中图法分类号:TP391　　文献标识码:A　　文章编号:1006-8961(2025)11-3617-17

论文引用格式:Wang T, Gao S B and Ren G. 2025. Two-stage vision transformer for fusing global and local features in distraction-driven driving behavior recognition. Journal of Image and Graphics, 30(11):3617-3633(王腾, 高尚兵, 任刚. 2025. 融合全局与局部特征的两阶段ViT分心驾驶行为识别方法. 中国图象图形学报, 30(11):3617-3633)[DOI:10.11834/jig.240533]

融合全局与局部特征的两阶段ViT分心驾驶行为识别方法

王腾[1,2],高尚兵[1,2*],任刚[3]

1. 淮阴工学院计算机与软件工程学院,淮安　223003; 2. 江苏省可信固件与智能软件重点实验室,淮安　223003;
3. 东南大学交通学院,南京　210018

摘　要:目的　针对基于端到端卷积神经网络(convolutional neural network,CNN)的分心驾驶行为识别模型缺乏全局特征提取能力以及视觉Transformer(vision Transformer,ViT)模型不擅长捕捉局部特征和模型参数量大的问题,提出一种融合全局与局部特征的两阶段ViT分心驾驶行为识别方法。**方法**　在第1阶段,为防止丢失先前层的信息,提出token信息补充模块,利用k层的class token来获得更全面的特征信息;在第2阶段,为解决特征复杂的图像识别问题,提出特征交互模块,通过交叉注意力机制和自注意力机制融合ViT全局特征和MobileNetV3局部特征。在提高识别准确率的基础上,提出两阶段注意力模块,用于缓解多头注意力可扩展性问题,从而进一步减少参数计算量。**结果**　实验表明,在State Farm数据集和课题组自建的客运车辆分心驾驶行为数据集上,本文方法准确率分别达到99.69%和96.87%,较主干网络ViT-B_16分别提升1.86%和1.65%;相比于TransFG(Transformer architecture for fine-grained recognition)模型,准确率分别提升0.98%和1.04%,浮点数运算次数(floating point operations,FLOPs)分别降低26.87%和17.23%。两个数据集上的整体性能均优于前沿的识别方法。**结论**　本文方法能够准确识别真实场景下的分心驾驶行为,具有更好的鲁棒性,为分类任务研究提供了新思路。

关键词:智能交通;分心驾驶行为识别;视觉Transformer(ViT);注意力机制;特征融合

Two-stage vision transformer for fusing global and local features in distracted driving behavior recognition

Wang Teng[1,2], Gao Shangbing[1,2*], Ren Gang[3]

1. College of Computer and Software Engineering, Huaiyin Institute of Technology, Huaian　223003, China;
2. Key Laboratory of Trusted Firmware and Intelligent Software of Jiangsu Province, Huaian　223003, China;
3. School of Transportation, Southeast University, Nanjing　210018, China

Abstract: Objective Distracted driving, as a primary cause of traffic accidents, claims over 1.2 million lives globally each year. Notably, nearly 90% of incidents involve high-risk behaviors, such as mobile phone usage. While existing end-to-end recognition methods based on convolutional neural networks (CNNs) effectively capture local features (e.g., hand

收稿日期:2024-09-24;**修回日期:**2025-03-24;**预印本日期:**2025-03-31

*通信作者:高尚兵　luxiaofen_2002@126.com

基金项目:国家自然科学基金项目(62076107);教育部人文社会科学研究规划项目(23YJAZH034);淮阴工学院研究生科技创新计划项目(HGYK202415)

Supported by: National Natural Science Foundation of China(62076107); Humanities and Social Sciences Research Project of the Ministry of Education(23YJAZH034); Huaiyin Institute of Technology Graduate Technological Innovation Program(HGYK202415)

posture and facial expressions), their performance is constrained by the limited receptive fields of convolutional operations. This limitation hinders the modeling of global spatial relationships in driving scenarios (e. g. , limb coordination and cabin environment interactions). This inadequacy results in significant accuracy degradation under complex scenarios. Although vision Transformers (ViTs) achieve global feature extraction through self-attention mechanisms, their practical deployment involves multiple challenges. For example, traditional ViT models overly rely on final-layer classification tokens for decision making, which leads to underutilization of fine-grained semantic information from shallow and middle layers. Moreover, their self-attention mechanisms, while modeling global dependencies, inadequately address localized detail features, which causes misclassification among behavior categories with similar spatial distributions (e. g. , phone calls vs. smoking). In addition, the quadratic computational complexity of standard multi-head attention relative to image resolution hinders real-time performance on resource-constrained vehicular platforms. Furthermore, existing studies predominantly rely on simulated driving datasets (e. g. , State Farm), which exhibit limitations such as single-view capture, highly standardized actions, and low background complexity. These datasets diverge markedly from real-world driving environments characterized by multi-view perspectives, dynamic lighting variations, and complex background interference. As a result, the generalizability of the model is restricted. To address the aforementioned issues, this study proposes a dual-stage ViT framework for distracted driving behavior recognition, which integrates global and local features. **Method** The proposed framework employs a two-stage collaborative architecture to enhance distracted driving detection through hierarchical feature integration and computational optimization. In the first stage, an improved ViT addresses the semantic loss in shallow layers by aggregating class token features from intermediate Transformer layers (9–12). This method is achieved via a weighted fusion mechanism, where learnable coefficients——optimized through end-to-end training——dynamically balance contributions from different layers. A hierarchical supervision strategy further reinforces feature preservation by applying cross-entropy loss at multiple network depths, which preserves early-stage visual cues (e. g. , preliminary hand positions or facial orientations) for global context modeling. The second stage focuses on challenging cases by synergizing the broad scene understanding of ViT with the localized precision of MobileNetV3. Spatial attention masks generated from the outputs of ViT first amplify driver-centric regions (e. g. , hand-screen interactions) while suppressing irrelevant elements such as window reflections and moving backgrounds through adaptive pixel-wise masking. The refined images then feed into MobileNetV3 to capture micro-scale patterns, such as finger flexion during screen operations or distinct shapes of held objects (e. g. , cigarettes vs. phones). A cross-attention mechanism bridges the two modalities: the global context of ViT (serving as query vectors) interacts with the localized key-value pairs of MobileNetV3. This mechanism enables bidirectional feature alignment where high-level scene semantics guide the interpretation of fine details, and vice versa. For optimized computational efficiency, a dynamic attention pipeline first segments images into 16 × 16 blocks, calculates region affinity scores to construct adjacency matrices, and retains Top-k high-relevance areas (e. g. , 68% of hand/face regions). The remaining zones undergo token-level attention with reduced spatial granularity, which cuts redundant calculations by 26. 87% while maintaining discriminative power. An adaptive threshold η governs sample routing——strict filtering (η = 1. 0) ——for standardized datasets such as State Farm ensures efficiency. Meanwhile, relaxed thresholds (η = 0. 9) for real-world data redirect ambiguous cases (e. g. , low-light conditions and occluded actions) to stage 2 for reevaluation. This dual-phase approach achieves 99. 69% mean average precision across datasets while operating at 5. 48 GFLOPs, which demonstrates robust performance in diverse scenarios from controlled lab environments to complex road conditions. By strategically combining multi-scale feature extraction with context-aware computation allocation, the method overcomes the inherent limitations of ViT in handling fine details and computational overhead. Therefore, it is a practical solution for in-vehicle safety systems. **Result** Experiments were conducted on the State Farm dataset and a custom dataset of dangerous driving behaviors in passenger vehicles built by the research team. These experimental results were then compared with those of a number of different algorithms, including CNN-based counterfactual attention learning(CAL), attentive pairwise interaction network (API-NET), Stacked-LSTM, progressive- attention convolutional neural network (PA-CNN), and MobileNetV3. In addition, Transformer model-based vision Transformer(ViT), cross-attention multi-scale vision transformer(CrossViT), and Transformer architecture for fine-grained recognition(TransFG) were compared. Results demonstrate that the method exhibits average accuracies (mAP) of 99. 69% and 96. 87% on the State Farm dataset and the self-

constructed dataset of hazardous driving behaviors of passenger vehicles by the group, respectively. Moreover, the numbers of floating-point operations (FLOPs) are 5.48 and 6.82 G, while the number of covariates is 93.31 M. The TransFG algorithm results in increases of 0.98% and 1.04% in mAP, reductions of 25.54% and 17.23% in FLOPs, and an increase of only 7.08 M in the number of parameters. Furthermore, the detection efficacy of ViT can be enhanced through the incorporation of a multimodal information interaction module, a token information supplementation module, and a two-stage attention module. The addition of the multimodal information interaction module results in improvements of 1.24% and 0.97% in the mAP of ViT. The token information supplementation module leads to improvements of 0.15% and 0.64% in the mAP of ViT, while the two-stage attention module yields improvements of 0.16% and 0.19% in the mAP. While the accuracy improvement is not statistically significant, the FLOPs decrease considerably. The test results demonstrate that the proposed method outperforms the state-of-the-art TransFG algorithm in terms of mAP, FLOPs, and the number of parameters. The visualization analysis demonstrates that the model can focus on key areas such as hands and faces while significantly suppressing window reflections and dynamic background interference. **Conclusion** This study takes dangerous driving behavior detection as the starting point, followed by the introduction of the ViT model and MobileNetV3 model. Moreover, a two-stage model based of ViT for dangerous driving behavior recognition called TS-ViT is proposed to address the large number of ViT parameters and high computational cost. The first stage of the inference process of the model inputs the input images into the ViT model, and the images with obvious features are recognized. In the second stage, the global features of ViT and the local features of MobileNet are fused by the cross-attention mechanism and the self-attention mechanism. This fusion re-recognizes the images with complex backgrounds and poor lighting in the first stage. This process also filters out the background information to retain the foreground information, which distinguishes the complex features of the images. The method proposed in this study can accurately identify dangerous driving behaviors in complex scenarios, which demonstrates superior robustness. It provides a new perspective for research in the field of classification tasks.

