

文章编号: 1671-7449(2026)04-0449-16

J. Test Measur. Technol., 2026, 40(4): 449-464.

基于双分支网络的动态知识蒸馏运动想象脑电解码

李乐恒, 乔晓艳*, 吴健民

(山西大学 无线通信与检测山西省重点实验室, 山西 太原 030006)

摘要: 针对多任务运动想象脑电解码模型存在解码精度低、计算复杂度高的问题, 提出一种基于双分支网络的动态知识蒸馏多任务运动想象脑电解码模型。该架构设计“先时间多尺度融合, 后空间统一处理”的多尺度时空特征提取模块, 采用Transformer与BiLSTM双分支结构作为教师模型, 并有效融合Transformer全局特征表示与BiLSTM精细时序捕捉优势, 以提高运动想象脑电解码精度。知识蒸馏阶段, 采用轻量化EEGNet作为学生模型, 提出双分支动态知识蒸馏策略和自适应选择最优分支生成软标签, 使学生模型学习到适应不同个体的脑电特征表示, 以实现在保持学生模型较小规模的同时获得接近教师模型的性能。通过BCI Competition IV 2a数据集, 实验验证该方法的有效性。双分支教师模型多任务运动想象脑电解码平均准确率达到91.32%; 知识蒸馏的学生模型脑电解码平均准确率达到88.85%, 相比于无蒸馏的学生模型平均准确率提升6.63%, 与教师模型相比参数量减少99.78%, 计算量降低89.02%。结果表明, 该方法在保持脑电高解码精度的同时, 可显著降低计算复杂度, 为在资源受限环境下的脑机接口系统应用提供切实可行的轻量化解决方案。

关键词: 运动想象脑电解码; 双分支教师模型; EEGNet学生模型; 动态知识蒸馏

中图分类号: TP391

文献标识码: A

DOI: 10.62756/csjs.1671-7449.2026054

引用格式: 李乐恒, 乔晓艳, 吴健民. 基于双分支网络的动态知识蒸馏运动想象脑电解码[J]. 测试技术学报, 2026, 40(4): 449-464.

Li Leheng, Qiao Xiaoyan, Wu Jianmin. Dynamic knowledge distillation based on dual-branch network for motor imagery EEG decoding[J]. Journal of Test and Measurement Technology, 2026, 40(4): 449-464.

Dynamic Knowledge Distillation Based on Dual-Branch Network for Motor Imagery EEG Decoding

Li Leheng, Qiao Xiaoyan*, Wu Jianmin

(Shanxi Key Laboratory of Wireless Communication and Detection, Shanxi University, Taiyuan 030006, China)

Abstract: To address low decoding accuracy and high computational complexity in multi-task motor imagery EEG decoding models, a dual-branch network-based dynamic knowledge distillation model is proposed. The principle of “multi-scale temporal fusion first, followed by unified spatial processing”, is followed in this architecture and a multi-scale spatiotemporal feature extraction module is designed. A dual-branch structure combining Transformer and BiLSTM are employed as the teacher model, which effectively integrates the Transformer’s strength in global feature representation with the BiLSTM’s advantage in fine-grained temporal modeling, thereby enhancing the decoding accuracy of motor imagery EEG signals. During the knowledge distillation phase, a lightweight EEGNet is utilized as the student model. A dual-branch dynamic knowledge

【收稿日期】2025-11-14

【修回日期】2026-02-11

【录用日期】2026-03-11

【基金项目】山西省研究生教育创新计划资助项目(2025SJ067)

【作者简介】李乐恒(2001—), 男, 硕士生, 主要从事运动想象脑机接口和机器学习算法研究。E-mail: 17891213085@163.com。

*【通信作者】乔晓艳(1969—), 女, 教授, 博士, 主要从事机器学习 and 生物医学信息处理研究。E-mail: xyqiao@sxu.edu.cn。

distillation strategy is proposed, where the optimal branch is adaptively selected to generate soft labels. This enables the student model to learn EEG feature representations adaptable to different individuals, allowing it to achieve performance close to that of the teacher model while maintaining a compact size. Experimental validation on the BCI Competition IV Dataset 2a demonstrates the effectiveness of the proposed method. The dual-branch teacher model achieves an average decoding accuracy of 91.32% for multi-task motor imagery. After knowledge distillation, the student model attains an average EEG decoding accuracy of 88.85%, representing an improvement of 6.63% compared to the student model without distillation. Furthermore, compared to the teacher model, the number of parameters in the student model is reduced by 99.78%, and the computational load is decreased by 89.02%. The results indicate that the proposed method maintains high EEG decoding accuracy while significantly reducing computational complexity, providing a practical and lightweight solution for brain-computer interface systems in resource-constrained environments.

Key words: motor imagery electroencephalogram (EEG) decoding; dual-branch teacher model; EEGNet student model; dynamic knowledge distillation

0 引言

脑机接口(Brain-Computer Interface, BCI)通过解码大脑活动产生的神经信号,将其转化为数字指令以直接控制计算机或其他外部设备,无需依赖外周神经与肌肉组织的正常传导,从而建立大脑与外部环境之间的交互连接^[1]。其中,基于运动想象的BCI系统因其无创性、低成本、无肢体动作的特性成为临床研究和应用的热点。其核心技术运动想象脑电解码是一种基于脑电图(Electroencephalogram, EEG)信号,通过分析大脑活动模式来解析和识别特定的运动想象任务,并将它们转化为控制指令^[2]。然而,EEG信号具有信噪比低、平稳性弱、个体差异大等特点,使得脑电解码模型需依赖复杂的结构和计算才能提升精度,严重制约了算法在资源受限设备的部署能力^[3]。

为了克服EEG信号的低信噪比与时空耦合问题,提升解码精度,研究者们不断探索机器学习与深度学习解码模型。早期研究主要依赖传统机器学习方法,包括线性判别分析(Linear Discriminant Analysis, LDA)、支持向量机(Support Vector Machines, SVM)和K近邻(K-Nearest Neighbors, KNN)方法^[4]。但是这类方法高度依赖于人工提取特征的质量和适用性。特征提取过程繁琐且需要领域知识,提取的特征也难以捕捉EEG信号中复杂、高维的时空特性和通道间的高阶关系,且泛化能力受限,精度提升难。为了应对传统机器学习对人工提取特征的依赖及其在捕捉复杂模式上的不足,研究者转向端到端学习

的深度学习模型。卷积神经网络(Convolutional Neural Networks, CNN)因能够通过多层非线性变换直接从原始的EEG信号中自动学习具有判别性的时空特征表示,而成为运动想象解码的主流选择。目前,多种专门为EEG信号设计的CNN架构已被提出。Schirrneister等^[5]设计了一系列深度卷积网络架构(Deep ConvNets),引入批量归一化(Batch Normalization, BN)和指数线性单元(Exponential Linear Unit, ELU)等优化技术,结合裁剪训练策略,在BCI competition IV 2a数据集上MI脑电解码的平均准确率达到72.53%,优于广泛使用的滤波器组共空间模式(Filter Bank Common Spatial Pattern, FBCSP)分类算法^[6]。然而Deep ConvNets模型参数量大,需要大量高质量数据才能完成有效训练并避免过拟合。MI-EEG实验通常受限于被试疲劳、实验时长等因素,单个被试的可用数据往往非常有限。在这种小样本场景下,Deep ConvNets极易过拟合。针对这个问题,Mane等^[7]提出了一种新型滤波器组卷积网络(FBCNet),通过融合滤波器组分解与空间滤波优化的思想,构建了一个即使训练数据有限也能高效训练的网络,取得了76.20%的平均准确率。虽然CNN在MI-EEG解码中取得了一定进展,但其局部感受野和缺乏显式时序建模能力,限制了捕捉EEG信号中跨越长时间尺度的复杂动态模式。为了更好地捕捉EEG信号的时序特征,研究者提出采用长短时记忆网络(Long Short-Term Memory, LSTM)^[8]。然而,LSTM模型依赖时间顺序传递隐藏状态,无法进行并行训练,且长程依赖建模能力较弱。近年来兴起的Transformer架构,凭借独特的自注意力机制,彻底摒弃了循环结构,实现了对整

个序列的全局建模与高效并行计算,成为脑电解码研究的新方向。Xie 等^[9]提出了一种基于 Transformer 的运动想象脑电信号分类模型,利用 Transformer 的注意力机制和脑电信号的时空特性,在 PhysioNet 四分类运动想象数据集上获得了 64.22% 的平均准确率。但该模型的解码精度并不高,因为其虽然使用 Transformer 模型提取了 EEG 信号的全局依赖关系,但忽略了局部特征的学习。为了解决单一网络模型的局限性,越来越多的 Transformer 混合架构被提出。Song 等^[10]设计 EEG Conformer 模型,通过融合 CNN 部时空特征提取与 Transformer 自注意力的长时序依赖建模,在运动想象四分类 BCI competition IV 2a 数据集上取得了 78.66% 的平均准确率。Hamidi 等^[11]提出了一种结合 Transformer 和图卷积网络(Graph Convolutional Networks, GCN)技术的混合方法,Transformer 模型用于预训练,以强调时间动态,GCN 则设计用来捕捉空间依赖性,在 PhysioNet 四分类运动想象数据集上实现了 97.43% 的平均分类准确率。这些混合模型架构能够协同利用不同模型的互补特性,从而在复杂 EEG 模式识别任务中提升性能。然而,这种性能的提升通常伴随着混合模型复杂度的急剧增加,融合多个模块会导致模型参数量庞大、计算复杂度高,对训练和推理所需的计算资源提出了严峻挑战。此外,复杂的模型结构也增加了优化难度,并限制其在资源受限的边缘设备上的实际部署能力。

