2025 年 第 25 卷 第 19 期
2025, 25(19):08117-10
科 学 技 术 与 工 程
Science Technology and Engineering
ISSN 1671—1815
CN 11—4688/T

DOI:10.12404/j.issn.1671-1815.2405740
引用格式:赵旭姣, 李瑶, 孙霖慧, 等. 基于多特征融合的脑电图高质量视频重建技术[J]. 科学技术与工程, 2025, 25(19): 8117-8126.
Zhao Xujiao, Li Yao, Sun Linhui, et al. High quality EEG video reconstruction technique based on multi-feature fusion[J]. Science Technology and Engineering, 2025, 25(19): 8117-8126.

自动化技术、计算机技术

基于多特征融合的脑电图高质量视频重建技术

赵旭姣[1], 李瑶[2], 孙霖慧[1], 杨艳丽[1], 郭浩[1*]
(1. 太原理工大学计算机科学与技术学院(大数据学院), 晋中 030600; 2. 太原理工大学软件学院, 晋中 030600)

摘 要 为探索大脑与视觉之间的联系,提高大脑活动重建视频的清晰度与准确性,提出了一种名为高质量脑电视频重建(high quality electroencephalogram video reconstruction, HQEEGVR)的方法进行脑电信号重建视频。首先,提出三分支脑电特征提取网络——掩蔽时空频融合网络(masking spatio-temporal frequency fusion network, MSTFFNet)从脑电信号中提取大脑活动信息,深入挖掘大脑活动变化背后的语义,同时提取时空频信息;其次,引入跨模态对比学习,对齐脑电、文本、图像特征,以便生成阶段使用;然后,提出级联视频扩散模型,具体来说,先利用稳定扩散模型以脑电特征为条件生成参考视频帧,接着以视频帧为参考,融入运动矢量,引入视频扩散模型捕捉视频时间特征;最终生成高质量视频。结果表明,该模型在重建视频的主体、动作、颜色、语义等方面表现较好。可见利用脑电信号可以捕获大脑活动的视觉与语义信息,从而重建高保真度和视觉真实性的视频。

关键词 脑电信号; 掩蔽时空频融合网络; 稳定扩散模型; 视频扩散模型; 运动矢量; 视频重建
中图法分类号 TP181; **文献标志码** A

High Quality EEG Video Reconstruction Technique Based on Multi-Feature Fusion

ZHAO Xu-jiao[1], LI Yao[2], SUN Lin-hui[1], YANG Yan-li[1], GUO Hao[1*]
(1. College of Computer Science and Technology (College of Data Science), Taiyuan University of Technology, Jinzhong 030600, China; 2. College of Software, Taiyuan University of Technology, Jinzhong 030600, China)

[**Abstract**] In order to explore the connection between brain and vision and improve the clarity and accuracy of brain activity reconstruction video, a new method called high quality electroencephalogram video reconstruction (HQEEGVR) was proposed to reconstruct video from EEG (electroencephalogram) signals. Firstly, the masking spatio-temporal frequency fusion network (MSTFFNet), a three-branch EEG feature extraction network, was proposed to extract brain activity information from EEG signals and dig deeper into the semantics behind brain activity changes, spatio-temporal frequency information was extracted at the same time. Secondly, cross-modal contrast learning was introduced to align EEG, text and image features for use in the generation stage. Then, a cascade video diffusion model was proposed, specifically, the stable diffusion model was used to generate reference video frames based on EEG features, and then the video frames were used as references, motion vectors were integrated, and the video diffusion model was introduced to capture the video time features. High quality videos were ultimately generated. The results show that the model performs well in the reconstruction of the subject, motion, color and semantics of the video. It can be seen that the EEG signal can be used to capture the visual and semantic information of the brain activity, so as to reconstruct the video with high fidelity and visual authenticity.

[**Keywords**] EEG; MSTFFNet; stable diffusion model; video diffusion model; motion vector; video reconstruction

收稿日期: 2024-07-31 **修订日期**: 2024-12-23
基金项目: 山西省科技厅基础研究计划(20210302123129, 20210302123099, 20210302124166);山西省科技厅基础研究计划青年项目(202203021222095, 202303021212166)

第一作者: 赵旭姣(1999—),女,汉族,山西临汾人,硕士研究生。研究方向:计算机视觉、人工智能、视觉信息处理。E-mail:qzp138698@163.com。
*通信作者: 郭浩(1981—),男,汉族,山西太原人,博士,教授,博士研究生导师。研究方向:视觉信息处理、人工智能、脑信息学。E-mail:feiyu_guo@sina.com。

视觉在复杂的大脑活动中具有重要地位,解码视觉信息一直是认知神经科学的关键目标[1]。在这一广泛目标中,利用大脑活动重建视频一直是一项备受关注的任务。近年来,研究人员已经成功利用功能性磁共振成像(functional magnetic resonance imaging, fMRI)技术将多维大脑活动数据转换为视频。传统研究[2-3]利用卷积神经网络、变分自编码器等方法与皮质响应相连接,利用动态自然视觉刺激诱发血氧水平依赖(blood oxygen-level dependent, BOLD)实现了对视频的真实重建。然而,这些研究虽然利用fMRI技术在脑活动视频重建方面取得一定的成果,但仍存在一些不足。一是fMRI受血流动力学影响[4],时间分辨率低,不适用于高时间要求任务,且采集成本高、操作难,阻碍其发展;二是从fMRI提取的视频动态特征及语义信息不足,导致视频模糊、伪影、闪烁,与真实场景差异大,难以准确反映真实细节[5]。与fMRI相比,一些研究人员也利用脑电图(electroencephalogram, EEG)技术来重建自然视觉,一些研究[6-7]利用长短期记忆网络(long short-term memory, LSTM),对比学习等深度学习模型提取EEG特征以生成图像,但在EEG特征提取方面仍有待改进。传统方法仅提取单一特征,未能全面捕捉多维信息,影响视觉重建的准确与真实感。

针对以上问题,现提出以下方法:

(1)运用EEG技术捕获大脑活动,并通过其重建视频。相较于fMRI,EEG携带便捷[8],操作简单,时间分辨率高,为生成高质量、流畅的视频内容提供坚实的技术基础。

(2)为了进一步提取EEG特征,提出一种名为掩蔽时空频融合网络(masking spatio-temporal frequency fusion network, MSTFFNet)的融合时域、频域和空域特征的EEG特征融合方法。与传统单一特征提取方法相比,融合了时域、频域以及空域特征等多维度信息,从而更为全面、深入地捕捉和解析大脑活动的复杂模式和动态变化,为后续视频重建任务提供更加丰富和准确的EEG特征。