Key words: intelligent transportation; dangerous driving behavior recognition; vision transformer(ViT); attention mechanism; feature fusion

0 引　言

分心驾驶是影响驾驶安全的重大问题，80%的车祸和65%的险些车祸都涉及分心驾驶(Li等，2022a)。美国国家公路交通安全管理局(National Highway Traffic Safety Administration，NHTSA)的统计数据显示，全球每年因分心驾驶导致的死亡人数超过120万，受伤人数高达5 000万。在分心驾驶相关的车祸中，近90%的驾驶员在车祸发生时正在玩手机或打电话，而开车时使用手机所引发的车祸风险比正常驾驶要高出约4倍(Ma等，2023)。因此，如何实现实时且可靠的分心驾驶行为识别成为一项具有重要意义的研究工作。

分心驾驶行为识别的核心任务在于对分心行为(例如打电话、玩手机、抽烟等)进行准确分类。当前，分心驾驶行为识别主要分为接触式识别与非接触式识别两大类。接触式识别技术主要依赖于驾驶员的生理信号，如心电图、脑电图和皮肤电导等(Nallaperuma等，2018；Li等，2018)，以及眼动数据(如眼睛离开道路时长、眨眼频率、视线广度等)(Vicente等，2015)。然而，该方法存在局限性，不同的分心驾驶行为可能产生相似的生理信号，导致结果混淆(Huang等，2022)；且眼动数据在驾驶员头部或身体大幅度动作时可能难以准确捕捉，同时该领域的研究尚不充分(杨巨成 等，2024)。相较于接触式检测技术，基于计算机视觉的非接触式识别方案逐渐成为研究热点。该技术依托视频流数据，采用卷积神经网络(convolutional neural network，CNN)等深度学习架构构建特征提取模块，通过手部姿态估计与面部关键点检测捕获驾驶员生物特征(Yuen等，2016；Le等，2017)。由于具有非接触式、实时性和准确性高等优点，计算机视觉方法被视为最具前景的分心驾驶行为识别方法之一(Li等，2020)。

在研究基于计算机视觉方法识别分心驾驶行为时，卷积神经网络是广泛应用的深度学习模型。常用的CNN架构包括AlexNet(Krizhevsky等，2017)、VGG(Visual Geometry Group)(Simonyan和Zisserman，2015)、GoogLeNet(Szegedy等，2015)、ResNet(residual network)(He等，2016)、MobileNet(Howard

等，2020）以及YOLO（you only look once）（蔡创新等，2020；汪长春 等，2022）。Ugli等人（2022）使用不同模型架构的迁移学习和微调方法进行分心驾驶行为识别，主要使用MobileNet、VGG16和ResNet50模型实现，使用各种预训练权重进行微调以提高准确性，实验证明MobileNet的迁移学习在3个模型中性能最好。柳长源等人（2022）改进ResNet50对驾驶员分心行为进行识别，通过对正常驾驶、玩手机、打电话、喝水、向后座拿东西和与乘客交谈等6种驾驶员的行为进行测试，改进后ResNet50模型的平均识别准确率达94.2%。将改进后的ResNet50与EfficientNet模型双线性特征融合，融合模型的平均识别准确率达96.7%。李少凡等人（2023）构建了基于姿态引导的实例感知学习框架，通过结合姿态信息和实例感知技术，有效提升了分心行为检测的准确性。Ma等人（2024）提出一种改进的YOLOv8分心驾驶行为和驾驶员情绪检测方法，结合MHSA（multi-head self-attention）注意力机制和CNN模块，其中MHSA用于分心驾驶行为识别，CNN用于检测驾驶员表情，并使用TensorRT和DeepStream方法对模型体积和运算速度进行优化。尽管基于CNN的分心驾驶行为识别方法取得了显著成果，并且在常规任务中表现出色，但其内在缺陷也带来了一些局限性。CNN模型在局部特征提取方面具有显著优势，但受限于卷积操作的局部感受野，难以建立跨区域的全局语义关联。这种特征表达的局限性易引发对局部细节的过度敏感，导致复杂场景下的误判风险增加。

Transformer在自然语言处理领域的成功应用，激发了其在计算机视觉（computer vision，CV）领域的应用，并且广泛地运用到图像分类任务中。其高度的灵活性使其能够适应不同的视觉任务和数据类型，并在多个基准测试中表现出色。传统的CNN主要依赖局部卷积操作，难以捕捉全局信息，而Transformer通过自注意力机制更好地捕捉图像中的全局信息，提高模型的表现。Dosovitskiy等人（2021）提出一种变体Transformer模型ViT（vision Transformer），该模型将图像划分为固定大小的图像块，并将其展平后线性嵌入到向量中。通过添加位置编码输入到标准的Transformer编码器中，利用注意力机制捕捉全局关系。最后在编码器输出上添加分类头，用于图像分类任务（Devlin 等，2019；Han 等，2023）。Wang等人（2021）提出PVT（pyramid vision Transformer）模型，通过引入CNN架构中的金字塔结构来改进ViT在处理高分辨率图像时的效果。PVT采用渐进收缩策略，利用补丁嵌入层和空间缩减机制的编码器，实现对特征图尺度的灵活调整，从而生成多尺度特征图并降低计算成本。Jeevan和Sethi（2022）对ViT架构进行双重优化：首先通过前置卷积层替代原有的分类令牌与位置编码嵌入模块；其次采用线性化自注意力计算方案取代标准自注意力机制。该改进策略通过降低算法复杂度有效减少图形处理器（graphics processing unit，GPU）显存占用，同时增强模型在小样本场景下的泛化能力。在图像分类等计算机视觉任务中，ViT模型展现出显著的性能优势，在复杂场景理解中优于基于CNN的模型（Li等，2022b）。尽管ViT在许多方面表现出色，但并非完美。ViT在局部特征提取方面存在固有缺陷，导致其难以实现前景信息与背景的有效区分。现有改进方案多采用参数扩展策略以提升模型性能，但此类方法将导致模型体积膨胀与推理延迟加剧（Chen等，2022）。这对车载嵌入式系统等资源受限场景构成严峻挑战，因其难以承载高参数量与高延迟的模型架构（Mehta和Rastegari，2022）。

为了解决上述问题，本文提出一种融合ViT的全局特征和CNN局部特征的两阶段分心驾驶行为识别方法，主要工作是：1）针对ViT忽略了先前层中学习到的类别标签，仅在最后一层class token学习类别标签的问题，提出token信息补充模块，充分利用不同层的信息关系，更好地捕捉细粒度信息、增强特征的表达能力和判别性能；2）针对Transformer和CNN特征融合过程中相关性不足以及对特征复杂图像识别效率低的问题，提出特征交互模块，通过交叉注意力机制和自注意力机制，实现Transformer全局特征与CNN局部特征的有效交互；3）基于稀疏注意力机制理论提出两阶段注意力，引入token筛选机制，在粗略级别过滤掉不相关的键值对，再对剩余部分使用细粒度的token-to-token attention，减小模型计算复杂度。

1 数据描述

1.1 State Farm数据集

当前分心驾驶研究多采用Kaggle平台发布的State Farm标准数据集（Kaggle，2019），该数据集包

含10类典型驾驶行为：安全驾驶、左/右手发短信、左/右手打电话、操作收音机、喝水、伸向后座、整理仪容和与乘客交谈，各类别示意图如图1所示。该数据集由分辨率为640×480像素的27 879幅RGB图像组成，训练集和测试集按8∶2进行划分。其中，训练集和测试集中每幅图像都标注了对应的驾驶员状态类别标签。

State Farm数据集采集自受控模拟驾驶环境，通过牵引装置模拟车辆运动。其预设的10类驾驶行为存在动作范式标准化程度过高、环境背景多样性匮乏及右前侧单视角采集等缺陷，如图1所示。典型表现为安全驾驶类仅包含双手握持方向盘的标准姿势，而实际驾驶场景中存在单手握盘协同挡位操作等复合行为模式。由于模拟数据与实际场景存在语义差异，基于仿真数据训练的模型在真实道路测试中呈现出泛化性差、准确率不高的问题。为了解决以上问题，本文在真实场景下构建了客运车辆分心驾驶行为数据集，通过使用真实场景的驾驶数据，可以更好地训练和评估模型，提高模型在实际场景中的泛化性和准确率。

(a) C0:安全驾驶 (b) C1:右手打电话 (c) C2:右手发消息 (d) C3:左手打电话 (e) C4:左手发消息

(f) C5:操作收音机 (g) C6:喝东西 (h) C7:伸手到后面 (i) C8:整理发型 (j) C9:与乘客交谈

图1 State Farm数据集类别示意图

Fig. 1 Schematic diagram of State Farm dataset categories ((a) C0: safe driving; (b) C1: making a call with the right hand; (c) C2: sending a message with the right hand; (d) C3: making a call with the left hand; (e) C4: sending a message with the left hand; (f) C5: operating the radio; (g) C6: drinking; (h) C7: reaching behind; (i) C8: fixing your hair; (j) C9: talking to passengers)

1.2 自建数据集

本文构建的客运车辆分心驾驶行为数据集整合公交平台红外监控与高清日间监控数据源，采集样本覆盖42辆货运车辆及16辆客运车辆在国道、高速公路等复合路况下的自然驾驶行为。数据集完整覆盖昼夜循环及晴/雨等多气象条件，视频流采用1 280×720像素分辨率标准化采集。原始视频流经多级质量筛选流程后，保留156段有效分心行为视频序列(平均时长230±5 s)。采用时空采样策略，以固定间隔20帧提取关键帧，并结合特征相似度分析(余弦相似度阈值<0.82)进行图像去重处理。经此优化流程，最终构建包含34 477幅高质量图像的客运车辆分心驾驶行为数据集。

由于客运车辆驾驶员通常遵循更严格的驾驶规范和行为要求，行为相对单一，很少出现喝东西、把手伸向后座和与乘客交谈等不规范的操作，但货车驾驶员抽烟的情况较为常见。因此，本文将客运车辆分心驾驶行为数据集划分为4类：安全驾驶、打电话、玩手机和抽烟。为增强数据表征能力，每个类别均有来自8个不同视角的数据，覆盖驾驶舱前视、侧视等角度，有效捕捉职业驾驶员在复杂路况下的行为特征，每一类驾驶行为的数量统计如表1所示，类别示意图如图2所示。图3为以C1打电话为例的多视角图。