针对混合模型带来的计算负担与部署难题,知识蒸馏作为一种将复杂“教师”模型知识迁移到轻量“学生”模型的通用模型压缩范式,为权衡精度-复杂度提供了有效路径^[12]。知识蒸馏方法的核心思想在于学生模型模仿学习教师模型输出的蕴含丰富知识的“软标签”(Soft Labels),以实现在学生模型保持结构简单高效的同时提升其性能。研究人员已将知识蒸馏技术应用于各种基于 EEG 的任务中。Liu 等^[13]提出了一种跨模态知识蒸馏框架,该框架同时建模皮肤电反应信号和 EEG 信号的异质性和交互性,在不显著降低性能的情况下减少了对多模态数据的依赖,使情感识别更加适用和可行。Wu 等^[14]提出了一种基于多级教师辅助的知识蒸馏框架,该框架在保持高性能的同时压缩深度学习模型用于 MI 分类任务,在运动想象四分类 BCI competition IV 2a 数据集上将基准学生模型的平均准确率提高了 6.61%,同时与教师模型相比,模型尺寸减小了近 90%。当前高性能的 MI 解码模型往往结构复杂、参数量大,难

以满足 BCI 系统对实时性、低功耗和嵌入式部署的迫切需求。因此,将知识蒸馏这种高效的模型压缩和迁移范式应用于 MI 脑电解码,具有重要的研究价值和广阔的应用前景。

为此,本研究提出一种基于双分支网络的动态知识蒸馏运动想象脑电解码方法,旨在解决高性能模型固有的高计算复杂度问题,实现精度与效率的有效平衡。首先,采用数据增强策略对脑电数据进行扩展;其次,设计一个高效的多尺度时空卷积模块,该模块并行使用 3 个不同时间尺度的卷积核提取多尺度时间特征,并将特征拼接后通过单一空间卷积层统一处理。这种“先时间多尺度融合,后空间统一处理”的设计,相比传统的“各分支独立时空处理再拼接”方式,不仅减少了参数量和计算量,还实现了跨尺度特征的融合;然后,设计了 Transformer 与双向 LSTM(BiLSTM)并行的双分支异构网络,共享多尺度时空卷积模块,通过双分支协同训练策略,结合 Transformer 的全局建模能力和 BiLSTM 的时序动态捕捉优势,设计加权联合损失函数来调节网络参数,实现特征互补,兼顾全局与局部的特征表达,极大提高了教师模型的解码精度;最后,在知识蒸馏阶段,选择参数量少,计算复杂度低的 EEGNet 作为轻量化的学生模型,其深度可分离卷积结构契合 EEG 信号具有的时空特性。在蒸馏过程中,本文提出了双分支动态知识蒸馏策略,在每轮训练前使用验证集计算教师模型各个分支准确率,并选择当前表现更优的分支作为该轮的知识源,用于指导学生模型训练。相比于固定使用静态融合和单分支教师模型,该策略依据模型学习状态选择最优知识源,可避免无效的知识迁移和固定多分支融合的计算冗余,在实现高性能运动想象脑电解码的同时,可以显著降低模型的参数量和计算复杂度。

1 模型架构与方法

本文设计的模型架构如图 1 所示。该架构主要包括 Transformer 与 BiLSTM 构成的并行双分支教师模型,轻量级的学生模型 EEGNet,以及双分支动态知识蒸馏策略 3 个关键部分。教师模型通过设计联合损失函数,对两个分支的损失进行加权求和调节模型参数,实现两个分支的特征深度互补,从而有效提高运动想象脑电信号解码的精度,并为学生模型提供丰富且鲁棒的迁移知识。学生模型采用了专门为处理 EEG 信号而设计的轻量级卷积神经网络

EEGNet, 该网络主要由时间卷积、深度卷积和可分离卷积3个部分组成, 在保持强大表达能力的同时, 大幅降低模型的参数量和计算复杂度^[15]。双分支动态知识蒸馏策略能够自适应地选取每次训练阶段表现最优的分支作为知识源, 指导学生模型学习, 从而确保知识传递的高质量与鲁棒性。该架构的实现过程如下: 首先, 将输入的脑电数据划分为训练集和验证集, 并对训练集进行数据增强, 以丰富训练样本; 其次, 预训练双分支教师模型获得高精度运动想象任务分类预测能力; 然后, 加载预训练好的教师模型并冻结其权重, 将训练数据同时输入到学生模型和教师模型中, 获取学生模型和教师模型两个分支的原始输出, 再通过验证集的数据评估教师

模型两个分支的解码准确率, 动态选择准确率更高的分支作为教师模型的原始输出; 最后, 将教师模型的原始输出通过温度为 $T=3$ 的 Softmax 进行软化, 生成教师模型的软标签, 同时将学生模型的原始输出分别通过 $T=3$ 的 Softmax 和 $T=1$ 的 Softmax 得到学生模型的软标签和硬标签。计算学生模型的软标签与教师模型的软标签之间的蒸馏损失, 以及学生模型的硬标签和真实标签之间的交叉熵损失, 将这两个损失函数加权求和得到总损失。通过最小化总损失, 学生模型学习拟合真实标签并且模仿教师模型丰富的知识表示, 从而实现知识迁移, 使学生模型在保持较少参数和较低计算复杂度的同时, 获得与教师模型接近的解码精度。

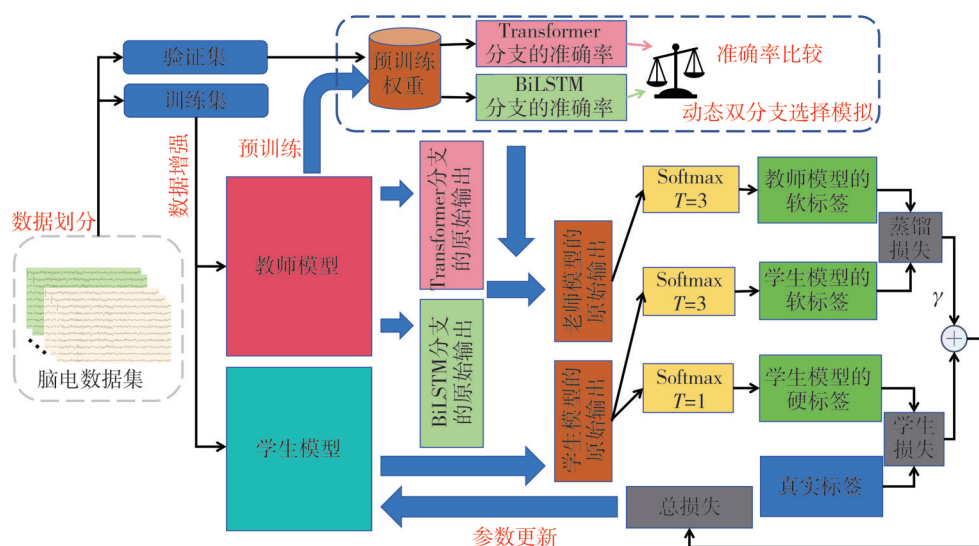


图1 模型整体架构图

Fig. 1 Overall architecture of model

1.1 双分支教师模型

双分支教师模型的结构如图2所示, 它主要由改进的多尺度时空卷积模块, Transformer与BiLSTM构成的并行双分支异构网络及联合训练分类模块三部分构成, 旨在从输入数据中提取全面且高质量的多层次特征表示, 为学生模型提供判别性强的先验知识。

首先, 改进的多尺度时空卷积模块负责提取局部细粒度的时空特征, 显示建模不同尺度特征间的交互关系, 为后续处理提供丰富且判别性强的特征。其次, 设计Transformer与BiLSTM并行的双分支异构网络, 旨在对脑电特征序列实现全面而深入的建模。Transformer凭借其强大的注意力机制, 能有效学习序列内部长距离的全局依赖

关系; 而BiLSTM以其门控机制可从前向与后向两个角度精准地捕捉序列的动态演化规律与上下文信息。这种并行结构使得局部与全局、静态与动态的时序信息能够被互补地捕获与融合。最后, 联合训练与分类模块用于协同优化上述两个分支, 将双通路汇聚的丰富特征进行高层次整合与决策, 使网络在学习过程能平衡并利用不同分支的优势, 从而避免单一结构的局限性, 最终有效提升模型对复杂场景的综合判别能力。

1.1.1 多尺度时空卷积

改进的多尺度时空卷积模块结构如图2所示, 输入数据通过3个并行的时间卷积层, 每层卷积核的大小都不同, 这种多尺度设计允许网络从输入数据中捕获各种与时间相关的特征, 获得更