(3)提出一种级联视频扩散模型。相较于传统的视觉重建技术,融合稳定扩散模型与视频扩散模型,不仅有效提取空间特征,更进一步增强时间特征的提取能力,尤其是引入运动矢量作为关键辅助条件,在一定程度上提升生成视频的视觉保真度与运动流畅性,确保画面的真实性与连贯性。

1 相关工作

1.1 利用fMRI重建视觉刺激

利用大脑活动进行视觉重建最常见的方法是使用机器学习方法来构建多体素fMRI模式与结构视觉特征之间的映射。通过这种方式,之前的研究已经利用诱发的fMRI活动重建二值图像[9-10]和复杂自然图像[11-12]。然而,由于fMRI的时间分辨率有限,难以对动态自然视觉进行建模。为了克服这一限制,一项早期研究[13]提出了一种运动能量编码模型,该模型使用贝叶斯解码器将估计的编码模型与采样的自然电影先验相结合来重建观看的电影。虽然该研究将重构问题简化为辨识问题,但证明了慢速BOLD信号也包含快速视觉信息。

在过去几年中,深度学习技术在视觉皮层建模领域中得到了广泛应用,主要因其能够有效提取多层次的视觉特征而备受瞩目。研究者们利用这些模型深入探讨了大脑如何精确地表征视觉信息。Wang等[14]提出了一种新方法,即功能磁共振成像条件视频生成对抗网络(fMRI-conditional video generative adversarial network, f-CVGAN),成功利用fMRI数据生成时空信息丰富的视频帧,但生成的视频在视觉保真度和语义方面存在较低水平。

与此同时,近年来,由于大脑信息与配对视觉信息数据集的匮乏,Kupershmidt等[15]提出了一种自监督的视频重建方法,利用自然时间视频先验来实现从fMRI大脑记录中重建自然视频。利用了大量无配对fMRI数据的外部自然视频,将适用的训练数据增加了几个数量级,引入自然视频先验解码网络,以及时间相干性。文献[16]提出一种名为MinD-Video的方法,通过掩蔽大脑建模[17]、多模态对比学习和增强稳定扩散模型[18]的联合训练,从连续fMRI数据中逐步学习时空信息。然而,由于fMRI成本高、数据规模大、处理复杂性高,这些方法在实际环境中的应用受到限制。

1.2 利用EEG重建视觉刺激

受到fMRI本身条件的限制,一些研究者利用深度学习技术探索了从EEG信号生成图像。EEG2image[19]采用了一种对比学习方法从EEG信号中提取特征,同时利用条件生成对抗网络(conditional generative adversarial network, cGAN)从这些提取的特征中重建图像。文献[20]利用编码的EEG信号来生成图像,尽管面临着训练数据有限的挑战。文献[21]将大脑反应作为监督信号,利用EEG数据学习语义特征,从而实现与手动标记相当的语义图像编辑性能。文献[22]介绍了一种名为DreamDiffusion的方法,可以直接从EEG信号生成高质量图像,无需转化为文本,实现“思想到图像”。然而,只捕获了EEG信号的时域特征,其模型的无监督特征嵌入能力依赖于对一个额外的大规模自

采集的 EEG 数据集(约 12 万个样本)的预训练,这表明其需要进一步增强 EEG 的特征表示能力。文献[23]通过将 EEG 信号的分类特征与图像生成模型结合,提高生成图像的质量。总的来说,这些方法展示了利用 EEG 重建视觉刺激并推进脑机接口领域的潜力,但是在使用 EEG 进行研究时,存在着 EEG 特征提取不全面的挑战,同时也存在语义信息捕获较少等问题,这些限制可能影响到对大脑活动的全面理解和精确分析。

2　研究方法

2.1　系统介绍

本文方法包括 3 个主要部分,如图 1 所示,一是融合时间-频率-空间特征的三分支 EEG 特征提取模块,如图 1(a)所示;二是利用跨模态对比学习对齐 EEG、文本、图像,如图 1(b)所示;三是使用级联视频扩散模型生成高质量视频,如图 1(c)所示。首先,利用预训练的掩蔽自动编码器(masked autoencoders, MAE)模型提取 EEG 上下文知识,同时利用小波变换将 EEG 特征从时域转为频域,利用轻量级卷积神经网络提取时频特征。接下来利用包含时间卷积与空间卷积的网络体系结构提取时空特征。最后使用自适应特征融合方法将三部分特征融合起来为视频生成提供条件。为了增强 EEG 特征与稳定扩散模型的兼容性,利用基于对比文本-图像对的预训练(contrastive language-image pre-training, CLIP)模型[24]将 EEG 和文本、EEG 和图像对齐。其中,对于文本特征,使用了 EEG 分类器,利用 EEG 信号进行分类,将得到的类别作为文本特征进行调整。最后,将经过对齐的 EEG 特征作为条件利用稳定扩散模型生成视频帧。为进一步提升视频质量,将生成的视频帧作为参考图像输入视频扩散模型并以运动矢量作为辅助条件生成高质量视频。

2.2　时空频特征融合

EEG 信号中包含了丰富的时域、空域以及频域信息[25],而以往利用 EEG 重建视觉内容的研究在提取 EEG 特征时往往集中在时域,忽略了空域、频域特征融合的优势,而空域、频域却可以更加全面的表征 EEG 的复杂特征[26]。为了解决这个问题,提出 MSTFFNet,这是一种基于 EEG 的时空频域融合的模型。MSTFFNet 以预处理后的 EEG 作为输入,包括 MAE 模型、时空特征提取、时频特征提取、特征融合 4 个部分。设 EEG 数据为 $\boldsymbol{X}=\{\boldsymbol{x}_i\}_{i=1}^{n}$,$\boldsymbol{x}_i\in\mathbf{R}^{C\times T}$,其中 n 为样本总数,C 为 EEG 通道数,T 为在每个通道上连续采集的时间点数量。下面将详细的描述 MSTFFNet4 个组成部分。

2.2.1　MAE 模型

以 MAE 模型为基础模块,用于提取 EEG 的时域特征。该模型属于一种可扩展的自监督学习计算机视觉模型,通过从剩余像素块中提取信息来恢复丢失像素以实现图片重建。其核心设计包括非对称的编码器-解码器架构和高比例的输入图像掩蔽。