表1 客运车辆分心驾驶行为数据集的详细信息

Table 1 Detailed information on the dataset of distracted driving behaviors in passenger vehicles

类别	类别标签	图像数量/幅
C0	安全驾驶	10 869
C1	打电话	11 102
C2	玩手机	4 544
C3	抽烟	7 962
总计		34 477

(a) C0:安全驾驶　(b) C1:打电话　(c) C2:玩手机　(d) C3:抽烟

图2　客运车辆分心驾驶行为数据集类别示意图

Fig. 2　Diagram of categories in a dataset of distracted driving behaviors for passenger vehicles

((a) C0: safe driving; (b) C1: making a phone call; (c) C2: playing with mobile phone; (d) C3: smoking)

(a) 正前上方　(b) 正左上方　(c) 左侧前方　(d) 左侧上方

(e) 正右方　(f) 正右上方　(g) 右侧前方　(h) 右侧上方

图3　客运车辆分心驾驶行为数据集多视角图

Fig. 3　Multi-view diagram of the dataset on distracted driving behaviors in passenger vehicles ((a) front and top; (b) top left; (c) front and left; (d) above and left; (e) directly to the right; (f) directly above the right; (g) front and right; (h) top right)

2　模型构建

本文在TransFG (Transformer architecture for fine-grained recognition)(He等,2022)的基础上作出了改进,提出一种融合全局与局部特征的两阶段ViT分心驾驶行为识别方法(two-stage ViT,TS-ViT),如图4所示。

第1阶段利用图像粗细粒度进行推理,使用较少计算量对特征明显的图像进行识别。具体做法是首先对图像先使用粗粒度patch过滤掉部分背景信息,再对筛选的patch细粒度分割。其次将分割后的结果编码成token输入ViT模型,以低成本获得推理结果。紧接着判断结果是否可信,采用预测置信度与一个预设阈值进行比较,若预测置信度大于阈值,则认为结果可信,结束推理;反之则认为结果不可信,进入第2阶段推理。

第2阶段是融合Transformer和CNN的特征进行推理,对背景复杂和光线不佳的图像进行识别。具体做法是将识别图像中重要的粗粒度patch对应的图像送入MobileNetV3网络中,并将粗粒度patch对应的特征通过卷积进行上采样。之后将上采样的特征和MobileNetV3卷积的特征送入特征交互模块以获得细粒度推理结果。

第1阶段ViT模型的参数与第2阶段共享,大大减少了推理的参数负担。为了更好地协同两个推理阶段,本文设计了特征交互模块、token信息补充模块和两阶段注意力模块。

2.1　特征交互模块

近年来,将Transformer的全局特征和CNN的局部特征进行融合受到了广泛关注。目前大部分的算法融合阶段都是采用平均融合、拼接融合、加权融

合、自注意力融合等，但在融合过程中没有充分考虑 Transformer 和 CNN 特征之间的相关性。Transformer 模型通过全局注意力机制计算输入图像中所有像素的关系，生成全局性强的特征，MobileNetV3 擅长捕捉局部特征，其卷积操作可以高效地聚焦于局部区域的信息。MobileNetV3 中的查询（Query，$\boldsymbol{Q}$）和 Transformer 中的键值（Key，$\boldsymbol{K}$）、值（Value，$\boldsymbol{V}$）进行交互，能够有效整合全局与局部特征，提升模型对细粒度特征的捕捉能力。因此，本文提出特征交互模块，实现 Transformer 全局性特征和 CNN 局部特征的交互。特征交互模块的模型结构如图 5 所示。

特征交互模块通过多阶段注意力融合机制实现跨模态特征筛选与对齐。具体而言，首先将 Transformer 编码器中带有位置编码的信息通过层间注意力权重聚合策略生成掩码表：采用 L 层多头注意力图的均值融合构建注意力图，随后对 N 层注意力图执行加权聚合、归一化和阈值二值化处理后形成空间注意力掩码表。将此二值掩码表与末层特征图进

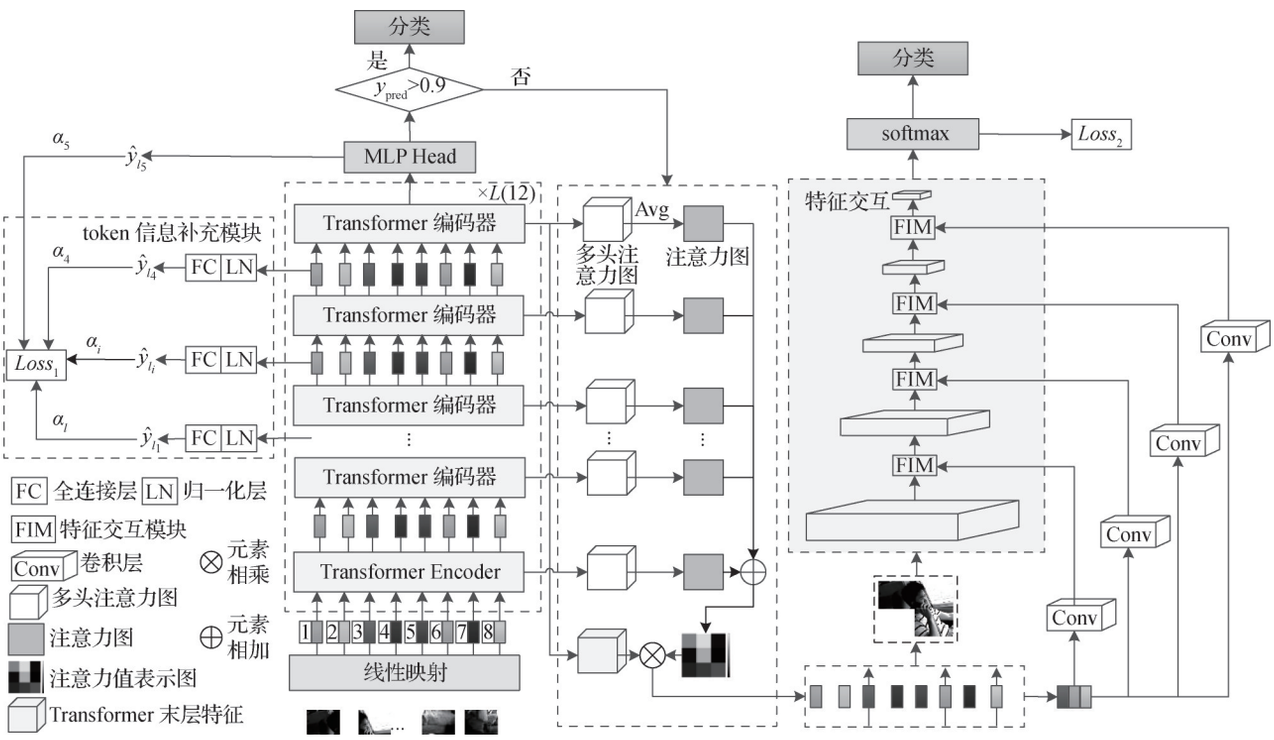


图 4　TS-ViT 网络结构

Fig. 4　TS-ViT network structure

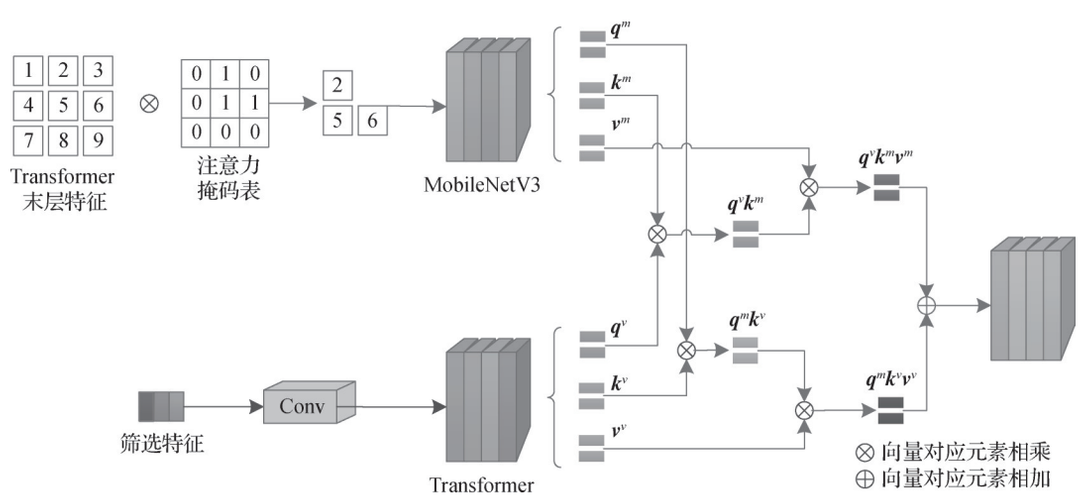


图 5　特征交互模块

Fig. 5　Feature interaction module

行Hadamard乘积运算，得到前景信息的特征，此方法能够有效抑制背景噪声并增强前景特征响应。

基于动态掩码的特征选择机制，模块执行双路径特征增强：主路径将掩码作用于原始图像特征，使MobileNetV3接收经注意力强化的结构化特征；辅助路径则针对Transformer筛选的class token特征，将筛选的class token特征进行反卷积或卷积操作，详见表2，所有转换层均采用深度可分离卷积优化计算效率，最终确保各分支输出与MobileNetV3对应阶段的特征图在空间维度精确对齐。

表2　特征转换操作

Table 2　Feature transformation operations

输入	卷积核	步长	通道	卷积	反卷积	输出
32×32×512	4×4	3	16	–	√	112×112×16
32×32×512	8×8	2	24	–	√	56×56×24
32×32×512	4×4	1	40	√	–	28×28×40
32×32×512	3×3	2	112	√	–	14×14×112

