基于多模态交叉注意力机制的目标检测框架

郑骁[1,2]

1. 中国电子科技集团公司光电研究院，中国·天津 300308

2. 河北工业大学电气工程学院，中国·天津 300401

摘要：脑电信号能直接捕获大脑皮层神经活动的电位变化，反应用户注意力分配与目标识别意图；眼动行为则通过记录眼球运动轨迹与注视区域，间接表征视觉注意力焦点。本文提出一种基于多模态交叉注意力机制的目标检测框架，利用生物信号在目标识别过程中高效选择与快速决策能力，设计了双分支 Transformer 编码器，捕捉脑电信号与眼动行为的依赖关系，并基于 Transformer-YOLO v5 网络利用生理信号注意力权重增强 YOLO v5 的视觉特征表达，构建了符合人类认知规律的混合增强系统。

关键词：脑电信号；眼动追踪；目标检测

Object detection framework based on mutil-source signals cross attention mechanism

Zheng Xiao[1,2]

1. Academy of Opto-Electronics, China Electronics Technology Group Corporation, China Tianjin 300308

2. Hebei University of Technology, College of Electrical Engineering, China Tianjin 300401

Abstract: Electroencephalogram can directly capture the electrical potential changes in cerebral cortical neural activity, reflecting users' attention allocation and target recognition intentions. Eye-tracking behavior indirectly characterizes visual attention focus by recoding eye movement trajectories and fixation regions. This study proposes a cross-modal attention mechanism-based object detection framework. Leveraging the efficient selection and rapid decision-making capabilities of biological signals in target recognition processes, we designed a dual-branch transformer encoder to capture the dependencies between EEG signals and eye movement behavior. Furthermore, based on a transformer-enchanced YOLOv5 network. We utilized attention weights derived from physiological signals to enhance the visual feature representation of YOLO v5, ultimately building a hybrid enhancement system that aligns with human cognitive principles.

Keywords: EEG; ET; Object detection

0 引言

随着人工智能与神经工程的快速发展，目标检测作为计算机视觉与脑机接口领域的关键任务，已在安防监控、自动驾驶、医疗辅助等场景得到广泛关注。然而，传统基于计算机视觉的方法在复杂环境下（如低光照、目标遮挡、动态干扰等）鲁棒性不足，且难以直接反应人类认知意图。与此同时，人类大脑在目标识别过程中展现出高效选择性注意力机制与快速决策能力，为提升目标检测系统性能，提供了新的思路。

近年来，自发脑电信号（Electroencephalogram,EEG）与眼电追踪技术（Eye Tracking,ET）的融合研究逐渐兴起。EEG 能够直接捕获大脑皮层神经活动的电位变化，反应用户的注意力分配与目标识别意图；ET 则通过记录眼球运动轨迹与注视区域，间接表征视觉注意力焦点与目标优先级。将二者与机器视觉算法协同使用可以突破单模态的局限性，构建更加符合人类认知规律的混合增强系统，从而提升目标检测的准确性与抗干扰能力。

当前基于 EEG 或 ET 的目标检测研究已取得一定进展。脑电信号单模态方法中利用事件相关电位 ERP 或频谱特征解码目标刺激诱发的神经响应，但其空间分辨率较低，且容易受非目标信号干扰。眼动追踪单模态方法通过注视点聚类或扫视路径分析定位目标区域，但其在多目标场景下易受自主眼动控制能力差异的影响。传统多模态融合技术采用特征拼接或权重融合，未解决 EEG 与 ET 的时空异步性与跨模态关联系数性问题。

尽管已有研究尝试融合多模态生理信号（如 EEG 与近红外光谱），但自发 EEG 与眼动追踪的协同机制尚未被充分挖掘，现有方法多采用简单的特征拼接，未深入探索

EEG 特征与眼动时空模式间的深层关联，需要设计动态融和策略，以弥补特征互补性的不足。

针对上述问题，本文提出一种基于自发脑电信号、眼动追踪与机器视觉技术的人机混合目标检测框架，主要创新如下：（1）设计双分支 Transformer 编码器：分别处理 EEG 特征与眼动序列，捕捉生理信号的依赖关系；（2）跨模态特征交互机制：将 EEG 特征与眼动特征作为机器视觉特征的 Query，实现了不同模态的动态交互；（3）生理信号驱动的注意力增强策略：基于 Transformer-YOLO v5 网络协同优化，通过生理信号注意力权重增强 YOLO v5 的视觉特征表达能力。