丰富的脑电特征信息;然后,将各时间卷积层的输出在通道维度拼接融合,形成多尺度时序特征

的集成表达。将融合后的特征通过空间卷积层整合空间信息,进一步细化特征提取过程。

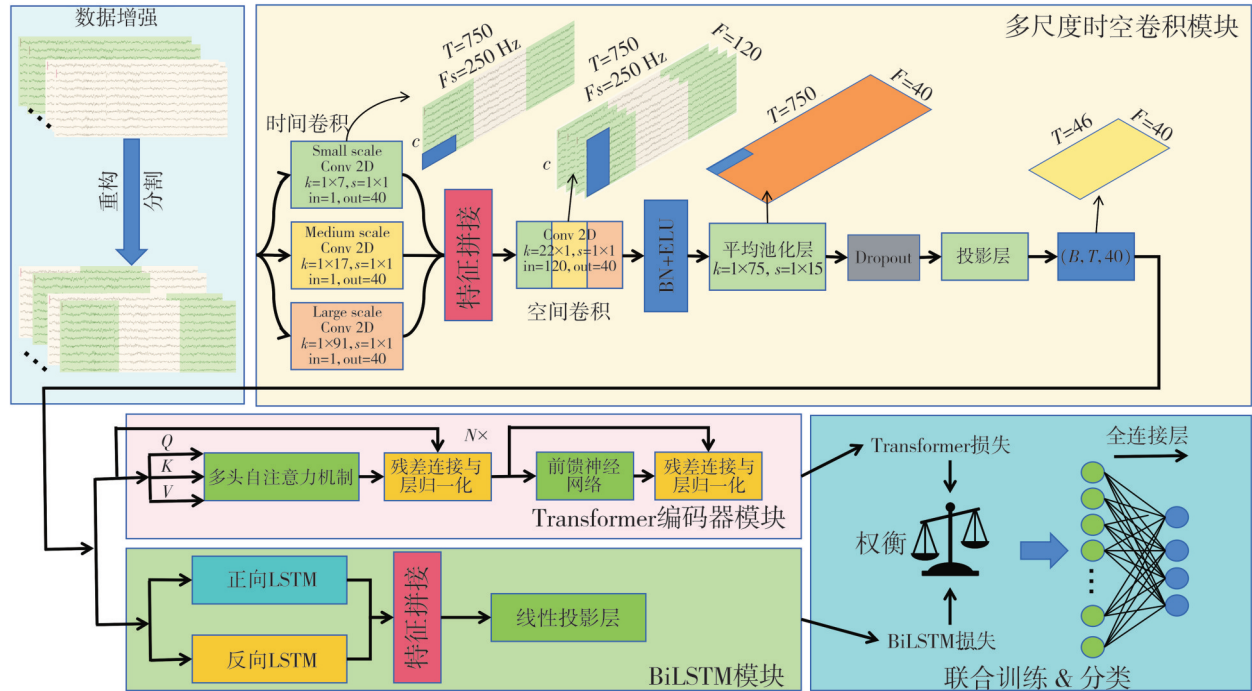


图 2 双分支教师模型的结构图

Fig. 2 Architectural diagram of dual-branch teacher model

随后,应用批归一化和指数线性单元激活函数,来稳定训练过程和缓解梯度消失问题。接着,用平均池化层进行下采样,降低时间维度,增强局部特征鲁棒性和减少后续计算复杂度。最后,使用投影层将4D时空特征张量通过通道调整和维度重组,转换为3D序列化表示,这种结构化转换既保留了原始时空关系,又适配了Transformer和BiLSTM对序列输入的建模要求。这种“先时间多尺度融合,后空间统一处理”的设计,相比传统的“各分支独立时空处理再拼接”的多尺度时空卷积方式,不仅可以降低参数量,实现精准的时序建模和高效的特征融合,还可以避免传统方法中多个独立分支在时空域的特征冗余与干扰,使模型能够学习到更具判别性和鲁棒性的时空特征表示。

1.1.2 异构双分支网络

异构双分支网络的结构如图3所示,该网络包含并行的Transformer和BiLSTM分支,每个分支都配有独立的分类头,并通过联合损失函数进行协同训练。

Transformer分支采用其编码器结构,通过多头注意力机制捕捉序列内部的全局依赖关系,形

成富含上下文信息的特征表示^[16]。而BiLSTM分支则通过其双向门控机制,从前向与后向两个角度建模序列的时序依赖关系,专注于捕捉序列中稳定存在的局部与长程时序模式^[17]。本研究设计了一个联合损失函数,通过对两分支的交叉熵损失加权求和得到总损失,利用总损失来调节模型参数。此外,通过贝叶斯优化来选择最优的损失权重,其关键步骤如下:

1) 计算各分支的交叉熵损失 L_{CE} , L_{CE} 通过衡量真实概率分布与预测概率分布之间的差异,并利用其优良的梯度特性,成为分类任务的首选损失函数,其定义为

$$L_{CE} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^M y \lg(\hat{y}), \quad (1)$$

式中: M 为脑电信号类别数量; y 与 $\hat{y}$ 分别为真实标签和预测标签; N 为一个批次的试验次数。

2) 计算联合损失函数,联合损失函数定义为两个分支损失的加权和,

$$L_{joint} = \lambda L_{trans} + (1 - \lambda) L_{bilstm}, \quad (2)$$

式中: L_{joint} 为双分支的联合损失; L_{trans} 为Transformer分支的损失; L_{bilstm} 为BiLSTM分支的损失; $\lambda \in [0, 1]$ 为权重系数。

3) 通过贝叶斯优化自动选择最优权重 λ^* , 优

化目标为

$$\lambda^* = \underset{\lambda \in [0,1]}{\operatorname{argmin}} E[f(\lambda)], \quad (3)$$

$$f(\lambda) = 1 - \operatorname{Accuracy}(\lambda), \quad (4)$$

式中: $f(\lambda)$ 为验证集上使用 λ 权重时的性能度量, 如准确率; $E[f(\lambda)]$ 为期望算子; argmin 为参数优化, 用于找到使目标函数最小的 λ 值。

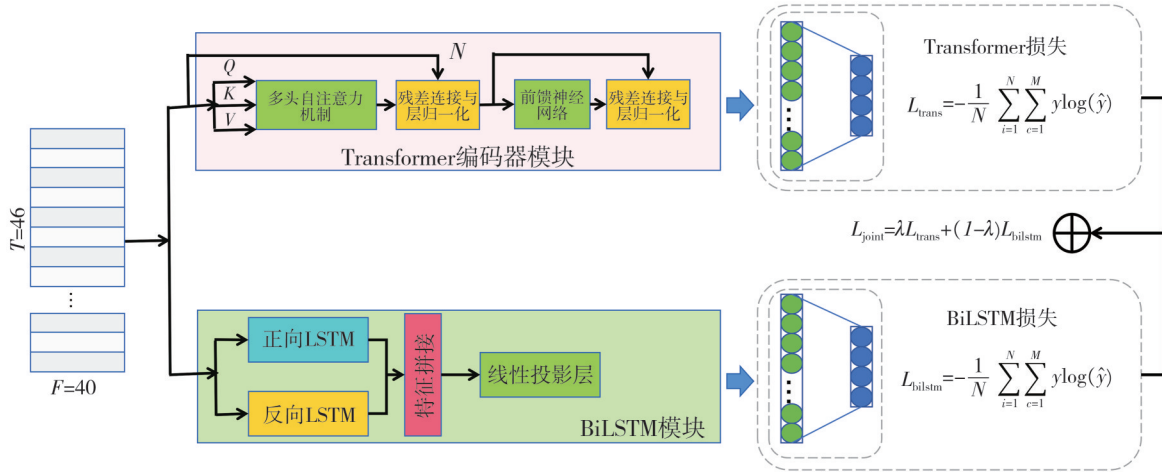


图3 异构双分支网络结构

Fig. 3 Architecture of heterogeneous dual-branch network

4) 使用高斯过程对 $f(\lambda)$ 建模,

$$f(\lambda) \sim gp(\mu(\lambda), k(\lambda, \lambda')), \quad (5)$$

式中: gp 为高斯过程, 用于对未知函数建模; $\mu(\lambda)$ 为均值函数, 预测 $f(\lambda)$ 的期望值; $k(\lambda, \lambda')$ 为高斯核函数, 衡量两个 λ 点和 λ' 之间的相关性。

5) 每次迭代中, 通过优化采集函数选择下一个评估点 λ_{next} ,

$$\lambda_{\text{next}} = \underset{\lambda}{\operatorname{argmax}} EI[\lambda], \quad (6)$$

$$EI[\lambda] = E[\max(f(\lambda)_{\min} - f(\lambda), 0)], \quad (7)$$

式中: $EI[\lambda]$ 为期望改进函数; $f(\lambda)_{\min}$ 为当前观察到的最佳性能值。选择使得 EI 最大的 λ 作为下一个评估点。

6) 网络参数 θ 通过梯度下降更新,

$$\theta_{i+1} = \theta_i - \eta \nabla_{\theta} L_{\text{joint}}(\theta_i), \quad (8)$$

式中: η 为学习率; $L_{\text{joint}}(\theta_i)$ 为第 i 个网络参数的联合损失项, 梯度计算公式为

$$\nabla_{\theta} L_{\text{joint}} = \lambda^* \nabla_{\theta} L_{\text{trans}} + (1 - \lambda^*) \nabla_{\theta} L_{\text{bilstm}}, \quad (9)$$

式中: ∇_{θ} 为梯度算子, 计算损失函数对网络参数的偏导数; $\nabla_{\theta} L_{\text{trans}}$ 为 Transformer 分支的梯度; $\nabla_{\theta} L_{\text{bilstm}}$ 为 BiLSTM 分支的梯度。

与单分支模型相比, Transformer 与 BiLSTM 构成的异构双分支网络既可以缓解 Transformer 在处理局部时序模式上的不足, 也能突破 BiLSTM 在捕捉长程全局依赖关系上的瓶颈, 实现全局与局部特征的互补。此外, 相比双分支特征拼接的

静态融合, 该网络通过联合损失函数在训练过程动态调节及平衡两个分支的学习, 促使二者充分发挥各自的特长并实现深度融合, 从而有效缓解传统融合中常出现的单分支主导而另一分支贡献不足的问题。

1.2 EEGNet 学生模型

EEGNet 是一个轻量级的卷积神经网络, 它通过结合脑电信号处理的领域知识深度学习方法来高效地处理 EEG 数据。其结构如图 4 所示, 模型主要由时间卷积、深度卷积和可分离卷积 3 个部分组成。首先, 使用多个一维的时间卷积核, 仅在时间轴上滑动以捕获信号的时间动态特性, 同时保持空间通道的独立性。此操作专注于单一维度特征提取, 避免了多维卷积带来的参数膨胀, 为后续处理提供高效的时序特征表示。接着, 采用深度卷积操作实现跨通道的空间特征整合, 深度卷积专门负责空间卷积, 通过一组覆盖所有电极通道的空间卷积核, 对前一模块输出的每个时间特征图进行独立的空间滤波, 每个卷积核独立处理单个输入通道, 避免了通道间的全连接参数, 大幅减少了参数数量。在获得时空联合特征后, 网络通过可分离卷积进一步提取深层抽象特征。依次执行深度卷积和逐点卷积, 深度卷积在特征维度上进行深层次的时间模式挖掘, 而逐点卷积则专注于通道维度特征组合, 通过轻量