注："√"与"–"分别表示使用对应模块与不使用。

将MobileNetV3与转换的Transformer特征进行特征交互，具体操作如下：将Transformer和CNN提取的特征进行重新塑形，获取各自的查询$\boldsymbol{Q}$、键值$\boldsymbol{K}$和值$\boldsymbol{V}$的表示。之后，利用自注意力和交叉注意力的机制通过计算得到融合后的特征表示$\boldsymbol{q}^v\boldsymbol{k}^m\boldsymbol{v}^m$，$\boldsymbol{q}^m\boldsymbol{k}^v\boldsymbol{v}^v$。最后，将这些融合后的特征逐元素相加，从而得到Transformer与CNN的融合特征，具体为

$$\boldsymbol{q}^v\boldsymbol{k}^m\boldsymbol{v}^m = f_{\text{softmax}}\left(\frac{\boldsymbol{q}^v(\boldsymbol{k}^m)^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{v}^m \tag{1}$$

$$\boldsymbol{q}^m\boldsymbol{k}^v\boldsymbol{v}^v = f_{\text{softmax}}\left(\frac{\boldsymbol{q}^m(\boldsymbol{k}^v)^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{v}^v \tag{2}$$

$$\boldsymbol{F} = \boldsymbol{q}^v\boldsymbol{k}^m\boldsymbol{v}^m + \boldsymbol{q}^m\boldsymbol{k}^v\boldsymbol{v}^v \tag{3}$$

式中，$\{\boldsymbol{q}^v\boldsymbol{k}^m\boldsymbol{v}^m, \boldsymbol{q}^m\boldsymbol{k}^v\boldsymbol{v}^v \in \mathbf{R}^{H\times W\times C}\}$代表注意力的映射结果；$d$表示缩放因子，设为1；$\boldsymbol{q}^v$，$\boldsymbol{k}^m$，$\boldsymbol{v}^m$分别代表Transformer重新塑形得到的$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$，而$\boldsymbol{q}^m$，$\boldsymbol{k}^v$，$\boldsymbol{v}^v$分别代表CNN重新塑形得到的$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$，$f_{\text{softmax}}(\cdot)$表示softmax函数，$\boldsymbol{F}$表示Transformer和CNN的融合特征。

2.2　token信息补充模块

在ViT中，将一幅图像分为$N = H \times W$个图像块，每一块的大小为$P \times P$，通道数C为3，使用线性层将其映射到token中。此外引入了可学习的class token $\boldsymbol{x}_{\text{cls}}$进行分类，并添加位置编码保留空间信息。因此，第1个Transformer层的输出为

$$\boldsymbol{z}_0 = \left[\boldsymbol{x}_{\text{cls}}; \boldsymbol{x}^1\boldsymbol{E}; \boldsymbol{x}^2\boldsymbol{E}; \cdots; \boldsymbol{x}^N\boldsymbol{E}\right] + \boldsymbol{E}_{\text{pos}} \tag{4}$$

式中，$\boldsymbol{E} \in \mathbf{R}^{(P^2\times C)\times D}$表示图像块的嵌入投影，$D$为token的维度，$\boldsymbol{E}_{\text{pos}}$为位置编码。如果有$L$层Transformer层，每层都使用多头自注意力机制MHSA和多层感知机（multilayer perceptron，MLP）模块，并进行归一化（instance normalization，LN）操作，每层的输出为

$$\boldsymbol{z}'_l = MHSA\left(LN\left(\boldsymbol{z}_{l-1}\right)\right) + \boldsymbol{z}_{l-1}, l = 1, 2, \cdots, L \tag{5}$$

$$\boldsymbol{z}_l = MLP\left(LN\left(\boldsymbol{z}'_l\right)\right) + \boldsymbol{z}'_l, l = 1, 2, \cdots, L \tag{6}$$

在ViT中将最后一层的class token送入到分类器中生成预测标签，计算为

$$\boldsymbol{y} = C_L\left(\boldsymbol{z}_{L,\text{cls}}\right) \tag{7}$$

式中，$\boldsymbol{y}$为预测标签向量，$\boldsymbol{z}_{L,\text{cls}}$为第$L$层的class token，$C_L$为分类器。

然而，标准的ViT忽略了先前层中学习到的类别标签，仅在最后一层class token学习类别标签，可能会丢失先前层的信息。不同层之间的信息可以相互补充，本文在实验中得到了验证。受此启发，利用k层（$\boldsymbol{l}_i, i \in \{1, \cdots, k\}$）class token来获得更全面的细粒度信息，而不是仅使用最后一层来获得。具体过程见图6，每个选定层的class token送入分类器中生成预测标签向量，具体为

$$\hat{\boldsymbol{y}}_{li} = C_i\left(\boldsymbol{z}_{l,\text{cls}}\right), i = 1, 2, \cdots, k \tag{8}$$

式中，$\hat{\boldsymbol{y}}_{li}$为第$li$层的预测标签，$C_i$为第$li$层的分类器，$\boldsymbol{z}_{l,\text{cls}}$为第$l$层的class token。通过这种方式可以获得输入图像在不同层的预测值，为了得到最终的预测结果，需要对所有预测值进行加权，损失为

$$Loss_1 = \sum_{i=1}^{k} \alpha_i CE\left(\boldsymbol{y}, \hat{\boldsymbol{y}}_{li}\right) \tag{9}$$

式中，α_i为权重系数，$CE(\cdot)$为交叉熵损失函数，$\boldsymbol{y}$为真实标签，$\hat{\boldsymbol{y}}_{li}$为预测标签向量。

本文提出的TS-ViT算法主要由两个阶段的损失函数构成，即第1阶段损失函数$Loss_1$和第2阶段损失函数$Loss_2$。第2阶段的损失函数具体表达式为

$$Loss_2 = -\sum_{i=1}^{k} \boldsymbol{y}\ln\left(\boldsymbol{y}_i\right) \tag{10}$$

式中，$\boldsymbol{y}$为真实标签，$\boldsymbol{y}_i$为模型对i个样本的预测输出。基于此，推导出整体损失函数$Loss$的表达式为

$$Loss = \xi \times Loss_1 + \varphi \times Loss_2 \tag{11}$$

式中，ξ、φ是超参数，分别为第1阶段和第2阶段损失函数的系数。

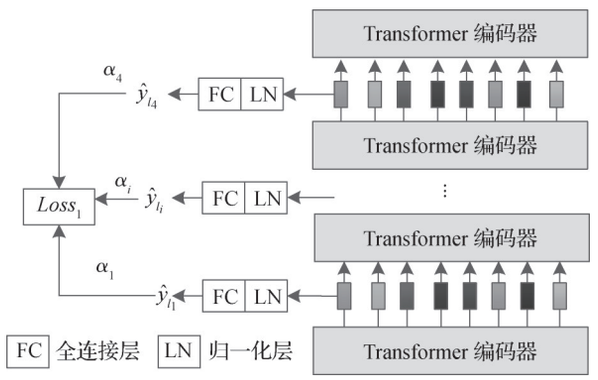


图6　token信息补充模块细节图

Fig. 6　Detail diagram of token information supplementation module

2.3　两阶段注意力

为了缓解多头注意力的可扩展性问题，本文提出一种动态查询感知的稀疏注意力机制，详细操作流程如图7所示。其主要思想是在粗略级别上剔除最不相关的键值对，随后针对剩余部分应用细粒度的token-to-token attention。

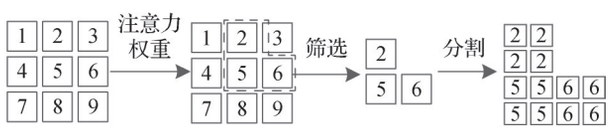


图7　两阶段注意力

Fig. 7　Two stages of attention

2.3.1　区域划分与生成邻接矩阵

首先进行区域划分，输入一个特征图$\boldsymbol{X} \in \mathbf{R}^{H \times W \times C}$，将其划分为$S \times S$的非重叠区域，使每个区域都包含$\frac{HW}{S^2}$个特征向量，将$\boldsymbol{X}$进行切片处理转换为$\boldsymbol{X}^r \in \mathbf{R}^{S^2 \times \frac{HW}{S^2} \times C}$，得到$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$，即

$$\boldsymbol{Q} = \boldsymbol{X}^r \boldsymbol{W}^q \tag{12}$$

$$\boldsymbol{K} = \boldsymbol{X}^r \boldsymbol{W}^k \tag{13}$$

$$\boldsymbol{V} = \boldsymbol{X}^r \boldsymbol{W}^v \tag{14}$$

式中，$\boldsymbol{W}^q$、$\boldsymbol{W}^k$、$\boldsymbol{W}^v \in \mathbf{R}^{C \times C}$分别为$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$的投影权重。然后通过有向图找到关注区域，对$\boldsymbol{Q}$、$\boldsymbol{K}$的第$i$个区域求平均值得到$\boldsymbol{Q}_i^r$，$\boldsymbol{K}_i^r$，通过$\boldsymbol{Q}^r$和转置$\boldsymbol{K}^r$的矩阵乘法，推导出具有区域语义相关度的邻接矩阵$\boldsymbol{A}_i^r$，即

$$\boldsymbol{A}_i^r = \boldsymbol{Q}_i^r \left(\boldsymbol{K}_i^r\right)^{\mathrm{T}} \tag{15}$$

2.3.2　稀疏化操作

为了降低计算复杂度，执行稀疏化操作，利用区域间的语义相关性，量化各区域之间的关系，筛选保留每个区域最相关的k个区域。具体而言，对$\boldsymbol{A}_i^r$使用topkIndex函数生成路由索引矩阵$\boldsymbol{I}_i^r$，即

$$\boldsymbol{I}_i^r = f_{\mathrm{topkIndex}}\left(\boldsymbol{A}_i^r\right) \tag{16}$$

式中，$\boldsymbol{I}_i^r \in \mathbf{N}^{S^2 \times k}$表示与第$i$个区域相关区域的索引。

2.3.3　token-to-token attention

在保留区域的基础上，通过矩阵收集相应的键和值特征，并对这些特征进行细粒度的注意力操作，输出结果是经过筛选与加权的区域相关特征。具体而言，首先根据索引矩阵收集所有相关的键($\boldsymbol{K}$)和值($\boldsymbol{V}$)张量$\boldsymbol{K}_i^g$、$\boldsymbol{V}_i^g$，即

$$\boldsymbol{K}_i^g = f_{\mathrm{gather}}\left(\boldsymbol{K}, \boldsymbol{I}_i^r\right) \tag{17}$$

$$\boldsymbol{V}_i^g = f_{\mathrm{gather}}\left(\boldsymbol{V}, \boldsymbol{I}_i^r\right) \tag{18}$$

