L3级自动驾驶工况下驾驶人非驾驶任务轻量级检测算法*

王畅[1] 郭家煜[1] 李朝[1] 张韡[1] 张杨[2]

（1.长安大学汽车学院，西安 710064；2.陕西交通控股集团有限公司宝鸡分公司，宝鸡 721000）

【摘要】针对驾驶人非驾驶任务的检测和评估问题，在实车座舱驾驶模拟器中搭建L3级自动驾驶工况，对40名被试者开展了真实非驾驶任务采集试验，设计了移动设备操作、转身取物品等10种非驾驶任务真实试验，构建了相应的真实操作数据集。通过将小波变换卷积融入MobileViT架构，结合主干网络优化、嵌入位置调整及卷积核参数重构，将多尺度处理与增强局部-全局注意力机制进行融合建立了MobileViT-WT模型。测试结果表明，该模型在自建数据集上对非驾驶任务实现了93.85%的分类准确率，泛化性测试集准确率达82.90%，较原模型提高5.15百分点，参数量较原模型减少11%，模型处理图像的推理延迟仅为8.0 ms。研究结果可为L3级自动驾驶的人机接管与安全预警提供优化。

主题词：L3级自动驾驶 非驾驶行为分类 MobileViT网络 小波变换卷积

中图分类号：U461.91 **文献标志码：**A **DOI:** 10.19620/j.cnki.1000-3703.20250493

Lightweight MobileViT-WT Detection Algorithm for Non-Driving Tasks under L3 Conditions

Wang Chang[1], Guo Jiayu[1], Li Zhao[1], Zhang Wei[1], Zhang Yang[2]

(1. School of Automobile, Chang'an University, Xi'an 710064; 2. Baoji Branch, Shaanxi Transportation Holding Group Co., Ltd., Baoji 721000)

【Abstract】Under the L3 level automatic driving condition, the driver is in the state of leaving the driving link. Real time detection of the driver's non-driving task is an important mechanism to ensure the safe operation of the L3 level automatic driving system for the automatic driving system to understand the driver's state and evaluate the driver's emergency ability according to the driver's state. For the detection and evaluation of drivers' non-driving tasks, this study built L3 level automatic driving conditions in the real vehicle cockpit driving simulator, carried out real non-driving task collection tests on 40 subjects, carried out 10 kinds of real non-driving tasks experiments such as mobile device operation and turning around to pick up items, and constructed the corresponding real operation data set. By integrating wavelet transform convolution into MobileViT architecture, combined with backbone network optimization, embedding position adjustment and convolution kernel parameter reconstruction, the MobileViT-WT model is established by integrating multi-scale processing and enhanced local global attention mechanism. The test results show that the model achieves 93.85% classification accuracy for non-driving tasks on the self-built data set, and the accuracy of the generalized test set is 82.90%, which is 5.15% higher than the original model, and the number of parameters is 11% lower than the original model. The reasoning delay of the model processing image is only 8.0 ms. The research results can provide optimization for man-machine takeover and safety early warning of L3 automatic driving.

Key words: L3 Autonomous Driving, Non-Driving Behavior Classification, MobileViT Network, Wavelet Transform Convolution

【引用格式】王畅, 郭家煜, 李朝, 等. L3级自动驾驶工况下驾驶人非驾驶任务轻量级检测算法[J]. 汽车技术, 2026(10): 25-36.

WANG C, GUO J Y, LI Z, et al. Lightweight MobileViT-WT Detection Algorithm for Non-Driving Tasks under L3 Conditions[J]. Automobile Technology, 2026(10): 25-36.

*基金项目：陕西省重点研发计划项目（2023-YBGY-035）；陕西省交通运输厅交通科研项目（23-69K）；国家自然科学基金项目（52362050）。

通信作者：张韡（1973—），男，博士学位，副教授，主要研究方向为道路交通安全，zhweijx@chd.edu.cn。

1 前言

L3级自动驾驶作为汽车智能化转型的核心突破口，其独特价值在于平衡了技术可实现性与商业落地潜力，正推动全球车企与科技公司加速关键技术攻坚。德、日、美中四大市场已通过专项法规推进商业化，奔驰DRIVE PILOT、本田SENSING Elite等系统的量产，标志着L3技术范式趋于成熟。根据标准SAE J3016《Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles》，L3属于“有条件自动化”，系统在设计运行域（Operational Design Domain，ODD）内主导全部动态驾驶任务，但极端场景下需驾驶人应急接管[1]。这种“系统主导-人类备用”的特性，使得驾驶人在多数时间脱离驾驶操作，转而参与非驾驶活动，导致接管阶段的人机协同存在风险。

而这一应急接管过程恰恰是L3级自动驾驶系统安全的核心挑战：认知神经科学研究揭示，风险根源在于“注意瞬脱”现象，从非驾驶相关任务（NDRTs）切换至驾驶任务时，大脑前额叶皮层需350~650 ms完成认知定向[2]，且视觉型NDRTs会进一步延长该过程（注视点熵增加15%~20%）[3]。研究表明，驾驶人从非驾驶状态切换至驾驶状态时，平均需2.1 s完成认知重构，标准差为0.4 s，事故发生率较人工驾驶显著提升[4-5]。本质上，L3级自动驾驶系统接管事故的核心问题在于：系统缺乏对驾驶人NDRTs状态的实时感知能力，导致接管评估失效与人机状态失配。解决这一矛盾的关键路径在于实时NDRTs检测，通过识别NDRTs类型，结合非驾驶任务时长及环境复杂度量化认知负荷，再通过预警前置化（5~8 s分级刺激）、信息精准投放（适配人机交互界面呈现）与认知预热（神经信号激活）三重机制优化接管反应时间[6-8]，从而系统性化解“人机失配”风险。因此，构建实时NDRTs检测体系成为保障L3级自动驾驶系统应急接管安全的核心前提。

然而，现有驾驶人分心行为研究主要基于State Farm Distracted Driver Dataset（SF-DDD）[9]和Distracted Driver's Dataset[10]等传统数据集，其记录的“分心行为”仍以驾驶人双手在转向盘上的短暂注意力分散为主，与L3级自动驾驶系统中“驾驶人深度参与NDRTs”的场景存在本质区别，且NDRTs与接管表现（Takeover Performance，TOP）存在显著相关性[11-13]。模型方面，传统CNN模型（AlexNet/GoogLeNet/ResNet50）在State Farm数据集上准确率为74.9%~81.6%[14-16]。轻量级模型（如SwiftFormer[17]、EfficientViT[18]）凭借低计算量的优势，更适合车载嵌入式部署，轻量级模型MobileViT[19]在State Farm数据集下能达到99.74%的准确率[20]，但这些模型未针对L3特有的“双手脱离+多任务复合”行为，其时空特征提取能力在L3场景下的适应性尚未验证。因此，构建L3级自动驾驶场景导向的多模态NDRTs基准数据集并开发适配的分类模型，对智能网联汽车安全技术发展具有关键意义。

针对上述L3场景NDRTs检测的核心挑战（专用数据集缺失、模型适配性不足），本文提出3方面方案：针对缺少数据集的问题，建立首个面向L3场景的NDRTs专用数据集，涵盖系统接管状态下的典型多模态行为；设计MobileViT-WT网络，以期在现有框架中集成小波变换卷积（Wavelet Transform Convolution）[21]增强多尺度特征提取能力以及减少模型参数；最后，建立非驾驶行为检测框架，以期在专用数据集和泛化性测试数据集上验证算法的准确率与场景适应性。

2 L3级自动驾驶非驾驶任务数据集构建

研究基于L3级自动驾驶模拟平台，通过设计典型接管场景试验，同步采集驾驶人视频数据，构建了包含接管全过程的人-车-路数据集。本节将从试验平台和试验设备、试验设计、NDRTs设计试验流程4个方面详细介绍数据集的构建。

