
- Schlegel, A., Kohler, P. J., Fogelson, S. V., Alexander, P., Konuthula, D., & Tse, P. U. (2013). Network structure and dynamics of the mental workspace. *Proceedings of the*

- National Academy of Sciences of the United States of America, 110(40), 16277–16282.
<https://doi.org/10.1073/pnas.1311149110>
- Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 07-12-June, 815–823.
<https://doi.org/10.1109/CVPR.2015.7298682>
- Schulz, C., Kaufmann, J. M., Kurt, A., & Schweinberger, S. R. (2012). Faces forming traces: Neurophysiological correlates of learning naturally distinctive and caricatured faces. NeuroImage, 63(1), 491–500. <https://doi.org/10.1016/j.neuroimage.2012.06.080>
- Schweinberger, S. R. (1996). How gorbachev primed yeltsin: Analyses of associative priming in person recognition by means of reaction times and event-related brain potentials. Journal of Experimental Psychology: Learning Memory and Cognition, 22(6), 1383–1407. <https://doi.org/10.1037/0278-7393.22.6.1383>
- Schweinberger, S. R., & Burton, A. M. (2003). Covert recognition and the neural system for face processing. In Cortex (Vol. 39, Issue 1, pp. 9–30). Masson SpA.
[https://doi.org/10.1016/S0010-9452\(08\)70071-6](https://doi.org/10.1016/S0010-9452(08)70071-6)
- Schweinberger, S. R., Kloth, N., & Jenkins, R. (2007). Are you looking at me? Neural correlates of gaze adaptation. NeuroReport, 18(7), 693–696.
<https://doi.org/10.1097/WNR.0b013e3280c1e2d2>
- Schweinberger, S. R., & Neumann, M. F. (2016). Repetition effects in human ERPs to faces. In Cortex (Vol. 80, pp. 141–153). Masson SpA.
<https://doi.org/10.1016/j.cortex.2015.11.001>
- Schweinberger, S. R., Pfütze, E. M., & Sommer, W. (1995). Repetition Priming and Associative Priming of Face Recognition: Evidence From Event-Related Potentials. Journal of Experimental Psychology: Learning, Memory, and Cognition, 21(3), 722–736.
<https://doi.org/10.1037/0278-7393.21.3.722>
- Serences, J. T., Ester, E. F., Vogel, E. K., & Awh, E. (2009). Stimulus-specific delay activity in human primary visual cortex. Psychological Science, 20(2), 207–214.
<https://doi.org/10.1111/j.1467-9280.2009.02276.x>

- Shatek, S. M., Grootswagers, T., Robinson, A. K., & Carlson, T. A. (2019). Decoding images in the mind's eye: The temporal dynamics of visual imagery. *Vision (Switzerland)*, 3(4), 53. <https://doi.org/10.3390/vision3040053>
- Simonyan, K., Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2013). Fisher vector faces in the wild. *BMVC 2013 - Electronic Proceedings of the British Machine Vision Conference* 2013. <https://doi.org/10.5244/C.27.8>
- Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *Computer Vision and Pattern Recognition (CVPR)*. <https://arxiv.org/abs/1409.1556>
- Smith, F. W., & Smith, M. L. (2019). Decoding the dynamic representation of facial expressions of emotion in explicit and incidental tasks. *NeuroImage*, 195, 261–271. <https://doi.org/10.1016/j.neuroimage.2019.03.065>
- Sobotka, S., & Ringo, J. L. (1994). Stimulus specific adaptation in excited but not in inhibited cells in inferotemporal cortex of Macaque. *Brain Research*, 646(1), 95–99. [https://doi.org/10.1016/0006-8993\(94\)90061-2](https://doi.org/10.1016/0006-8993(94)90061-2)
- Sprague, T. C., Ester, E. F., & Serences, J. T. (2016). Restoring Latent Visual Working Memory Representations in Human Cortex. *Neuron*, 91(3), 694–707. <https://doi.org/10.1016/j.neuron.2016.07.006>
- Stevenage, S. V., Hale, S., Morgan, Y., & Neil, G. J. (2014). Recognition by association: Within- and cross-modality associative priming with faces and voices. *British Journal of Psychology*, 105(1), 1–16. <https://doi.org/10.1111/bjop.12011>
- Stokes, M., Thompson, R., Cusack, R., & Duncan, J. (2009). Top-down activation of shape-specific population codes in visual cortex during mental imagery. *Journal of Neuroscience*, 29(5), 1565–1572. <https://doi.org/10.1523/JNEUROSCI.4657-08.2009>
- Sun, Y., Wang, X., & Tang, X. (2014). Deep learning face representation from predicting 10,000 classes. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1891–1898. <https://doi.org/10.1109/CVPR.2014.244>
- Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). DeepFace: Closing the gap to human-level performance in face verification. *Proceedings of the IEEE Computer*