1 研究现状

1.1 脑机接口在目标检测的研究现状

脑机接口在目标检测的应用主要通过解码大脑活动信号与计算机视觉技术结合，其核心在于利用神经信号增强或代替传统视觉输入，从而提升检测效率或适应特殊场景需求。

在医疗领域，脑机接口系统通过非侵入式脑电信号或侵入式皮层电极 ECoG 捕捉用户意图。Meta 团队尝试直接从视觉皮层信号中重建目标信息，利用功能性磁共振成像 fMRI 解码大脑对图像的反应，通过生成对抗网络重建用户观察的场景，再利用目标检测算法提取物体信息，为盲人提供导航反馈。

智能驾驶领域探索了脑机接口在注意力检测中的应用，通过 EEG 识别驾驶员对危险目标（如突然出现的行人）的神经响应，若检测到注意力分散或反应延迟，系统提前触发自动驾驶模块介入进行刹车和避障。

1.2 眼动追踪在目标检测的研究现状

眼动追踪技术在目标检测中通过捕捉用户的视觉注意力焦点，与计算机视觉技术结合，优化检测效率，实现自然地交互逻辑。

医疗领域，利用眼动追踪辅助疾病诊断，例如记录自闭儿童对社交场景中的人脸目标注视模式（如回避眼神接触），结合目标检测算法统计其注视特定面部区域的频率，为早期筛查提供量化依据。

智能驾驶领域，眼动追踪系统通过红外摄像头记录驾驶员注视点，结合车载摄像头画面实时分析其视线落点区域内的目标（如行人、交通标志、异常障碍物等），若检测到驾驶员未及时注视关键危险目标，系统可触发预警或自动纠偏机制。特斯拉的驾驶员监控系统通过计算停留时长判断是否注意力分散，联动目标检测的算法验证危险是否存在。

1.3 脑眼融合技术研究现状

脑眼融合技术通过整合脑电信号与眼动追踪的双模态数据，构建人机交互系统。其核心目标是利用脑神经活动的意图解码与视觉注意力的空间定位互补性，突破单一模态的信息局限，在目标检测、意图推断与环境交互中实现“1+1>2”的效能。

在医疗康复领域，脑眼融合系统被用于运动障碍患者的交互效率，如渐冻症患者通过眼动追踪快速锁定目标物体，同时结合运动想象脑电信号确认意图选择，避免传统眼动交互中因无意注视导致的误触发。

在智能驾驶场景中，脑眼融合技术通过驾驶员视线焦点与脑电注意力指标的联合分析，增强危险目标检测的可靠性。当眼动追踪显示驾驶员未注意前方行人，但脑电 γ 波能量骤增时，系统可越过视线数据直接触发紧急制动，特斯拉 2024 年公开的专利显示器正在开发此类“神经冗余安全模块”。

1.4 机器视觉目标检测相关发展

机器视觉目标检测技术的发展历经了从传统手工特征到深度学习驱动的范式变革。YOLO 系列单阶段模型实现了特征融合、损失函数改进及轻量化等突破。当前技术挑战集中于小目标检测（低分辨率特征语义缺失）与复杂环境鲁棒性（遮挡、光照突变）。本文提出的多模态目标检测框架即通过多模态融合手段寻求复杂环境下微弱目标检测的突破。

2 基于脑 – 眼 – 机器视觉融合的人机混合目标检测框架

本节提出一种基于多模态异构数据深度融合的目标检测框架，通过跨模态注意力机制实现脑电信号、眼动信号与机器视觉的动态协同推理。其核心架构采用双分支 Transformer 编码器分别建模脑电信号的的神经响应特征与眼动行为的时空注意力特征，并将其抽象为高阶语义 Query 向量[1]；同时提取 YOLO v5l Backbone 网络的深层视觉特征作为 Key 和 Value，构建跨模态交叉注意力融合模块。通过自适应的特征交互机制，算法将脑眼认知先验与视觉感知特征在隐空间进行动态权重分配，利用多头注意力实现认知注意力的软性空间映射，最终通过特征重构层实现多模态信号对目标检测任务的联合驱动。