式中，$f_{\mathrm{gather}}(\cdot)$为根据索引矩阵$\boldsymbol{I}_i^r$从$\boldsymbol{K}$中选择相应的数据点，并组合成新的向量；$\boldsymbol{K}$和$\boldsymbol{V}$分别为原始的键和值向量；$\boldsymbol{K}_i^g$，$\boldsymbol{V}_i^g \in \mathbf{R}^{H \times W \times C}$为索引矩阵收集后的新向量。然后对第$i$区域的查询向量$\boldsymbol{Q}_i$和从索引矩阵提取的键向量$\boldsymbol{K}_i^g$，计算注意力权重，具体为

$$\boldsymbol{W}_i = f_{\mathrm{softmax}}\left(\frac{\boldsymbol{Q}_i (\boldsymbol{K}_i^g)^{\mathrm{T}}}{\sqrt{d_k}}\right) \tag{19}$$

式中，$\boldsymbol{W}_i \in \mathbf{R}^{1 \times C}$表示第$i$个区域与其关联区域之间的注意力分布，反映其与每个区域的相关程度。之后利用注意力权重$\boldsymbol{W}_i$和值向量$\boldsymbol{V}_i^g$，生成第i个区域的输出特征，即

$$\boldsymbol{O}_i = \boldsymbol{W}_i \times \boldsymbol{V}_i^g \tag{20}$$

最后，将输出特征$\boldsymbol{O}_i$重新组装为全局特征，即

$$\boldsymbol{O} = f_{\mathrm{Concat}}\left(\boldsymbol{O}_1, \boldsymbol{O}_2, \cdots, \boldsymbol{O}_k\right) \tag{21}$$

式中，k为划分后的区域总数，$\boldsymbol{O} \in \mathbf{R}^{H \times W \times \mathrm{C}}$。

3　实验结果及分析

3.1　实验平台及训练细节

本文实验一阶段采用官方ViT-B_16作为主干网络，使用动量为0.9的SGD(stochastic gradient descent)优化器进行优化，初始学习率设为3×10^{-3}，权重衰减设为5×10^{-4}，与TransFG一致。在实验阶段，将图像的最短边缩放到512像素，进而调整输入图像的大小，同时保持长宽比，随机裁剪一个大小为448 × 448像素的区域用于训练；在测试阶段，使用

中心裁剪替代随机裁剪，图像大小仍为448×448像素。在两阶段注意力模块中，大图像块尺寸为64×64像素，小图像块尺寸为16×16像素。在token信息补充模块中，除了第12层外，还选择了9、10和11层，即$k=4$，$l_1=9$，$l_2=10$，$l_3=11$，$l_4=12$。超参数根据对训练数据的评估而确定，在State Farm数据集中，$\alpha_1=\alpha_2=0.05$，$\alpha_3=\alpha_4=1$；在自建数据集中，$\alpha_1=0.01$，$\alpha_2=0.25$，$\alpha_3=0.87$，$\alpha_4=0.96$。二阶段采用MobileNetV3-Large作为主干网络，与官方提供的训练参数一致。本文采用PyTorch框架完成整个模型的构建，并在Tesla V-100 GPU上进行所有实验。

3.2 模型性能评价指标

为了评估模型的性能，本文实验使用了准确率（accuracy，Acc）和浮点数运算次数（floating point operations，FLOPs）两种评价指标。准确率是指模型正确分类的样本数占总样本数的比例，能够直观地反映出模型的分类能力，定义为

$$Acc=\frac{TP+TN}{TP+TN+FP+FN} \tag{22}$$

式中，TP（true positive）为真正例，即模型正确地将正例预测为正例的数量；TN（true negative）为真负例，即模型正确地将负例预测为负例的数量；FP（false positive）为假正例，即模型错误地将负例预测为正例的数量；FN（false negative）为假负例，即模型错误地将正例预测为负例的数量。

FLOPs是一个衡量模型复杂度的指标，表示在执行模型时需要完成的浮点数运算次数。FLOPs的计算主要涉及两个部分：CNN和ViT。CNN部分又分为卷积层和全连接层。卷积层的FLOPs计算式为

$$FLOPs=\left(2C_{\text{in}}K^2-1\right)HWC_{\text{out}} \tag{23}$$

式中，C_{in}为输入通道数，K为卷积核尺寸，HW为输出特征图的尺寸，C_{out}为输出通道数。全连接层的FLOPs计算为

$$FLOPs=(2I-1)O \tag{24}$$

式中，I为输入层神经元的个数，O为输出神经元的个数。

ViT的FLOPs计算为

$$FLOPs=2\times N\times D_{\text{hidden}}\times D \tag{25}$$

式中，N为一幅图像划分的块数，D_{hidden}为隐藏层维度，D为图像块嵌入后的维度。

3.3 对比实验

为了验证本文方法的性能优势，在State Farm数据集与客运车辆分心驾驶行为数据集上与其他方法进行对比实验。对比实验方法包括基于CNN的方法CAL（counterfactual attention learning）（Rao等，2021）、API-Net（attentive pairwise interaction network）（Zhuang等，2020）、Stacked-LSTM（Ge等，2019）、PA-CNN（progressive-attention convolutional neural network）（Zheng等，2020）、MobileNet-V3（Howard等，2020）；基于Transformer模型的方法ViT（Dosovitskiy等，2021）、CrossViT（cross-attention multi-scale vision transformer）（Chen等，2021）、TransFG（He等，2022）、Co-Td-ViT（deep convolution and tokens downscaled vision transformer）（赵霞等，2023）以及MobileViT-CA（mobilevit coordinate attention）（贺宜等，2024）；基于CNN和Transformer融合的算法CNN-Transformer（Mou等，2023）。不同方法的准确率对比实验结果见表3和表4。

根据表3和表4所展示的实验结果，可以观察到，基于Transformer模型的主干网络ViT-B_16在State Farm数据集和客运车辆分心驾驶行为数据集分别取得了97.83%和95.22%的准确率。这一表现与基于卷积神经网络的方法相比具有显著优势，从而有力地验证了ViT在处理一阶段数据任务时的出色能力。此外，TransFG模型作为首个将Transformer应用于细粒度图像分类的开创性尝试，在表3和表4中均展现了其作为最佳基线模型的强大实力。与ViT-B_16相比，尽管本文方法TS-ViT在浮点数运算次数FLOPs上略有增加，但增幅并不明显。相比之下，在State Farm数据集中TransFG的FLOPs高达ViT-B_16的1.6倍。这一差异主要源于第1阶段引入两阶段注意力模块，精准筛选前景信息，减少QKV的匹配次数，使得FLOPs值大幅降低；而在第2阶段，通过特征交互模块，有效地结合CNN局部特征和Transformer全局特征，导致FLOPs有所增加，但无疑为模型带来了更为丰富的特征表示，使其准确率大幅提升。

为了进一步了解本文方法的优点与不足，图8展示了在State Farm数据集和客运车辆分心驾驶行为数据集的混淆矩阵，混淆矩阵中的数字代表测试样本的数量以及每个类别的准确率。图8（a）为采用TS-ViT模型在State Farm数据集上实验所得识别

表 3　本文方法与其他方法在 State Farm 数据集上的效果对比

Table 3　Performance comparison of our method and baseline methods on the State Farm dataset

方法	主干网络	准确率/% ↑	FLOPs/G ↓	参数量/M ↓
CAL(Rao 等,2021)	ResNet-101	96.45	3.80	42.52
API-Net(Zhuang 等,2020)	DenseNet-161	96.31	2.83	27.84
Stacked-LSTM(Ge 等,2019)	GoogleNet	96.46	1.55	10.32
PA-CNN(Zheng 等,2020)	VGG-19	93.82	15.48	138.36
MobileNet-V3(Howard 等,2020)	MobileNet-V3-Large	96.84	**0.39**	**8.92**
CNN-Transformer(Mou 等,2023)	ResNet-50 + ViT	98.40	12.50	85.60
ViT(Dosovitskiy 等,2021)	ViT-B_16	97.83	4.60	85.76
CrossViT(Chen 等,2021)	ViT-B_16	98.06	4.13	54.52
TransFG(He 等,2022)	ViT-B_16	98.71	7.36	86.23
Co-Td-ViT(赵霞 等,2023)	ViT-B_16	99.21	5.21	87.53
MobileViT-CA(贺宜 等,2024)	MobileViT	99.42	2.50	11.40
TS-ViT(本文)	ViT-B_16 + MobileNetV3	**99.69**	5.48	93.31

