深度学习算法重建大脑感知人脸情绪图像

北京师范大学　谭泽宏 贺耀仪 王祖煜

1． 北京师范大学人工智能学院，北京市 海淀区 100875；
樊亚春 高级工程师

摘　要：通过对 fMRI 数据以及 RGB 图片的特定算法深度学习，从而在仅获得人脑活动 fMRI 数据的情况下重构人眼所看到的图像。已经通过利用不同数据集对神经网络模型进行训练，通过模型在验证集上的准确率以及对神经网络预测不同情感类型图片时 softmax 层的输出值进行配对样本 t 检验，得到了明确的结果，表明了神经网络模型对于图像情感的预测是有可解释性的。还对本项目训练神经网络所使用的数据集进行了说明。简要介绍了关于 fMRI 的基本知识，以及本组为何要收集 fMRI 数据、如何收集 fMRI 数据、收集了什么样的 fMRI 数据，简要介绍了通过 fMRI 数据与 resnet18 模型的 RSA 矩阵进行 kendall 分析并进行 FDR 多重比较校正的例子。在文章的最后展示了通过深度学习算法重建大脑感知图像的成果，并且讲述了将如何实现以及优化 fMRI 数据自监督算法。

关键词：神经网络 fMRI 脑成像 图像重构 情绪识别 解码器

引言

神经网络的基础结构是以人的大脑为基础的复杂系统，因而作者认为对于神经网络的研究切不可只在数学算法或计算机程序上做出改进，想要在神经网络上取得创新成果，对于人脑的研究是必不可少的。本项目组通过核磁共振技术，取得人脑活动的 fMRI 数据后进行处理，将其与神经网络模型的某部分参数共同映射至高维空间，然后再在高维空间中找到对应关系，从而编写解码器代码，最后希望达成能够获取到人脑在接受图像刺激下的活动信号后通过训练好的神经网络模型来解码出刺激图像本身或者刺激图像的部分信息，然后通过重构技术还原图像。本项目组提出了一种新的方法，除了稀缺的标记数据（训练对）外，还允许在“未标记”数据（即没有 fMRI 记录的图像和没有图像的 fMRI 记录）上训练 fMRI 到图像重建网络，从而完成了 fMRI 的无监督学习网络的设计。

1. 模型训练

在认知心理学领域，图像对人的情绪激活可以分为三种：正性（positive）、中性（neutral）、负性（positive）。作为项目的初期研究，本组用不同数据集和不同网络训练了不同的图像情绪三分类神经网络模型，使用的网络包括 resnet18、resnet50、vision transformer huge14、vision transformer large16。使用的。数据集的标注基准基于被试的 valence 值，将 1 到 4 的值标注为负性（positive），将 4 到 6 的值标注为中性（neutral），将 6 到 9 的值标注为负性（positive），共三类标注。通过将每个网络模型部署在 Nvidia Tesla V100GPU 上并且分别基于不同的数据集进行 100 个 epoch 的训练后，得到了部分模型在验证集上准确率大于 90%的结果。

以下为数据集信息。

Dataset	Introduction	number	label
IAPS	国际情感图片系统 (International Affective Picture System)	822 张	valence, arousal,dominance 值
CAPS	中国情感图片系统	852 张	valence,

	(Chinese Affective Picture System)		arousal,dominance 值
MAPS	军事情感图片系统 (Military Affective Picture System)	240 张	civilian and military populations
ALL	将 iaps 数据集部分去重后与 caps 数据集合并所得的数据集	1916 张	valence, arousal,dominance 值

1.1 IAPS 数据集

IAPS 是一种广泛使用的刺激材料，用于评估情感和心理过程，IAPS 全称是“国际情感图片系统”（International Affective Picture System）。它包括一组经过标准化和验证的图片，这些图片被设计成能够引发广泛的情感反应，如高兴、恐惧、厌恶、愤怒和悲伤等。

每幅图片都被赋予了代表情绪价值的三个数字，即主观情感价值、唤起程度和控制程度。其中主观情感价值反映了这幅图片引发情感反应的程度和方向，唤起程度反映了这幅图片能够引发情感反应的程度，而控制程度反映了这幅图片的情感反应的唯一性程度，即该图片是否在引发情感反应方面有所限制或具有独特的引发特征。

这些图片可用于研究者探索和了解人类情感、认知和行为过程，从而推动心理学领域的研究进展。IAPS 已被广泛应用于许多领域，包括临床心理学、神经科学、社会心理学、发展心理学等。

1.2 CAPS 数据集

中国情感图片系统（CAPS）是中国心理研究中使用的一组标准化和规范化情感图片。它包括一系列的图片，这些图片专门设计用来唤起不同的情绪，包括积极的、消极的情绪以及中性状态。

CAPS 的图片经过广泛的验证，并被广泛应用于情绪、认知和情感神经科学、心理病理学和其他相关领域的研究中。这些图片由大量参与者进行评分，以建立它们的价值（积极的、消极的或中性的）和唤起程度（情绪的强度）。

CAPS 是研究中国人群情感的重要工具，已经被广泛应用于各种研究中。它的使用有助于推动我们对情绪在大脑中如何加工和调节以及情绪在心理健康和福祉中的作用的理解。

CAPS 数据集共有 852 张图片，本项目按照 Valence 值将其分为了 negative、neutral、positive 三类，边界值图片被去掉，最后 negative 图片有 237 张，neutral 图片有 316 张，positive 图片有 298 张。