该框架突破传统单模态视觉检测的局限，通过神经认知信号对视觉特征的引导式增强，可显著提升对微弱目标的特征分辨力与复杂场景的认知鲁棒性。

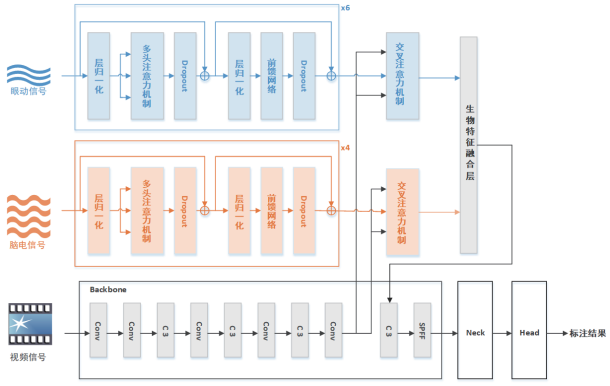

图1 多模态异构数据深度融合的目标检测框架结构图

2.1 数据采集与预处理

2.1.1 数据采集流程

涉及的目标检测素材可选择小型无人机目标的遥感视频，帧率 50 帧 /s，视频中目标形态较小，背景复杂，已完成人工标注工作。实验前给予充足的时间使被试熟悉试验流程。实验所涉及的遥感素材分辨率可处理为 512*512 像素，按照 1s 长度对视频进行分割。

信号采集设备采用七鑫易维 aSee Pro 250Hz 眼动仪、定制多通道 1000Hz 脑电信号采集装置。脑电信号和眼动信号通过多参数同步器实现同步。

采集流程分为若干个序列，每个序列包含 50 段视频片段。采集过程不会出现目标框引导，要求被试自由的进行目标查找。采集开始前需要对眼动数据进行校准，过程中每个序列结束后会给被试 1min 左右的调整时间，两个序列结束后会有 5-10min 的休息时间，并对眼动数据进行再次校准。

数据采集时，屏幕下方固定眼动仪，实验者面对屏幕，佩戴好脑电信号采集装置。每个 block 开始前会呈现指导语，指导被试进行下一步实验，然后进入视频刺激呈现的循环直到每个 block 中包含的 50 段视频全部呈现完毕。采集过程中，每段视频刺激出现之前会呈现 1 个 1s 的注视提示符，视频刺激出现时同步器发送 TTL 电平信号到两种信号采集设备，被试自由的对目标进行搜索，视频在呈现 1 秒后消失，受试按键进入下一段视频。以目标出现时刻为 0 时刻，截取[0s,1s] 时段数据。

2.1.2 EEG 信号采集与预处理

在目标识别任务中，脑电信号的电极选择十分重要，不同电极位置对应大脑不同功能区域。基于国际 10-20 脑电信号采集系统，选择与目标识别认知任务相关的位置，包括前额叶、顶叶、颞叶和枕叶的部分电极。前额叶负责高级认知功能，如注意力和决策，选择电极 F3、F4、Fz。顶叶负责注意力分配和目标定位，选择电极 P3、P4、Pz。颞叶负责视觉听觉信息处理，选择电极 T7、T8。枕叶负责视觉信息处理，尤其是目标感知，选择电极 O1、O2、Oz。

将上述 11 导联脑电信号，通过主成分分析去除眼电伪迹和工频干扰，通过标准化处理，作为 Transformer 编码器的输入，构建与视觉注意力强相关的高区分度脑电信号表征，为多模态融合提供可靠地神经先验信息。