2.1 试验设备

本文采用驾驶模拟系统开展试验（见图1），该系统由实车座舱、运动模拟平台、视景显示系统和数据采集模块组成。实车改装的两厢轿车驾驶舱可提供真实的操纵反馈，被试者可在舱内完成驾驶操作及次任务活动。系统配备的车辆底部的垂向电动平台能提供较为真实的路面细微冲击感，其与真实高速行驶下的乘坐感受较为接近。环幕投影系统通过180°曲面屏呈现高仿真交通场景，支持后视镜虚拟成像和抬头显示功能。试验人员可通过交互界面设置自动驾驶车辆模式。采用单一视频采集系统，通过海康威视高清摄像头（DS-2CD3系列，1 920像素×1 080像素分辨率，30 Hz帧率）专门采集驾驶人面部及躯干动态信息，由视频采集软件进行同步记录。试验过程中，操作人员统一触发采集设备，确保数据获取的同步性和准确性。

2.2 试验设计

本研究招募40名健康驾驶人（男25名，女15名），平均年龄29.35岁，年龄范围23~37岁，平均驾龄8.12年，所有受试者视力正常且无心脑血管疾病。试验前要求受试者保持良好作息，避免摄入影响生理状态的食物或药物。

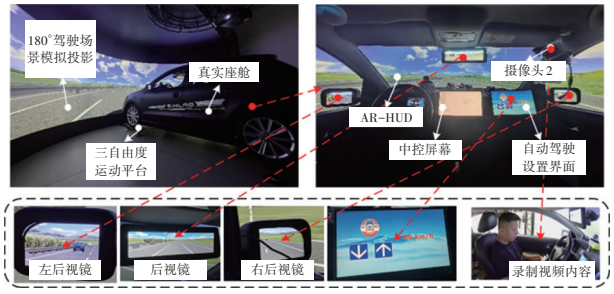


图1　驾驶仿真试验平台及其座舱布置

基于SILAB驾驶仿真软件构建高速公路L3级自动驾驶场景，包括接管过程。采用双向6车道、20 km直线路段，天气与路况良好，如图2所示。试验在L3级自动驾驶模式下进行，参照其他研究将接管请求前置时间设置为7 s，车辆沿中间车道匀速行驶（100 km/h），系统会在30~50 s的随机时间间隔后触发障碍物场景。当检测到前方障碍物时，基于预设的7 s接管请求时间，系统立即启动包含抬头显示器（Head-Up Display，HUD）红色闪烁警示和立即接管车辆语音指令的多模态接管提示。被试者需要在7 s时间窗内通过转动转向盘或踩下制动踏板完成接管操作。试验要求在5.5~6 km的循环测试路段中重复进行上述接管场景测试。

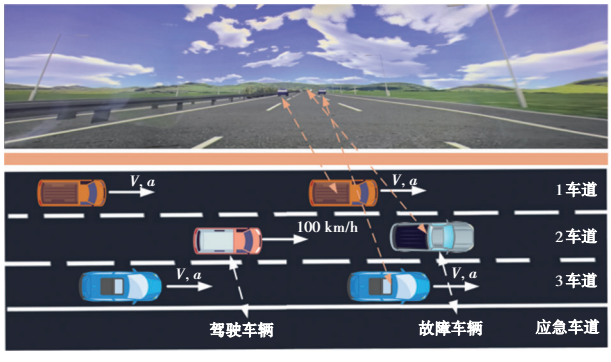


图2　试验场景示意

2.3　NDRTs设计

根据L3级驾驶过程的实际特点，参考现有辅助驾驶系统工作过程中驾驶人的行为，以及车载环境下驾驶人非驾驶任务的实际发生概率，试验确定了以下非驾驶任务。包括单模态任务（如观看中控屏视频、单手刷短视频、吃零食、喝水、转身拿东西、打电话、双手打游戏）、复合型任务（如边进食边操作触屏、边喝水边刷短视频）及系统交互任务（如车载游戏操作）共10种典型的L3级自动驾驶非驾驶任务。具体NDRTs细节如图3所示。

为降低被试者自动驾驶经验差异的影响，试验前进行了标准化接管培训。培训内容包括：驾驶系统操作说明（启停方法）、车辆控制装置适应性训练、试验场景熟悉、接管操作练习（含NDRTs任务）。培训结束后设置5 min休息时间，待被试者确认掌握操作后进入正式试验。



图3　L3级自动驾驶典型非驾驶相关任务（NDRTs）行为示例

2.4　试验流程

被试者需在正式试验开始之前了解整个试验过程并签署知情同意书，填写姓名、驾龄以及性别等基本信息，进行接管培训。驾驶试验分为3个阶段：先进行人工驾驶，驾驶人驾驶车辆到中间车道，并缓慢加速到规定车速，随后手动开启自动驾驶；系统接管后执行10类预定义NDRTs；响应系统触发接管请求（Take-Over Request，TOR），驾驶人在接管车辆后需完成避障操作并安全返回中间车道，经试验人员确认后停车并填写问卷。每名受试者完成10次完整循环（10个任务×10次重复），将录制好的完整视频提取纯自动驾驶阶段的视频片段（剔除人工驾驶/TOR响应阶段），获得400段有效NDRTs视频，随后通过多阶段数据精炼构建专用数据集：首先进行关键帧提取，为避免每一类别中驾驶人行为姿态过于相似，按每秒提取1帧图像并通过人工检验去除相似图像；最后通过人工检验确保数据集质量。最终，形成了驾驶人L3级自动驾驶非驾驶任务数据集，共计25 253张，数据集中各任务所占比例如图4所示。

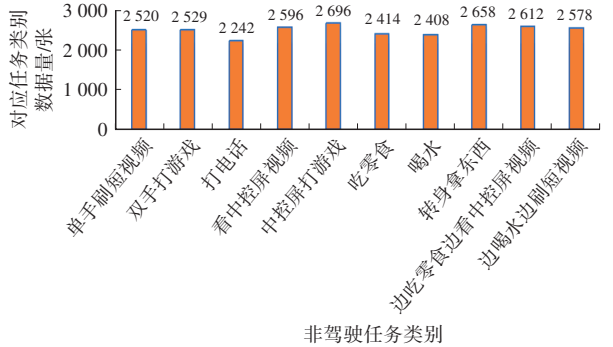


图4　L3级自动驾驶非驾驶任务数据集非驾驶行为统计

3 L3级自动驾驶NDRTs检测方法

3.1 MobileViT

MobileViT是一种创新的轻量级混合架构，通过深度融合卷积神经网络（Convolutional Neural Network，CNN）的局部归纳偏置与ViT（Vision Transformer）的全局上下文建模能力，实现了计算效率与表征能力的平衡。其核心架构由三个关键组件构成：标准卷积层、MobileNetV2倒残差模块（Inverted Residual Block）和MobileViT Block。其中，MobileViT Block采用三阶段特征处理流程：

a. 局部特征编码阶段：采用深度可分离卷积（Depthwise Separable Convolution）进行局部感受野内的特征提取，捕获低层次视觉特征（如Gabor-like边缘滤波器和纹理模式）。通过1像素×1像素点卷积（Pointwise Convolution）实现通道维度的线性投影，满足后续模块的输入维度要求。

b. 全局关系建模阶段：将特征图划分为$n \times n$的非重叠块（Patch），通过多头自注意力机制（Multi-Head Self-Attention, MHSA）建立跨区域依赖关系。该过程采用位置编码（Positional Encoding）保留空间信息，其计算复杂度通过块划分策略降低至$O(N^2/P^2)$，其中O为时间复杂度，N为输入特征图的空间尺寸，P为划分块大小。

c. 特征融合阶段：采用跳跃连接（Skip Connection）实现残差学习，通过通道拼接（Channel Concatenation）和1像素×1像素卷积完成局部-全局特征聚合。

通过这种局部与全局信息相结合的设计，MobileViT网络不仅提高了特征提取的能力，还保持了较低的计算开销，适用于移动端和嵌入式设备等资源受限的场景。

3.2 融合小波变换卷积的轻量化NDRTs检测网络

本文提出的MobileViT-WT网络通过引入小波变换卷积模块对原始MobileViT架构进行了优化改进。在网络输入端，将传统的3 × 3深度卷积替换为5 × 5小波变换卷积模块，同时在每个MobileViT Block中的第一个深度卷积层也采用5 × 5小波变换卷积模块进行替代。这种改进设计充分发挥了小波变换在特征提取方面的优势，通过其多尺度分析能力有效捕捉图像中的关键细节特征。