- Society Conference on Computer Vision and Pattern Recognition, 1701–1708.
<https://doi.org/10.1109/CVPR.2014.220>
- Tanaka, J. W., & Curran, T. (2001). A neural basis for expert object recognition. *Psychological Science*, 12(1), 43–47. <https://doi.org/10.1111/1467-9280.00308>
- Tanaka, J. W., Curran, T., Porterfield, A. L., & Collins, D. (2006). Activation of preexisting and acquired face representations: The N250 event-related potential as an index of face familiarity. *Journal of Cognitive Neuroscience*, 18(9), 1488–1497.
<https://doi.org/10.1162/jocn.2006.18.9.1488>
- Tanaka, J. W., & Farah, M. J. (1993). Parts and Wholes in Face Recognition. *The Quarterly Journal of Experimental Psychology Section A*, 46(2), 225–245.
<https://doi.org/10.1080/14640749308401045>
- Teichmann, L., Quek, G. L., Robinson, A. K., Grootswagers, T., Carlson, T. A., & Rich, A. N. (2020). The Influence of Object-Color Knowledge on Emerging Object Representations in the Brain. *Journal of Neuroscience*, 40(35), 6779–6789.
<https://doi.org/10.1523/JNEUROSCI.0158-20.2020>
- Tsao, D. Y., & Livingstone, M. S. (2008). Mechanisms of face perception. In *Annual Review of Neuroscience* (Vol. 31, Issue 1, pp. 411–437). Annual Reviews.
<https://doi.org/10.1146/annurev.neuro.30.051606.094238>
- Urgen, B. A., Pehlivan, S., & Saygin, A. P. (2019). Distinct representations in occipito-temporal, parietal, and premotor cortex during action perception revealed by fMRI and computational modeling. *Neuropsychologia*, 127, 35–47.
<https://doi.org/10.1016/j.neuropsychologia.2019.02.006>
- Van de Nieuwenhuijzen, M. E., Backus, A. R., Bahramisharif, A., Doeller, C. F., Jensen, O., & van Gerven, M. A. J. (2013). MEG-based decoding of the spatiotemporal dynamics of visual category perception. *NeuroImage*, 83, 1063–1073.
<https://doi.org/10.1016/j.neuroimage.2013.07.075>
- Van Der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: A structure for efficient numerical computation. *Computing in Science and Engineering*, 13(2), 22–30.
<https://doi.org/10.1109/MCSE.2011.37>

- Vetter, P., Smith, F. W., & Muckli, L. (2014). Decoding sound and imagery content in early visual cortex. *Current Biology*, 24(11), 1256–1262.
<https://doi.org/10.1016/j.cub.2014.04.020>
- Vida, M. D., Nestor, A., Plaut, D. C., & Behrmann, M. (2017). Spatiotemporal dynamics of similarity-based neural representations of facial identity. *Proceedings of the National Academy of Sciences of the United States of America*, 114(2), 388–393.
<https://doi.org/10.1073/pnas.1614763114>
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., ... Vázquez-Baeza, Y. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3), 261–272. <https://doi.org/10.1038/s41592-019-0686-2>
- Wakeman, D. G., & Henson, R. N. (2015). A multi-subject, multi-modal human neuroimaging dataset. *Scientific Data*, 2(1), 150001. <https://doi.org/10.1038/sdata.2015.1>
- Walther, C., Schweinberger, S. R., Kaiser, D., & Kovács, G. (2013). Neural correlates of priming and adaptation in familiar face perception. *Cortex*, 49(7), 1963–1977.
<https://doi.org/10.1016/j.cortex.2012.08.012>
- Wang, Y., Wang, P., & Yu, Y. (2018). Decoding English Alphabet Letters Using EEG Phase Information. *Frontiers in Neuroscience*, 12(FEB), 62.
<https://doi.org/10.3389/fnins.2018.00062>
- Weiner, K. S., & Grill-Spector, K. (2015). The evolution of face processing networks. In *Trends in Cognitive Sciences* (Vol. 19, Issue 5, pp. 240–241). Elsevier Ltd.
<https://doi.org/10.1016/j.tics.2015.03.010>
- Weiner, K. S., Sayres, R., Vinberg, J., & Grill-Spector, K. (2010). fMRI-adaptation and category selectivity in human ventral temporal cortex: Regional differences across time scales. *Journal of Neurophysiology*, 103(6), 3349–3365.
<https://doi.org/10.1152/jn.01108.2009>
- Wiese, H., Kachel, U., & Schweinberger, S. R. (2013). Holistic face processing of own- and