级的 1×1 卷积实现特征重新加权。这种分离策略确保网络在保持强大表达能力的同时,通过功能

专一化的卷积操作大幅降低参数量和计算复杂度。

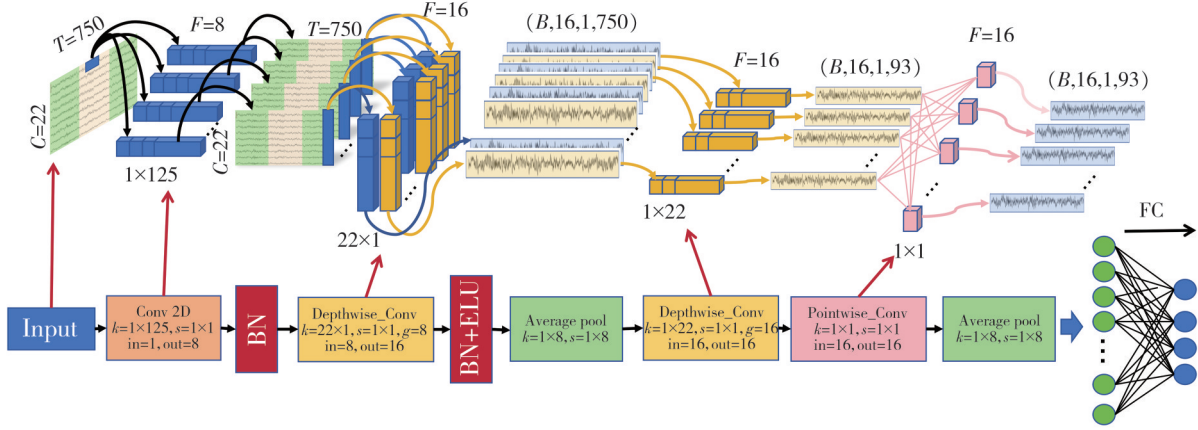


图 4 EEGNet 的模型结构图

Fig. 4 Structure of EEGNet model

1.3 动态知识蒸馏策略

针对运动想象任务分类中脑电信号存在个体差异性,本文提出一种双分支动态知识蒸馏策略。该策略的核心思想是利用教师模型的双分支结构生成多样化的知识表示,并根据当前各分支在验证集上的表现动态选择最优知识源进行蒸馏。每轮训练开始前,在验证集上评估教师模型两个分支的当前性能,如式(10)和(11)所示。

$$Acc_{trans} = Accuracy(y, \hat{y}_{trans}), \quad (10)$$

$$Acc_{bilstm} = Accuracy(y, \hat{y}_{bilstm}), \quad (11)$$

式中: $\hat{y}_{trans}$ 和 $\hat{y}_{bilstm}$ 分别为 Transformer 和 BiLSTM 两个分支的预测值; Acc_{trans} 和 Acc_{bilstm} 分别为 Transformer 和 BiLSTM 两个分支的准确率。根据评估结果,选择当前性能最优的分支作为本轮蒸馏的知识源,如式(12)所示。

$$B^* = \begin{cases} \text{Transformer}, & \text{if } Acc_{trans} \geq Acc_{bilstm}, \\ \text{BiLSTM}, & \text{otherwise.} \end{cases} \quad (12)$$

该动态选择机制确保在蒸馏过程始终使用最可靠的知识源,从而适应脑电信号的动态特性。基于选择的最优分支 B^* , 构建动态蒸馏损失函数,如式(13)所示。

$$L_{KD} = (1 - \gamma) \times L_{CE}(y, \sigma(z_s)) + \gamma \times T^2 \times L_{KL}(\sigma(z_t^B/T) // \sigma(z_s/T)), \quad (13)$$

式中: z_s 和 z_t 分别为教师模型和学生模型的输出; σ 为 softmax 函数; T 为温度参数,用于软化概率分布; γ 为平衡蒸馏损失和学生硬标签损失的权重系数; L_{KD} 为知识蒸馏的总损失; L_{KL} 为 Kullback

Leibler 散度。

该损失函数将传统知识蒸馏的单一知识源扩展为动态选择的最优知识源。与传统知识蒸馏相比,所提出的双分支动态知识蒸馏策略具有三重优势:知识多样性、动态适应性和脑电特异性。传统方法依赖单一静态教师模型,知识来源受限且固定,难以充分捕捉脑电信号的复杂时空特性。通过并行的 Transformer 和 BiLSTM 构建双路径知识库,在每轮训练前基于验证集性能实时选择最优分支作为知识源,实现动态知识传递。这种策略不仅能根据被试的脑电特性自动适配最佳知识表示,还能响应信号的非平稳变化。在保证提高计算效率的同时,提升教师模型知识迁移的针对性和完备性,使学生模型具有更强的泛化能力。

2 模型实现与评价

2.1 数据集

实验数据集是 2008 年 BCI 竞赛 IV 的数据集 2a,由奥地利格拉茨工业大学提供^[18]。该数据集由 9 名被试的 EEG 数据组成,每个被试在实验过程中需要按照屏幕箭头所指方向分别想象左手、右手、双脚和舌头的运动,每次想象任务持续约 3 s。每名被试在不同日期记录了两组运动想象脑电数据,每组试验中每名被试进行 6 轮运动想象,每轮包括 4 种运动想象任务各 12 次,共得到 288 个实验样本。将每个被试不同日期记录的两组脑电数据进行合并得到 576 个实验样本。该数据集

按照图 5 的实验范式使用 25 个 Ag/AgCl 电极采集脑电数据, 采样的频率为 250 Hz, 其中 EEG 电极有 22 个, 眼电(EOG)电极有 3 个。实验中, 3~6 s 是大脑进行运动想象的时间段。

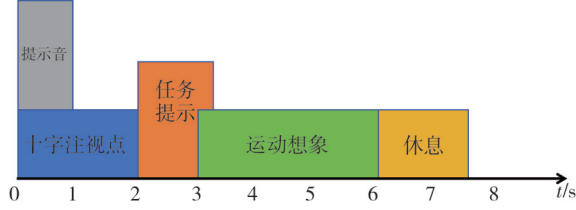


图 5 BCI competition IV 2a 运动想象实验范式

Fig. 5 BCI competition IV 2a motor imagery experimental paradigm

对采集记录的脑电数据进行预处理, 预处理的流程如下: 移除 3 个 EOG 通道并保留 EEG 通道, 进行电极定位; 使用 0.5~40 Hz 带通滤波以抑制噪声和干扰, 并进行平均参考; 通过独立成分分析去除眼电、心电等伪迹; 截取运动想象任务时段(3~6 s), 每个样本对应相应任务标签。最终每位被试获得 576 个样本, 按 7:2:1 划分为训练集、验证集和测试集, 分别用于模型训练、调优和最终评估。此外, 为解决运动想象脑电训练数据量少的问题, 采用时域分割与重构(S&R)的方法对训练数据进行实时的扩充^[19]。该方法将同类型任务的 EEG 样本按时间分割为 N 个片段(如 $N=3$, 则分为 A、B、C 三段), 并在保持时序的前提下随机重组这些片段, 每次迭代生成与批次等量的新数据。

2.2 模型训练

双分支网络的动态知识蒸馏算法的整体训练流程如下。

双分支网络的动态知识蒸馏算法

输入: Transformer-BiLSTM 双分支教师模型 $f_T = [f_T^{\text{trans}}, f_T^{\text{bilstm}}]$, 学生模型 f_S , 温度参数 τ , 蒸馏损失权重 γ , 训练数据 D_{train} , 验证数据 D_{val} , 教师学习率 η_T , 学生学习率 η_S , 联合损失权重 λ 。

预训练教师模型:

- 1 初始化 Transformer 分支参数 θ_T^{trans} , 初始化 BiLSTM 分支参数 θ_T^{bilstm}
- 2 While not converged do:
- 3 $X, Y \sim D_{\text{train}}$
- 4 $X_{\text{aug}}, Y_{\text{aug}} \sim \text{S\&R_Augment}(X, Y)$
- 5 $X_{\text{batch}} = [X, X_{\text{aug}}], Y_{\text{batch}} = [Y, Y_{\text{aug}}]$
- 6 $z_{\text{trans}} = f_T^{\text{trans}}(X_{\text{batch}}; \theta_T^{\text{trans}})$
- 7 $z_{\text{bilstm}} = f_T^{\text{bilstm}}(X_{\text{batch}}; \theta_T^{\text{bilstm}})$
- 8 $L_{\text{trans}} = \text{CrossEntropy}(z_{\text{trans}}, Y_{\text{batch}})$

$$9 \quad L_{\text{bilstm}} = \text{CrossEntropy}(z_{\text{bilstm}}, Y_{\text{batch}})$$

$$10 \quad L_{\text{teacher}} = \lambda L_{\text{trans}} + (1 - \lambda) L_{\text{bilstm}}$$

$$11 \quad \nabla_{\theta_T} L_{\text{teacher}}, \text{ 其中 } \theta_T = \{\theta_T^{\text{trans}}, \theta_T^{\text{bilstm}}\}$$

$$12 \quad \theta_T \leftarrow \theta_T - \eta_T \nabla_{\theta_T} L_{\text{teacher}}$$

13 end while

14 保存预训练教师模型参数 θ_T

知识蒸馏训练:

1 加载预训练教师模型参数 θ_T , 冻结所有参数, 初始化学生模型 EEGNet 参数 θ_S

2 While not converged do:

$$3 \quad f_T^* = \begin{cases} f_T^{\text{trans}}, & \text{if } \text{perf}_{\text{trans}} \geq \text{perf}_{\text{bilstm}}, \\ f_T^{\text{bilstm}}, & \text{otherwise.} \end{cases}$$

$$4 \quad X, Y \sim D_{\text{train}}$$

$$5 \quad X_{\text{aug}}, Y_{\text{aug}} \sim \text{S\&R_Augment}(X, Y)$$

$$6 \quad X_{\text{batch}} = [X, X_{\text{aug}}], Y_{\text{batch}} = [Y, Y_{\text{aug}}]$$

$$7 \quad p_T = \text{softmax}(f_T^*(X_{\text{batch}})/\tau)$$

$$8 \quad p_S = \text{softmax}(f_S(X_{\text{batch}})/\tau)$$

$$9 \quad Y_S = \text{softmax}(f_S(X_{\text{batch}}))$$

$$10 \quad L_{\text{KD}} = \text{KLDivergence}(p_S // p_T)$$

$$11 \quad L_{\text{student}} = \text{CrossEntropy}(Y_S, Y_{\text{batch}})$$

$$12 \quad L_{\text{total}} = \gamma \tau^2 \cdot L_{\text{KD}} + (1 - \gamma) \cdot L_{\text{student}}$$

$$13 \quad \nabla_{\theta_S} L_{\text{total}}$$

$$14 \quad \theta_S \leftarrow \theta_S - \eta_S \nabla_{\theta_S} L_{\text{total}}$$

15 end while

输出: 训练好的学生模型 f_S

本研究的模型训练过程分为两个核心阶段: 双分支教师模型的预训练和基于知识蒸馏的学生模型训练。实验环境为 pycharm2020.1.3, 采用 pytorch 框架搭建模型, 硬件环境为 NVIDIA GeForce RTX 3080Ti GPU。教师模型和学生模型采用相同的优化器和批次大小(batchsize), 训练轮次(epochs)都设置为 1 000, 并使用早停法防止模型过拟合。模型的超参数设置如表 1 所示, 这些超参数的具体数值, 包括学习率和权重系数, 均是通过贝叶斯优化方法自动寻优确定。利用贝叶斯优化框架, 以验证集上的性能作为优化目标, 经过多轮迭代, 获得了表 1 中所列的最优超参数组合, 从而确保模型能够达到最佳性能潜力。模型在被试 A03 的准确率曲线和损失曲线如图 6 所示。

表 1 模型超参数设置

Tab. 1 Model hyperparameter settings

超参数	最优值
教师学习率(η_T)	0.000 623
学生学习率(η_S)	0.001
批量大小(batchsize)	16
联合损失权重 λ	0.44
蒸馏损失权重 γ	0.7
温度参数 τ	3
优化器	Adam

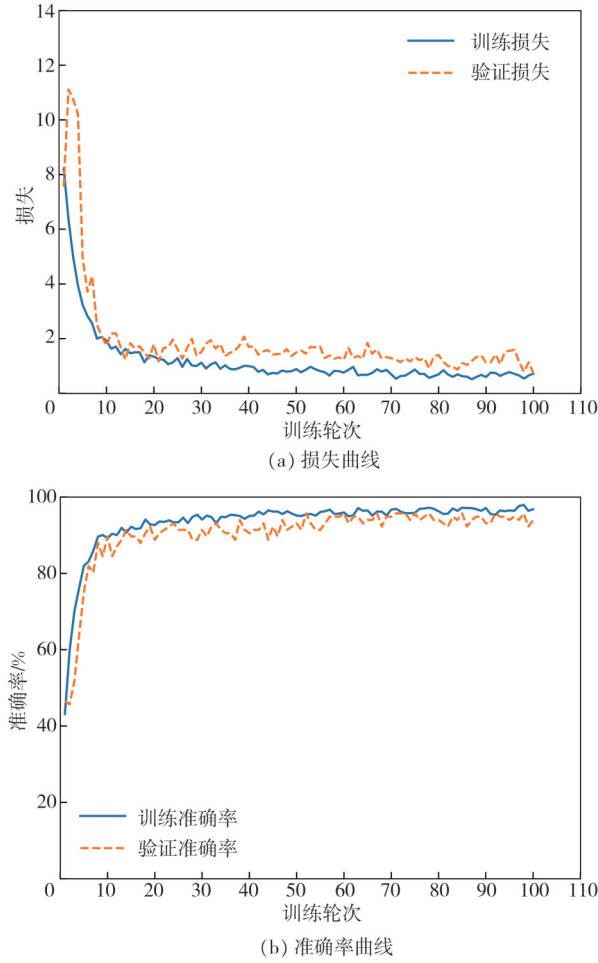


图 6 被试 A03 学生模型的准确率曲线和损失曲线

Fig. 6 Accuracy and loss curves of student model for subject A03

2.3 评价指标

为全面评估模型的性能,采用多种指标对模型进行分类性能、计算效率的综合评价。

运动想象任务分类准确率(Accuracy)是指通过训练集训练得到的模型在测试集上进行分类时,能够被正确分类的样本数与总样本数的比值,如式(14)所示。

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \quad (14)$$

式中: TP 为将正类预测为正类; TN 为将负类预测为负类; FP 为将负类预测为正类; FN 为将正类预测为负类。在多任务分类问题中,把某一类别作为正类,其余类别均被视为负类。

Kappa 系数是一种用于衡量分类结果与真实标签之间一致性的指标。它考虑了随机猜测可能带来的正确率,比单纯的准确率更为鲁棒,其值可基于混淆矩阵计算,

$$\begin{cases} Kappa = \frac{P_0 - P_e}{1 - P_e}, \\ P_e = \frac{\text{实际数量} \times \text{预测数量}}{(\text{样本总数})^2}, \end{cases} \quad (15)$$

式中: P_0 为分类准确率; P_e 为随机一致性。Kappa 系数的取值范围通常在 $[-1, 1]$ 之间,值越接近 1,表示一致性越高,模型的性能越好。以 BCI IV 2a 数据集中被试 A03 为例,其混淆矩阵如图 7 所示。

真实标签	左手	右手	脚	舌头
	100.0	0.0	0.0	0.0
	0.0	96.6	0.0	3.4
	3.4	0.0	96.6	0.0
预测标签	左手	右手	脚	舌头
	0.0	3.4	10.3	86.2

图 7 被试 A03 的混淆矩阵

Fig. 7 Confusion matrix for subject A03

此外,使用参数量(Parameter)和计算量(FLOPs)作为模型的效率指标。*Parameter* 指模型中所有需要学习的权重和偏置的总数量,该指标直接反映了模型的大小和内存占用情况,是评估模型轻量化程度的关键指标。*FLOPs* 通常用前向传播过程中所需的浮点运算次数来衡量,体现了模型的计算复杂度和推理速度。模型的 *Parameter* 和 *FLOPs* 取决于模型的结构和类型,通常集中在全连接层、卷积层及注意力层。全连接层的 $Parameter_{FC}$ 和 $FLOPs_{FC}$ 主要取决于输入特征维度 D 、输出特征维度 E 及偏置项,每一个输出节点都有一个偏置量,计算公式为

$$Parameter_{FC} = (D \times E) + E, \quad (16)$$

$$FLOPs_{FC} = 2 \times D \times E + E. \quad (17)$$

卷积层的 $Parameter_{cov}$ 和 $FLOPs_{cov}$ 取决于输入通道数 c_{out} 、输出通道数 c_{in} 、卷积核的大小 $k_h \times k_w$ 及输出特征的尺寸 $k_{out} \times W_{out}$, 计算公式为

$$Parameter_{cov} = (c_{in} \times c_{out} \times k_h \times k_w) + c_{out}, \quad (18)$$

$$FLOPs_{cov} = 2 \times H_{out} \times W_{out} \times c_{in} \times c_{out} \times k_h \times k_w + H_{out} \times W_{out} \times c_{out}. \quad (19)$$

注意力层的 $Parameter_{MHA}$ 和 $FLOPs_{MHA}$ 取决

于序列长度 L 、模型的维度 d_{model} 、注意力头数 h 、每个注意力头的维度 d_k 及偏置项,计算公式为

$$\text{Parameter}_{\text{MHA}} = 3 \times d_{\text{model}} \times d_k \times h + d_{\text{model}} \times d_{\text{model}} + 4 \times d_{\text{model}}, \quad (20)$$

$$\text{FLOPs}_{\text{SMHA}} = 2 \times L \times (4 \times d_{\text{model}}^2 + 2 \times L \times d_{\text{model}}) + 3 \times L^2 + 4 \times L \times d_{\text{model}}. \quad (21)$$

3 实验与分析

通过设计不同的实验验证所提方法的有效性与优越性。实验主要包括3个部分:首先,通过消融实验验证模型核心组件的贡献;其次,与当前主流模型进行对比,证明所提方法的综合性能优势;最后,通过特征可视化定性分析模型的特征学习能力。

3.1 消融实验

为系统性地评估所提方法中各核心模块的有效性,在BCI competition IV 2a数据集上对9名被试进行了消融实验研究。所有实验均使用相同训练/验证/测试集划分,以及数据预处理与增强的策略,以确保结果的公平可比性。消融实验分为教师模型和学生模型两部分。

3.1.1 双分支教师模型的消融实验

该消融实验旨在逐步验证高性能教师模型的3个核心优势:“先时间多尺度融合,后空间统一处理”多尺度时空卷积模块的效率与性能优势;Transformer与BiLSTM并行的异构双分支网络架构的特征表征能力;改进的联合损失函数相较于普通特征融合策略的优越性。

实验一为基础实验,使用传统的多尺度时空卷积模块,即3个独立的时空卷积分支并行处理,然后进行特征融合,再将融合后的特征输入到Transformer模型进行分类。此实验旨在建立性能基线,用于对比后续所有改进。