如图5所示的网络结构中，改进后的MobileViT-WT首先通过5 × 5的1级小波变换卷积模块对输入图像进行特征提取，随后依次经过layer1至layer5的改进模块处理。每个改进的MobileViT Block都包含1级小波变换卷积模块，能够更好地提取多尺度特征。最后网络通过卷积层、全局池化层和全连接层完成10类NDRTs的分类。试验表明，这种结构改进在降低模型参数量的同时提高了分类准确率，实现了轻量化与性能的平衡优化。

小波变换卷积模块计算过程如下：将输入特征图$\boldsymbol{X}^{(C\times H\times W)}$首先通过小波变换核进行多尺度分解，对输入的低频和高频内容进行滤波和下采样。其中C为通道数，H为特征图高度，W为特征图宽度。在对不同的频

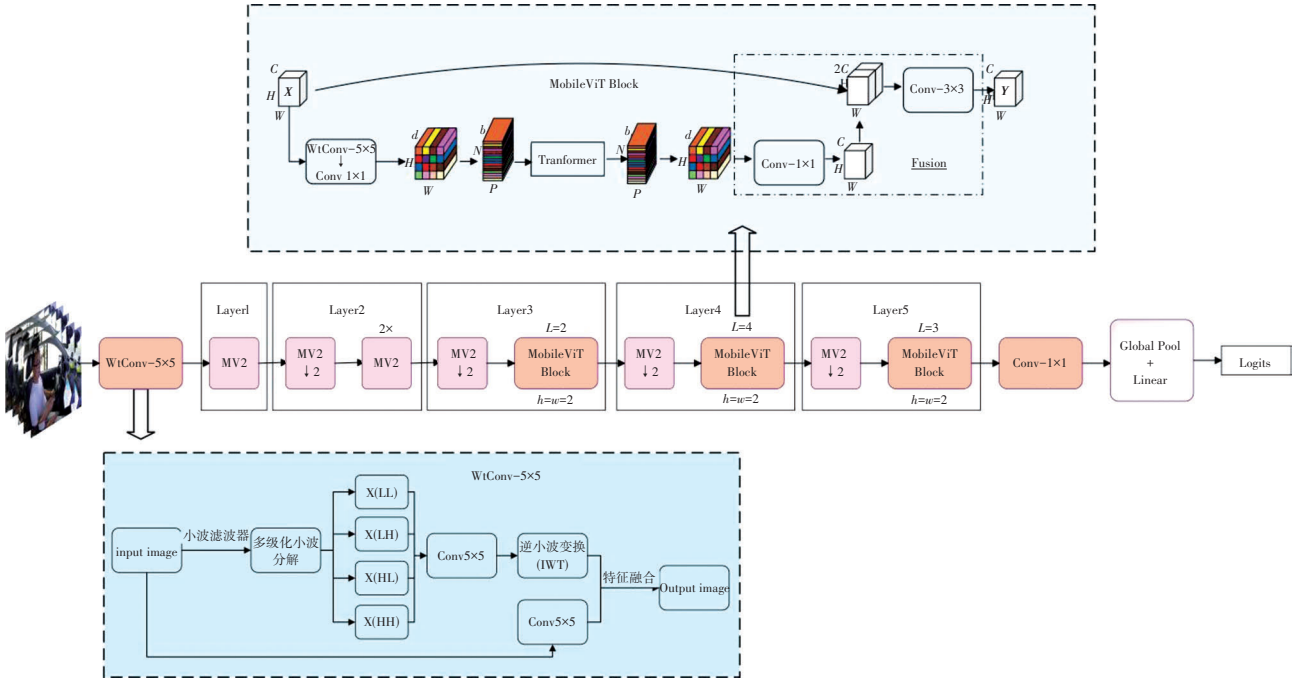


图5 MobileViT-WT网络结构

率图进行小核深度卷积之后，使用逆小波变换构建输出，其过程可表示为：

$$[\boldsymbol{X}_{\mathrm{LL}},\boldsymbol{X}_{\mathrm{LH}},\boldsymbol{X}_{\mathrm{HL}},\boldsymbol{X}_{\mathrm{HH}}]=\mathrm{Conv}([f_{\mathrm{LL}},f_{\mathrm{LH}},f_{\mathrm{HL}},f_{\mathrm{HH}}],\boldsymbol{X}) \quad (1)$$

$$\boldsymbol{Y}=IWT(\mathrm{Conv}(W,WT(\boldsymbol{X}))) \quad (2)$$

式中：$\boldsymbol{X}$为输入张量；f_{LL}为一个低通滤波器；f_{LH}、f_{HL}、f_{HH}为一组高通滤波器；$\boldsymbol{X}_{\mathrm{LL}}$为$\boldsymbol{X}$的低频分量，$X_{\mathrm{LH}}$、$X_{\mathrm{HL}}$、$X_{\mathrm{HH}}$分别为其水平、垂直和对角高频分量，经过逆小波变换（Inverse Wavelet Transform，IWT）、通过子带的线性组合，重建原始特征图；$\boldsymbol{Y}$为输出特征图；WT为小波交换；Conv为卷积层。

此操作不仅分离了不同频率分量之间的卷积，还允许较小的核在原始输入的较大区域上操作，即相对于输入增加了其感受野。其中上标i为第i个网络层（或第i个阶段/模块）的索引：

$$\boldsymbol{X}_{\mathrm{LL}}^{(i)},\boldsymbol{X}_{\mathrm{H}}^{(i)}=WT(\boldsymbol{X}_{\mathrm{LL}}^{(i-1)}) \quad (3)$$

$$\boldsymbol{Y}_{\mathrm{LL}}^{(i)},\boldsymbol{Y}_{\mathrm{H}}^{(i)}=\mathrm{Conv}\left(W^{(i)},(\boldsymbol{X}_{\mathrm{LL}}^{(i)},\boldsymbol{X}_{\mathrm{H}}^{(i)})\right) \quad (4)$$

为了组合不同频率的输出，利用了小波变换及其逆是线性操作这一事实，即：

$$IWT(\boldsymbol{X}+\boldsymbol{Y})=IWT(\boldsymbol{X})+IWT(\boldsymbol{Y}) \quad (5)$$

$$\boldsymbol{Z}^{(i)}=IWT(\boldsymbol{Y}_{\mathrm{LL}}^{(i)}+\boldsymbol{Z}^{(i+1)},\boldsymbol{Y}_{\mathrm{H}}^{(i)}) \quad (6)$$

执行式(6)可以将不同级别的卷积相加，其中$\boldsymbol{Z}^{(i)}$为从第i级开始的累积输出。将2个不同大小卷积的输出相加作为最终输出。最终输出特征图$\boldsymbol{Y}$保持了与输入相同的空间维度，具有更强的多尺度特征表征能力。传统卷积层通过密集的卷积核处理输入数据，导致参数量随通道数和卷积核尺寸平方增长。而小波卷积利用小波基函数的局部性和多分辨率分析特性，将输入分解为不同频带的子信号，仅需对关键频带进行卷积操作，从而显著减少冗余参数。1级小波卷积处理流程如图6所示。

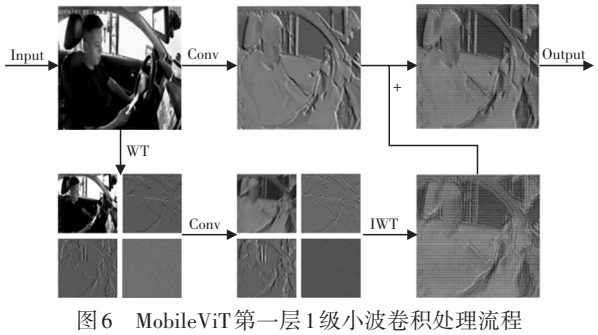


图6　MobileViT第一层1级小波卷积处理流程

特征图$\boldsymbol{Y}$依次通过MV2和MobileViT Block的组合模块，在MV2的倒残差结构中进行下采样；在MobileViT Block中，利用5×5小波变换卷积（Wavelet Transform Convolution）模块对特征图进行重构，并通过1像素×1像素卷积层将通道数调整为Transformer所需的特征维度，得到高维空间特征表示。随后，将该特征图划分为不重叠的局部特征块（Patch），并进行展平操作，生成N个特征块：

$$P=wh \quad (7)$$

$$N=\frac{WH}{P} \quad (8)$$

式中：P为单个特征块（Patch）包含的特征数量，w、h分别为单个特征块的宽度和高度，N为输入特征图划分后的特征块数量。

随后，通过多个Transformer的编码模块计算相似像素块之间的注意力信息，即对输入的像素向量进行位置编码，计算每个相似向量自注意力，获取向量元素之间的注意分数进行加权平均，通过多头注意力机制将自注意力的输出进行拼接，最后通过add&Norm层和Feed Forward输出特征图$\boldsymbol{Y}_G\in R^{P\times N\times d}$。其中，$p$为第$p$个特征块的索引，$P$为特征块总数量；$YU(p)$为第$p$个输入特征块，即未经Transformer编码的局部特征；Transformer$Y_G(p)$为经过Transformer编码后的第p个全局特征表示；Transformer( )为Transformer编码器对输入特征进行自注意力计算和特征融合的映射操作。其定义式为：

$$\boldsymbol{Y}_G(p)=\mathrm{Transformer}(\boldsymbol{Y}_U(p)),\quad 1\leqslant p\leqslant P \quad (9)$$