注:加粗字体表示各列最优结果。“↑”表示值越高越好,“↓”表示值越低越好。

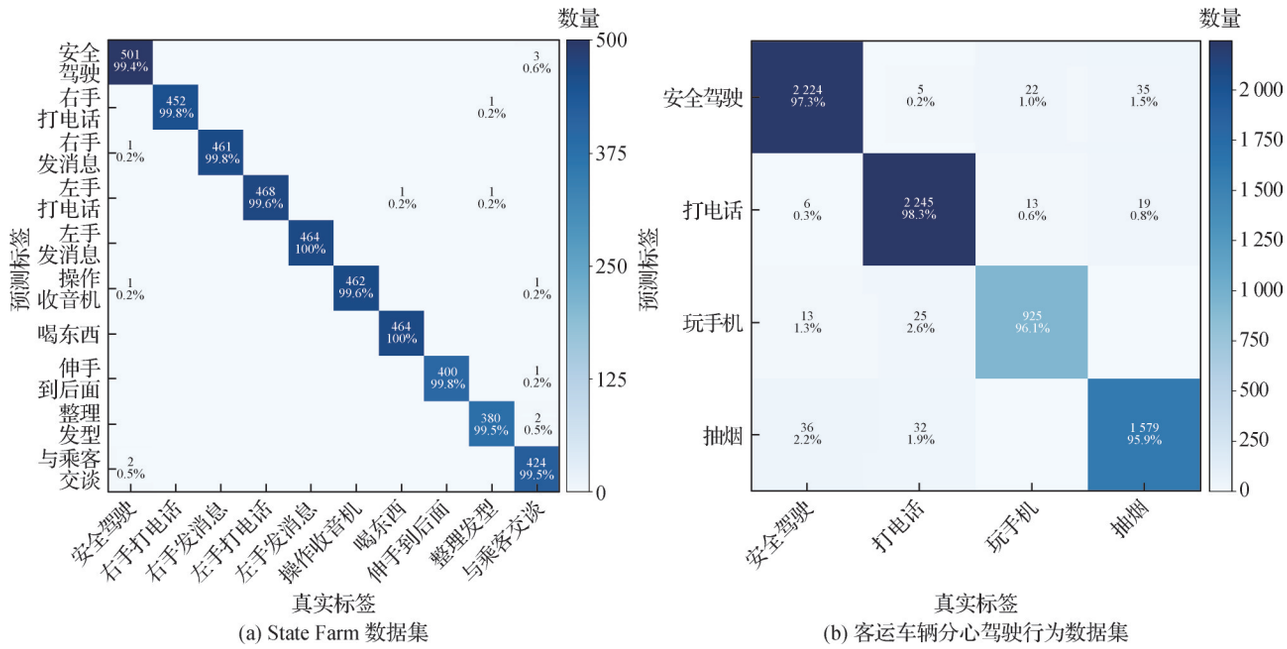
表 4　本文方法与其他方法在客运车辆分心驾驶行为数据集上的效果对比

Table 4　Performance comparison of our method and baseline methods on the passenger vehicles distracted driving dataset

方法	主干网络	准确率/% ↑	FLOPs/G ↓	参数量/M ↓
CAL(Rao 等,2021)	ResNet-101	90.85	4.61	42.52
API-Net(Zhuang 等,2020)	DenseNet-161	92.55	3.96	27.84
Stacked-LSTM(Ge 等,2019)	GoogleNet	93.80	2.75	10.32
PA-CNN(Zheng 等,2020)	VGG-19	90.21	18.06	138.36
MobileNet-V3(Howard 等,2020)	MobileNet-V3_Large	95.47	**0.53**	**8.92**
CNN-Transformer(Mou 等,2023)	ResNet-50 + ViT	95.81	13.80	85.60
ViT(Dosovitskiy 等,2021)	ViT-B_16	95.22	5.49	85.76
CrossViT(Chen 等,2021)	ViT-B_16	95.76	5.31	54.52
TransFG(He 等,2022)	ViT-B_16	95.83	8.24	86.23
Co-Td-ViT(赵霞 等,2023)	ViT-B_16	96.16	6.27	87.53
MobileViT-CA(贺宜 等,2024)	MobileViT	96.21	3.41	11.40
TS-ViT(本文)	ViT-B_16 + MobileNetV3	**96.87**	6.82	93.31

注:加粗字体表示各列最优结果。“↑”表示值越高越好,“↓”表示值越低越好。

结果混淆矩阵,图 8(b)为采用 TS-ViT 模型在客运车辆分心驾驶行为数据集实验所得识别结果混淆矩阵,用以描述不同驾驶行为的准确率和测试样本数量。由图 8(a)可知,本文提出的 TS-ViT 模型能精准识别 10 个类别的分心驾驶行为,平均正确识别率为 99.69%,对于左手发消息和喝东西动作的正确识别率达到了 100%;与乘客交谈和安全驾驶两个类别存在相互误判,主要原因是驾驶员双手握方向盘,扭头不明显,导致与正常驾驶行为相似,进而导致相互误判;整理发型正确分类的准确率仍有提升空间。由图 8(b)可知,打电话和安全驾驶的行为识别效果较好,但打电话和抽烟两个类别存在误判。主要原因是打电话和抽烟的人体姿态相似,特征相似度较高,因此模型的误判率增加。抽烟识别的准确率相对较



(a) State Farm 数据集　　(b) 客运车辆分心驾驶行为数据集

图 8　识别结果的混淆矩阵

Fig. 8　Confusion matrices of recognized results ((a) State Farm dataset; (b) passenger vehicle distracted driving behavior dataset)

低,其原因在于真实场景中受光线以及左侧车窗背景的影响,部分香烟特征不明显,导致识别准确率相对较低。虽然客运车辆分心驾驶行为数据集上准确率有所降低,但总体保持了较好的泛化性能。

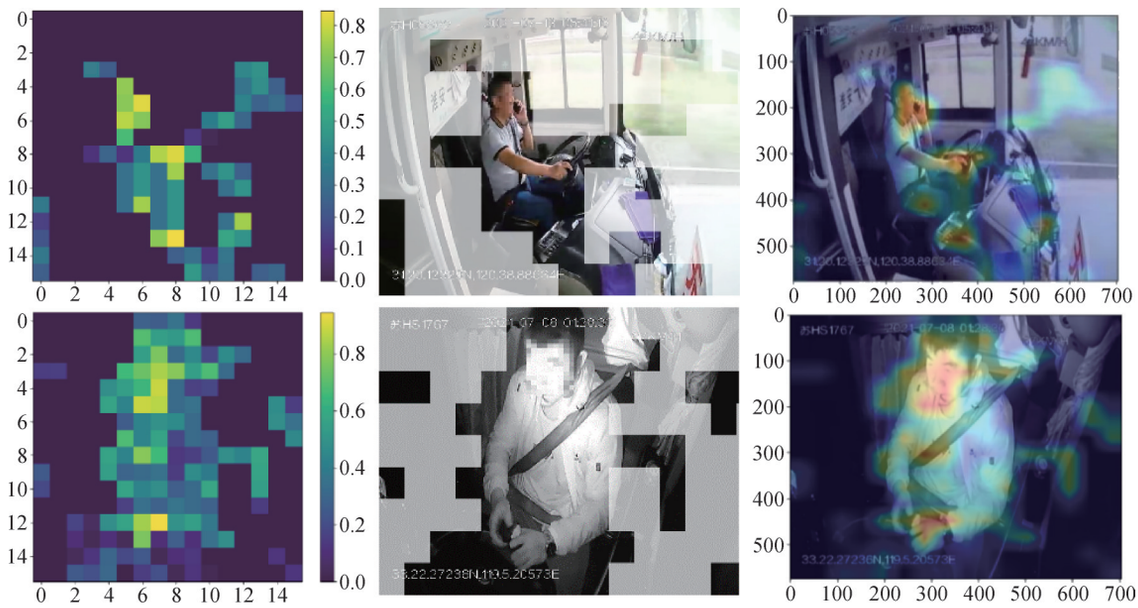
为了更好地验证第 2 阶段的效果,本文在客运车辆分心驾驶行为数据集上展示了模型两阶段注意力可视化信息和第 2 阶段可视化信息。图 9(a)展示了经过第 1 阶段注意力筛选后的特征图。基于特征图对各个 patch 块进行了精确的拼接,从而得到了图 9(b)所示的筛选图。利用第 2 阶段注意力机制对筛选图进一步处理,生成图 9(c)热力图。从热力图中可以清晰地观察到,经过两阶段注意力处理后,特征主要集中在脸部和手部区域。因此,两阶段注意力方法能够极为高效地剔除背景信息的干扰,降低了后续识别任务的计算负担。从图 10 可以观察到,第 1 阶段主要识别的图像大多是“容易”的区域,即白天视角的正前上方、正左上方、左侧前方、左侧上方以及右侧上方。由于正面及左侧区域驾驶员视角突出,因此这部分的手部和脸部特征明显,并且在第 1 阶段就能达到较好的识别效果。而对于夜晚场景、正右上方和右侧前方视角的图像,场景较复杂、部分图像模糊,则需要进一步分割背景信息提取前景信息才能正确识别图像类别。

3.4　模型参数设置

为了验证第 2 阶段的模型的性能并确定阈值 η 的最佳值,本文分别在 State Farm 数据集和客运车辆分心驾驶行为数据集上进行实验分析,以验证本文提出的特征交互模块的有效性。

由表 5 和表 6 的实验结果可知,在主干网络为 ViT-B_16 的情况下,本文方法 TS-ViT 显著提高准确性,随着 η 值的增大,模型的准确率和 FLOPs 都有所增加,主要原因是随着 η 的增大,更多场景复杂和光线不佳的图像送入到模型第 2 阶段进行识别,捕捉更细节的图像信息,从而增加识别的准确率,但也增加了计算量。由表 5 可知,通过调整 η 值,TS-ViT 模型在保持相似计算量(FLOPs)的同时,显著提高了准确率。特别是当 η = 1.0 时,模型的准确率达到了 99.69%,而计算量仅增加了约 8%。由表 6 可知,当 η = 0.9 时,模型的准确率为 96.87%,计算量为 6.82 G,相较其他值,η 为 0.9 在保持较高准确率的同时,计算量的增加相对较小。在计算资源有限的情况下,在 State Farm 数据集上可以选择 η = 0.95 或更小的值,本文采用的 η 值为 1.0;在客运车辆分心驾驶行为数据集上选择 η = 0.7 或更小的值,本文采用的 η 值为 0.9。