实验二在实验一的基础上将传统的多尺度时空卷积模块替换为“先时间多尺度融合,后空间统一处理”模块,旨在评估该模块在减少计算复杂度和提升特征表达方面的有效性。

实验三在实验二的基础上将Transformer替换为BiLSTM,评估BiLSTM模型在本任务中的独立性能,旨在与后续的双分支架构进行对比。

实验四在实验二的基础上引入实验三中的BiLSTM,构建一个Transformer与BiLSTM并行

的异构双分支网络。该实验采用常见的特征融合策略:将两个分支输出的特征序列在特征维度进行拼接,然后输入到一个共同的分类头。此实验旨在验证双分支架构的有效性。

实验五在实验四的基础上,摒弃特征拼接,采用加权联合损失函数进行协同训练,旨在验证联合损失函数作为一种更灵活、更高效的协同机制,相较于仅进行特征拼接的静态融合的优越性。

实验结果如图8和表2所示,提出的双分支教师模型其各个核心模块均对最终脑电解码性能的提升起到了关键作用。

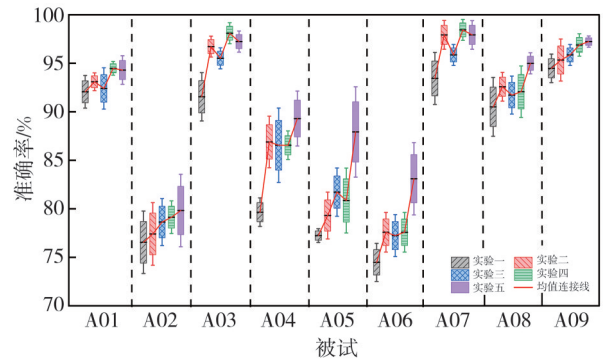


图8 各个被试的教师模型消融实验对比

Fig. 8 Comparison of ablation studies on teacher model for each subject

表2 教师模型消融实验的平均解码准确率对比

Tab. 2 Comparison of average decoding accuracy in teacher model ablation studies

	平均准确率/%	Kappa	Parameter/M	FLOPs/M
实验一	85.55	0.816 9	1.262 9	158.72
实验二	88.54	0.847 2	0.709 9	154.56
实验三	88.39	0.848 0	0.660 1	152.38
实验四	89.35	0.858 0	1.257 8	154.25
实验五	91.32	0.884 3	1.257 8	158.25

首先,对比实验一和实验二可以看出,将传统的多尺度时空卷积模块替换为“先时间多尺度融合,后空间统一处理”结构,其平均分类准确率从85.55%提升至88.54%, $Kappa$ 系数从0.816 9提升至0.847 2。更重要的是,在性能提高的同时,模型的 $Parameter$ 从1.262 9 M下降至0.709 9 M, $FLOPs$ 也略有降低。这一结果表明,本文提出的多尺度模块不仅在结构上更为高效,还增强了对时空特征的特征能力。实验三仅使用BiLSTM作为任务分类骨干网络,取得了88.39%的平均准确率,低于实验二中使用的Transformer结构,说明Transformer在捕获全局时序依赖方面表现更优。实验四将Transformer与BiLSTM并

行组合成双分支网络结构, 准确率提升至 89.35%, $Kappa$ 值达到 0.858 0, 这一改进证实了异构双分支架构的有效性。实验五采用联合损失函数替代实验四的特征拼接策略, 在保持 $Parameter$ 与实验四基本一致的情况下, 平均准确率进一步提升至 91.32%, $Kappa$ 值提高至 0.884 3。该结果表明联合损失函数能够更有效地调节模型参数, 达成特征深度融合, 避免特征拼接带来的信息冲突或冗余, 从而实现了更优的协同训练效果。

3.1.2 学生模型的消融实验

为了验证双分支动态知识蒸馏策略在学生模型训练中的有效性, 在 BCI competition IV 2a 数据集上针对同一预训练的双分支教师模型, 设计多种蒸馏策略进行消融实验。所有实验均使用相同的 EEGNet 学生模型架构、初始化权重、训练集/验证集/测试集划分, 以及优化器设置(学习率、批次大小等), 以确保对比的公平性。实验设置如下:

实验 A: 基线实验, 无蒸馏的独立训练。学生模型 EEGNet 不接收教师模型指导, 单独训练, 此实验旨在建立学生模型不借助知识蒸馏的基线性能。

实验 B: 静态 Transformer 分支蒸馏。使用预训练教师模型中固定的 Transformer 分支作为知识源进行知识蒸馏, 该实验用于验证单一 Transformer 教师模型的知识迁移效果。

实验 C: 静态 BiLSTM 分支蒸馏。使用预训练教师模型中固定的 BiLSTM 分支作为知识源进行知识蒸馏, 该实验用于验证单一 BiLSTM 教师模型的知识迁移效果。

实验 D: 静态平均融合蒸馏。将教师模型两个分支输出的软标签进行平均, 得到一个融合的软标签, 然后用其指导学生模型训练, 这是一种常见的多教师蒸馏策略。

实验 E: 双分支动态知识蒸馏。在每轮训练前, 根据两个分支在验证集上获得的当前准确率, 动态选择表现更优的分支作为该轮的知识源。图 9 和表 3 展示了采用不同蒸馏方法时, 学生模型在各个被试达到的分类准确率对比结果。

由图 9 可知, 双分支动态知识蒸馏在大多数被试上取得了最优或接近最优的分类性能, 其平均准确率为 88.85%, 显著高于基线实验 A 的 82.22%, 也高于其他静态蒸馏策略。该结果表明双分支动态蒸馏策略能更有效地从教师模型中提取特征并迁移

判别性知识, 进一步提升学生模型的分类准确率。由两种单一分支的静态蒸馏实验结果可知, 不同结构教师分支的知识迁移能力存在明显被试依赖性。实验 B 在被试 A01、A03、A04、A05 和 A07 上平均准确率高于实验 C, 说明 Transformer 分支对这些被试的 EEG 特征具有更强的任务类别表示能力; 而在被试 A02、A06、A08 和 A09 上, 实验 C 的表现优于实验 B, 表明 BiLSTM 分支在这些被试上提供更具指导性的软标签。这一差异凸显了单一分支蒸馏策略的局限性: 其效果高度依赖于被试特有的脑电数据分布和特征模式。实验 D 通过融合双分支软标签, 在一定程度上缓解了单一分支的不稳定性。然而, 其静态平均的融合方式无法适应不同被试和训练阶段的动态需求。相比之下, 双分支动态知识蒸馏方法能够自适应选择表现更优的教师分支输出软标签, 从而实现更好的知识迁移。

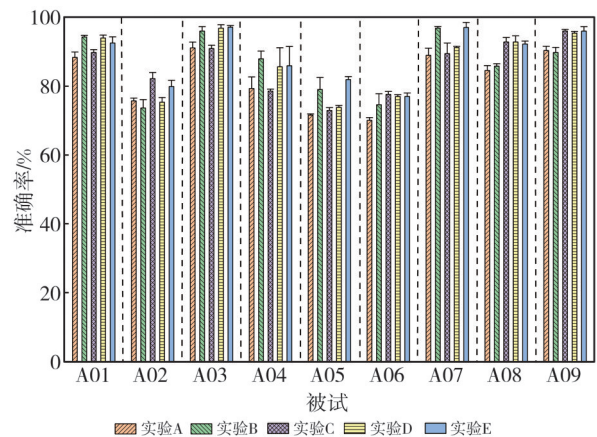


图 9 各个被试的学生模型消融实验对比

Fig. 9 Comparison of ablation studies on student model for each subject

表 3 学生模型消融实验的平均解码准确率对比

Tab. 3 Comparison of average decoding accuracy in student model ablation studies

	平均准确率/%	$Kappa$
实验 A	82.22	0.763 0
实验 B	86.43	0.819 0
实验 C	85.58	0.807 5
实验 D	86.00	0.825 3
实验 E	88.85	0.851 6

3.1.3 知识蒸馏效率验证

为验证知识蒸馏方法对模型效率的影响, 设计实验对比无蒸馏学生模型、双分支教师模型以及知识蒸馏模型, 对运动想象脑电解码的各种性能参数, 结果如表 4 所示。

表4 蒸馏前后的性能与效率评估结果

Tab. 4 Evaluation of performance and efficiency before and after distillation

实验	平均准确率/%	<i>Kappa</i>	<i>Parameter</i> /M	<i>FLOPs</i> /M	吞吐量/(样本·s ⁻¹)	峰值GPU内存/M	模型体积/M
双分支教师模型	91.32	0.884 3	1.257 8	158.25	6 481.35	1 498.28	4.85
学生模型(无蒸馏)	82.22	0.763 0	0.002 8	17.40	51 751.38	101.46	0.02
学生模型(蒸馏后)	88.85	0.851 6	0.002 8	17.40	51 751.38	101.46	0.02

图9和表4展示了知识蒸馏在模型压缩与性能提升方面的作用。在性能上,未经蒸馏的学生模型平均准确率为82.22%,而通过知识蒸馏后,其准确率提升至88.85%,提高了6.63百分点。这表明从复杂的教师模型中有效迁移了运动想象脑电特征信息,弥补了学生模型性能的不足。在效率上,学生模型的 $Parameter$ 仅为0.002 8 M, $FLOPs$ 为17.40 M。与复杂的教师模型相比,学生模型仅牺牲2.47百分点的准确率,参数量减少了99.78%, $FLOPs$ 降低了89.02%,模型体积也降低了99.6%。结果表明,知识蒸馏使得学生模型在保留较低 $Parameter$ 和 $FLOPs$ 的同时,可获

得接近教师模型的性能,达到精度与效率之间的良好平衡,为运动想象脑电解码在可穿戴设备应用提供解决方案。

3.2 与其他方法对比

近年来,多种深度学习模型被应用于EEG运动想象解码任务,为了评估模型对于运动想象任务四分类的性能,将其与当前具有代表性的5种先进模型在BCI Competition IV 2a数据集上进行了全面的对比实验。参与对比的模型包括:Conformer^[10]、EEG-TCNet^[20]、MCLANet^[21]、MBFT^[22]和AMFTCNet^[23],结果如表5所示。