其次，将Y_G折叠回原高维度得到$\boldsymbol{Y}_F\in\boldsymbol{R}^{W\times H\times d}$，在1像素×1像素的卷积之后将其与改进后MobileViT Block的输入特征图$\boldsymbol{Y}$进行拼接。最后经过3×3的卷积层进行全局特征和局部特征的融合，使得MobileViT具有更有效的全局感受野。

将小波变换（WT）模块嵌入到MobileViT网络的第一层卷积层之后，并替换MobileViT Block中的首个3×3深度卷积层，达到在算法初期通过多尺度频域分解增强图像关键特征提取能力的效果。通过小波变换的低频-高频分离特性，网络能够同时捕获全局结构信息和局部细节特征，与算法后期MobileViT Block的全局自注意力机制形成互补，进一步提升特征表征的丰富性。这种设计显著提高了MobileViT-WT网络的多尺度特征提取能力，在保证模型轻量化的同时有效提升了分类准确率。

4　试验设置与结果分析

试验采用自主构建的L3-NDB Dataset（SAE Level 3 Non-Driving Behaviors Dataset）。该数据集包含10类典型NDRTs，共计25 253张图像，能够有效支持相关研究。在网络模型的训练过程中，将数据集以8∶2的比例划分为训练样本和测试样本。为提高模型泛化能力，对训练集进行图像增强，包括随机裁剪、水平翻转

和归一化处理。验证集仅进行了尺寸调整和归一化处理，使用动量为0.9、初始学习率为0.01的随机梯度下降（Stochastic Gradient Descent，SGD）优化器来优化模型的参数，Batch Size设置为32，共训练150轮次。试验使用Windows 11专业版系统、Intel® Core™ i7 13代处理器、NVIDIA RTX 4070 12G GPU，基于PyTorch 2.1的深度学习框架，配套环境为CUDA 11.8。模型架构搭建具体试验流程如图7所示。

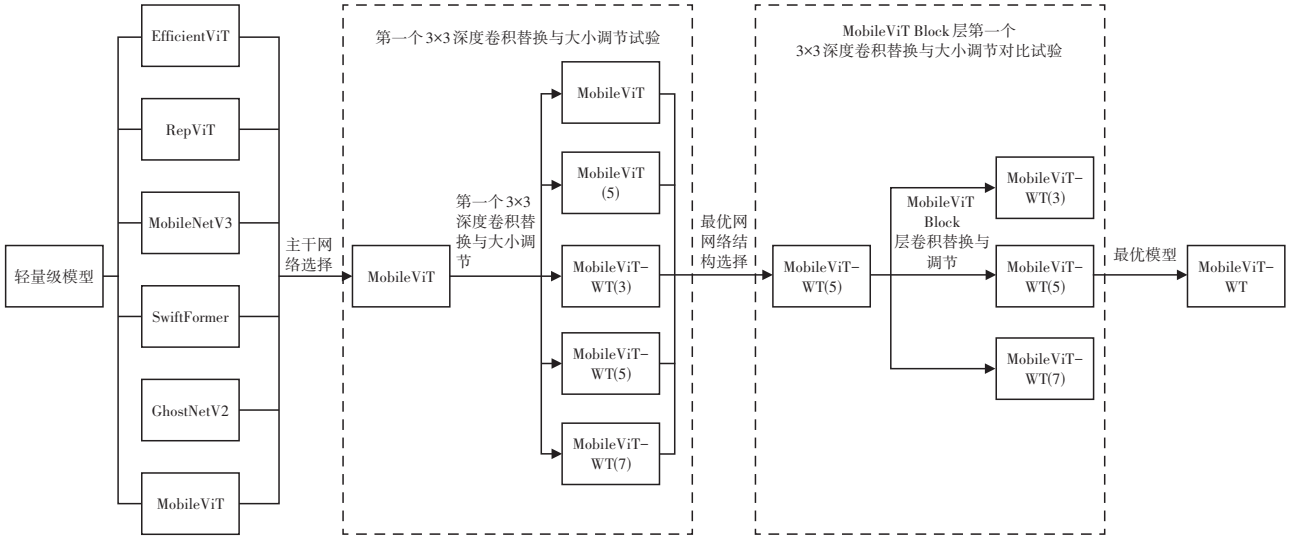


图7　模型试验分析与验证流程

4.1　轻量级分类模型对比试验

本研究系统对比了当前主流的轻量级CNN架构（MobileNetV3[22]、GhostNetV2[23]）与Transformer架构（MobileViT、RepViT[24]、EfficientViT）在L3自动驾驶非驾驶行为分类任务中的性能表现。试验采用统一训练设置，模型均从头开始训练以避免预训练权重的影响。不同模型性能对比如图8所示。

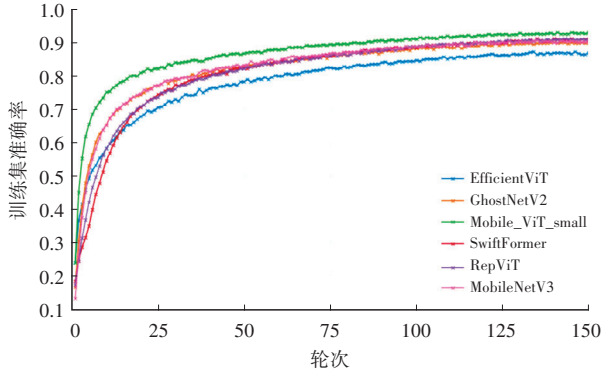


图8　不同轻量级分类网络模型效果对比

6个轻量级分类网络的检测准确率与模型参数对比如表1所示。试验结果表明，MobileViT以93.03%的准确率位列第一，显著优于其他模型，较第2名的SwiftFormer提升1.2百分点。虽然EfficientViT和MobileNetV3在参数量上较小，但检测准确率低于MobileViT。将SwiftFormer、RepViT和MobileViT横向对比可见，MobileViT不仅准确率高，参数量也小于SwiftFormer。因此，综合权衡检测准确率与计算效率，MobileViT更能满足L3级自动驾驶驾驶人行为分类任务对高精度和实时性的双重要求，因此被选定为本研究的基础架构。

表1　不同分类网络准确率及参数量对比

模型	Top准确率/%	参数量/×10⁶
EfficientViT	87.31	2.16
RepViT	91.23	4.72
MobileNetV3	90.66	3.91
SwiftFormer	91.2	5.65
GhostNetV2	90.19	4.89
MobileViT	93.03	4.95

4.2　基于小波变换卷积的深度卷积替换与核大小调节对比试验

本研究针对MobileViT网络架构进行了小波变换卷积（WTConv）的模块化改进，重点探究了不同层级和卷积核尺寸对模型性能的影响。设计了两种改进方案：MobileViT-WT（n）（替换输入层深度卷积）和MobileViT-WK（n）（替换MobileBlock首层深度卷积），n为卷积核大小，并通过控制试验验证了不同小波核尺寸（3×3、7×7）的性能差异。在模块选择方面，排除对InvertedResidual结构的改造，该结构采用深度可分离卷积（Depthwise Separable Convolution），并通过通道扩展比例（Expansion Ratio）控制倒残差结构中扩展层的通道数量，本文设置扩展比例为6。InvertedResidual模块本身具备高效的特征提取能力，若在此处替换为WTConv，会增加计算复杂度，且未必带来显著的性能提升。