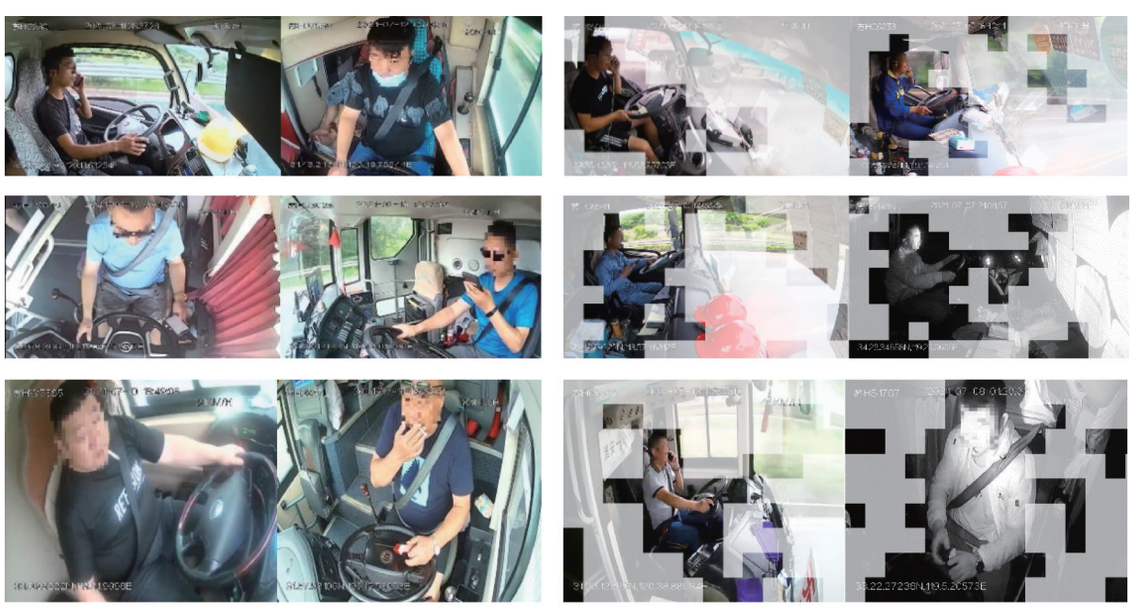
为了验证 token 信息补充模块的有效性,本文分别在 State Farm 数据集和客运车辆分心驾驶行为数



(a) 特征图　(b) 筛选图　(c) 热力图

图9　两阶段注意力可视化信息

Fig. 9　Two-stage attention visualization information ((a) feature maps; (b) filter diagrams; (c) heat maps)



(a) 第1阶段　(b) 第2阶段

图10　两阶段可视化信息

Fig. 10　Two-stage visualization information ((a) stage-one; (b) stage-two)

据集上进行实验证明其性能。ViT-B_16模型具有$2^{12}-1$种不同的层与层之间的组合方式，由于其数量众多，无法一一验证。为了有效选择最优组合，本文提出步长采样法：从ViT-B_16的最后一层（第12层）开始，以步长（stride）s向前跨步选取k层特征。具体而言，层索引计算为

$$S_L = L - s \times (m-1),\quad m = 1, 2, \cdots, k \tag{26}$$

式中，L为Transformer编码器总层数，s为步长，k为选取的层数，系统探索不同跨层组合的特征互补性，实验结果如图11和图12所示。

由图11和图12的实验结果可知，当步长s较小时，模型的性能较好。特别是当$k=4$，$s=1$时，模型的性能显著优于$s=2$和$s=3$的情况。进一步观察图11发现，当$s=1$时，$k=1$为ViT-B_16的基础网

表5　State Farm数据集不同η值的准确率与计算量

Table 5　Accuracy and computation of different η values for the State Farm dataset

模型	η	FLOPs/G ↓	准确率/% ↑
ViT-B_16	–	**5.49**	97.83
TS-ViT	0.92	5.99	98.39
TS-ViT	0.95	6.25	99.68
TS-ViT	1.00	6.82	**99.69**

注:加粗字体表示各列最优结果。“–”表示未对该项指标提供相关信息。

表6　客运车辆分心驾驶行为数据集不同η值的准确率与计算量

Table 6　Accuracy and computation of different η values for the passenger vehicles dangerous driving behavior dataset

模型	η	FLOPs/G ↓	准确率/% ↑
ViT-B_16	–	**5.49**	95.22
TS-ViT	0.50	5.86	95.85
TS-ViT	0.70	6.49	96.13
TS-ViT	0.90	6.82	96.87
TS-ViT	0.95	7.55	96.91
TS-ViT	1.00	8.32	**96.92**

注:加粗字体表示各列最优结果。“–”表示未对该项指标提供相关信息。

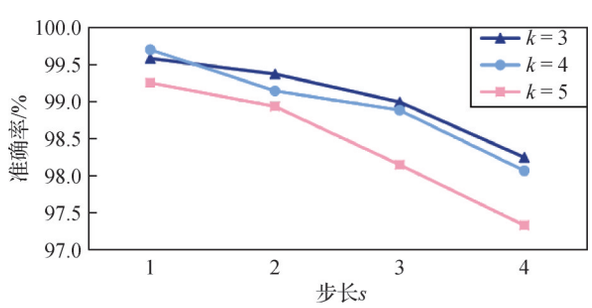


图11　State Farm数据集不同k值的准确率

Fig. 11　Accuracy of State Farm dataset with different k values

络。随着k值的增加,模型准确率呈现出先上升后下降的趋势,在$k=4$时达到峰值后开始降低。其原因可能在于,随着层数的增加,模型能够积累更多层的特征信息,因此,适当增加层数可以更好地补充不同层之间的信息,但选择过多的层数(k)时,模型预测准确率有所下降。例如,当$s=2$时,$k=5$的性能甚至不如$k=4$的性能,主要原因在于过多的层数可

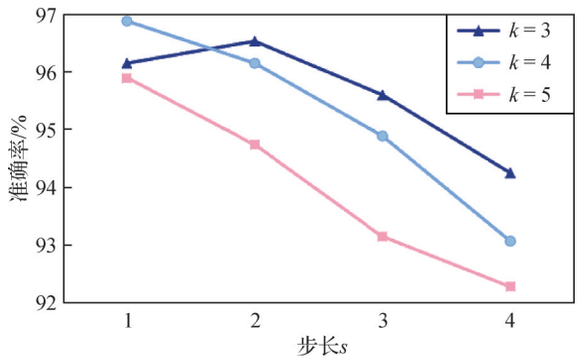


图12　客运车辆分心驾驶行为数据集不同k值的准确率

Fig. 12　Accuracy of the dataset on distracted driving behaviors in passenger vehicles with different k values

能引入冗余或不相关的信息,使得模型的注意力分布难以聚焦。此外,参数优化的难度也会随之增加,导致训练过程更容易受到梯度噪声的影响,从而削弱顶层的特征提取能力。因而,本文在State Farm数据集和客运车辆分心驾驶行为数据集上选择$k=4$、$s=1$的设置,此时模型的性能最佳。

3.5　消融实验

本文提出3个新的模块,即特征交互模块、token信息补充模块和两阶段注意力模块。为了验证各个模块的有效性,本文在公开数据集State Farm和客运车辆分心驾驶行为数据集上进行了消融实验研究,消融实验结果如表7和表8所示。

表7　State Farm数据集消融实验结果

Table 7　Results of ablation experiments on the State Farm dataset

ViT-B_16	特征交互模块	token信息补充模块	两阶段注意力	准确率/% ↑
√	–	–	–	97.83
√	√	–	–	99.07
√	–	√	–	98.34
√	–	–	√	97.99
√	√	√	√	**99.69**

注:加粗字体表示最优结果。“√”和“–”表示使用和不使用对应模块。

由表7和表8的实验结果可知,本文提出的3个模块在State Farm数据集上分别将ViT-B_16提高1.24%、0.51%和0.16%,在客运车辆分心驾驶行为数据集上分别将ViT-B_16提高0.97%、0.64%和0.19%。这表明,通过有效结合3个模块可以获得

表8　客运车辆分心驾驶行为数据集消融实验结果

Table 8　Results of ablation experiments on the dataset of distracted driving behaviors in passenger vehicles

ViT-B_16	特征交互模块	token信息补充模块	两阶段注意力	准确率/%↑
√	–	–	–	95.22
√	√	–	–	96.19
√	–	√	–	95.86
√	–	–	√	95.41
√	√	√	√	**96.87**

注：加粗字体表示最优结果。“√”和“–”表示使用和不使用对应模块。

更好的识别效果。其中，token信息补充模块和两阶段注意力模块对性能提升相对较小。由于特征交互模块使用交叉注意力机制和自注意力机制有效地结合Transformer的全局特征和CNN的局部特征，增强了对局部特征的全局感知能力，并学习了全局表示的局部细节，使得模型准确率超过了单独使用任何一个模块的情况，从而达到了较好的实验效果。

4　结　论

本文以分心驾驶行为识别为出发点，引入ViT模型和MobileNetV3模型，针对ViT参数量大、计算成本高的问题，提出一种融合全局与局部特征的两阶段ViT分心驾驶行为识别模型TS-ViT。该模型推理过程的第1阶段将输入的图像输入到ViT模型，对特征明显的图像进行识别；在第2阶段，通过交叉注意力机制和自注意力机制，将ViT的全局特征和MobileNet的局部特征融合，对第1阶段背景复杂和光线不佳的图像进行再识别，此过程过滤掉背景信息保留前景信息，从而更好地获取图像的复杂特征。实验结果表明，本文提出的TS-ViT模型能够在性能和效率之间取得良好的权衡。此外，针对实验中出现的部分类别误判的问题，未来将引入姿态估计算法进行改进。

参考文献（References）

Cai C X, Gao S B, Zhou J and Huang Z H. 2020. Freeway anti-collision warning algorithm based on vehicle-road visual collaboration. Journal of Image and Graphics, 25(8): 1649-1657 (蔡创新, 高尚兵, 周君, 黄子赫. 2020. 车路视觉协同的高速公路防碰撞预警算法. 中国图象图形学报, 25(8): 1649-1657) [DOI: 10.11834/jig.190633]

Chen C F R, Fan Q F and Panda R. 2021. CrossViT: cross-attention multi-scale vision transformer for image classification//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 347-356 [DOI: 10.1109/ICCV48922.2021.00041]

Chen Y P, Dai X Y, Chen D D, Liu M C, Dong X Y, Yuan L and Liu Z C. 2022. Mobile-former: bridging MobileNet and transformer// Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 5270-5279 [DOI: 10.1109/CVPR52688.2022.00520]

Devlin J, Chang M W, Lee K and Toutanova K. 2019. BERT: pre-training of deep bidirectional Transformers for language understanding//Proceedings of the North American Chapter of the Association for Computational Linguistics. Minneapolis, USA: ACL: 4171-4186 [DOI: 10.18653/v1/N19-1423]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houlsby N. 2021. An image is worth 16 × 16 words: transformers for image recognition at scale//Proceedings of the 9th International Conference on Learning Representations. [s.l.]: OpenReview.net