利用该模型从剩余数据中恢复掩蔽数据,以学习 EEG 的一般特征,模型框架如图 2 所示。具体而

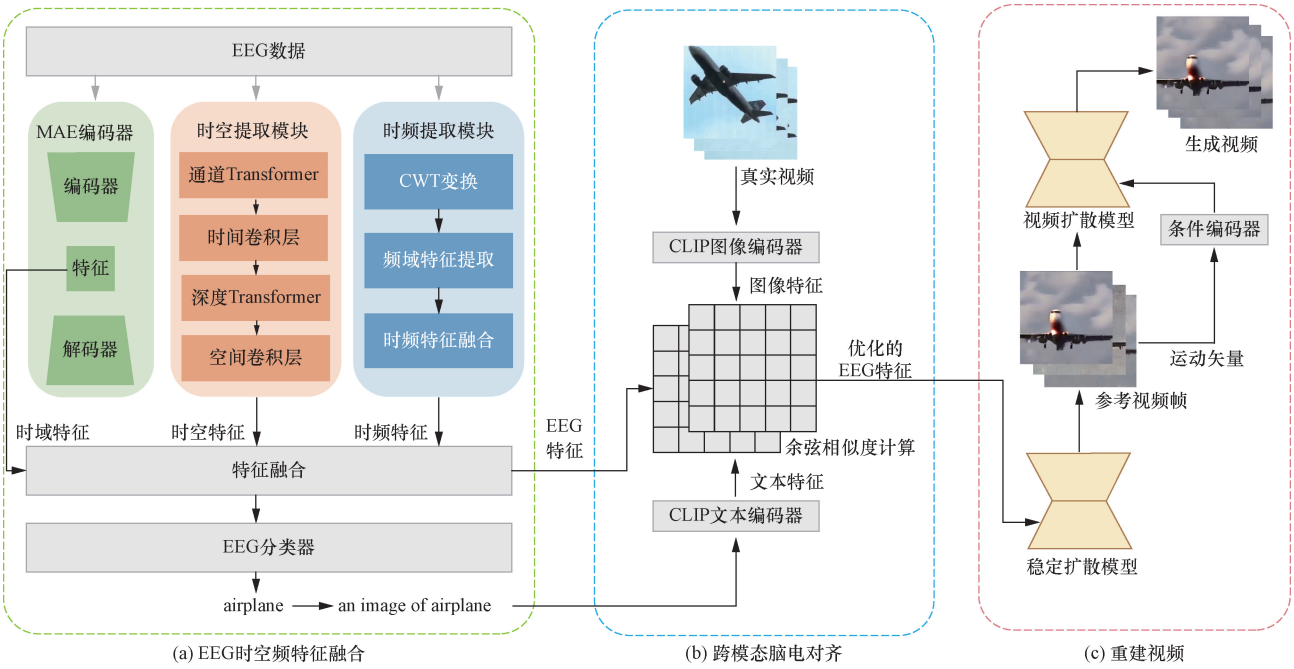

图 1　总体框架

Fig. 1　General framework

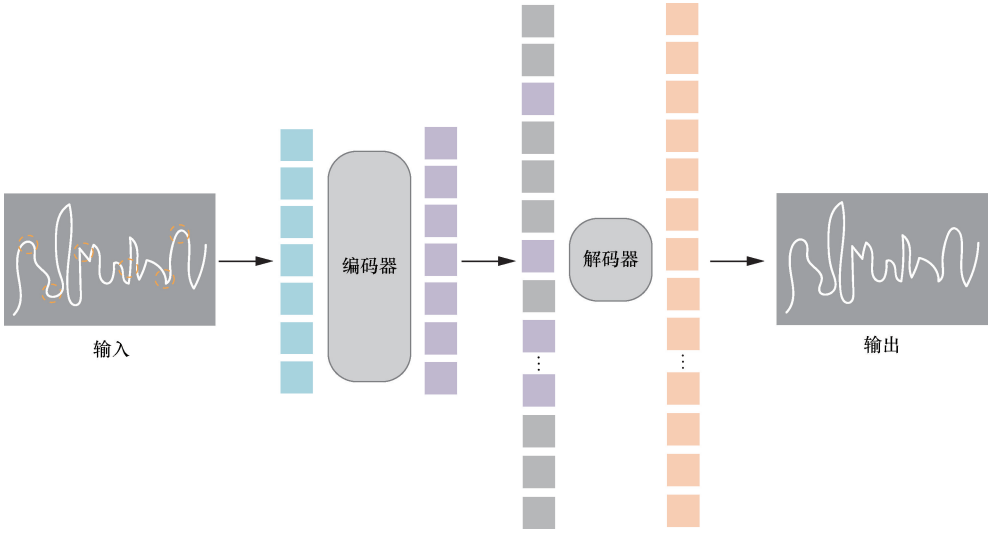

图 2 MAE 模型框架

Fig. 2 MAE model framework

言,首先将 EEG 数据在时间域上划分为大小相同的数据块,并随机地掩蔽一部分数据,掩码比率为 75%。然后,将未掩蔽数据送入编码器,经过编码的数据与加入位置信息的掩蔽数据按照原先的次序拼接在一起送入解码器中。最后,解码器会对掩蔽数据进行预测,并将预测结果与原始 EEG 数据进行比较,这里用到的损失函数是均方误差(mean squared error, MSE)函数。

通过重构掩蔽数据,预训练的 EEG 编码器可以深入理解不同个体和不同大脑活动的 EEG 数据,从而提取 EEG 特征。虽然该模型可以有效地从噪声和可变数据中获取上下文信息,但是该模型需要大量的训练数据,并且难以精准地表示复杂的 EEG 信号,因此在此基础上添加了时空提取模块以及时频提取模块。

2.2.2 时空提取模块

在时空特征提取方面,本文所提出特征提取网络分为 4 个部分:通道 Transformer 模型、时间卷积层、深度 Transformer 模型和空间卷积层,如图 3 所示。该网络是一种轻量级的网络体系结构,在时间卷积层和空间卷积层中都采用了可分离的卷积方法。而且通过在 EEG 通道中引入 Transformer 模块,并结合通道注意力机制将输入信号映射到深度维度,使其更能适应复杂 EEG 信号数据的分析,从而与后续时间卷积层相结合。

深度 Transformer 模块作为连接时间卷积层和空间卷积层之间的桥梁,强调了时间序列 EEG 信号中时域和空间域特征之间的关联性。其通过自适应池化层对输入数据进行初步处理,并借助 Transformer 模块对信号进行深度和空间信息的精细化处理,从而有效地提升了 EEG 信号的处理能力。这种结构的设计充分考虑了 EEG 数据的特点,为后续任务的顺利执行提供了更可靠的基础。

2.2.3 时频提取模块

在时频域特征方面,由于训练数据规模有限,通过复杂的网络学习 EEG 频域特征可能会导致具有挑战性的过拟合问题[27]。因此,本文中使用轻量级的网络[28]提取时频特征,具体框架如图 4 所示。