- other-age faces in young and older adults: ERP evidence from the composite face task. NeuroImage, 74, 306–317. <https://doi.org/10.1016/j.neuroimage.2013.02.051>
- Wolff, M. J., Ding, J., Myers, N. E., & Stokes, M. G. (2015). Revealing hidden states in visual working memory using electroencephalography. *Frontiers in Systems Neuroscience*, 9(september), 123. <https://doi.org/10.3389/fnsys.2015.00123>
- Wolff, M. J., Jochim, J., Akyürek, E. G., & Stokes, M. G. (2017). Dynamic hidden states underlying working-memory-guided behavior. *Nature Neuroscience*, 20, 864–871. <https://doi.org/10.1038/nn.4546>
- Xie, S., Kaiser, D., & Cichy, R. M. (2020). Visual Imagery and Perception Share Neural Representations in the Alpha Frequency Band. *Current Biology*, 30(13), 2621-2627.e5. <https://doi.org/10.1016/j.cub.2020.04.074>
- Xing, Y., Ledgeway, T., McGraw, P. V., & Schluppeck, D. (2013). Decoding working memory of stimulus contrast in early visual cortex. *Journal of Neuroscience*, 33(25), 10301–10311. <https://doi.org/10.1523/JNEUROSCI.3754-12.2013>
- Xu, S., Zhang, Y., Zhen, Z., & Liu, J. (2020). The face module emerged in a deep convolutional neural network selectively deprived of face experience. *BioRxiv*, 2020.07.06.189407. <https://doi.org/10.1101/2020.07.06.189407>
- Xue, G., Dong, Q., Chen, C., Lu, Z., Mumford, J. A., & Poldrack, R. A. (2010). Greater neural pattern similarity across repetitions is associated with better memory. *Science*, 330(6000), 97–101. <https://doi.org/10.1126/science.1193125>
- Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. In *Nature Neuroscience* (Vol. 19, Issue 3, pp. 356–365). Nature Publishing Group. <https://doi.org/10.1038/nn.4244>
- 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 of the United States of America*, 111(23), 8619–8624. <https://doi.org/10.1073/pnas.1403112111>
- Yokoi, A., & Diedrichsen, J. (2019). Neural Organization of Hierarchical Motor Sequence Representations in the Human Neocortex. *Neuron*, 103(6), 1178-1190.e7.

<https://doi.org/10.1016/j.neuron.2019.06.017>

Young, A. W., Hellawell, D., & Hay, D. C. (1987). Configurational information in face perception. *Perception*, 16, 747–759. <https://doi.org/10.1068/p160747n>