1.3 MAPS 数据集

军事情感图片系统(MAPS)是一套标准化和验证的图片，用于研究军人对各种刺激的情感反应。MAPS 包括描绘军人在其服务期间可能遇到的各种事件、情况和物品的图片，例如战斗场面、军事装备和军用车辆。这些图片被设计成能引发不同的情感，包括恐惧、愤怒、悲伤和快乐，可以用于评估军人对不同类型刺激的情感反应。MAPS 已经被用于各种研究中，以更好地了解军人的情感体验，并为与军事服务相关的心理健康问题的干预和治疗的发展提供指导。

本数据集共 240 张图片，本数据集被五种类型的被试进行了评分，分别是 'Civilian_Males','Civilian_Females','Military_Combat_Males','Military_Noncombat_Males','Military_Noncombat_Females'。本项目按照 Valence 值，对于每一类被试的 valence 值评分 ，将其分为了 negative、neutral、positive 三类。

本组还基于预训练的深度卷积神经网络 ResNet50 模型，对比了模型在 4 类图片（高威胁军事相关图片、高威胁非军事相关图片、低威胁军事相关图片、低威胁非军事相关图片）辨识能力上的性

能表现。根据模型 ResNet50 对军事相关图片和非军事相关图片的高威胁预测概率和低威胁预测概率绘制的箱式图，图中标注的 p 值是军事相关图片预测概率和非军事相关图片预测概率配对样本 t 检验的结果。说明了神经网络对于图像的情感进行预测是存在可解释性的。

关于 p 值的补充说明如下：

a. 无显著差异：当 p 值大于显著性水平（一般为 0.05）时，可以认为样本之间不存在显著差异，即差异很可能由随机因素引起的，而不是由其他因素引起的

b. 显著差异：当 p 值小于显著性水平（一般为 0.05）时，可以认为样本之间存在显著差异，即差异不是由随机因素引起的，而是具有统计学意义的

c. 非常显著差异：当 p 值小于 0.01 时，差异非常显著，即样本之间的差异非常明显，而且具有非常高的统计学意义

d. 极显著差异：当 p 值小于 0.001 时，差异极其显著，即可以非常确信地认为样本之间的差异不是由随机因素引起的，而是由其他因素引起的。

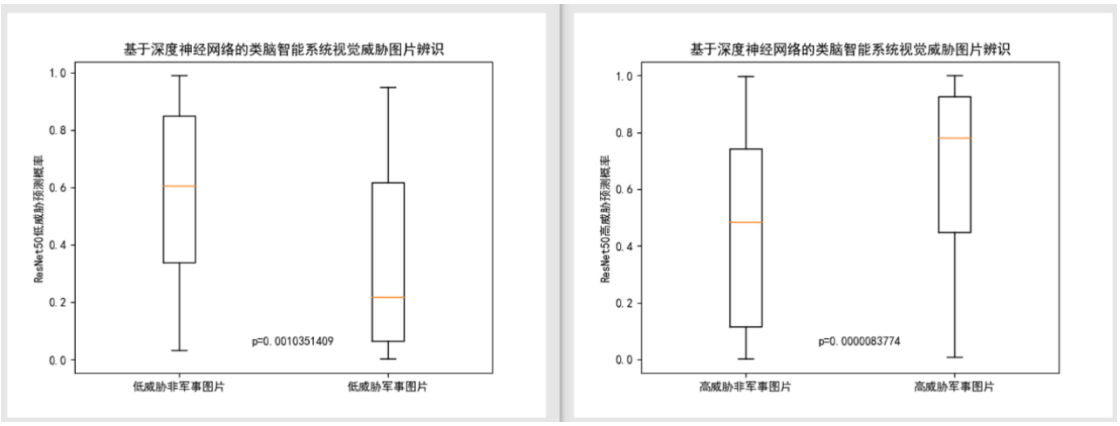

图 1.1 基于 resnet50 的图片情感辨识

2. fMRI 技术简介、收集过程及相关分析

fMRI (functional magnetic resonance imaging)是一种神经影像技术，可以通过测量大脑不同区域的血氧水平变化，来研究大脑在特定任务或行为中的功能活动。fMRI 利用强磁场和无害的无线电波，扫描被试者的大脑，通过检测血红蛋白的氧合状态的变化，来反映不同脑区的代谢活动水平。

fMRI 技术的原理是基于 BOLD 信号（血氧水平依赖性磁共振成像），即在大脑活动时，氧气供应不足导致神经元需要更多的氧气，血流会增加来满足需求，因此相应的血液中的血红蛋白的含氧量会有所不同，这种变化可以通过 fMRI 扫描仪检测和测量。

fMRI 技术的应用非常广泛，包括神经科学、心理学、心理治疗、精神病学、神经疾病的诊断和治疗等领域。fMRI 技术已经被用于研究认知和感知功能、情绪和情感的处理、学习和记忆、决策和风险评估、疼痛和压力等方面的神经机制。

在脑成像中，体素（Voxel）是指图像中的三维像素，是构成三维图像的最小单位，类似于二维图像中的像素。在脑成像中，每个体素都对应着大脑的一个小区域，通常为立方体或长方体形状。

体素的大小可以根据扫描的分辨率进行调整。例如，一个 1 mm^3 的体素表示该体素对应的脑区域大小为 1 毫升。随着技术的发展，现代的脑成像技术可以获得越来越高的空间分辨率，从而可以更精细地刻画脑区域之间的关系和功能活动。

