自动驾驶技术的蓬勃发展为社会带来了前所未有的变革，自动驾驶汽车也逐渐普及，同时交通安全问题更是受到空前关注。根据世界卫生组织 (WHO) 2018 年全球道路安全状况报告，道路交通死亡人数持续攀升，每年死亡人数达 135 万人【1】，可见，人们对道路安全受到的重视远远不够。根据自然驾驶研究中的警察据报道，有 65-80% 的车祸和相关事故与驾驶员分心有关【2】。分心驾驶是指驾驶员将注意力从驾驶任务转移到其他活动的任何行为，主要表现在驾驶员行为【3】和情绪上【4】，但这并不是全部影响因素，外界路况场景因素也会直接或间接地影响驾驶员的驾驶状况【5,6】。例如，愤怒的司机环顾四周的原因很可能是由于交通堵塞或恶意超车【7】，而大多数吸烟或说话行为发生在交通畅通的情况下。因此车内外状况均会对交通驾驶产生影响。幸运的是，借助精确的辅助驾驶感知算法，道路上发生车辆事故的可能性可以降低10%-20%【8】。因此，对驾驶员表现、车辆状况和场景环境的全面了解对实现更有效的辅助驾驶感知是重要且必要的，开发一种高效可靠的辅助驾驶算法模型具有重要意义。//展开介绍自动驾驶四个任务领域大背景及意义//以计算机视觉技术为基础的监测能够对驾驶车辆内外情况进行全天候监测、并对采集的大量数据进行全天候自动处理和分析，具有精准性和实时性，且其非侵入式监测方式、和较低的成本在高级驾驶辅助系统中也具有显著优势，为提高驾驶安全性、降低事故风险和改善驾驶体验提供了重要支持，因此得到广泛应用。//应用计算机视觉的意义

//四个任务的文献调研//目前，为了实现对自动驾驶状况进行全方位检测，研究者专注于驾驶员行为识别、情绪识别、交通状况识别等任务。起初，研究者多结合CNN和LSTM对驾驶员行为和表情进行识别【9，10】，以实现深层时空特征的提取；为了进一步提取辨别性特征，方法【11,12，13】加入光流或高低频特征对特征进行进一步精细化处理，并嵌入注意力机制关注关键信息，进一步提升特征的有效性；CNN和LSTM等模型需要通过滑动窗口或循环来捕捉远距离的依赖关系，处理复杂长序列的能力有限，因此方法【14，15,16,17】引入Transformer模型在不同位置间建立全局连接，灵活建立全局依赖，得到更加稳健的特征，但同时也消耗了更多的计算资源；为了寻求实时性和精确度的平衡，方法【18,19】利用知识蒸馏代替传统的模型压缩方式将网络进行轻量化处理，通过创建高性能的师生网络，以构建更准确、快速和轻量化的模型以实现对驾驶员的分心行为识别。//以驾驶员为中心的识别为了监测复杂且动态的道路环境对驾驶员分心的影响，方法【20】通过对比驾驶员真正和预测的注意力焦点的相似度或差异完成驾驶员分心判定。//以场景为单一任务的识别//方法【21】通过引入关节点模态，构建CNN特征、关节点成对高级特征、关节点与物交互的关键特征，设计多流深度融合网络进一步提升驾驶员动作的准确性，但难以平衡模型精度与实时性。//多模态的识别//为了突破自动驾驶中单任务的识别上限，方法【22】设计了一个可以同时识别驾驶员行为和情绪的多任务系统，增强了智能驾驶车辆驾驶员分心的推理能力，但缺乏对车外交通场景及车辆自身状况的检测和识别。//情绪和行为多任务的识别//

目前的自动驾驶任务感知算法在行为识别、情绪识别、交通状况识别或车辆状态识别等任务上取得一定进展，在特定条件下能够根据路况进行实时处理和响应，提升车辆对周围环境的感知能力，从而提前预测潜在危险，在一定程度上增强了汽车的智能性和自动化水平，但依旧存在一些挑战：首先，现有方法仅专注于解决如驾驶员行为识别、情绪识别或者交通状态识别等特定单一问题，忽略了任务间的关联性，例如交通拥堵往往引起司机的愤怒，导致司机可能会频繁变换车道，因此，这些任务虽然可以独立进行处理，但在复杂动态的交通驾驶环境中，单一任务模型具有局限性且泛化性较差，多个相关任务之间协同工作对实现自动驾驶车辆内外驾驶状况的全面捕捉和综合评判至关重要。其次，现有方法仍以CNN为主流特征提取模型，其局部性使其专注于局部关键信息，但丢失了长范围的上下文，在复杂场景、恶劣天气条件或异常情况下，算法容易失效或出现误判，而采用LSTM虽然能够建立具有全局或长距离依赖的模型，捕捉良好的时序特征，但是在处理长序列数据时容易出现梯度消失或梯度爆炸的问题。此外，现有方法大部分仅依赖车内驾驶员或车外场景的RGB视频单一模态信息，忽略了车内外信息结合对自动驾驶状况的全面表征能力，且RGB视频/图像信息在复杂驾驶场景中存在冗杂信息多、冗杂信息干扰有效信息等问题，从而导致对人体面部、身体姿态、车外的多视角信息等关键细节关注度不足，所以如何有效整合多模态信息，以提高模型准确度也是一大挑战。

1. 为了克服以上问题，本文提出一种XX算法，该算法通过对自动驾驶过程中车内外多模态数据的强大表征能力和深层次多模态间的交互关系进行建模，实现了驾驶行为检测、情绪分析、交通流量监控或车辆状态评估四个任务的精准识别。此外，为了充分提取驾驶场景下车内-外图像的关键特征，我们设计了一个XX子网络（4个场景图像子网络），该网络通过引入嵌入位置敏感的轴向注意力实现对车内外图像局部和全局特征的关注，从而提高网络对于关键特征的捕获能力。同时，为了提高多模态特征的多尺度融合能力，，我们设计了XX融合方法，该方法通过动态注意力机制沿通道维度聚合上下文信息，兼顾了全局大对象和局部小对象，从而增强了对识别对象的极端尺度变化的应对能力。为了充分验证XXnet的有效性，我们在公开数据集AIDE上进行实验，与最先进的网络模型对比，结果显示，我们的模型在驾驶行为检测、情绪分析、交通流量监控或车辆状态评估四个任务中精度分别达到76.56%、71.88%、95.31%、82.81%、81.64%，均高于其它模型，证明了本文提出xxnet的优越性。本文主要贡献为：我们提出了xxxxnet，该网络首次将行为识别、情绪识别、交通流量识别和车辆状况识别四个任务进行联合学习，利用多任务间的优势互补，提升了对自动驾驶车辆车内外状况的综合感知能力，解决了模型仅对单一任务敏感的问题，提升了模型的准确性和泛化性。
2. 我们提出了一个位置敏感轴向注意力子网络，该网络通过嵌入位置敏感的轴向注意力模块对车内外图像进行特征提取，为了提升对复杂驾驶场景中关键特征的精准识别能力。
3. 为了充分融合多模态特征信息，我们设计了一种动态注意力特征融合方法，通过沿通道维度聚合多尺度上下文信息，有效应对识别对象的极端尺度变化，增强了极端情况下的识别精度。
4. 本研究在AIDE数据集上实现了sota效果（最优性能），实验结果表明，我们的方法在每个任务上的精度提升了1.69%-3.19%，四个任务的平均精度提升了3.89%，证明了模型的优越性能。