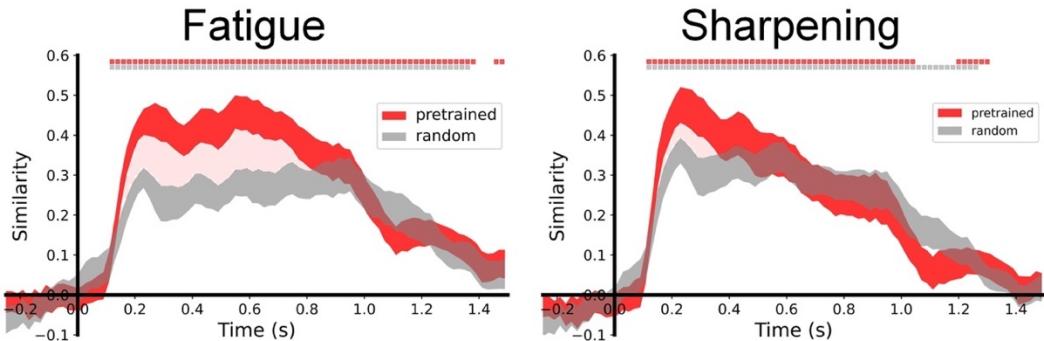
Yu, Z., & Zhang, C. (2015). Image based static facial expression recognition with multiple deep network learning. *ICMI 2015 - Proceedings of the 2015 ACM International Conference on Multimodal Interaction*, 435–442.

<https://doi.org/10.1145/2818346.2830595>

Zheng, X., Mondloch, C. J., & Segalowitz, S. J. (2012). The timing of individual face recognition in the brain. *Neuropsychologia*, 50(7), 1451–1461.

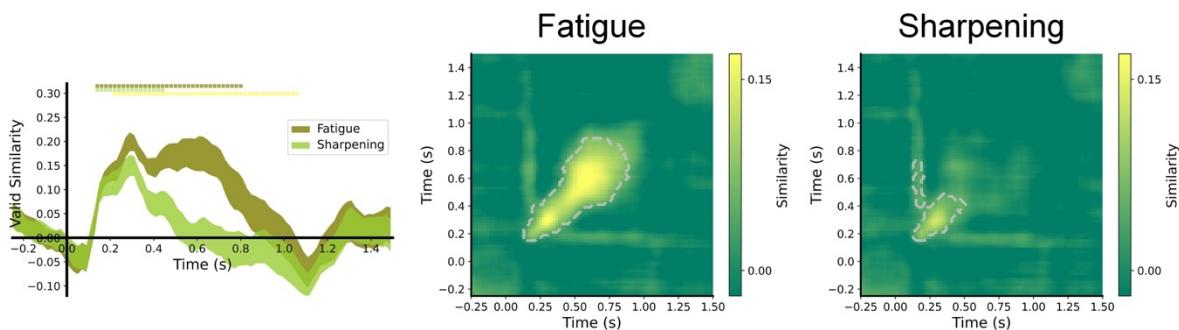
<https://doi.org/10.1016/j.neuropsychologia.2012.02.030>