通过将大量的体素组合起来，可以形成整个脑图像。这些图像可以用来分析脑结构、脑功能以及脑区域之间的联系，为神经科学研究和临床应用提供重要的信息。

从现有技术水平上来说，fMRI 技术是在非直接接触条件下获取人体大脑活动的最精准技术。本

项目组与北京师范大学认知神经科学与学习国家重点实验室秦绍正组展开合作，共同设计开展实验，从被试身上采集了项目所需 fMRI 材料。这些材料是后续模型训练的材料。

收集过程：

（1）目的概述

基于深度神经网络对视觉威胁图像激活与人脑对相同视觉图像辩识时威胁程度评分进行拟合（此为北京师范大学认知与神经科学实验室秦绍正组研究项目）。同时记录人脑活动的 fMRI 影像供后续训练。

（2）测试步骤

招募大学生作为研究被试，通过计算机按键反应获取他们辨识视觉威胁图片时的威胁程度（即图片的情感激活）评分，同时通过核磁共振仪获取被试的脑活动并记录。

（3）实验过程

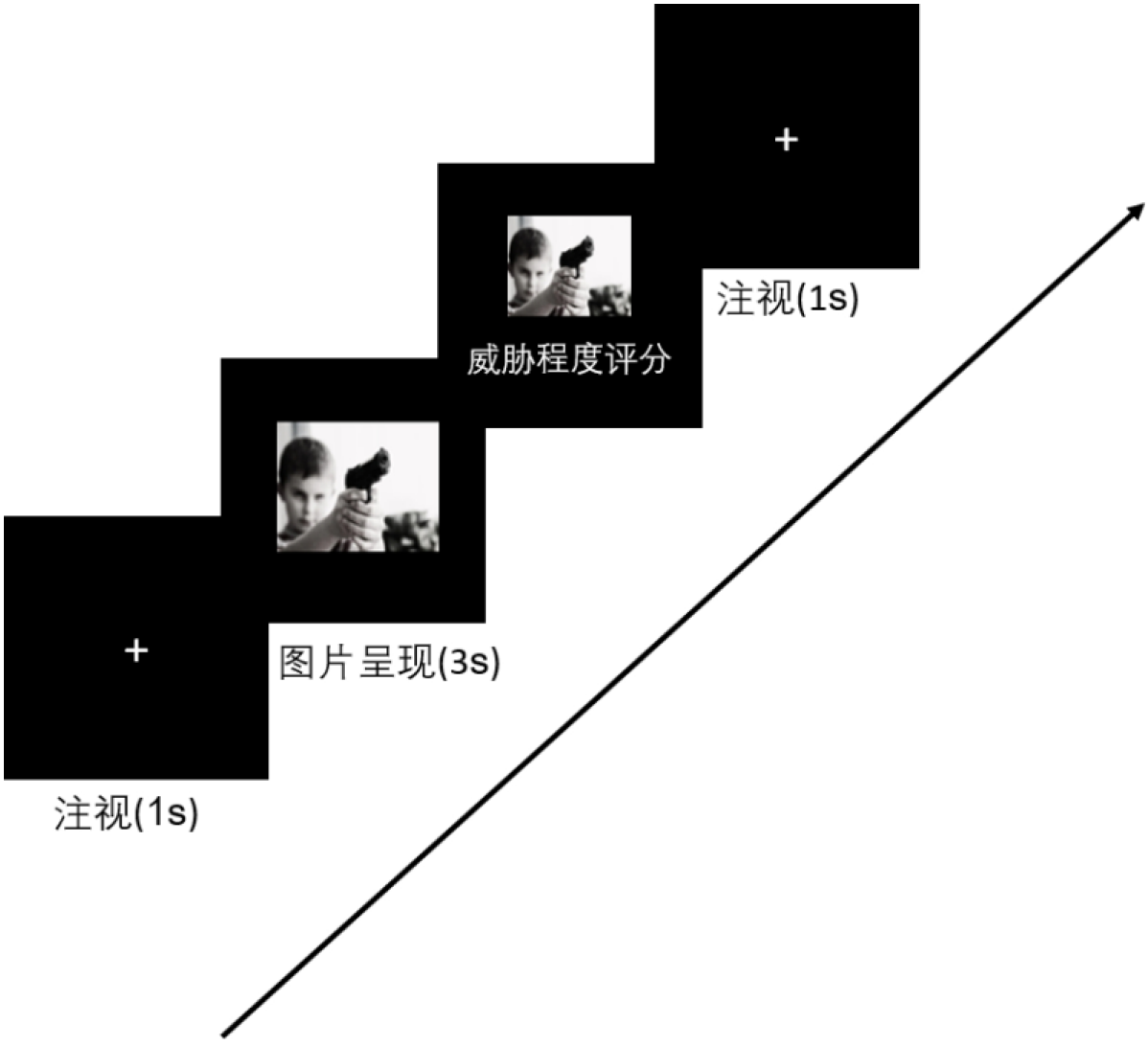

图 2.1　视觉威胁图片的威胁程度测评工具流程图

本组还利用 fMRI 数据进行了人脑层级化神经分布式表征的数据分析与特征提取，基于所收集的 fMRI 人脑数据，提取 19 个视觉和情绪相关脑区的特征向量，以及从预训练的深度卷积神经网络 ResNet18 中提取卷积层（Conv）和全连接层（FC）的特征向量共计共计 22 个层级。然后进行 Kendall 分析并进行 FDR 多重比较校正。

步骤：

（1） 将从磁共振设备上采集的脑像 DICOM 文件转换成可以进行后续数据处理和分析的 nii 文件。

（2） 脑像数据矫正。主要使用的矫正方法分为：时间层矫正（Slicing Timing）、头动矫正（Realign）、配准（Coregister）、分割（Segment）、标准化（Normalize）、平滑（Smooth）。