表5 本文方法与其他方法对比

Tab. 5 Comparison of proposed method with other methods

被试	分类准确率/%						
	Conformer	EEG-TCNet	MCLANet	MBFT	AMFTCNet	本文方法 教师模型	本文方法 学生模型
A01	88.19	89.32	89.62	94.30	89.58	94.31	92.53
A02	61.46	72.44	78.46	72.36	76.39	79.83	79.89
A03	93.40	97.44	92.05	98.22	97.92	97.24	97.12
A04	78.13	75.87	82.91	90.71	87.85	89.31	85.92
A05	52.08	83.69	77.62	65.89	82.99	87.93	81.90
A06	65.28	70.69	87.15	75.42	78.47	83.10	77.01
A07	92.36	93.14	91.33	96.30	96.88	97.93	97.04
A08	88.19	86.71	87.07	93.58	88.19	95.00	92.24
A09	88.89	85.23	89.24	95.60	91.67	97.24	95.98
均值	78.66	83.84	86.16	86.93	87.77	91.32	88.85
<i>Kappa</i>	0.72	0.78	0.82	—	0.84	0.88	0.85
<i>Parameter</i> /M	0.615 74	0.020 5	—	—	0.040 4	1.257 8	0.002 8
<i>FLOPs</i> /M	47.33	24.2	—	—	—	158.25	17.4

表5中的Conformer方法结合卷积操作局部特征提取与Transformer的全局建模能力,旨在同时捕捉EEG信号的局部时空特征和长程依赖关系。然而,该模型的平均分类准确率仅为78.66%,表明其混合架构在EEG解码任务中存在局限性。EEG-TCNet作为一种轻量时序卷积网络, $Parameter$ 仅0.020 5 M, $FLOPs$ 为24.2 M,注重特征提取的高效性,但结构简单限制了特征表征能力,平均准确率为83.84%。MCLANet采用多分支结构,并通过注意力机制增强特征融合,在9个被试上平均准确率达到86.16%,体现了多分支与注意力机制的有效性,但模型复杂度较高。MBFT提出一种多分支融合

Transformer框架,利用自注意力融合全局时空频谱信息,平均准确率达到86.93%,但在不同被试间性能波动较大,且计算开销显著。AMFTCNet通过多尺度特征提取与动态注意力融合,将平均准确率进一步提升至87.77%,但模型结构较为复杂, $Parameter$ 和 $FLOPs$ 较高。与上述方法相比,本研究的方法在解码精度和模型效率上展现出明显优势。通过双分支动态知识蒸馏策略,轻量化的学生模型仅有0.002 8 M $Parameter$ 和17.4 M $FLOPs$ 的极低资源消耗,运动想象多任务平均分类准确率达到88.85%,为资源受限条件下的脑电解码提供了有效方法。

为了进一步验证所提出方法的泛化能力,本研究在另一个经典的四分类运动想象数据集 BCI Competition III 3a 上进行了对比实验,结果如表 6 所示。实验结果表明,本文提出的基于双分支动态知识蒸馏的框架,在不同的数据集与不同被试个体上均能实现稳定且优越的解码性能。本文的教师模型的平均分类准确率达到 94.77%,其 *Kappa* 系数为 0.884 3,优于其他对比方法。这充分证明了教师模型架构在运动想象解码任务中的

强大表征与泛化能力。此外,作为学生模型的轻量化 EEGNet,在严格继承原始 EEGNet 极低 *Parameter* (0.003 7 M) 和 *FLOPs* (31.55 M) 的前提下,通过知识蒸馏获得了性能的提升,其平均准确率提升至 92.21%,*Kappa* 系数达到 0.851 6,优于各类对比模型。这清晰地表明,本文所采用的蒸馏策略有效地将教师模型的知识迁移到了轻量级学生网络中,在几乎不增加推理开销的情况下提升了模型的性能。

表 6 本文方法与其他方法对比

Tab. 6 Comparison of proposed method with other methods

	分类准确率/%					
	EEGNet ^[15]	DCN-ResNet18 ^[24]	DCNv3-ASAM-ResNet18 ^[25]	HWA ^[26]	本文方法 教师模型	本文方法 学生模型
K3b	96.39	94.78	95.26	96.67	98.76	97.45
K6b	76.72	83.65	86.92	87.50	89.71	85.42
L1b	80.17	85.83	89.24	90.83	95.83	93.75
均值	84.48	88.08	90.47	91.67	94.77	92.21
<i>Kappa</i>	0.787 3	—	—	—	0.884 3	0.851 6
<i>Parameter</i> /M	0.003 7	—	—	—	1.387 3	0.003 7
<i>FLOPs</i> /M	31.55	—	—	—	289.22	31.55

从不同被试个体的结果分析,本文方法对于解码难度不同的被试均表现出良好的适应性。在解码难度较大的被试 K6b 上,学生模型的准确率达到了 85.42%,相较于原始 EEGNet 的 76.72% 有了较大的提升,并接近于教师模型的性能。在被试 L1b 上,学生模型更是取得了 93.75% 的高准确率,与教师模型的差距进一步缩小。这些结果说明,所提方法能够有效缓解脑电信号显著的个体差异性所带来的挑战,在不同被试上均能实现鲁棒的特征解码。

3.3 特征可视化

为了进一步探究双分支教师模型各模块的特征提取与表征能力,采用 t-SNE (t-Distributed Stochastic Neighbor Embedding) 降维技术对不同阶段的 EEG 特征分布进行了可视化分析。以被试 A01 为例进行展示,其结果如图 10 所示。图 10(a) 为原始 EEG 信号的特征分布,4 种运动想象任务的特征类别重叠严重、无明显聚类趋势。图 10(b) 为经过多尺度时空卷积模块提取后的特征分布,特征空间中的类别可分性得到改善,不同类别的特征点开始形成局部聚集区域,但仍存在部分重叠。图 10(c) 和图 10(d) 分别为双分支联合训练的 Transformer 输出和 BiLSTM 输出的

特征分布。Transformer 分支的 4 种类别特征簇不仅分离度极高,而且簇内凝聚力强、簇间距离大;同样地,联合训练的 BiLSTM 分支的输出特征也呈现出良好的聚类效果,且 4 种类别可分性强。该结果表明所提出的动态双分支教师模型具有较好的运动想象脑电特征提取与表征能力。

此外,为探究知识蒸馏对学生模型特征表征学习的影响,进一步对比了蒸馏前、后学生模型的脑电特征分布。如图 11 所示,经知识蒸馏后的 4 个类别特征簇变得更加紧凑和内聚,而且簇与簇之间重叠区域大幅减少。该结果表明通过双分支教师模型提供的“软标签”所蕴含的丰富信息,学生模型较好地学习到更具判别性的特征表示,它不仅是记忆标签,而且是更深入地理解数据中内在的结构和模式,从而使其特征空间的组织方式更接近于高性能的教师模型。

通过特征可视化可以直观地对比所提出的多尺度时空卷积模块、异构双分支架构、联合损失函数协同训练策略,以及双分支动态知识蒸馏方法的有效性和必要性,同时直观展示了模型能够学习到高度判别性的特征表示和知识蒸馏向轻量级学生模型迁移的过程,从而证实了所提方法对提高运动想象脑电解码精度的贡献。