首先使用连续小波变换(continuous wavelet transform, CWT)将 EEG 数据从时域转换为频域。本文使用的是 Morlet 小波[29],Morlet 小波是一种常用的小波函数,适用于许多信号处理应用,包括时频分析和特征提取。经过小波变换后返回的数据是一个矩阵,其中的每个元素代表了 EEG 在不同频率和时间上的变化。这个矩阵记录了信号的频域信息,将其输入时频提取网络进行特征提取。

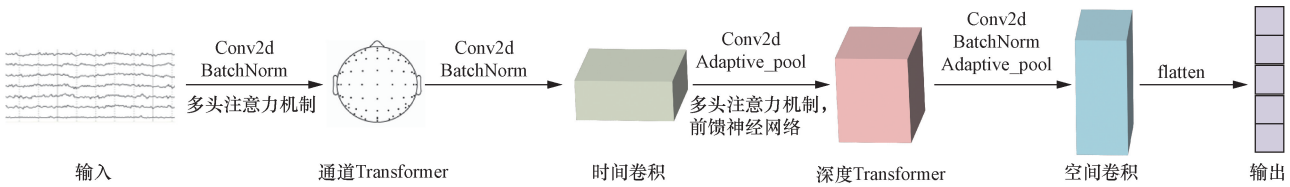

图 3 时空网络框架

Fig. 3 Space-time network framework

随后,通过卷积神经网络实现对经过预处理的原始数据与频域数据的特征提取。接着,将通过处理的频域数据转化为引导信号,用以指导数据在网络中的流向和处理方式。最终,将引导信号与原始数据进行拼接,实现了时频数据的融合。与自然图像特征提取网络架构中常用的最大池化操作相比,本文所提出网络中的池化层由于 EEG 信号的低信噪比和易受噪声污染而使用了平均池化。为了减少参数数量,网络最终的卷积层也采用了可分离卷积。

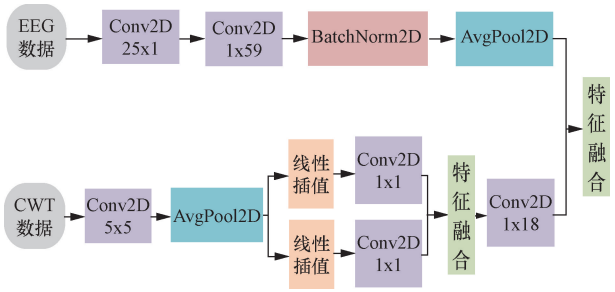

图 4　时频网络框架

Fig. 4　Time-frequency network framework

2.2.4　特征融合

时间-频率-空间域的特征融合包括两个组成部分:对时频特征提取层和时空特征提取层的输出特征进行一种自适应加权融合。融合策略结构通过自动学习获得权重系数矩阵 V 来衡量和整合两个维度特征的重要性。在操作上,对两个特征进行池化和全连接处理,然后通过 Softmax 函数生成权重系数矩阵 $\boldsymbol{V}$。最终特征 Z_1 和 Z_2 与 $\boldsymbol{V}$ 相乘得到融合输出特征 F_1。然后,将融合得到的特征与 MAE 模型得到的特征连接起来,得到一个新的三域特征。

$$F_1 = [Z_1, Z_2]\boldsymbol{V} \tag{1}$$

$$F = [F_1, F_2] \tag{2}$$

式中:Z_1 为时空特征;Z_2 为时频特征;F_1 为时频与时空融合特征;F_2 为经过 MAE 模型得到的特征;F 为融合 3 个模块得到的最终 EEG 特征。

2.3　跨模态脑电对齐

为了提高视频重建的质量,采用跨模态对比学习,并将预训练的级联视频扩散模型作为生成器。最近的研究表明,多模态模型预先训练的对比损失更能预测大脑活动,这表明跨模态对比学习更符合大脑学习方式[30]。受此启发,本文中采用 CLIP 模型作为连接大脑活动、视频帧和文本描述的桥梁,在训练过程中应用了跨模态对比学习。

2.3.1　跨模态对比学习

预先训练的稳定扩散模型是为文本到图像的生成任务而设计的,但是 EEG 数据具有独特的特征,其潜在空间与文本和图像的潜在空间有显著差异。而且由于 EEG-图像配对数据有限,直接端到端地微调稳定扩散模型可能无法准确地将 EEG 特征与预先训练的模型中的文本特征对齐[22]。因此,为使稳定扩散模型适用于视频重建的任务,本文中引入 CLIP 模型作为桥梁进行跨模态学习来对齐 EEG、文本、图像,并将其输入到稳定扩散模型中进行视频帧的生成。

CLIP 是一种预先训练好的图像-文本相似度度量模型,通过预测图像与字幕的对应关系进行预训练,来学习视觉表示。模型能够迁移到多数任务上,并且通常能与完全监督模型竞争。首先,将上一阶段得到的 EEG 时空频特征通过投影层转换为与 CLIP 相同维度的特征。然后,在固定 CLIP 文本编码器和图像编码器的情况下,使用损失函数来最小化 EEG 特征与图像特征和 EEG 特征与文本特征的距离,损失函数为

$$L_{\mathrm{EI}} = 1 - \frac{l_{\mathrm{EEG}} l_{\mathrm{image}}}{\| l_{\mathrm{EEG}} \| \| l_{\mathrm{image}} \|} \tag{3}$$

$$L_{\mathrm{ET}} = 1 - \frac{l_{\mathrm{EEG}} l_{\mathrm{text}}}{\| l_{\mathrm{EEG}} \| \| l_{\mathrm{text}} \|} \tag{4}$$

$$L_{\mathrm{clip}} = \frac{L_{\mathrm{EI}} + L_{\mathrm{ET}}}{2} \tag{5}$$

式中:l_{EEG}、l_{image}、l_{text} 分别为 EEG、图片、文本所提取的特征;L_{EI}、L_{ET} 分别是 EEG 与图片、EEG 与文本的余弦相似度;L_{clip} 为三者的损失。

2.3.2　视频帧文本语义

对于文本语义,之前的研究通常利用引导语言图像预训练(bootstrapping language image pretraining,BLIP)模型或图片分类器从图片中提取语义信息。文献[31]将两种方法做了对比,实验结果显示使用标签标题的方法优于使用 BLIP 标题的方法。推断出 EEG 信号可能只能捕获类别级别的信息,导致 BLIP 的预测不准确,从而降低了扩散模型的性能。