2.1.3 眼动信号采集与预处理

实验开始前，通过九点校准法剔除注释点漂移误差。使用卡尔曼滤波器平滑眼动轨迹，并将眼动数据从采样的 250Hz 上采样到与脑电信号同频 1000Hz，便于对齐处理[2]。眼动行为特征选择时，结合视觉认知机制与目标检测需求，提取能够表征注意力分配和目标搜索策略的时空特征。特征通道共 4 个维度，包括注视点坐标（两个维度：x 和 y）、瞳孔直径、扫视速度。注视点坐标表征了实验者的动态注意力，扫视速度代表了目标导向性，瞳孔直径表征了实验者认知负荷情况。

将 4 维眼动特征标准化，作为 Transformer 编码器的输入，构建具有明确认知解释性与目标搜索强相关的眼动行为表征，为跨模态融合提供可靠地视觉先验信息。

2.2 端到端的目标检测框架

2.2.1 特征编码

（1）EEG 模态特征编码。EEG 信号具有复杂时频特征，需要深层网络建模长程依赖。EEG 模态特征编码器包含输入嵌入层、位置编码层、Transformer 编码层和时间压缩层。

输入嵌入层用于时域特征提取及特征维度提升，包括 1 层 Conv1d 通过跨步卷积降低时间分辨率，GELU 非线性激活函数和全连接层。输入嵌入层的输入形状为[11，1000], 输出形状为[200,1024]，为后续的跨模态注意力提供统一的特征维度。

位置编码层使用参数可学习的位置编码。Transformer 编码层为多头自注意力的网络结构，包括 6 层 Transformer 块，d_model 为 1024。注意力头数为 16，每头维度为 64。前馈网络维度选用 4096（即 4*d_model）以增强非线性表征能力，捕捉不同脑电节律之间的调制关系。Dropout 参数为 0.1，用以抑制电极间随机噪声及个体差异。

时间压缩层将 200 的时间步长线性插值至 256，以匹配后续机器视觉特征的形状。并通过 Conv1d 进行特征平滑处理，此时输出特征形状为[256,1024]，记为 E_{EEG}。

（2）眼动模态特征编码。输入特征为注视点坐标、瞳孔直径、扫视速度，形状为[1000,4],1000 代表时间步长，即上采样后的频率 1000Hz，4 为特征通道。眼动模态特征编码器同样包含输入嵌入层、位置编码层、Transformer 编码层和时间压缩层。

输入嵌入层用于时域特征提取及特征维度提升，包括一层线性输入层用于特征维度的扩展，一层 Conv1d 通过跨步卷积降低时间分辨率，GELU 非线性激活函数和全连接层。输入嵌入层的输入形状为[4，1000], 输出形状为[500,1024]，为后续的跨模态注意力提供统一的特征维度。

位置编码层使用参数可学习的位置编码。

Transformer 编码层为多头自注意力的网络结构，包括 4 层 Transformer 块，d_model 为 1024。注意力头数为 8，每头维度为 128。前馈网络维度选用 2048（即 2*d_model）以增强非线性表征能力，捕捉不同脑电节律之间的调制关系。Dropout 参数为 0.1，用以抑制随机噪声及个体差异。

时间压缩层将 500 的时间步长线性插值至 256，以匹配后续机器视觉特征的形状。并通过 Conv1d 进行特征平滑处理，此时输出特征形状为[256,1024]，记为 E_{Eye}。

（3）机器视觉特征编码。YOLO v5l 模型更适合微弱目标检测场景，从 Backbone 的最后一个 C3 层的输入提取机器视觉特征[3]。图像帧分辨率为 512*512，因此经过一系列卷积计算，Backbone 的最后一个 C3 层的视觉特征形状为[16,16,1024]。将特征展平为[256,1024]，记为 $E_{Machine}$。

2.2.2 跨模态交叉注意力设计

（1）脑电 – 机器视觉交叉注意力设计。以 EEG 特征 E_{EEG} 为 Query，机器视觉特征 $E_{Machine}$ 为 Key 和 Value，实现 EEG 为机器视觉特征进行注意力引导。包括 2 个 Transformer 块，d_model 为 1024。注意力头数为 8，每头处理 128 维信息。8 个并行头，独立计算注意力后进行拼接，输出形状为[256,1024]。