Ge W F, Lin X R and Yu Y Z. 2019. Weakly supervised complementary parts models for fine-grained image classification from the bottom up//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 3034-3043 [DOI: 10.1109/CVPR.2019.00315]

Han K, Wang Y H, Chen H T, Chen X H, Guo J Y, Liu Z H, Tang Y H, Xiao A, Xu C J, Xu Y X, Yang Z H, Zhang Y M and Tao D C. 2023. A survey on vision transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(1): 87-110 [DOI: 10.1109/TPAMI.2022.3152247]

He J, Chen J N, Liu S, Kortylewski A, Yang C, Bai Y T and Wang C H. 2022. TransFG: a transformer architecture for fine-grained recognition//Proceedings of the 36th AAAI Conference on Artificial Intelligence. [s.l.]: AAAI: 852-860 [DOI: 10.1609/aaai.v36i1.19967]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

He Y, Lu M K, Gao S, Cao B and Li J P. 2024. Distracted behavior detection of commercial vehicle drivers based on the MobileViT-CA model. China Journal of Highway and Transport, 37(1): 194-204 (贺宜, 鲁曼可, 高嵩, 曹博, 李继朴. 2024. 基于MobileViT-CA模型的营运车辆驾驶人分心行为检测. 中国公路学报, 37(1): 194-204) [DOI: 10.19721/j.cnki.1001-7372.2024.01.016]

Howard A, Sandler M, Chen B, Wang W J, Chen L C and Tan M X. 2020. Searching for MobileNetV3//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 1314-1324 [DOI: 10.1109/ICCV.2019.00140]

Huang T, Fu R, Chen Y X and Sun Q Y. 2022. Real-time driver behavior detection based on deep deformable inverted residual network with an attention mechanism for human-vehicle co-driving system. IEEE Transactions on Vehicular Technology, 71(12): 12475-12488 [DOI: 10.1109/TVT.2022.3195230]

Jeevan P and Sethi A. 2022. Convolutional Xformers for vision [EB/OL]. [2024-09-10]. https://arxiv.org/pdf/2201.10271.pdf

Kaggle. 2019. State farm distracted driver detection [EB/OL]. [2024-09-10]. https://www.kaggle.com/rightway11/state-farm-distracted-driver-detection

Krizhevsky A, Sutskever I and Hinton G E. 2017. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6): 84-90 [DOI: 10.1145/3065386]

Le T H N, Zhu C C, Zheng Y T, Luu K and Savvides M. 2017. DeepSafeDrive: a grammar-aware driver parsing approach to driver behavioral situational awareness (DB-SAW). Pattern Recognition, 66: 229-238 [DOI: 10.1016/j.patcog.2016.11.028]

Li L, Zhong B X, Hutmacher C Jr, Liang Y L, Horrey W J and Xu X. 2020. Detection of driver manual distraction via image-based hand and ear recognition. Accident Analysis and Prevention, 137: #105432 [DOI: 10.1016/j.aap.2020.105432]

Li P H, Yang Y F, Grosu R, Wang G D, Li R, Wu Y H and Huang Z. 2022a. Driver distraction detection using octave-like convolutional neural network. IEEE Transactions on Intelligent Transportation Systems, 23(7): 8823-8833 [DOI: 10.1109/TITS.2021.3086411]

Li S F, Gao S B and Zhang Y Y. 2023. Pose-guided instance-aware learning for driver distraction recognition. Journal of Image and Graphics, 28(11): 3550-3561 (李少凡, 高尚兵, 张莹莹. 2023. 用于驾驶员分心行为识别的姿态引导实例感知学习. 中国图象图形学报, 28(11): 3550-3561) [DOI: 10.11834/jig.220835]

Li Y, Wang L F, Mi W, Xu H, Hu J Y and Li H. 2022b. Distracted driving detection by combining ViT and CNN//Proceedings of the 25th IEEE International Conference on Computer Supported Cooperative Work in Design. Hangzhou, China: IEEE: 908-913 [DOI: 10.1109/CSCWD54268.2022.9776082]

Li Z J, Bao S, Kolmanovsky I V and Yin X. 2018. Visual-manual distraction detection using driving performance indicators with naturalistic driving data. IEEE Transactions on Intelligent Transportation Systems, 19(8): 2528-2535 [DOI: 10.1109/TITS.2017.2754467]

Liu C Y, Hu H Y and Bi X J. 2022. Driver distraction recognition using bilinear fusion networks. Journal of Beijing University of Posts and Telecommunications, 45(2): 79-84 (柳长源, 虎浩媛, 毕晓君. 2022. 双线性融合网络的驾驶员分心行为识别. 北京邮电大学学报, 45(2): 79-84) [DOI: 10.13190/j.jbupt.2021-171]

Ma B, Fu Z J, Rakheja S, Zhao D F, He W B, Ming W Y and Zhang Z G. 2024. Distracted driving behavior and driver's emotion detection based on improved YOLOv8 with attention mechanism. IEEE Access, 12: 37983-37994 [DOI: 10.1109/ACCESS.2024.3374726]

Ma Y S, Yuan L Q, Abdelraouf A, Han K, Gupta R, Li Z H and Wang Z R. 2023. M2DAR: multi-view multi-scale driver action recognition with vision transformer//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Vancouver, Canada: IEEE: 5287-5294 [DOI: 10.1109/CVPRW59228.2023.00557]

Mehta S and Rastegari M. 2022. MobileViT: light-weight, general-purpose, and mobile-friendly vision transformer//Proceedings of the 10th International Conference on Learning Representations. [s.l.]: OpenReview.net

Mou L T, Chang J L, Zhou C, Zhao Y Y, Ma N, Yin B C, Jain R and Gao W. 2023. Multimodal driver distraction detection using dual-channel network of CNN and Transformer. Expert Systems with Applications, 234: #121066 [DOI: 10.1016/j.eswa.2023.121066]

Nallaperuma D, De Silva D, Alahakoon D and Yu X H. 2018. Intelligent detection of driver behavior changes for effective coordination between autonomous and human driven vehicles//Proceedings of the 44th Annual Conference of the IEEE Industrial Electronics Society. Washington, USA: IEEE: 3120-3125 [DOI: 10.1109/IECON.2018.8591357]

Rao Y M, Chen G Y, Lu J W and Zhou J. 2021. Counterfactual attention learning for fine-grained visual categorization and re-identification//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 1025-1034 [DOI: 10.1109/ICCV48922.2021.00106]

Simonyan K and Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition//Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: [s.n.]

Szegedy C, Liu W, Jia Y Q, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V and Rabinovich A. 2015. Going deeper with convolutions//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 1-9 [DOI: 10.1109/CVPR.2015.7298594]

Ugli I K, Hussain A, Kim B S, Aich S and Kim H C. 2022. A transfer learning approach for identification of distracted driving//Proceedings of the 24th International Conference on Advanced Communication Technology. PyeongChang Kwangwoon_Do, Korea (South): IEEE: 1-4 [DOI: 10.23919/ICACT53585.2022.9728846]

Vicente F, Huang Z H, Xiong X H, De la Torre F, Zhang W D and Levi D. 2015. Driver gaze tracking and eyes off the road detection system. IEEE Transactions on Intelligent Transportation Systems, 16(4): 2014-2027 [DOI: 10.1109/TITS.2015.2396031]

Wang C C, Gao S B, Cai C X and Chen H L. 2022. Vehicle-road visual cooperative driving safety early warning algorithm for expressway

scenes. Journal of Image and Graphics, 27(10): 3058-3067 (汪长春, 高尚兵, 蔡创新, 陈浩霖. 2022. 高速公路场景的车路视觉协同行车安全预警算法. 中国图象图形学报, 27(10): 3058-3067) [DOI: 10.11834/jig.210290]

Wang W H, Xie E Z, Li X, Fan D P, Song K T, Liang D, Lu T, Luo P and Shao L. 2021. Pyramid vision transformer: a versatile backbone for dense prediction without convolutions//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 568-578 [DOI: 10.1109/ICCV48922.2021.00061]

Yang J C, Wei F, Lin L, Jia Q X and Liu J Z. 2024. A research survey of driver drowsiness driving detection. Journal of Shandong University (Engineering Science), 54(2): 1-12 (杨巨成, 魏峰, 林亮, 贾庆祥, 刘建征. 2024. 驾驶员疲劳驾驶检测研究综述. 山东大学学报(工学版), 54(2): 1-12) [DOI: 10.6040/j.issn.1672-3961.0.2023.279]

Yuen K, Martin S and Trivedi M M. 2016. Looking at faces in a vehicle: a deep CNN based approach and evaluation//Proceedings of the 19th IEEE International Conference on Intelligent Transportation Systems. Rio de Janeiro, Brazil: IEEE: 649-654 [DOI: 10.1109/ITSC.2016.7795622]

Zhao X, Li Z, Fu R, Ge Z Z and Wang C. 2023. Real-time detection of distracted driving behavior based on deep convolution-tokens dimensionality reduction optimized visual transformer model. Automotive Engineering, 45(6): 974-988, 1009 (赵霞, 李朝, 付锐, 葛振振, 王畅. 2023. 基于深度卷积—tokens降维优化视觉Transformer的分心驾驶行为实时检测. 汽车工程, 45(6): 974-988, 1009) [DOI: 10.19562/j.chinasae.qcgc.2023.06.008]

Zheng H L, Fu J L, Zha Z J, Luo J B and Mei T. 2020. Learning rich part hierarchies with progressive attention networks for fine-grained image recognition. IEEE Transactions on Image Processing, 29: 476-488 [DOI: 10.1109/TIP.2019.2921876]

Zhuang P Q, Wang Y L and Qiao Y. 2020. Learning attentive pairwise interaction for fine-grained classification//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI: 13130-13137 [DOI: 10.1609/aaai.v34i07.7016]

作者简介

王腾，男，硕士研究生，主要研究方向为智能交通与计算机视觉。E-mail：1310820260@qq.com

高尚兵，通信作者，男，教授，主要研究方向为计算机视觉、模式识别和数据挖掘。E-mail：luxiaofen_2002@126.com

任刚，男，教授，主要研究方向为应急交通系统建模与仿真、交通大数据分析及应用、轨道交通运营风险防控。

E-mail：rengang@seu.edu.cn