因此,在实验中使用类别级别信息,与之前工作不同的是,本文从 EEG 数据中提取出类别语义信息,具体来说,将 EEG 数据输入 EEG 分类器[32],数据结合递归神经网络,学习阅读思维中视觉类别的大脑活动流形。接着,通过训练基于卷积神经网络的回归器,将这种学习能力转移到机器上,使机器能够将图像投影到这个学习到的流形上,从而使用基于人类大脑特征的自动视觉分类,分类器框架如图 5 所示。对于每个实例得到类别注释,例如:“flower”,然后将其扩充为文本提示,如“an image of a flower”。

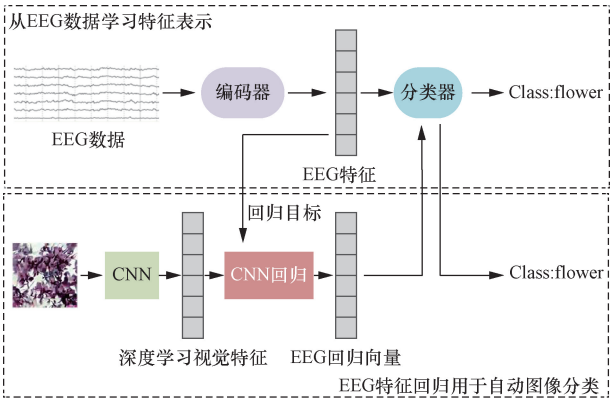

图 5 EEG 分类器框架

Fig. 5 EEG classifier framework

2.4 级联视频扩散模型

鉴于 EEG 信号的空间分辨率有限,噪声大,不可避免地出现信息不完全对齐的问题,生成视频时可能出现不流畅或视频伪影等问题。因此,本文提出一种级联视频扩散模型,如图 6 所示。该模型主要分成两部分,第一部分是稳定扩散模型,用来生成视频帧。第二部分是视频扩散模型,以运动矢量为条件进一步增强视频帧间一致性。

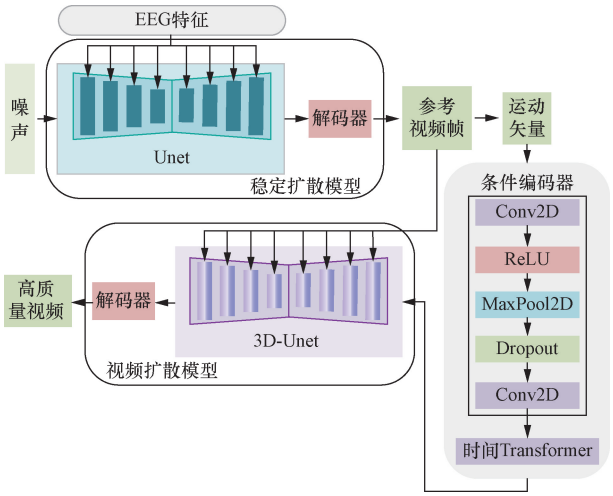

图 6 级联视频扩散模型

Fig. 6 Cascading video diffusion model

2.4.1 稳定扩散模型

稳定扩散模型是一种文本到图像的生成模型,通过将图像生成过程分解为去噪自编码器的连续应用来实现图像数据处理。本文利用该模型生成参考视频帧,捕获视频帧内信息,模型通过逐步去噪正态分布变量来了解数据分布,同时增加了交叉注意机制以 EEG 特征为条件生成参考视频帧。条件信号是由 UNet 网络中的交叉注意力机制引入的,具体来说,经过对齐的 EEG 特征通过投影层被投影到合适的维度中,然后,通过交叉注意力层将该 EEG 特征合并到 UNet 网络中。

通过结果观察到,由于 EEG 固有的噪声问题,重建视频帧之间强时间一致性不高,视频存在伪影。因此,为进一步提高视频质量,将生成的视频帧作为条件,作为视频扩散模型的新输入。

2.4.2 视频扩散模型

在这一部分,将视频分解为两种不同类型的条件,即空间条件和关键的时间条件,二者共同决定视频中的空间和时间模式。视频是由连续的图像组成的,一个图像通常显示这个视频的内容和结构。因此,将稳定扩散模型生成的视频帧作为参考输入到视频扩散模型中,具体来说,选择视频片段对应的第一个参考视频帧作为空间条件来执行图像到视频的生成。为了在时间维度上实现更精细的控制,引入运动矢量作为时间条件。这些二维向量表示视频中像素级别的水平和垂直运动关系。通过对连续帧之间的运动进行编码,利用这些运动矢量作为时间参考,可以保证视频帧的连续和无缝过渡。参考文献[33-34]的方法,从标准 MPEG-4 格式的压缩视频中提取出来运动矢量。

同时,为了更好地融合添加的条件,设计一个轻量级的条件编码器来合并时空关系。具体来说,对于提取空间信息,网络中有两个二维卷积层提取输入数据的特征,一个最大池化层通过选择每个池化窗口内的最大值来提取最显著的特征,Dropout 层用于减少过拟合风险。在训练过程中,会随机丢弃一部分神经元,使得网络更加健壮,更好的提取局部空间信息。随后,将得到的特征输入到时间 Transformer 提取时间特征,时间模块有多头注意力,前馈神经网络帮助模型有效地学习输入数据的表示,归一化层提高模型鲁棒性,以及残差网络构成。运动矢量增强了帧间一致性,促进高质量视频生成。

3 实验结果与分析

3.1 数据采集

利用博睿康 64 导湿电极 EEG 采集设备记录了 4 位被试观看 23 个时长为 8 min 的视频所诱发的 EEG 数据。参与实验的被试为两名女健康志愿者(S1、S2),以及两名男健康志愿者(S3、S4),年龄在 23~25 周岁,正常视力,并且获得了所有被试的书面知情同意书,被试具体信息如表 1 所示。

表 1 被试信息

Table 1 Subject information

被试编号	性别	年龄/岁	视力
S1	女	23	正常
S2	女	24	正常
S3	男	24	正常
S4	男	25	正常

实验所使用的视频刺激材料从 Videoblocks 和 YouTube 中剪辑而来的[16],视频刺激包括室内、室外、人、脸、鸟、昆虫、水生动物、陆生动物、花卉、水果、自然景观、汽车、飞机、船舶和运动等,是多样化但代表现实生活中的视觉体验。视频不含语音,每个视频包括多个片段,不是一整段连续的电影视频,视频片段的顺序是随机的和平衡的。视频分辨率为 600 × 600 像素,帧率为 30 fps,视频格式为 mp4。