（2）人眼 – 机器视觉交叉注意力设计。以眼动特征 E_{Eye} 为 Query，机器视觉特征 $E_{Machine}$ 为 Key 和 Value，实现眼动特征为机器视觉特征进行注意力引导，实现眼动行为特征为机器视觉特征进行注意力引导。包括 2 个 Transformer 块，d_model 为 1024。注意力头数为 8，每头处理 128 维信息。8 个并行头，独立计算注意力后进行拼接，输出形状为[256,1024]

（3）生物特征融合。将脑电信号和眼动行为特征进行融合，两种特征拼接融合为[256,2048]，使用线性层将融合特征映射到[256,1024]，进而重塑为[16,16,1024]，与 YOLO v5l 的特征维度对齐。将融合特征输入到 YOLO v5l 剩余网络中，输出目标检测结果，即边界框和目标概率。

2.3 训练策略与损失函数

2.3.1 分阶段训练策略

（1）双分支 Transformer 编码器预训练。这一阶段目的是预训练双分支 Transformer 编码器，以使得编码器能够充分表征脑电信号和眼动行为特征。此时，需要单独设计解码器，以解码重建原始信号，通过输入均方误差损失最小化与重建后的误差，迫使编码器能够捕捉生物信号中的关键信息。这一预训练过程大约训练 20 epoch。

（2）端到端参数联合微调。这一阶段解冻 YOLO v5l 中 backbone 的最后一层 C3 和 SPPF 层参数，与双分支 Transformer 编码器、生物特征融合前馈网络，同时进行微调。微调过程可加入梯度均衡、梯度裁剪策略，避免梯度爆炸。端到端参数联合微调耗时较长，根据损失下降情况、验证数据的验证效果保存优异的网络。

2.3.2 损失函数

损失函数设置包括主损失函数和辅助重建损失函数。主损失函数即 YOLO v5l 自身用于目标检测的损失，包括分类损失、定位损失、置信度损失。辅助重建损失函数在编码器预训练和端到端微调都需要使用，迫使编码器能够捕捉生物信号中的关键信息。

3 分析与讨论

本文提出了一种基于多模态数据交叉注意力机制的目标检测框架，通过脑电信号的注意力编码与眼动信号的空间先验引导，可突破传统视觉检测的信噪比瓶颈[4]。脑电信号成分可提供毫秒级目标识别事件标记，配合眼动注视的像素级空间定位，可提升微弱目标的特征区分度。同时，交叉注意力机制通过动态权重分配实现多模态特征的软性对齐，在保留 YOLO v5 原有 40–50 FPS 实时性的基础上，使得检测性能得到提升，尤其在目标遮挡率过大或对比度较小的极端场景。与传统多模态融合方法（特征拼接、决策级投票）相比，本文提出的框架通过 Transformer 交叉注意力机制实现模态交互的定向强化，而非简单的加权平均。特征维度上，本文提出的框架通过脑电特征和眼动特征对机器视觉特征进行注意力加权，针对性增强目标相关通道的激活强度。

本文提出的目标检测框架，重新定义了人机协同智能范式，通过认知信号为机器视觉算法注入了亚秒级预测性决策，在动态微弱目标感知场景具有独特优势，为高风险场景提供了兼具生物感知与机器执行的目标检测方案。在

城市智能安防系统中，该框架可利用安保人员生物特征实现主动式威胁感知。在自动驾驶系统中，该框架通过驾驶员神经认知信号，可增强系统环境感知的意图预判能力。

参考文献：

[1] Vaswani A, Shazeer N, Parmar N, et al. Attention is all your need[J]. Neural information processing systems, 2017,30.

[2] Matran-Fernandez A, Poli R. Brain-Computer Interfaces for Detection and Localisation of Targets in Aerial Images[J]. IEEE transactions on bio-medical engineering, 2016, 64(4): 959-969.

[3] Zhang N. Detection using YOLO v5n and YOLO v5s with small balls[C]. Preceedings of the IEEE conference on computer vision and pattern recognition. Las Vagas: IEEE, 2016:779-788.

[4] 谭嘉宁，罗方亮，张馨元等. 视觉刺激事件相关电位及其研究进展[J]. 中国法医学杂志，2017,32(1):44-47.