（3） 脑像群体特征提取与分析。从人脑数据提取与视觉和威胁相关脑区的特征向量，共计 19 个，脑区包括：初级视觉皮层：BA17_V；次级视觉皮层：BA18_V2、BA19_V345；高级视觉皮层：颞下回（BA20_ITG）、颞中回（BA21_MTG）、梭状回（BA37_Fusiform_Gyrus）、背外侧前额叶皮层（BA09_DLPFC、BA46_DLPFC）、前扣带回（BA33_ACC）、背侧前扣带回（BA32_DACC）、腹侧前扣带回（BA24_VACC；与情绪相关的脑区：杏仁核（Amygdala）、前额叶前部（BA10_APFC）、眶额皮层（BA11_OFA）、脑岛（BA13_Insular）、背侧后扣带回（BA31_DPCC）、腹侧前扣带回（BA24_VACC）、颞极（BA38_Temporopolar）、海马（Hippocampus）；之后对这些脑区进行 RSA 分析。

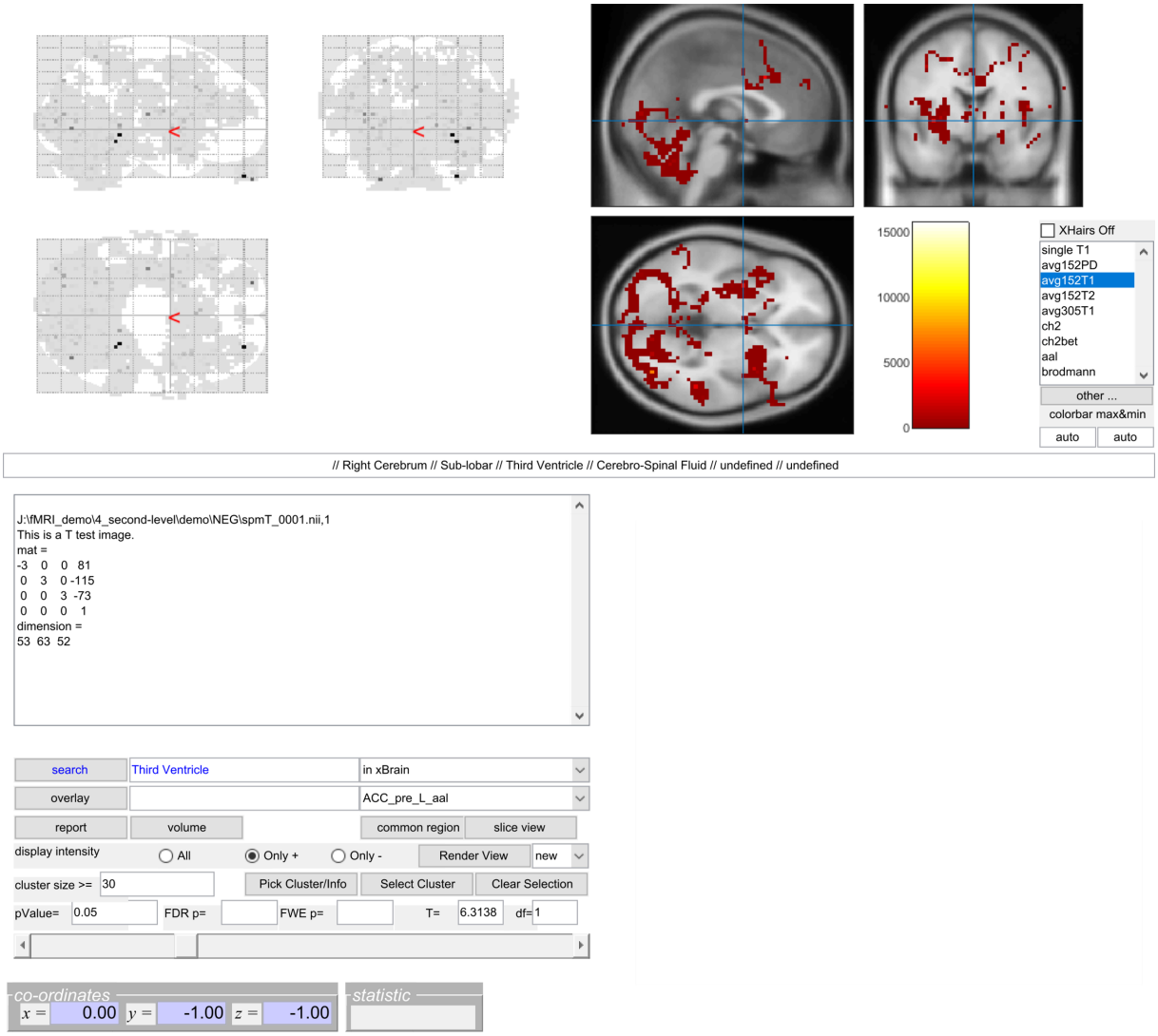

图 2.2 fMRI 数据处理中间过程例

（4） 基于预训练的深度卷积神经网络 ResNet，提取 ResNet18 中卷积层（Conv）和全连接层（Flatten）的特征向量并可视化（22 个层级）。从预训练的深度卷积神经网络 ResNet18 中提取卷积层（Conv）和全连接层（FC）的特征向量，共计 22 个层级，并进行 RSA 分析。

（5） 对深度卷积神经网络 ResNet18 中的 22 个层级特征向量与人脑 19 个脑区特征向量在处理不

确定条件下军事相关视觉信息的相关分析。将深度卷积神经网络 ResNet18 中 22 个层级的 RSA 矩阵和人脑 19 个脑区的 RSA 矩阵进行 Kendall 分析并进行 FDR 多重比较校正（False Discovery Rate）。