实验在照明可控的实验室环境中进行,EEG 采集设备采用国际通用的 10-20 系统,实验软件采用 E-Prime 软件。实验时,每个 session 包括 4 次 8 min24 s 的视频刺激环节,视频播放顺序采用拉丁方方法。首先,安排受试者坐在舒适的椅子上,距离屏幕约 1 m,确保其舒适度和放松度的最佳状态[35],调整角度,正对视频。在每个环节中,受试者首先注视中央注视交叉点 10 s,引导其进入状态并恢复到基线状态。注视点由一个黑色背景和一个中间带有白色十字的标志组成。然后,将会呈现一个 8 min 视频,在视频放映之前,第一个视频帧以静态图片形式显示 12 s;视频结束后,最后一个视频帧也以静态图片形式显示 12 s。视频观看完成后,被试需在 45 s 内回答关于视频内容的问卷,进行数据采集的质量控制。最后,为确保被试充分休息和恢复认知疲劳,在其完成问卷后安排一个 5 min 的休息期,以便进行下一个视频数据的采集。

3.2　数据预处理

EEG 信号中存在大量伪影,包括电力系统引起的工频干扰,受试者眨眼带来的眼电伪迹和受试者头部及身体带来的肌电伪迹等,需通过预处理来去除以获取较纯净的 EEG 信号。因此,使用 EEGLab 工具包(https://sccn.ucsd.edu/eeglab/download.php),对采集的 EEG 数据进行处理。

预处理主要包括滤波、去伪迹等操作。将数据导入 MATLAB 的 EEGLab 工具包中,导入通道信息。采集数据时,记录了一些与本研究无关的通道信息,因此本文删掉了无用电极。在滤波阶段,本文使用了两种滤波方法,一是使用了 1 ~ 100 Hz 的带通滤波。二是进行陷波滤波(50 Hz)消除直流噪声、电源噪声和其他伪迹。最后,使用独立成分分析法去除眼动、肌电、心电等伪迹。

3.3　实施细节

在本文所使用的硬件环境中,CPU 是 Intel(R) Xeon(R) Gold 5222 CPU @ 3.80GHz,GPU 是 Tesla V100S-PCIe。为实现更高效的训练与测试,将视频从 30 fps 下采样到 3 fps。对于本文中使用的主要模型,MAE 模型的编码器深度为 24,特征维度为 1 024,在实验中以 0.75 的掩码比率进行了 500 批次的预训练,经过时-空-频 EEG 特征提取网络后,特征将被映射到 77 ×768 的尺寸以便下一步使用。在生成阶段,使用 1.5 版本的稳定扩散模型在512 ×512 的分辨率下进行了另外 300 个批次的微调。

3.4　评估指标

利用结构相似性指数(structural similarity index, SSIM)、N-way Top-k 分类精度来评估视频重建结果。SSIM 是一种用于衡量两幅图像相似性的指标,常用于评估图像质量和图像处理算法的效果。N-way Top-k 分类精度将真实视频帧与重建视频帧输入一个预先训练好的 ImageNet1k 分类器[36],该分类器决定了真实视频帧与重建视频帧是否属于同一类。然后检查重建图像在 N 个选定类的前 k 分类中是否与真实图像类匹配。本文中选择 2 路和 50 路的前 1 位精度作为评价度量。

3.5　对比实验

本文方法与 3 个视频基线进行了比较:Wang 等[14]的 f-CVGAN 模型、Kupershmidt 等[15]的自监督模型以及 Chen 等[16]的 MinD-Video 模型,定性比较如图 7 所示,定量比较如表 2 所示。从图 7 来看,本文方法生成了更加真实的视频,语义信息更加丰富,对于视频主体的颜色、运动、数量等信息的生成优于其余 3 种方法。从表 2 来看,在视频帧的像素级别上,本文方法所取得的 SSIM 的分数为 0.229,高于其余的 3 种基线方法。使用语义级别度量,本文的方法在 2 路和 50 路 top-1 准确率分类测试中分别实现了 0.850 和 0.238 的成功率,高于 Chen 等[16]的 0.792 与 0.172,说明本文方法在大脑活动语义信息方面的提取更加丰富、全面。

表 2　与基准比较

Table 2　Comparison with benchmark

模型	评估指标		
	SSIM	2-way	50-way
Kupershmidt	0.136	—	—
f-CVGAN	0.118	—	—
MinD-Video	0.171	0.792 ±0.03	0.172 ±0.01
Ours	0.229	0.850 ±0.02	0.238 ±0.01

3.6　参数讨论

在提取 EEG 数据的时间特征时,使用了 MAE 模型,在该模型中,掩码比率的设置至关重要,过高或过低的掩码比率都会对模型的性能产生不利影响。如图 8 所示,比较分别设置掩码比率为 0.50、0.75、0.85,其他设置保持不变时所生成视频的质量。从图 8 中可以看出,掩码比率为 0.75 时,生成的视频在 SSIM 得分为 0.229,2 路分类得分为 0.85