4.2.1　输入层深度卷积替换与核大小调节对比试验

将小波变换卷积替换MobileViT中的第一个3×3卷积并调节大小。不同模型性能对比如图9所示。

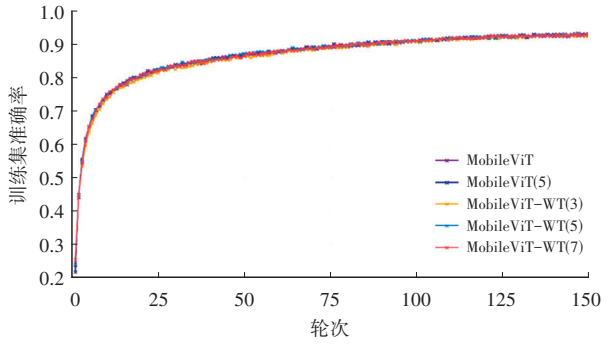


图9　MobileViT中的第一个3×3深度卷积替换与大小调节效果对比

深度卷积替换与大小调节对比如表2所示，研究系统评估了不同尺寸小波卷积核在MobileViT输入层的性能表现。试验结果表明，5×5小波卷积表现出最优的性能，在仅增加0.01×10^6参数量的情况下，将Top-1准确率提升0.48百分点至93.51%，Top5平均准确率提升0.35百分点。7×7小波卷积可以捕捉更广泛的特征，但是其精度提升不明显，且计算量增加。3×3小波卷积性能略有下降（-0.04百分点），这表明过小的感受野限制了多尺度特征的提取能力。结果证明，5×5小波卷积对原有卷积的替换在L3级自动驾驶场景中实现了精度与效率的最佳平衡，既有效捕捉驾驶人的细粒度行为特征，又保持了模型的轻量化特性。

表2　输入层模型改进性能对比表　%

模型	Top准确率	Top5平均准确率
MobileViT	93.03	92.97
MobileViT(5)	93.25	93.15
MobileViT-WT(3)	92.99	92.89
MobileViT-WT(5)	93.51	93.32
MobileViT-WT(7)	93.38	93.21

4.2.2　Mobile Block层首层深度卷积替换与核大小调节对比试验

将MobileViT Block层局部表征提取位置对深度卷积进行替换和尺寸调节，模型性能对比如图10所示。

不同模型网络检测准确率对比如表3所示。基于4.2.1节得到的MobileViT-WT(5)模型，对MobileBlock层的首层深度卷积进行了结构优化试验：将原始深度卷积替换为5×5大核深度卷积以扩大感受野，对比了3×3常规核的基准效果；引入小波卷积替代传统深度卷积并进行核大小调节。

试验结果如表3所示，将3×3深度卷积替换为5×5深度卷积时，模型准确率提升了0.63百分点，但参数量增加了17%；而用3×3小波卷积替换3×3深度卷积时，准确率提升0.31百分点；当小波卷积核为5×5时准确率提升0.73百分点，Top5准确率提升0.62百分点，且模型参数量为4.43×10^6，较原始MobileViT网络参数量降低11%，较5×5深度卷积的MobileViT网络参数量降低23%；当小波卷积核为7×7时，准确率提升0.43百分点。3×3小波卷积由于感受野较小，特征提取能力有限；而5×5深度卷积虽然提升了特征提取能力，却显著增加了参数量；7×7小波卷积核虽然进一步扩大了感受野，但特征提取过于粗糙。试验证实，5×5小波卷积替换MobileBlock层首层深度卷积，通过其多尺度分析特性，在扩大感受野的同时避免了过大卷积核带来的特征粗糙化问题，实现了精度（+0.73百分点）与参数量为4.43×10^6的最佳平衡。

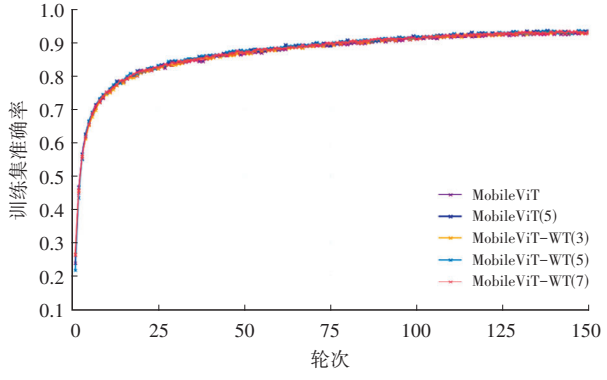


图10　MobileViT Block层首层3×3卷积的替换与大小调节效果对比

表3　MobileViT Block首层卷积改进的性能对比

模型	准确率/%	Top5准确率/%	参数量×10^6
MobileViT	93.03	92.97	4.95
MobileViT(5)	93.66	93.54	5.76
MobileViT-WK(3)	93.34	93.24	4.41
MobileViT-WK(5)	93.85	93.69	4.45
MobileViT-WK(7)	93.46	93.33	4.47

综上可知，本文所提出的MobileViT-WT模型在L3级自动驾驶非驾驶行为分类任务中展现出显著优势，相较于原模型在准确率上有一定提高，且参数量减少，优于原MobileViT模型。相较于其他轻量级模型，MobileViT-WT模型体积更小，准确率更高。

4.3　MobileViT-WT试验结果分析

为验证模型的泛化能力，构建了独立的泛化性测试数据集。包含6位新受试者，共计60段L3级自动驾驶视频。对视频进行处理形成区别于上述试验中L3级自动驾驶非驾驶任务数据集的一个泛化性验证数据集，

共计2 977张图片。泛化性数据集中驾驶人的外形特征与L3级自动驾驶非驾驶任务数据集存在差异。

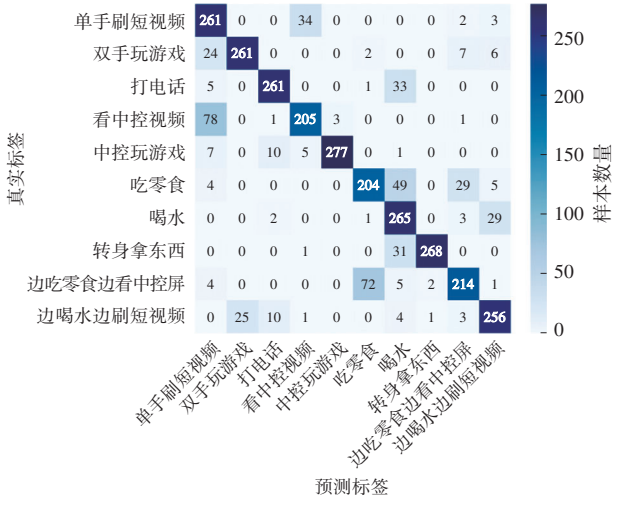
将不同的轻量级模型在L3驾驶人非驾驶任务泛化性验证数据集进行非驾驶行为分类检测试验，得到不同模型对于非驾驶任务的分类准确率，不同模型对于非驾驶任务分类准确率如表4所示。