（6） 使用 Kendall 相关系数研究 ResNet18 中的 22 个层级结构和人脑 19 个脑区之间的相关关系，如下图所示。Kendall 相关系数为正值表示，人脑脑区与 ResNet18 的模型层之间存在正相关关系，这可以解释为人脑脑区的激活模式能够很好地解释 ResNet18 模型层的激活模式，例如：与视觉威胁高度相关的脑区杏仁核（Amygdala，下图中的深青色柱线）与 ResNet18 后几层的网络结构具有更强的相关性，这可以解释为 ResNet18 后几层对威胁感知的作用更强；视觉对象识别和分类的脑区颞下回（ITG，下图中的黄色柱线）与多个模型层具有很强的正相关，这可以解释为 ResNet18 中多个模型层都在发挥着物体识别和分类的作用。

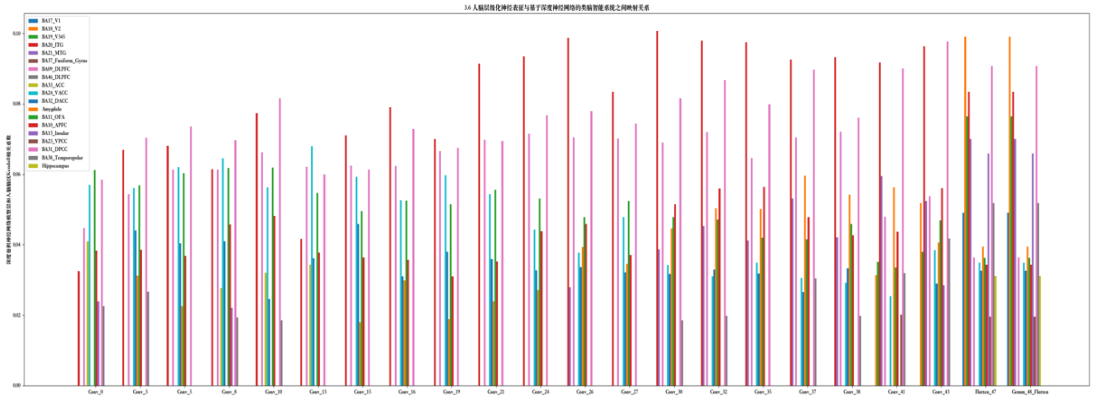

图 2.3 人脑层级化神经表征与基于深度神经网络的类脑智能系统之间映射关系

3. 通过 fMRI 数据预测图像的初步神经网络模型

受 Roman Beliy[1]在 NeurIPS 2019 的论文启发，本组利用 Kamitani Lab[2]发布的 fMRI 数据集以及 imagenet 2011 validation[3]数据集作为辅助材料，进行了 fMRI 数据预测图像的初步神经网络模型的训练。在 GPU 工作站上训练后得到了初步的成果。

开发一种从相应的大脑活动中高质量重建所见图像的方法，是解码梦和心理图像内容的重要里程碑。在这项任务中，人们试图使用许多“标记的”{图像，fMRI}对（即图像和它们对应的 fMRI 响应）来解决 fMRI 记录与其对应的自然图像之间的映射。一个好的 fMRI 到图像解码器可以很好地推广到从新的 fMRI 记录中构建从未见过的新图像（我们将其称为“测试数据”或“测试 fMRI”）。然而，缺乏“标记”的训练数据限制了当今 fMRI 解码器的推广能力。由于人类在 MRI 扫描仪中花费的时间有限，获取大量标记对{图像，fMRI}是抑制性的。因此，大多数数据集被限制为几千对这样的数据对。这样有限的样本无法跨越自然图像的巨大空间，也无法跨越其 fMRI 记录的空间。此外，fMRI 信号的时空分辨率差，以及信噪比低，降低了已经标记的训练数据的可靠性。最后，fMRI 数据的训练集和测试集往往在统计特性上有所不同，特别是在信噪比上。这种 SNR 差异是由于每个图像的重复记录的平均数不同（许多 fMRI 数据集的典型值）。因此，它引入了“域转移/自适应”的额外挑战，这使得泛化更加困难，并影响了当前解码方法的性能。

先前在 fMRI 图像重建方面的工作。从 fMRI 重建视觉刺激的任务已经通过多种方法进行，这些方法可以大致分为三类：（i）fMRI 数据和手工制作的图像特征（例如 Gaborwavelets）之间的线性回归[4，5，6]，（ii）fMRI 数据和深度（基于 CNN）图像特征之间的线性回归（例如，使用预训练

的 AlexNet）[7，8，9]，以及（iii）端到端深度学习[10，11，12，13]

受试者被要求注视位于图像中心的十字架， 图像数据集由从 200 个选定类别中提取的 1250 幅不同的 ImageNet 图像组成。训练和测试 fMRI 数据分别由每个刺激的 1 个和 35 个（重复记录）组成。50 个图像类别提供了 50 个测试图像，每个类别一个。剩下的 1200 被定义为训练集（只有一次 fMRI 记录）。

我们的培训包括两个阶段。在第一阶段，我们应用了 Encoder Ealone 的监督训练。我们使用图像 fMRI 训练对对对其进行训练，以预测输入图像的 fMRI 响应。在第二阶段，我们使用预训练的编码器（来自第一阶段）并训练解码器 D，同时使用标记数据和未标记数据来保持固定的权重。每个训练批次由三种类型的训练数据组成：（i）来自训练集的标记图像 fMRI 对，（ii）未标记的自然图像，和（iii）未标记 fMRI。