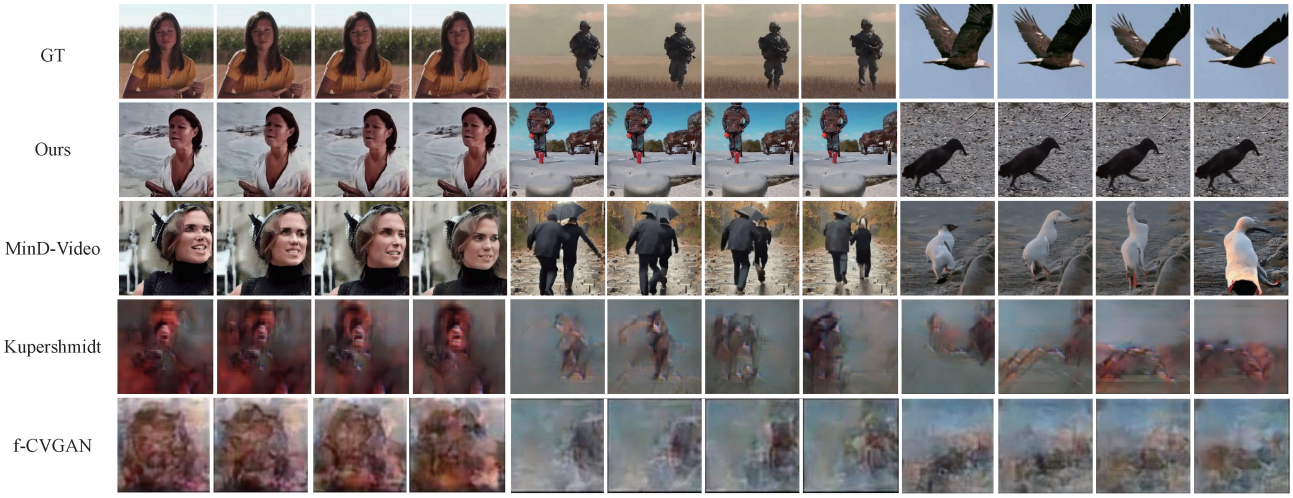

图 7 对比实验

Fig. 7 Contrast experiment

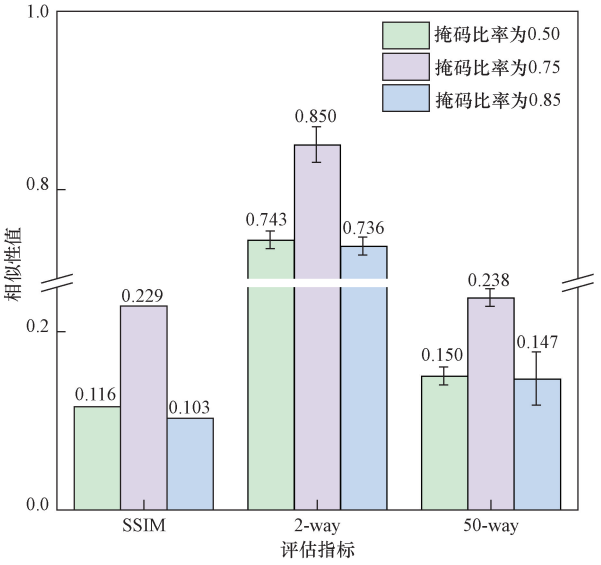

图 8 掩码比率对比

Fig. 8 Mask ratio comparison

以及 50 路分类得分为 0.238 均高于掩码比率为 0.50 和 0.85 的得分,实现了更好的总体效果。该结果表明,与较低或较高的掩码比率相比,在 EEG 上执行 MAE 模型时,0.75 的掩码比率是更好的选择。

3.7 消融实验

3.7.1 时空频特征融合

在 MAE 模型的基础上添加时空特征提取模块与时频特征提取模块来加强对 EEG 数据特征的提取。通过将直接使用 MAE 模型与使用时空频融合模块来提取 EEG 特征作对比,验证添加空域与频域的作用,结果如表 3 所示。从表 3 中可以看出,添加了 EEG 的空域与频域特征提取后,视频质量有一定提升,说明本文的方法提取的 EEG 特征更加丰富、全面,对于重建视频质量的提高有一定的作用。

表 3 EEG 特征提取模型对比

Table 3 Comparison of EEG feature extraction models

模型	评估指标		
	SSIM	2-way	50-way
MAE 模型	0.190	0.792 ±0.01	0.209 ±0.01
MAE + 时频	0.217	0.819 ±0.02	0.218 ±0.01
MAE + 时空	0.205	0.804 ±0.01	0.214 ±0.02
MSTFFNet 模型	0.229	0.850 ±0.02	0.238 ±0.01

3.7.2 运动矢量条件

提出一种轻量级的条件编码器,旨在整合时空关系。具体而言,在生成高质量视频的过程中,进一步融合了运动矢量条件,以捕捉视频帧间的时间信息。为验证运动矢量的作用,对运动矢量条件做了消融实验,结果如图 9 所示,从图 9 中可以看出,添加运动矢量之后生成的视频在各个指标中均高于不添加条件而生成的视频。因此可以得出运动矢

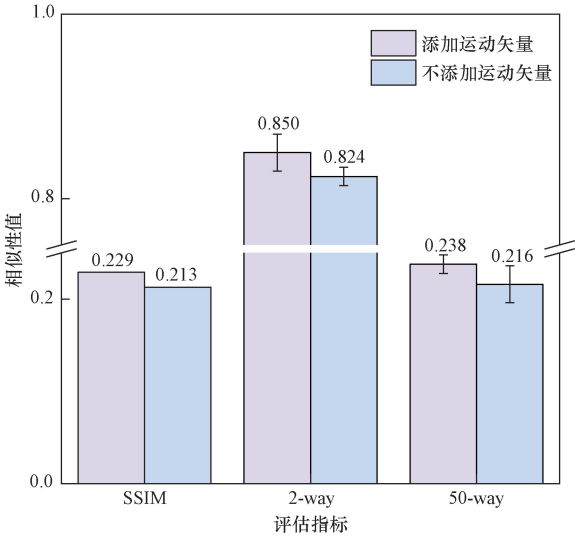

图 9 运动矢量条件消融

Fig. 9 Motion vector conditional ablation

量对于增强视频帧间一致性有一定价值,进而生成高质量视频。

4　结论

本文提出了一种新方法,用于从 EEG 信号重建高质量视频,EEG 信号是一种非侵入式、操作简单且容易采集的大脑活动数据。所提出的方法首先提取 EEG 的时-空-频域融合特征,增强 EEG 特征。接着,利用对比学习,提供 EEG 类别信息,进一步优化 EEG 特征,通过预训练和微调方案,可以使用稳定扩散模型将 EEG 数据编码为适合视频帧生成的表示形式。最后,为进一步增强视频帧间一致性,融入运动矢量作为辅助条件,捕获视频时间信息,进一步提升视频质量。结果表明,与真实视频相比,本文重建的视频在语义方面表现出一定的准确性,同时成功捕捉到了视频主体的运动、数量以及颜色等特征。这暗示了从 EEG 信号中捕获时间与空间信息来重建快速感知内容是可行的。

参　考　文　献

[1] 王冲. 基于功能磁共振成像的视觉信息编解码研究[D]. 成都: 电子科技大学, 2023.
Wang Chong. Visual information encoding and decoding based on functional magnetic resonance Imaging [D]. Chengdu: University of Electronic Science and Technology of China, 2023.

[2] Wen H G, Shi, J X, Zhang Y Z, et al. Neural encoding and decoding with deep learning for dynamic natural vision[J]. Cerebral Cortex, 2018, 28(12): 4136-4160.

[3] Han K, Wen H, Shi J, et al. Variational autoencoder: an unsupervised model for encoding and decoding fMRI activity in visual cortex [J]. Neuroimage, 2019, 198: 125-136.

[4] 姜小梅, 王萍芝. 基于静息态功能磁共振成像的慢性意识障碍脑功能网络研究进展[J]. 中西医结合心脑血管病杂志, 2024, 22(11): 1981-1984.
Jiang Xiaomei, Wang Pingzhi. Research progress of brain functional networks in chronic consciousness disorders based on resting state functional magnetic resonance imaging[J]. Chinese Journal of Integrative Medicine on Cardio-/Cerebrovascular, 2024, 22(11): 1981-1984.

[5] Li X, Chu W, Wu Y, et al. VideoGen: a reference-guided latent diffusion approach for high definition text-to-video generation[J]. ArXiv, 2023: 2309.00398.

[6] Jiang J M, Ahmed F, Zhong S H. A brain-media deep framework towards seeing imaginations inside brains[J]. IEEE Transactions on Multimedia, 2020, 23: 1454-1465.