表4　泛化性验证数据集分类检测试验结果对比　%

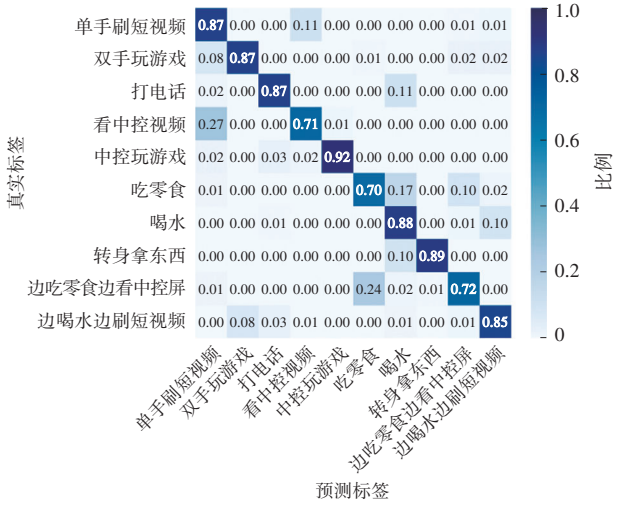
行为类别	EfficientViT	RepViT	MobileNetV3	SwiftFormer	GhostNetV2	MobileViT	MobileViT-WT
单手刷短视频	51.0	52.3	49.7	63.3	60.7	85.0	87.0
双手玩游戏	70.3	72.0	72.0	74.7	92.3	72.3	87.0
打电话	59.0	78.3	96.0	59.0	86.0	78.0	87.0
看中控视频	87.5	69.4	96.5	64.2	89.2	75.0	71.2
中控玩游戏	90.0	96.0	94.3	88.7	37.3	64.3	92.3
吃零食	68.7	62.2	63.2	66.0	54.6	76.3	70.1
喝水	88.3	95.7	77.3	91.0	64.3	91.0	88.3
转身拿东西	76.0	72.3	83.7	74.0	96.7	71.7	89.3
边吃零食边看中控屏	82.2	71.5	71.8	67.8	89.6	66.8	71.8
边喝水边刷短视频	88.0	74.3	89.0	75.0	91.7	93.3	85.3
平均准确率	76.10	74.40	79.35	72.37	76.24	77.75	82.90

表4的试验结果表明，MobileNet-WT与EfficientViT、RepViT、MobileNetV3等轻量级模型进行对比，在单手刷短视频、吃零食这两个任务上均有显著提升，且在中控屏玩游戏、喝水、转身拿东西、边喝水边刷短视频、单手刷短视频、双手玩游戏、打电话这些任务上的分类仍有不错的效果，总体平均准确率达到82.90%，高于其他轻量级模型在L3驾驶人非驾驶任务泛化性验证数据集上的表现。与原始的MobileViT模型对比，改进后的MobileViT-WT模型，除看中控屏视频、吃零食、边喝水边刷短视频3个任务上表现略低外，其他任务处理上MobileViT-WT模型更具优越性。

图11为采用MobileViT-WT模型在泛化性数据集上试验所得的分类混淆矩阵。MobileViT-WT模型在独立验证集上展现出良好的泛化能力，10类非驾驶行为的平均识别准确率达82%。具体而言：高性能类别的中控游戏（92%）和转身取物（89%）识别率最高，单手操作设备等5类行为准确率均超过85%；低性能类别的看中控视频（70%）、进食（73%）及其复合行为（71%）表现欠佳。相较于传统的分心检测，在L3级自动驾驶的情况下，驾驶人的行为更加灵活并且会产生复合任务如边吃零食边看中控和边喝水边刷短视频，导致了非驾驶行为在姿态上与个别行为有着较大的相似性，加上行为种类的增加，模型分类难度增加。在全新的泛化性验证数据集中，MobileViT-WT模型虽在个别任务中识别准确率略低，但总体表现出较好的泛化性。



（a）原始数量混淆矩阵



（b）归一化混淆矩阵

图11　MobileViT-WT模型对于L3驾驶人非驾驶任务泛化性验证数据集分类混淆矩阵

图12为通过Grad-CAM方法对图像特征提取的注意力分布特性为可视化结果。如图12第1行所示，MobileViT-WT模型可以端到端准确预测得到动作类别（class）和置信度（prod）。通过对比，提出的MobileViT-WT模型可精准定位关键行为特征，在手持移动设备类检测任务中，可以聚焦手机持握区域，对于中控交互任务，可以准确捕捉触屏位置（置信度>90%）；在复合型任务（如边喝水边刷短视频）中，可同步识别手部与面部区域。其余模型在不同非驾驶任务中都有错误的位置识别，MobileViT模型识别转身动作漏检躯干区域，MobileNetV3模型在中控玩游戏中错误关注非交互区域，GhostNetV2模型在饮水动作及边进食边看中控任务中忽略手部特征。对比基线模型，MobileViT-WT模型在单手刷短视频和双手打游戏等非驾驶任务中关注的区域更加全面，在转身拿东西、中控打游戏和打电话非驾驶任务中，MobileViT-WT模型关注的区域更加准确。

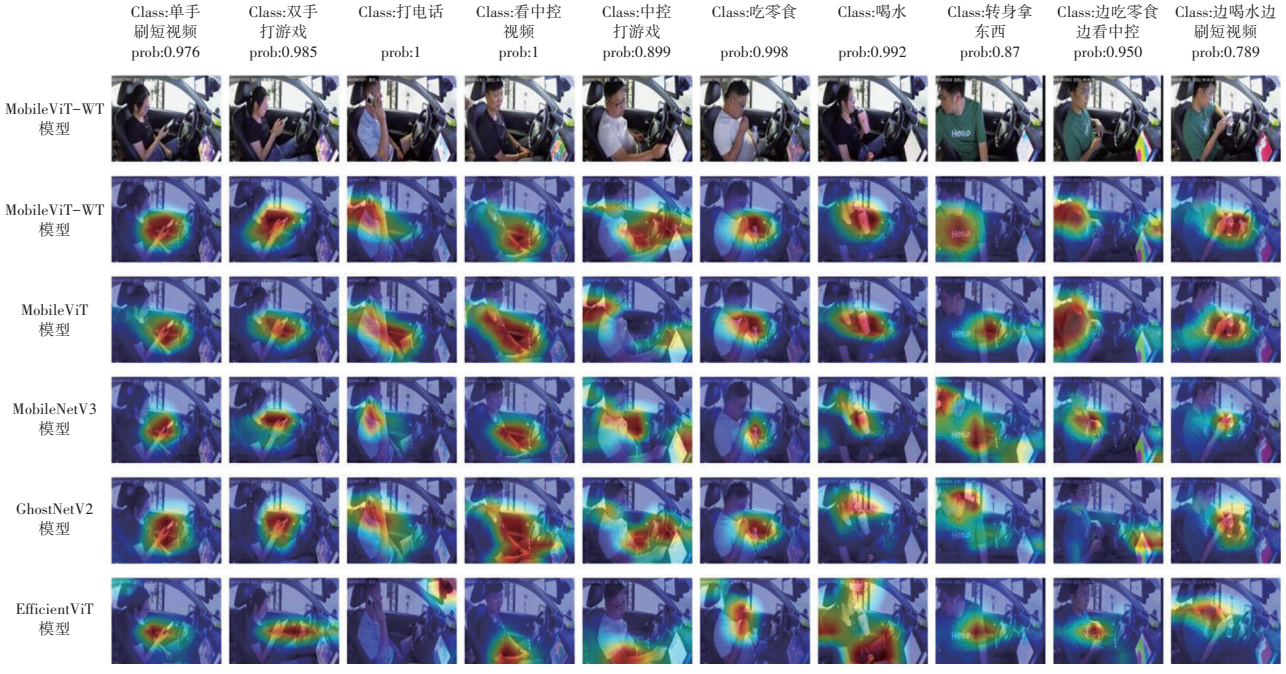


图12　热力图的模型注意力区域可视化

如图11b所示，MobileViT-WT模型在L3级自动驾驶条件下对看中控视频、吃零食及边进食边看视频3类行为的识别准确率较低（约70%）。看中控视频任务的误判主要表现为与单手刷短视频行为的混淆，有27%被错误识别为单手刷短视频。吃零食任务存在16%的误判率，主要混淆类型为饮水动作及复合型任务边吃零食边看中控视频。观看中控视频、吃零食这项复合型任务中，有24%被错误识别为吃零食。如图12第2行所示，看中控视频任务识别率低源于模型错误关注手部特征而忽略关键的面部和眼部特征；如图13所示，对于吃零食任务，热力图均能关注到手与口的特征，但是吃零食和喝水这两项非驾驶任务在运动学特征上具有相似性，存在手-口之间的轨迹相似，头部的伴随运动也存在相似，导致识别混淆；在观看中控视频、吃零食这项复合型任务，模型对复合型任务中的特征区域未能关注，且两项任务存在任务重叠的原因，导致识别效果不佳。且所有轻量级模型在吃零食任务上的识别性能均不理想，其中MobileViT-WT模型的准确率为70%（见表4）。