在结果方面，本组一共对三个不同被试的脑像进行了训练，训练完成的三个模型都在相同被试的脑像范围能进行推理。（图 3.1~3.3）

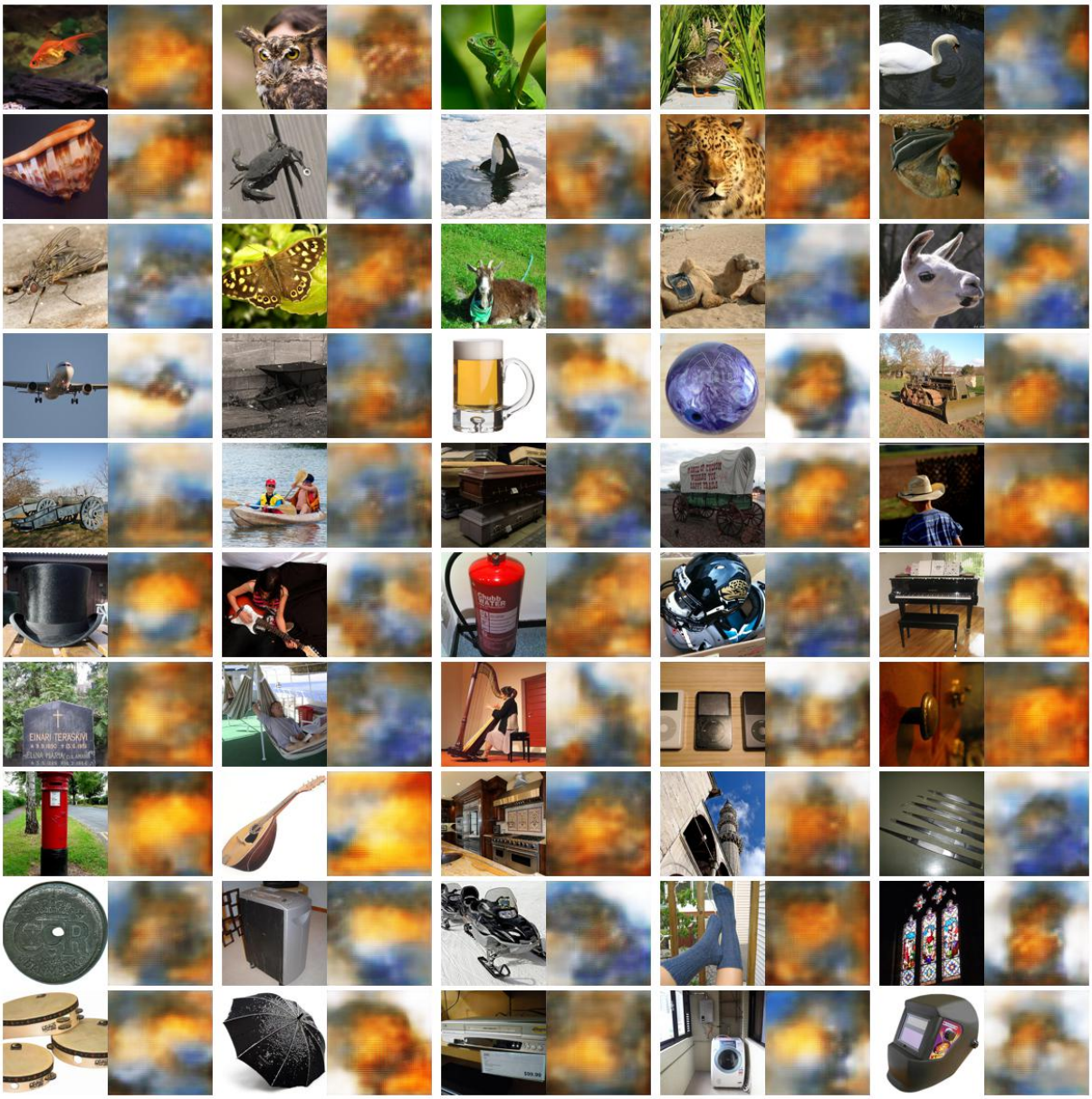

图 3.1 第一位被试 fMRI 脑像模型预测结果

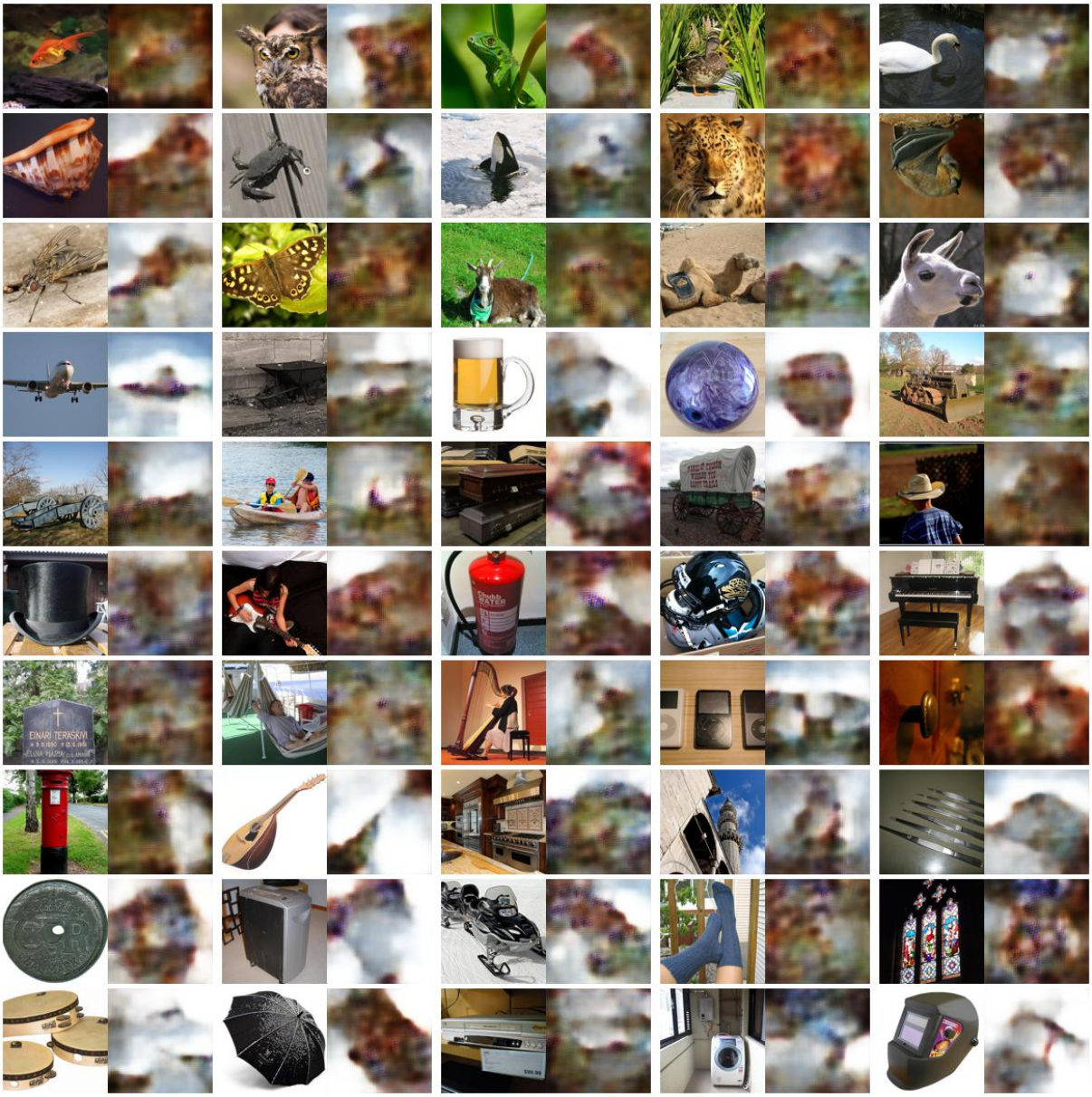

图 3.2 第一位被试 fMRI 脑像模型预测结果

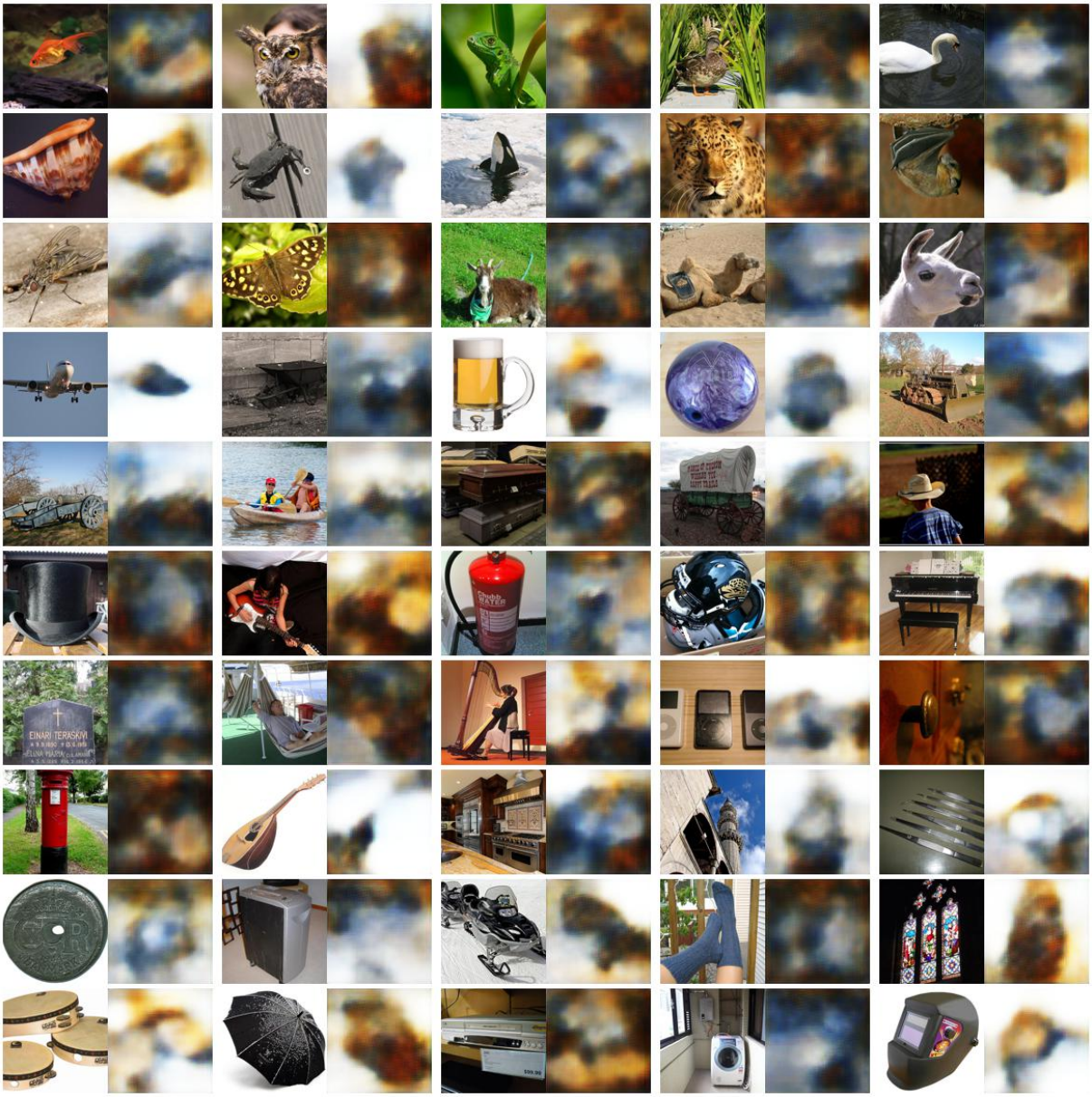

图 3.3 第一位被试 fMRI 脑像模型预测结果

可以从结果图直观的看见，在训练使用相同参数和模型结果的情况下，不同被试的模型的预测结果不同，这也说明了 fMRI 解码器目前的泛化效果不是特别理想。我们还曾经将特定被试的解码器运用在其他被试的脑像解码上，结果非常地不理想，说明模型仍然可能存在过拟合问题，但是也有可能是因为人脑的活动是特定于不同个体的，想做到精细的解码需要对个体数据进行训练。未来的方向也可能是开发一种快速低成本的训练网络，可以便捷地对于每个被试进行训练。从结果上来看，部分图像能够解码出近似的轮廓以及颜色，本项目作为这一领域的定性研究工作，无疑是成功的。

4. 总结

本研究旨在通过深度学习算法重建大脑感知人脸情绪图像。我们通过结合 fMRI 数据和 RGB 图片，并采用特定的算法进行深度学习，成功地实现了仅凭借人脑活动 fMRI 数据重构人眼所看到的图像。我们训练了多个神经网络模型，使用不同的数据集和网络结构进行情绪分类预测，并验证了模型在图像情感预测方面的可解释性。通过对模型预测输出值进行配对样本 t 检验，我们验证了模型对于不同情感类型图片的预测结果具有显著差异，进一步证明了模型的可解释性。我们还介绍了

收集和构建 fMRI 数据集的过程，这为模型的训练提供了多样性和广泛性。在图像重建方面，我们展示了通过训练好的深度学习算法成功重建大脑感知图像的成果。此外，我们还探讨了如何通过优化算法进一步改进 fMRI 数据的自监督学习，为未来的研究和应用提供了有价值的思路。总之，本研究通过深度学习算法在大脑感知图像重建领域取得了重要进展。我们的研究不仅验证了神经网络模型在图像情感预测方面的可解释性，还展示了通过融合 fMRI 数据和图像进行重建的潜力。这项研究对于深入理解大脑感知过程，以及在医学和人机交互领域的应用具有重要意义，为相关领域的未来研究提供了有价值的参考和启示。