[7] Singh P, Dalal D, Vashishtha G, et al. Learning robust deep visual representations from EEG brain recordings[C]//IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa: IEEE, 2024: 7538-7547.

[8] 张可新, 曲洪权, 李洋. 基于加权相位滞后指数热力图的脑力负荷识别[J]. 科学技术与工程, 2024, 24(28): 12055-12064.
Zhang Kexin, Qu Hongquan, Li Yang. Mental workload recognition based on weighted phase lag index heat map[J]. Science Technology and Engineering, 2024, 24(28): 12055-12064.

[9] Miyawaki Y, Uchida H, Yamashita O, et al. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders[J]. Neuron, 2008, 60(5): 915-929.

[10] Huang W, Yan H M, Ran L, et al. F-score feature selection based Bayesian reconstruction of visual image from human brain activity[J]. Neurocomputing, 2018, 316: 202-209.

[11] Du C, Du C, Huang L, et al. Reconstructing perceived images from human brain activities with Bayesian deep multiview learning [J]. IEEE Transaction on Neural Networks and Learning Systems, 2019, 30(8): 2310-2323.

[12] Huang W, Yan H, Wang C, et al. Long short-term memory-based neural decoding of object categories evoked by natural images[J]. Human Brain Mapping, 2020, 41(15): 4442-4453.

[13] Nishimoto S, Vu A T, Naselaris T, et al. Reconstructing visual experiences from brain activity evoked by natural movies[J]. Current Biology, 2011, 21(19): 1641-1646.

[14] Wang C, Yan H M, Huang W, et al. Reconstructing rapid natural vision with fMRI-conditional video generative adversarial network [J]. Cortex, 2022, 32(20): 4502-4511.

[15] Kupershmidt G, Beliy R, Gaziv G, et al. A penny for your (visual) thoughts: self-supervised reconstruction of natural movies from brain activity[J]. ArXiv, 2022: 2206.03544.

[16] Chen Z, Qing J, Zhou J H. Cinematic mindscapes: high-quality video reconstruction from brain activity[C]//Proceedings of the 37th International Conference on Neural Information Processing Systems. California: Neural Information Processing Systems, 2023: 24841-24858.

[17] He K M, Chen X L, Xie S N, et al. Masked autoencoders are scalable vision learners[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). California: IEEE Computer SOC, 2022: 15979-15988.

[18] Wu J Z, Ge Y X, Wang X T, et al. Tune-A-Video: one-shot tuning of image diffusion models for text-to-video generation[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). California: IEEE Computer SOC, 2023: 7589-7599.

[19] Singh P, Pandey P, Miyapuram K, et al. EEG2IMAGE: image reconstruction from EEG brain signals[C]// 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, Greece: IEEE, 2023: 1-5.

[20] Tirupattur P, Rawat Y S, Spampinato C, et al. ThoughtViz: visualizing human thoughts using generative adversarial network[C]// Proceedings of the 2018 ACM Multimedia Conference. New York: Assoc Computing Machinery, 2018: 950-958.

[21] Davis K M, Torre-Ortiz C D L, Ruotsalo T. Brain-supervised image editing[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). California: IEEE Computer SOC, 2022: 18459-18468.

[22] Bai Y, Wang X, Cao Y P, et al. DreamDiffusion: high-quality EEG-to-image generation with temporal masked signal modeling and CLIP alignment[C]//Computer Vision-ECCV 2024. Cham: Springer Nature Switzerland, 2025: 472-488.

[23] Yang G Y, Liu J G. A new framework combining diffusion models and the convolution classifier for generating images from EEG sig-

nals[J]. Brain Sciensce, 2024, 14(5): 478.

[24] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision [C]//International Conference on Machine Learning (ICML). California: JMLR-Journal Machine Learning Research, 2021: 8748-8673.

[25] 杜扶遥, 姜囡, 刘浠辰. 基于时频空多维融合特征的脑电情感识别[J]. 科学技术与工程, 2024, 24(18): 7769-7775.
Du Fuyao, Jiang Nan, Liu Xichen. EEG emotion recognition based on time-frequency-space multi-dimensional fusion features [J]. Science Technology and Engineering, 2024, 24(18): 7769-7775.

[26] 王雨佳, 鞠翔宇, 于扬, 等. 基于脑电频谱时空特征的认知负荷评估[J]. 控制理论与应用, 2025, 42(1): 50-58.
Wang Yujia, Ju Xiangyu, Yu Yang, et al. Cognitive workload assessment based on temporal and spatial characteristics of electroencephalogram spectrum[J]. Control Theory & Applications, 2025, 42(1): 50-58.

[27] Miao Z, Zhao M, Zhang X, et al. LMDA-Net: a lightweight multi-dimensional attention network for general EEG-based brain-computer interfaces and interpretability[J]. NeuroImage, 2023, 276: 120209.

[28] Liu X, Lü L, Shen Y, et al. Multiscale space-time-frequency feature-guided multitask learning CNN for motor imagery EEG classification[J]. Journal of Neural Engineering, 2021, 18(2): 026003.

[29] Cohen M X. A better way to define and describe Morlet wavelets for time-frequency analysis[J]. NeuroImage, 2019, 199: 81-86.

[30] Ma Y, Liu Y, Chen L, et al. BrainCLIP: brain representation *via* CLIP for generic natural visual stimulus decoding [J]. IEEE Transactions on Medical Imaging, 2025,2025: 3537287.

[31] Lan Y T, Ren K, Wang Y, et al. Seeing through the brain: image reconstruction of visual perception from human brain signals[J]. ArXiv, 2023: 2308.02510.

[32] Spampinato C, Palazzo S, Kavasidis I, et al. Deep learning human mind for automated visual classification[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2017: 4503-4511.

[33] Shou Z, Lin X D, Kalantidis Y, et al. DMC-Net: generating discriminative motion cues for fast compressed video action recognition [C]//Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). California: IEEE Computer SOC, 2019: 1268-1277.

[34] Wu C Y, Zaheer M, Hu H, et al. Compressed video action recognition [C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. California: IEEE Computer SOC, 2018: 6026-6035.

[35] 马金旭, 陶庆, 刘景轩, 等. 改进 FBCSP 和 CNN 的运动想象脑电信号分类[J]. 科学技术与工程, 2024, 24(27): 11726-11732.
Ma Jinxu, Tao Qing, Liu Jingxuan, et al. Classification of motor imagery EEG signals based on improved FBCSP and CNN[J]. Science Technology and Engineering, 2024, 24(27): 11726-11732.

[36] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale[J]. ArXiv, 2020: 2010.11929.