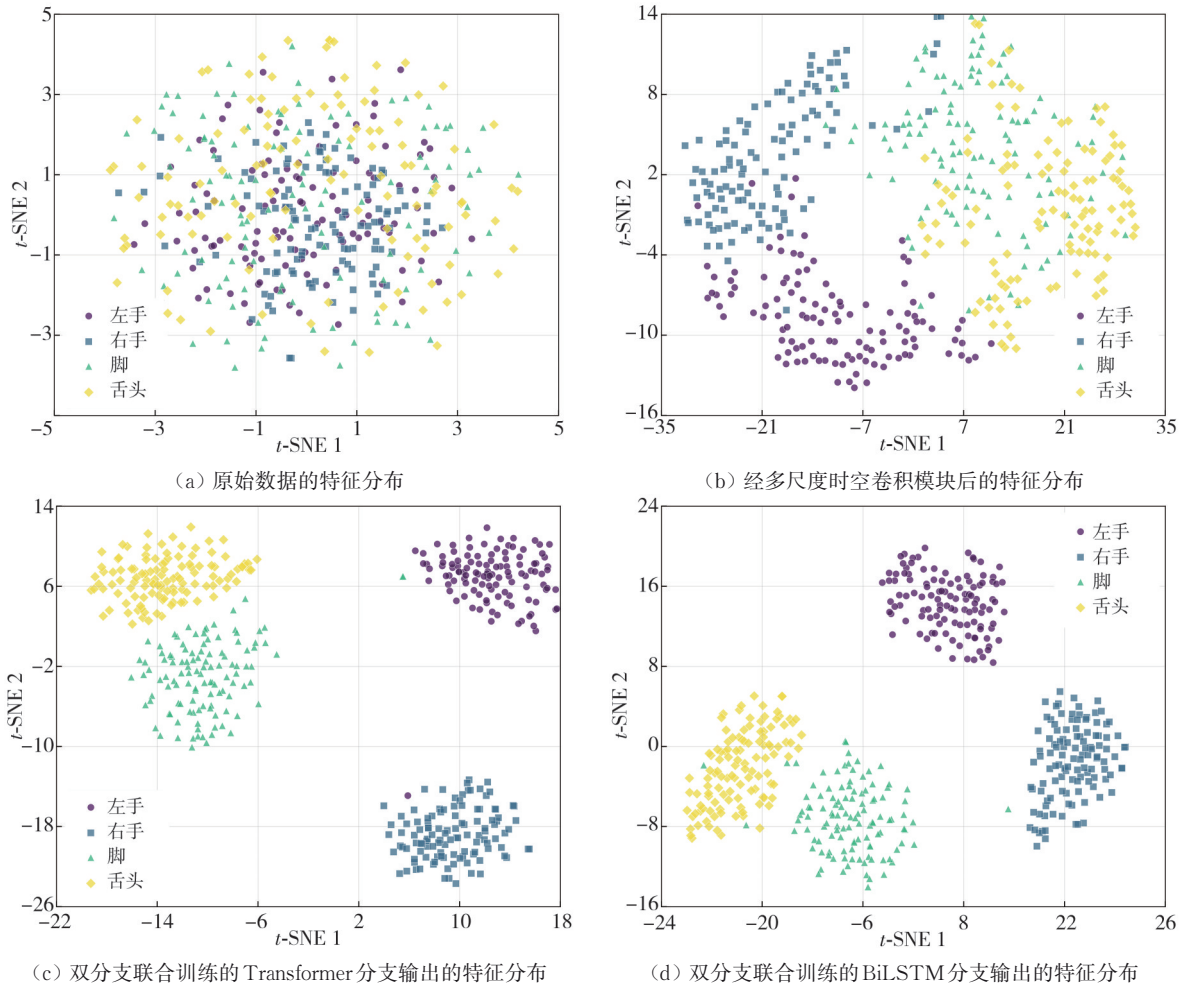


图 10 双分支教师模型的特征可视化

Fig. 10 Feature visualization of dual-branch teacher model

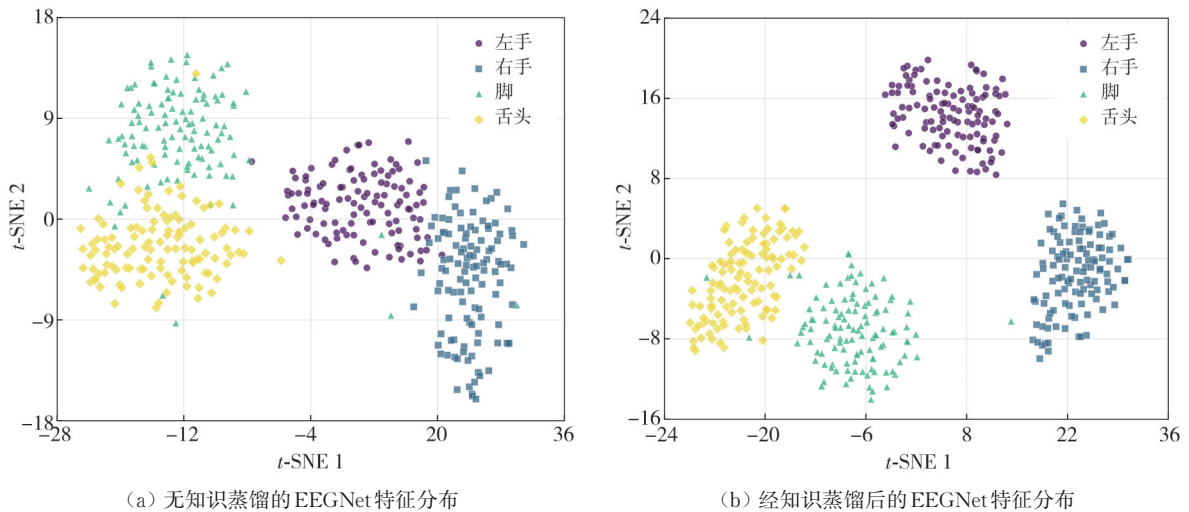


图 11 学生模型EEGNet蒸馏前后的特征可视化

Fig. 11 Feature visualization of student model EEGNet before and after distillation

4 结 论

本研究针对运动想象脑电解码任务中高精度

模型计算复杂度高的问题,提出了一种基于双分支网络的动态知识蒸馏多任务运动想象脑电解码方法。该方法通过构建融合多尺度时空特征提取

模块与 Transformer-BiLSTM 双分支结构的教师模型,有效整合全局建模与局部时序动态表征能力,显著提升了脑电解码精度。进一步地,本文提出双分支动态知识蒸馏策略,自适应选择教师分支生成软标签,以指导轻量化学生模型 EEG-Net 的训练。实验结果表明,该方法在显著降低模型参数量和计算复杂度的同时,使学生模型获得了接近教师模型的脑电解码性能,并验证了该轻量化模型框架在运动想象脑电解码任务中的有效性与实用性,为资源受限的脑机接口应用提供切实可行的解决方案。然而,模型所展现的自适应能力在跨被试场景中的有效性与稳定性仍有待验证,未来研究将致力于增强模型对不同个体差异的适应能力,以提升其在复杂真实场景中的鲁棒性。

利益冲突声明/Conflict of Interests

所有作者声明不存在利益冲突。

All authors declare no relevant conflict of interests.

参考文献

- [1] Wolpaw J R, Birbaumer N, Mcfarland D J, et al. Brain-computer interfaces for communication and control [J]. *Clinical Neurophysiology*, 2002, 113 (6): 767-791.
- [2] Altaheri H, Muhammad G, Alsulaiman M, et al. Deep learning techniques for classification of electroencephalogram (EEG) motor imagery (MI) signals: a review [J]. *Neural Computing and Applications*, 2023, 35(20): 14681-14722.
- [3] 骆睿鹏, 冯铭科, 黄鑫, 等. 脑电信号预处理方法研究综述[J]. *电子科技*, 2023, 36(4): 36-43.
Luo Ruipeng, Feng Mingke, Huang Xin, et al. A review of research on EEG signal preprocessing methods[J]. *Electronic Science and Technology*, 2023, 36 (4): 36-43. (in Chinese)
- [4] Lotte F, Congedo M, Lécuyer A, et al. A review of classification algorithms for EEG-based brain-computer interfaces [J]. *Journal of Neural Engineering*, 2007, 4(2): R1-R13.
- [5] Schirrneister R T, Springenberg J T, Fiederer L D J, et al. Deep learning with convolutional neural networks for EEG decoding and visualization[J]. *Human Brain Mapping*, 2017, 38(11): 5391-5420.
- [6] Ang K K, Chin Z Y, Zhang H, et al. Filter bank common spatial pattern (FBCSP) in brain-computer interface [C]//2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 2008: 2390-2397.
- [7] Mane R, Robinson N, Vinod A P, et al. A multi-view CNN with novel variance layer for motor imagery brain computer interface[C]//2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2020: 2950-2953.
- [8] Tortora S, Ghidoni S, Chisari C, et al. Deep learning-based BCI for gait decoding from EEG with LSTM recurrent neural network[J]. *Journal of Neural Engineering*, 2020, 17(4): 046011.
- [9] Xie J, Zhang J, Sun J, et al. A transformer-based approach combining deep learning network and spatial-temporal information for raw EEG classification [J]. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2022, 30: 2126-2136.
- [10] Song Y, Zheng Q, Liu B, et al. EEG conformer: convolutional transformer for EEG decoding and visualization [J]. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2023, 31: 710-719.
- [11] Hamidi A, Kiani K. Motor imagery EEG signals classification using a Transformer-GCN approach [J]. *Applied Soft Computing*, 2025, 170: 112686.
- [12] Hinton G, Vinyals O, Dean J. Distilling the Knowledge in a Neural Network [PP/OL]. arXiv (2025-05-02) [2025-11-14]. <https://doi.org/10.48550/arXiv.1503.02531/stat>, 2015, 1050: 9.
- [13] Liu Y, Jia Z, Wang H. EmotionKD: a cross-modal knowledge distillation framework for emotion recognition based on physiological signals [C]//Proceedings of the 31st ACM International Conference on Multimedia, 2023: 6122-6131.
- [14] Wu J, Tang B, Wang Y, et al. A multi-level teacher assistant-based knowledge distillation framework with dynamic feedback for motor imagery EEG decoding [J]. *Neural Networks*, 2026, 194: 108180.
- [15] Lawhern V J, Solon A J, Waytowich N R, et al. EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces [J]. *Journal of Neural Engineering*, 2018, 15(5): 056013.
- [16] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 6000-6010.
- [17] Siarni-Namini S, Tavakoli N, Namin A S. The performance of LSTM and BiLSTM in forecasting time series [C]//2019 IEEE International Conference on

- Big Data (Big Data), 2019: 3285-3292.
- [18] Tangermann M, Müller K R, Aertsen A, et al. Review of the BCI competition IV [J]. *Frontiers in Neuroscience*, 2012, 6: 55.
- [19] Lotte F. Signal processing approaches to minimize or suppress calibration time in oscillatory activity-based brain-computer interfaces [J]. *Proceedings of the IEEE*, 2015, 103(6): 871-890.
- [20] Ingolfsson T M, Hersche M, Wang X, et al. EEG-TCNet: an accurate temporal convolutional network for embedded motor-imagery brain-machine interfaces [C]//2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2020: 2958-2965.
- [21] Huang J, Zou J, Yu Z, et al. EEG-MCLANet: a multi-branch CNN-LSTM network for motor imagery EEG decoding [C]//2025 37th Chinese Control and Decision Conference (CCDC), 2025: 1972-1977.
- [22] Luo J, Cheng Q, Wang H, et al. MI-MBFT: superior motor imagery decoding of raw EEG data based on a multibranch and fusion transformer framework [J]. *IEEE Sensors Journal*, 2024, 24(21): 34879-34891.
- [23] Huang Q, Yang Y, Li J, et al. AMFTCNet: a multi-level attention-based multi-scale fusion temporal convolutional network for decoding MI-EEG signals [J]. *Biomedical Signal Processing and Control*, 2025, 108: 107916.
- [24] Du X, Li K, Lü Y, et al. Motor imaging EEG signal recognition of ResNet18 network based on deformable convolution [J]. *Electronics*, 2022, 11(22): 3674.
- [25] Du X, Xi M, Ding X, et al. Motor imagery EEG signal classification based on deformable convolution v3 and adaptive spatial attention mechanism [J]. *Biomedical Signal Processing and Control*, 2025, 99: 106905.
- [26] Mou H, Yang W, Zhang S, et al. Hierarchical window attention for motor imagery EEG classification [C]//2024 IEEE International Symposium on Biomedical Imaging (ISBI), 2024: 1-5.