附录



附图 1 DCNN 模型经过两参数精细化重复抑制模型修改后第 16 层的表征

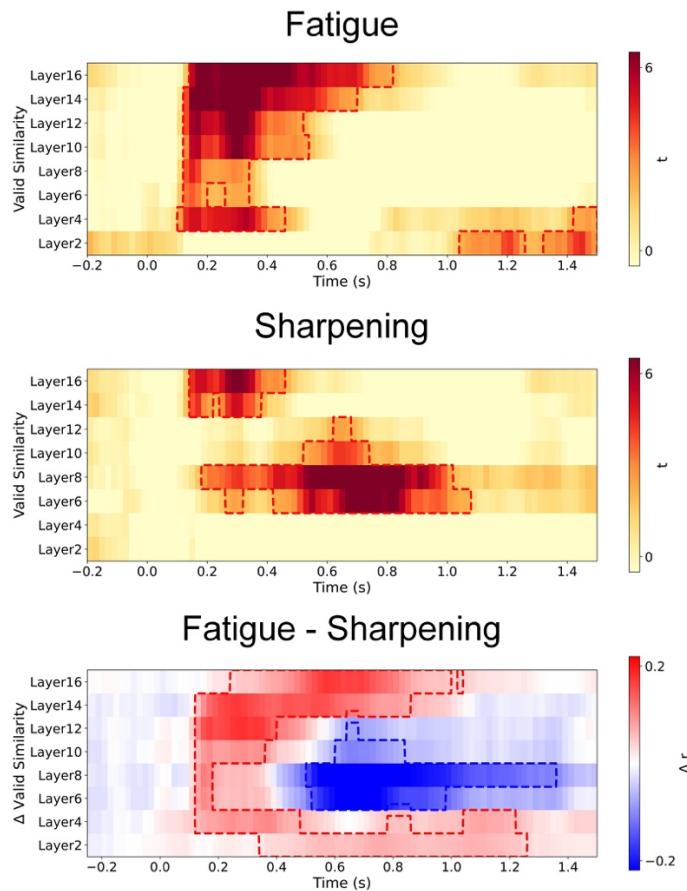
左侧：基于衰减模型修改的两 DCNNs 第 16 层 RDMs 与神经 RDMs 之间的相似性；右侧：基于锐化模型修改的两 DCNNs 第 16 层 RDMs 与神经 RDMs 之间的相似性。衰减模型的参数为： $\beta=0.9$ ，FE 条件下 $\alpha=0.5$ ，UE 条件下 $\alpha=0.45$ ，SE 条件下 $\alpha=0.4$ ，FL 条件下 $\alpha=0.05$ ，UL 条件下 $\alpha=0.01$ ，SL 条件下 $\alpha=0.01$ ；锐化模型的参数为：FE 条件下 $\theta=0.5$ ，LE 条件下 $\theta=0.49$ ，SE 条件下 $\theta=0.48$ ，FL 条件下 $\theta=0.05$ ，UL 条件下 $\theta=0.01$ ，SL 条件下 $\theta=0.01$ 。曲线上的红色方块与灰色方块分别代表 VGG-Face 模型与随机权重的 VGG-16 模型与人脑表征间存在显著相似的时间，浅红色阴影区域对应 VGG-Face 模型与人脑表征的相似性显著高于随机权重的 VGG-16 模型与人脑表征的相似性的时间，曲线的宽度反映的是加减一个标准误。



附图 2 DCNN 模型经过参数精细化模型修改后第 16 层与人脑对面孔的有效表征相似性

左：时序上基于两重复抑制模型得到的 DCNN 与人脑对面孔的有效表征相似性结果；中：跨时域上基于衰减模型得到的 DCNN 与人脑对面孔的有效表征相似性结果；右：跨时域上基于锐化模型得到的 DCNN 与人脑对面孔的有效表征相似性结果。逐时间点结果中：曲线上的橄榄绿色方块、青绿色方块和黄绿色方块分别代表基于衰减模型计算得到有效相似性的显著时间、基于锐化模型计算得到有效相似性的显著时间和前者显著高于后者的时间，曲线的宽度反映的是加减一个标准误。跨时域结果中：灰色虚线勾勒的区域表示与人脑的表征具有显著的有效相似性。衰减模型的参数为： $\beta=0.9$ ，FE 条件下 $\alpha=0.5$ ，UE 条件下 $\alpha=0.45$ ，SE 条件下 $\alpha=0.4$ ，FL 条件下 $\alpha=0.05$ ，UL 条件下 $\alpha=0.01$ ，SL

条件下 $\alpha=0.01$; 锐化模型的参数为: FE 条件下 $\theta=0.5$, LE 条件下 $\theta=0.49$, SE 条件下 $\theta=0.48$, FL 条件下 $\theta=0.05$, UL 条件下 $\theta=0.01$, SL 条件下 $\theta=0.01$ 。



附图 3 人脑与基于参数精细化模型修改的 DCNN 模型的时序分层表征相似性

从上到下依次是基于参数精细化的衰减模型修改后的 DCNN 模型与人脑表征相似性结果、基于参数精细化的锐化模型修改后的 DCNN 模型与人脑表征相似性结果以及基于两参数精细化的重复抑制模型修改后跨模态表征的相似性差异的结果。衰减模型的参数为: $\beta=0.9$, FE 条件下 $\alpha=0.5$, UE 条件下 $\alpha=0.45$, SE 条件下 $\alpha=0.4$, FL 条件下 $\alpha=0.05$, UL 条件下 $\alpha=0.01$, SL 条件下 $\alpha=0.01$; 锐化模型的参数为: FE 条件下 $\theta=0.5$, LE 条件下 $\theta=0.49$, SE 条件下 $\theta=0.48$, FL 条件下 $\theta=0.05$, UL 条件下 $\theta=0.01$, SL 条件下 $\theta=0.01$ 。

致谢