参考文献

[1] Roman Beliy, Guy Gaziv, Assaf Hoogi, Francesca Strappini, Tal Golan, Michal Irani. From voxels to pixels and back: Self-supervision in natural-image reconstruction from fMRI. NeurIPS 2019，arXiv:1907.02431 [eess.IV]，Submitted on 3 Jul 2019.

[2] Horikawa, T., Kamitani, Y. Generic decoding of seen and imagined objects using hierarchical visual features. Nat Commun 8, 15037 (2017).

[3] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255).

[4] K. N. Kay, T. Naselaris, R. J. Prenger, and J. L. Gallant, “Identifying natural images from human brain activity,” Nature, vol. 452, pp. 352–355, 3 2008.

[5] T. Naselaris, R. J. Prenger, K. N. Kay, M. Oliver, and J. L. Gallant, “Bayesian Reconstruction of Natural Images from Human Brain Activity,” Neuron, vol. 63, pp. 902–915, 9 2009.

[6] S. Nishimoto, A. T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J. L. Gallant, “Reconstructing visual experiences from brain activity evoked by natural movies.,” Current biology : CB, vol. 21, pp. 1641–6, 10 2011.

[7] H. Wen, J. Shi, Y. Zhang, K.-H. Lu, J. Cao, and Z. Liu, “Neural Encoding and Decoding with Deep Learning for Dynamic Natural Vision,” Cerebral Cortex, vol. 28, pp. 4136–4160, 12 2018.

[8] U. Guclu, M. A. J. van Gerven, U. Güçlü, and M. A. J. van Gerven, “Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream,” Journal of Neuroscience, vol. 35, pp. 10005–10014, 7 2015.

[9] G. Shen, T. Horikawa, K. Majima, and Y. Kamitani, “Deep image reconstruction from human brain activity,” PLOS Computational Biology, vol. 15, p. e1006633, 1 2019.

[10] G. Shen, K. Dwivedi, K. Majima, T. Horikawa, and Y. Kamitani, “End-to-end deep image reconstruction from human brain activity,” bioRxiv, p. 272518, 2018.

[11] G. St-Yves and T. Naselaris, “Generative Adversarial Networks Conditioned on Brain Activity Reconstruct Seen Images,” bioRxiv, p. 304774, 2018.

[12] K. Seeliger, U. Güçlü, L. Ambrogioni, Y. Güçlütürk, and M. A. van Gerven, “Generative adversarial networks for reconstructing natural images from brain activity,” NeuroImage, vol. 181, pp. 775–785, 2018.