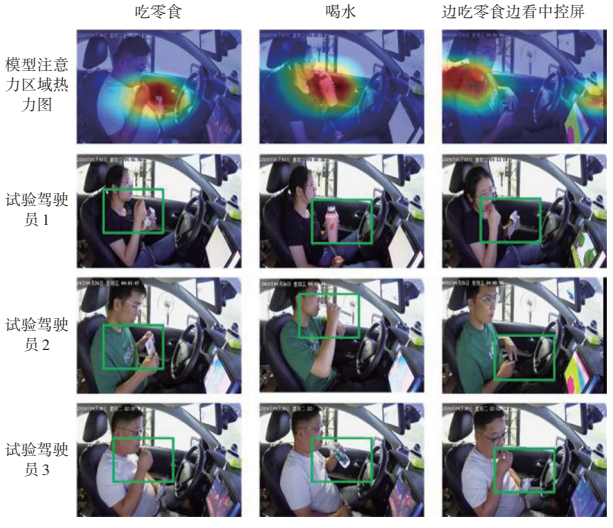
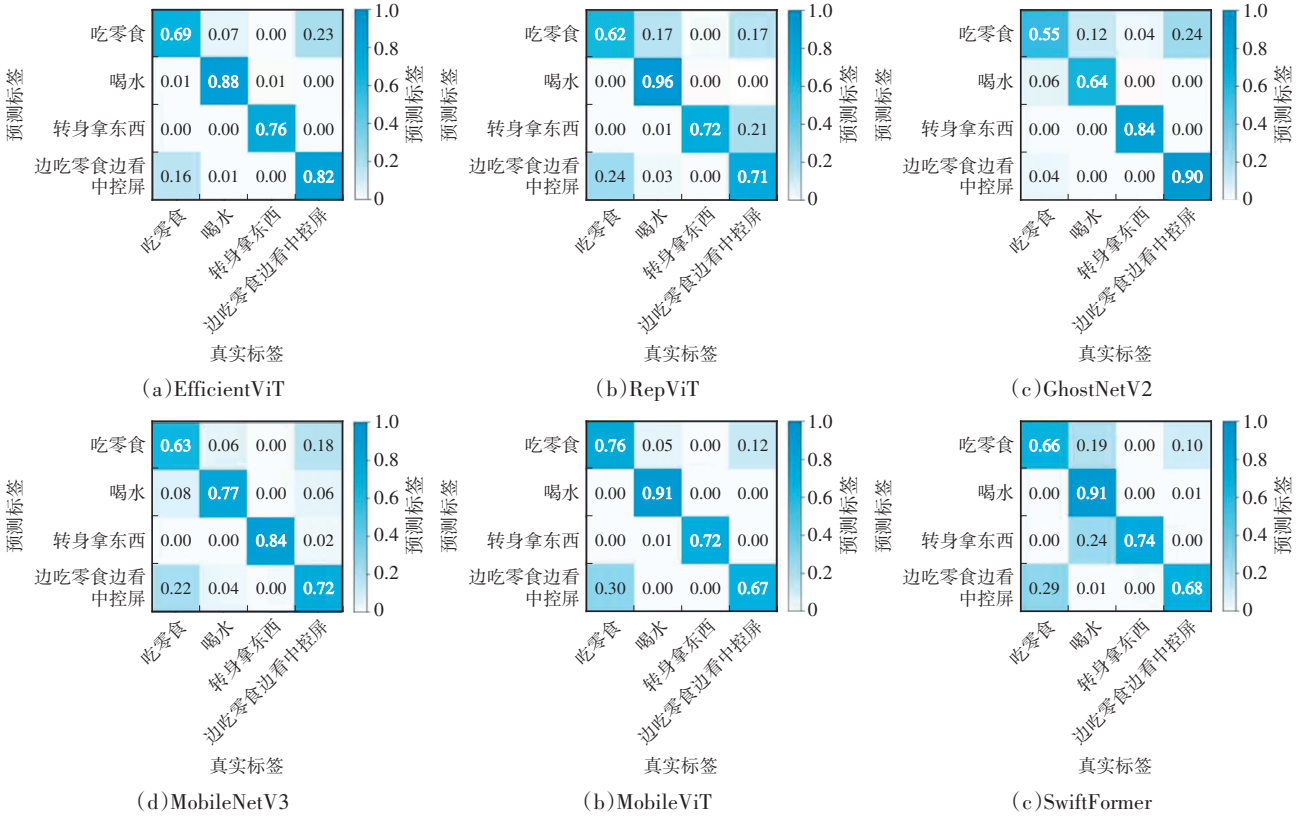


图13　混淆任务非驾驶行为特征细节

图14为轻量级模型在吃零食任务中的混淆矩阵，通过对图14的分析可以发现，6个轻量级模型在该任务上普遍存在识别混淆现象，主要表现为与饮水动作以及“观看中控视频+吃零食”复合型任务之间的

混淆，产生混淆的比例在15%~35%范围内。而导致这一现象发生的主要原因则是上文中提到的吃零食和喝水这两项非驾驶任务的运动学特征相似性以及复合型任务的影响。因此，吃零食行为的识别及多模态特征重叠任务的识别在当前轻量级模型架构下仍存在普遍性挑战。



(a)EfficientViT　(b)RepViT　(c)GhostNetV2

(d)MobileNetV3　(b)MobileViT　(c)SwiftFormer

图14　轻量级模型在吃零食任务上的混淆矩阵(含喝水、边吃零食边看中控视频)

本次试验将模型部署于NVIDIA RTX 4070桌面级GPU。测试结果表明，MobileViT-WT模型的端到端平均延迟(包含图像预处理、模型推理及后处理)仅为8 ms，显著优于原始MobileViT模型的11 ms。这一性能提升验证了所提出改进的有效性。就算法实时性而言，该速度等效125 帧/s的处理吞吐量，远超实时视频分析(30 帧/s)的基准要求。进一步地，从人机协同接管的角度分析，该延迟仅为人类驾驶员完成认知定向所需时间(约350~650 ms)的2%~5%，更是远小于完整的2~3 s接管过程。因此，无论在算法层面还是在实际应用层面，本模型的推理速度都为实时预警与干预留下了极其充裕的时间。尽管一个完整的车载监控系统还涉及图像采集、数据传输等额外开销，但本模型极低的延迟确保了它不会成为整个系统的性能瓶颈。在30 帧/s的帧率预算下，模型计算仅占用了不到1/4的时间，为其他系统环节留下了充足余地。

5　结束语

针对L3级自动驾驶驾驶人非驾驶任务的接管问题，本文在现有MobileViT网络框架中集成小波变换卷积，构建了L3级自动驾驶工况下驾驶人非驾驶任务分类检测模型MobileViT-WT。

a. 针对现有L3级自动驾驶缺乏驾驶人非驾驶任务数据集的问题，构建了基于驾驶模拟器的专用数据集，特别设计了包含单一任务和复合任务(如“边吃零食边打游戏”“边喝水边刷短视频”)的多场景数据，以充分反映L3级自动驾驶的典型使用特征。

b. 本文对比了不同的轻量级网络模型在L3级自动驾驶非驾驶任务数据集上的表现，最终选择了MobileViT作为主干网络。为了进一步提升性能，在MobileViT网络中适当引入了小波变换卷积，并调节了卷积核的大小，构建了最优的网络模型——MobileViT-WT。

c. 试验结果显示，现有轻量级模型对自动驾驶中的吃零食行为识别效果欠佳(准确率仅70.1%)，且其他轻量级模型普遍存在15%~35%的饮水与复合任务误判，主要由于手-口运动轨迹相似性及多模态特征重叠所致。

d. 本文提出将MobileViT-WT模型应用于L3级自动驾驶场景下的驾驶人非驾驶任务检测，试验表明该模

型在保持轻量化（参数量4.45×10^6，较MobileViT减少11%）的同时显著提升了检测性能：在标准测试集上准确率达93.85%（提升0.82百分点），在泛化性测试集上达到82.9%的准确率（提升5.15百分点），验证了其在模拟驾驶环境中的优越分类能力。

本文提出的MobileViT-WT模型在L3级自动驾驶非驾驶任务检测中展现出显著优势，但仍存在以下改进方向：吃零食任务准确率仅为70.1%，对手-口交互任务及复合任务存在识别混淆问题，未来将引入时空注意力机制和频-空域联合建模来增强特征区分能力；实时性方面可通过采取激活函数优化、知识蒸馏、模型卷积调整等试验来优化模型的实时性检测能力，将模型部署到高端平台增加其推理速度；在应用落地方面建议补充真实路面数据集来训练检验模型并进行实车部署上路检测，使模型能更好地应用于路面上的真实环境。

参 考 文 献

[1] SAE International. Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles: SAE J3016[S]. Warrendale, PA: SAE International, 2021.

[2] MÜLLER A L, FERNANDES E N, HETFLEISCH R, et al. Effects of Non-Driving Related Tasks on Mental Workload and Take-over Times During Conditional Automated Driving[J]. European Transport Research Review, 2021, 13(1): 1-15.

[3] PAN H, PAYRE W, WANG Y, et al. How Do Visual and Cognitive Non-Driving Related Tasks Affect Drivers' Visual Attention and Takeover Performance in Conditionally Automated Driving?[J]. Journal of Transportation Safety & Security, 2025, 17(1): 97-123.

[4] GOLD C, HAPPEE R, BENGLER K. Modeling Take-Over Performance in Level 3 Conditionally Automated Vehicles[J]. Accident Analysis & Prevention, 2018, 116: 3-13.

[5] ERIKSSON A, STANTON N A. Takeover Time in Highly Automated Vehicles: Noncritical Transitions to and from Manual Control[J]. Human Factors, 2017, 59(4): 689-705.

[6] MA S, ZHANG W, YANG Z, et al. Take Over Gradually in Conditional Automated Driving: The Effect of Two-Stage Warning Systems on Situation Awareness, Driving Stress, Takeover Performance, and Acceptance[J]. International Journal of Human - Computer Interaction, 2021, 37(4): 352-362.

[7] 马艳丽，董方琦，秦钦，等. 基于行车风险场的自动驾驶接管风险评估模型[J]. 哈尔滨工业大学学报，2024，56(9)：106-112.
MA Y L, DONG F Q, QIN Q, et al. Driving Risk Field-Based Risk Assessment Model for Autonomous Driving Takeover[J]. Journal of Harbin Institute of Technology, 2024, 56(9): 106-112.

[8] 王文军，李清坤，曾超，等. 自动驾驶接管绩效的影响因素、模型与评价方法综述[J]. 中国公路学报，2023，36(9)：202-224.
WANG W J, LI Q K, ZENG C, et al. Review of Influencing Factors, Models, and Evaluation Methods of Takeover Performance in Automated Driving[J]. China Journal of Highway and Transport, 2023, 36(9): 202-224.

[9] Statefarm. Distracted Driver Detection[DB/OL]. https://www.kaggle.com/c/state-farm-distracted-driver-detection, 2015.

[10] ABOUELNAGA Y, ERAQI H M, MOUSTAFA M N. Real-Time Distracted Driver Posture Classification[EB/OL]. https://arxiv.org/abs/1706.09498.

[11] HU W, ZHANG T, ZHANG Y, et al. Non-Driving-Related Tasks and Drivers' Takeover Time: A Meta-Analysis[J]. Transportation Research Part F: Traffic Psychology and Behaviour, 2024, 103: 623-637.

[12] 张磊，彭金栓，陈晓利. 接管场景驱动下非驾驶任务对驾驶人视觉与生理特征的影响[J]. 交通运输系统工程与信息，2024，24(3)：290-298.
ZHANG L, PENG J S, CHEN X L. Influence of Non-Driving Tasks on Drivers' Visual and Physiological Characteristics Under Takeover Scenarios[J]. Journal of Transportation Systems Engineering and Information Technology, 2024, 24(3): 290-298.

[13] BAI J, SUN X, CAO S, et al. Exploring the Timing of Disengagement from Non-Driving Related Tasks in Scheduled Takeovers with Pre-Alerts: An Analysis of Takeover-Related Measures[J]. Human Factors, 2024, 66(12): 2669-2690.

[14] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet Classification with Deep Convolutional Neural Networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[15] SZEGEDY C, LIU W, JIA Y Q, et al. Going Deeper with Convolutions[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE, 2015: 1-9.

[16] HU J, SHEN L, SUN G. Squeeze-and-Excitation Networks[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 7132-7141.

[17] SHAKER A, MAAZ M, RASHEED H, et al. SwiftFormer: Efficient Additive Attention for Transformer-Based Real-Time Mobile Vision Applications[C]// Proceedings of the

IEEE/CVF International Conference on Computer Vision. 2023: 17425-17436.

[18] KOONCE B. MobileNetV3[M]// Convolutional Neural Networks with Swift for TensorFlow: Image Recognition and Dataset Categorization. Berkeley, CA: Apress, 2021: 125-144.

[19] MEHTA S, RASTEGARI M. MobileViT: Light-Weight, General-Purpose, and Mobile-Friendly Vision Transformer[J]. https://arxiv.org/abs/2110.02178.

[20] 贺宜, 鲁曼可, 高嵩, 等. 基于MobileViT-CA模型的营运车辆驾驶人分心行为检测[J]. 中国公路学报, 2024, 37(1): 194-204.
HE Y, LU M K, GAO S, et al. Distracted Behavior Detection of Commercial Vehicle Drivers Based on the MobileViT-CA Model[J]. China Journal of Highway and Transport, 2024, 37(1): 194-204.

[21] FINDER S E, AMOYAL R, TREISTER E, et al. Wavelet Convolutions for Large Receptive Fields[C]// European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 363-380.

[22] HOWARD A, SANDLER M, CHU G, et al. Searching for MobileNetV3[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 1314-1324.

[23] TANG Y, HAN K, GUO J, et al. GhostNetV2: Enhance Cheap Operation with Long-Range Attention[J]. Advances in Neural Information Processing Systems, 2022, 35: 9969-9982.

[24] WANG A, CHEN H, LIN Z, et al. RepViT: Revisiting Mobile CNN from ViT Perspective[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 15909-15920.

（责任编辑　天下溪）

修改稿收到日期为2025年9月1日。

《汽车技术》征稿启事

《汽车技术》杂志是中国第一汽车集团有限公司主办的国内外公开发行的汽车前瞻与应用技术类月刊，为我国高质量科技期刊分级目录入选期刊、中国科学引文数据库（CSCD）来源期刊、中文核心期刊、中国科技核心期刊、RCCSE中国核心学术期刊（A）、Scopus数据库收录期刊、俄罗斯《文摘杂志》（AJ）收录期刊、日本科学技术振兴机构数据库入选期刊、EBSCO学术数据库收录期刊、欧洲学术出版中心（EuroPub）数据库收录期刊。

《汽车技术》杂志以报道汽车整车及其零部件设计、研究、试验等方面的前瞻与应用技术为主，并兼有理论研究内容，是中国汽车行业核心学术和知识传播与共享的平台。

《汽车技术》将在国家提出的“创新、协调、绿色、开放、共享”发展理念的指引下，把握《节能与新能源汽车技术路线图》和“低碳化、信息化、智能化”的汽车技术主流发展趋势，努力在传统内燃机汽车高效动力系统、轻量化、低阻力领域，新能源汽车和互联智能汽车技术领域，大力吸收优质稿源，为广大科研和工程技术人员服务，为我国汽车工程技术创新能力提升贡献力量。

《汽车技术》欢迎高等院校师生、研发工程技术人员、技术管理人员及相关人员不吝赐稿，反映国家重点扶持项目、自然科学基金项目和其他重点项目等研究成果的稿件将被优先选择刊登。

投稿要求：

1. 文章字数最好控制在6 000~8 000字范围之内；

2. 请按科技论文要求撰写文章摘要，摘要中文字数控制在180字左右；

3. 文章必须附有公开发表的、体现本领域最新研究成果的参考文献，且在文中应标注文献引用处；

4. 文章主要作者应提供其简介，包括出生年、性别、职称、学历、研究方向及技术成果等；

5. 来稿的保密审查工作由作者单位负责，确保署名无争议，文责自负；

6. 请勿一稿多投；

7. 本刊使用网站投稿，请先登陆网站注册成功后投稿，详细投稿要求见本刊网站中“下载中心”栏的“作者指南”，网址：http://qcjs.cbpt.cnki.net。

《汽车技术》编辑部